Proceedings of the Tenth Global Wordnet Conference

July 23–27, 2019, Wrocław (Poland)

Oficyna Wydawnicza Politechniki Wrocławskiej
Wrocław 2019

Editors
Christiane Fellbaum, Piek Vossen, Ewa Rudnicka, Marek Maziarz, Maciej Piasecki

Reviewers
Eneko Agirre, Sina Ahmadi, Mihael Arcan, Timothy Baldwin, Francis Bond, Sonja Bosch, Paul Buitelaar, Bharathi Raja Chakravarthi, Hannah Choiyunjung, Janos Csirik, Bruno Cuconato, Luis Morgado Da Costa, Leonel Figueiredo de Alencar, Gerard de Melo, Michael Wayne Goodman, Valeria de Paiva, Thierry Declerck, Bento C. Dias-Da-Silva, Tomaž Erjavec, Christiane Fellbaum, Darja Fišer, Ales Horak, Shu-Kai Hsieh, Filip Ilievski, Kyoko Kanzaki, Diptesh Kanojia, Neeme Kahusk, David Lindemann, Ahti Lohk, Isa Maks, John P. McCrae, Verginica Mititelu, Sanni Nimb, Haldur Oim, Hugo Gonçalo Oliveira, Sussi Olsen, Heili Orav, Adam Pease, Bolette Pedersen, Maciej Piasecki, Marten Postma, Alexandre Rademaker, German Rigau, Ewa Rudnicka, Shikhar Kr. Sarma, Kevin Scannell, Alberto Simões, Pia Sommerauer, Hennie van der Vliet, Chantal van Son, Umamaheswari Vasanthakumar, Kadri Vider, Piek Vossen, Shan Wang

Printed in the camera ready form

OFICYNA WYDAWNICZA POLITECHNIKI WROCŁAWSKIEJ
wyb. Stanisława Wyspiańskiego 27, 50-370 Wrocław
http://www.oficyna.pwr.edu.pl
e-mail: oficwyd@pwr.edu.pl

ISBN 978-83-7493-108-3

Program Committee

Eneko Agirre	University of the Basque Country
Sina Ahmadi	Insight Centre for Data Analytics
Mihael Arcan	Insight Centre for Data Analytics, National University of Ireland Galway
Timothy Baldwin	The University of Melbourne
Francis Bond	Nanyang Technological University
Sonja Bosch	Department of African Languages, University of South Africa
Paul Buitelaar	Insight Centre for Data Analytics, National University of Ireland Galway
Bharathi Raja Chakravarthi	Insight Centre for Data Analytics, National University of Ireland, Galway
Hannah Choiyunjung	Nanyang Technological University
Janos Csirik	University of Szeged
Luis Morgado Da Costa	Nanyang Technological University
Leonel Figueiredo de Alencar	UNIVERSIDADE FEDERAL DO CEARÁ
Gerard de Melo	Rutgers University
Valeria de Paiva	Samsung Research America and University of Birmingham
Thierry Declerck	DFKI GmbH
Bento C. Dias-Da-Silva	UNESP
Tomaž Erjavec	Dept. of Knowledge Technologies, Jožef Stefan Institute
Christiane Fellbaum	Princeton University
Darja Fišer	University of Ljubljana
Hugo Gonçalo Oliveira	University of Coimbra
Ales Horak	Masaryk University, Faculty of Informatics
Shu-Kai Hsieh	National Taiwan Normal University
Filip Ilievski	Vrije Universiteit Amsterdam
Diptesh Kanojia	IIT Bombay
Kyoko Kanzaki	Toyohashi University of Technology
David Lindemann	UPV-EHU University of the Basque Country
Ahti Lohk	Tallinn University of Technology
Isa Maks	Vrije Universiteit Amsterdam
John P. McCrae	National University of Ireland, Galway
Verginica Mititelu	Romanian Academy Research Institute for Artificial Intelligence
Sanni Nimb	Det Danske Sprog- og Litteraturselskab (DSL)
Haldur Oim	University of Tartu
Sussi Olsen	UCPH
Heili Orav	University of Tartu
Adam Pease	Infosys
Bolette Pedersen	University of Copenhagen
Maciej Piasecki	Wrocław University of Technology
Marten Postma	Vrije Universiteit Amsterdam
Alexandre Rademaker	IBM Research Brazil and EMAp/FGV
German Rigau	IXA Group, UPV/EHU
Ewa Rudnicka	Wrocław University of Technology
Shikhar Kr. Sarma	Gauhati University

Kevin Scannell	Saint Louis University
Pia Sommerauer	Vrije Universiteit Amsterdam
Hennie van der Vliet	Vrije Universiteit Amsterdam
Chantal van Son	Vrije Universiteit Amsterdam
Umamaheswari Vasanthakumar	Research Fellow NTU Singapore
Kadri Vider	University of Tartu
Piek Vossen	Vrije Universiteit Amsterdam
Shan Wang	University of Macau

Additional Reviewers

Chakravarthi, Bharathi Raja
Cuconato, Bruno
Goodman, Michael Wayne
Kahusk, Neeme
Simões, Alberto

Preface

The tenth Global WordNet Conference took place in Wroclaw (Poland) July 23-27, 2019. Fifty papers were presented by authors from four continents covering a wide range of topics and languages. New wordnets were introduced for Swiss German, siSwati, Coptic, Tatar, Cantonese and Mongolian as well as for different modalities (Spoken WordNet and ASLNet for American Sign Language).

Several authors reported on crosslingual wordnet alignment. Work on WordNet extensions covered ontology, gloss corpus annotation and the inclusion of geographical named entities. Applications of wordnets included sense alignment, semantic annotation, sentiment analysis, cognate detection, coreference resolution, document classification, alignment with wikipedia, reasoning, pedagogy and translation. The current focus on embeddings, an approach to semantics that considers syntagmatic rather than WordNet's paradigmatic perspective, was reflected in several presentations.

The present proceedings testify to the continuing growth of wordnet research and development and its place within the broader communities of colleagues in Natural Language Processing and computational and theoretical linguistics.

July 27, 2019
Wroclaw

Christiane Fellbaum
Piek Vossen
Ewa Rudnicka
Marek Maziarz
Maciej Piasecki

Table of Contents

Making Sense of schema.org with WordNet

Csaba Veres
Department of Information Science and Media Studies
The University of Bergen, Norway
csaba.veres@uib.no

Abstract

The *schema.org* initiative was designed to introduce machine readable metadata into the World Wide Web. This paper investigates conceptual biases in the *schema* through a mapping exercise between *schema.org* types and WordNet synsets. We create a mapping ontology which establishes the relationship between *schema* metadata types and the corresponding everyday concepts. This in turn can be used to enhance metadata annotation to include a more complete description of knowledge on the Web of data.

1 Introduction

Schema.org is an initiative to introduce machine readable metadata into HTML Web pages. It was launched on June 2, 2011, under the auspices of a consortium consisting of Google, Bing, and Yahoo!. The *schema.org* web site initially described the project as one that "provides a collection of schemas, i.e., html tags, that webmasters can use to markup their pages in ways recognized by major search providers .... making it easier for people to find the right web pages." (schema.org web site, 2011). The incentive for using the schema was that web sites that contained markup would appear with informative details in search results which in turn enables people to judge the relevance of the site more accurately. This could lead to higher user engagement and higher search ranking, which is the ultimate incentive for web masters.

The initial release contained 297 classes and 187 relations, but by 2016 had grown to 638 classes and 965 relations (Guha et al., 2016). It is important to note, however, that the expansion of the *schema* consists entirely in adding subclasses and properties to the core classes through the allowed extension mechanism[1]. From the outset the immediate sub classes of *Thing* were stiulated as *Action, CreativeWork, Event, Intangible, Organization, Person, Place* and *Product*. These high level conceptual divisions with their implicit ontological commitments are not, and never were open to discussion.

(Guha et al., 2016) explain that the primary driving force behind the design of the *schema*, and ultimately the reason for its success, was its simplicity. Previous efforts to introduce large scale metadata failed, in part because each standard was too narrow in terms of domain coverage. The result was too many standards for too few applications. On the other hand the *schema* offered a single, unified and broad vocabulary that could be used across several verticals and promised a benefit for perhaps the most important driving force, search rankings. As a part of this simplicity, the *schema* taxonomy and classes were intended more as an "organisational tool to help browse the vocabulary" than a definitive ontology of world (Guha et al., 2016). In other words, the *schema* was designed as an intuitive set of metadata classes that could be used to describe the majority of items people would search for on the Web.

Together these factors ensured that the *schema* has enjoyed a significant amount of success. (Guha et al., 2016) report that in a sample of 10 billion web pages, 31.3% of the pages had *schema.org* markup, a growth of 22% from a year earlier. The markup is used by many different data consumers for various tasks involving enhanced search results (rich snippets), populating the Google Knowledge Graph, exchange of transaction details in email, support for automatic formatting of recipes, reviews, etc., and advanced search features in Apple's Siri. The fifteen most popular implemented classes were *WebSite, SearchAc-*

[1] https://is.gd/HdnHkp

tion, WebPage, Product,ImageObject, Person, Offer, BlogPosting, Organization, Article, PostalAddress, Blog, LocalBusiness, AggregateRating, WPFooter. Many of these refer to elements of the web page itself rather than the content. The top fifteen content bearing classes were *Product, ImageObject, Person, Offer, Organization, PostalAddress, LocalBusiness, AggregateRating, CreativeWork, Review, Place, Rating, Event, GeoCoordinates*, and *Thing*. These are sun types of *Product, CreativeWork, Person, Intangible, Organization, Place,* and *Event*. Although the coverage was intended to be broad, it is clear that the use of the *schema* covers its range of types well, but that the types favour a particular view of web content, in the interests of the search providers.

The motivation for this paper was to try and characterize the conceptual biases of the *schema* top level categories, by mapping the types to their corresponding meanings in WordNet. To the extent that we believe WordNet captures the ontological commitments inherent in human language, it should provide insights about where the two conceptualisations diverge. The further aim, however, is to use the mappings to enrich the valuable human provided metadata towards the aim of providing general but rich meaning annotations to a large portion of Web content.

It is important to note that we are not advocating WordNet as a *gold standard* for ontologies and knowledge representation. On the contrary, we agree with (Hirst, 2004) who argues that WordNet contains modeling decisions which differentiate it from formal ontologies. As an example, there are cases where synsets have overlapping hyponyms whereas ontologies have disjoint subclasses. Consider the first noun sense of *mistake: {mistake, error, fault}* which includes the following hyponyms (among others): *{slip, slip-up, miscue, parapraxis}, {oversight, lapse}, {faux pas, gaffe, solecism, slip, gaucherie}, and {failure}*. A single act can be both a *slip* and a *faux pas*. The first implies the act was inadvertent, and the second that it possibly had a social component such as a mistake in etiquette. A *lapse* is also a *slip*, but it involves some sort of forgetfulness or inattention on top of the mere *slip*. A lapse can also be a *faux pas*, of course. If the *faux pas* is sufficiently severe, it can become a complete *failure*. These hyponyms contain more information that that they are a *kind-of mistake*, they also contain information about likely causes and implications, and these can be overlapping. Nevertheless, our interest is that people **do** consider these as kinds of mistake in everyday discourse. For the same reason we think it is beside the point to try and restrucutre WordNet by some formal methodology such as DOLCE (Gangemi et al., 2003a). We are interested here in intuitive relations, not formal ones.

2 The WordNet Mappings

The mapping involved two stages. First the *schema.org* types were aligned with WordNet synsets, while retaining the structure of the *schema*. This stage can be seen as adding information to the *schema*, namely, the corresponding WordNet synsets. Then, a new hierarchy of concepts was constructed from the synsets involved in the mapping. That is, by promoting the mapped synsets to be the central classes, we could get a better idea what sorts of concepts are in the *schema*, in relation to the WordNet taxonomy.

In order to distinguish between the concepts in the two taxonomies, WordNet names will be prefixed with *wn:* and the *schema* with the prefix *schema:*. In addition when necessary the WordNet name will be qualified with part of speech and sense tag, as in *wn:dog#n#1*.

To summarize, we constructed two artefacts at the end of the process:

- The WordNet to schema.org mapping ontology. This retains the *schema* class structure. The mappings were manually constructed and available on GitHub[2].
- The WordNet taxonomy for the synsets that have been mapped to the schema. This shows an alternative taxonomy of the words in the *schema*.

2.1 The Mapping Ontology

In this ontology the original *schema.org* taxonomy was retained, and the WordNet synsets were simply inserted into this taxonomy. In fig. 1 we see some example mappings, showing schema:Beach mapped to wn:beach. Since schema:Beach is a subtype of schema:CivicStructure, by implication so too is wn:beach. Similarly, the other WordNet synsets in the example become subclasses of schema:CivicStructure through their respective

[2] https://is.gd/XF0bJe

alignments. The mapping provides the immediate benefit that web sites which contained any of the WordNet synsets in the alignment, could automatically be connected to their corresponding *schema* types. This suggests a method for automatic metadata creation, which will be dicussed subsequently.

Notice that the mapping is not straightforward and in this example synsets of quite distinct types are grouped under the one *schema* type. For example wn:bus_terminal <is-a> wn:facility, wn:cinema <is-a> wn:theater <is-a> wn:building, and a wn:parking_lot <is-a> wn:tract,piece_of_land. Yet they all map to subclasses of schema:CivicStructure.

The second taxonomy was created precisely to reveal the *schema* conceptualisation in terms of the WordNet hierarchy. In other words, *"what IS a schema:CivicStucture in everyday language?"*

2.2 The WordNet Ontology

The full WordNet hypernym tree is quite deep, and quickly leads to a very complex taxonomy. For this reason we made use of a simple tool which uses an algorithm to eliminate low information nodes from a taxonomy (Veres et al., 2013). The algorithm prunes the tree by counting the number of outward links at each node, and eliminating any node that has fewer than a certain number of (user specified) hyponyms. When this is performed on every node in the graph, what remains is a number of intermediate synsets which are the maximally informative hypernyms of any leaf node. In the graphs reported here, the lower threshold was set at 3. The tool essentially implements the algorithm used by (Stoica and Hearst, 2004), but our interface has the advantage that the parameters can be dynamically adjusted and visually inspected to give the most intuitively pleasing result. A similar procedure was followed in (Izquierdo et al., 2006) to identify basic level concepts. Our work differs in that we do not distinguish between nodes above the basic threshold.

A part of the inferred hierarchy involving wn:beach is shown in figure 2. Note that wn:beach is a sibling of wn:mountain, whereas the *schema* choice to model the *civic structure* aspect of *beach* puts them in different subclasses; schema:Beach is a schema:CivicStructure while schema:Mountain is a schema:Landform. However, since wn:beach is a hyponym of wn:geological_formation, which in turn is an equivalent class of schema:LandForm, it could be inferred that schema:Beach could also be a schema:LandForm. The benefit of the alignment is that a new and sensible *schema* type could be added to any markup involving *beach*. Figure 3. shows how the WordNet hierarchy connects wn:beach to schema:Landform and potentially other subclasses. A web site about a geographical area with mountains and beaches could then be appropriately annotated.

Looking at the taxonomy itself, we can see what kind of WordNet synsets appear in the *schema*. The major division in fig. 4 is between wn:physical_entity and wn:abstraction, which is an ontological distinction that is typically considered fundamental (e.g. (Niles and Pease, 2001), (Gangemi et al., 2003b)). On this view the *schema* describes the world as populated by *physical entities* and *abstractions*, where the *physical entities* are predominantly *objects*, and *abstractions* are diverse sorts of *events* or *roles* which the entities engage in. For example wn:measure is *how much there is of something you can quantify*, and wn:state is *the way something is with respect to its attributes*. Other sub types of wn:abstraction, like wn:organization and wn:tourist_attraction apply to concepts that are typically human centered, functional collections of objects (Wierzbicka, 1984). Wierzbicka argues that putatively taxonomic concept hierarchies are in fact the majority of the time made up of a mixture of supercategory types, with the most prominent two being taxonomic and functional. (Pustejovsky, 1991) draws a similar distinction with the mechanism of *formal* and *telic* roles in his lexical structures.

The ontological commitment adopted by *schema.org* becomes clear if we compare the two taxonomies. The *schema* divides schema:Thing into: schema:CreativeWork, schema:Event, schema:Intangible, schema:Organization, schema:Person, schema:Place, and schema:Product. The focus is immediately on the functional categories: telic roles dominate the top level categories of the *schema*, and physical entities are sub types of these abstractions.

The most obvious example of a top-level purely functional type is schema:Product. Almost anything can be a *product*, and there is no property which *products* have in common except the telic

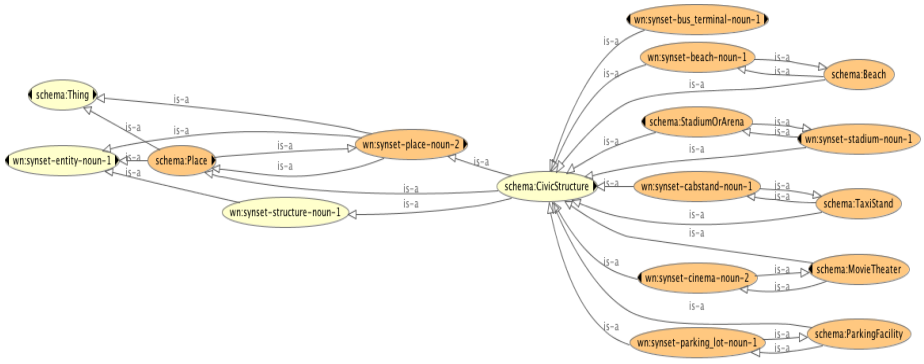


Figure 1: Example mappings between WordNet and schema.org, for the corresponsing concepts *beach*. The ovals in darker shading represent concepts which have equivalent classes in the two namespaces.

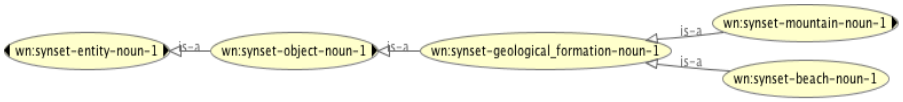


Figure 2: Part of the WordNet taxonomy

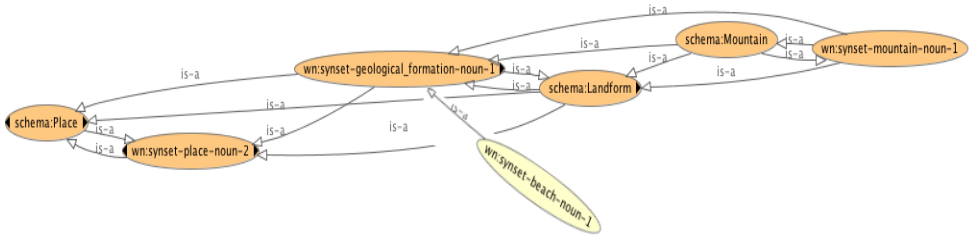


Figure 3: wn:beach inherits schema:Landform

Figure 4: Part of the WordNet taxonomy from SynsetTagger

role that they are "made available for sale". One can sell a sewing needle or a Saturn V rocket. Actually the situation is even more complicated because Products don't even have to be individuated "things". The documentation of schema:Product reads: "a pair of shoes; a concert ticket; the rental of a car; a haircut; or an episode of a TV show streamed online".

The fact that there are in fact a number of functional categories at the highest level helps explain the strange tangle of types at the lower levels of the hierarchy, where many different kinds of things (in the formal, taxonomic sense) can appear if they serve particular functions. To see how this becomes problematical, consider the common functional category *weapon* which can include items such as *crossbow, flamethrower, gun, knife, poison gas, anthrax bacillus, novichok, boomerang, and hydrogen bomb*. Clearly as individual objects these would have quite different sets of properties. The problem for the *schema* is that different *formal* objects are forced to coexist as siblings in a taxonomy dominated by *telic* roles. This results in examples such as schema:Beach having opening hours, schema:Continent with a telephone number and review, and other *strange and wonderful things*. One is forced to assume that schema:Beach was designated as a schema:CivicStructure, for example, because the emphasis is on the facilities available at the beach, not the beach itself.

The inclusion of telic roles such as schema:Product at such a high level of generality has the additional consequence that the *schema* does not contain a type which corresponds to the simple notion of a *physical object*. There is no option in *schema.org* for the structured markup of cars, boats, computer chips, barbells, antiques, or any of the other hundred million human artefacts ancient and modern, except as a "Product", because the schema lumps these into the class of "sellable things". Neither does there seem to be any proper place for natural objects like *cats* or *dogs*[3] or *tree* amd *forest*, which simply have no place.

Finally it should be noted that the hierarchy in WordNet does also include purely functional types among its hypernyms. For example in the *weapon* example above we see that wn:gun is-a wn:weapon is-a wn:object. George Miller (in (Fellbaum, 1998)) explains that this was perhaps an unfortunate problem that might have been avoided had the importance of Wierzbicka's work been realized earlier. However, the structure of WordNet ensures that, whenever such a confusion exists, the formal properties of the word are still recorded. One mechanism is that words can appear in more than one hierarchy. For example *anthrax bacillus* is both a wn:microorganism, and a wn:weapon. Another possibility is that words with both roles are listed twice. For example wn:chicken#n#1 <is-a> wn:meat#n#1, and wn:chicken#n#2 <is-a> wn:bird#n#1. The *schema* only offers one choice for the poor chicken, schema:MenuSection.

3 Finding correct mappings

There are a number of potential pitfalls in defining appropriate mappings between the two taxonomies. One of the most important is to avoid introducing unwanted inferences from the semantics of the mapping axioms. A prevalent example of this is the use of owl:sameAs to represent equivalence between individuals, or classes in OWL-Full. owl:sameAs asserts full equivalence between the individuals such that all of their properties are automatically shared, even though most commonly this is not the desired consequence (Halpin et al., 2010). To avoid this problem we used the weaker owl:equivalentClass axiom, which does not imply complete equality. What is required instead is the weaker condition that every instance of one class must also be an instance of the other.

Even with a weaker semantics we found that equivalent classes could not always be found. One reason is that schema.org includes concepts which involve various sorts of compounding of simple concepts, and WordNet contains only common, lexicalized compounds. For example LandmarksOrHistoricalBuildings is a compound concept that includes any kind of general *landmark* as well as the specific concept of *buildings with historical significance*. There is no such lexical entry in English. Most likely there is no such compound in any language, because the concept is un-natural, mixing different levels of generalization. It is analogous to a concept for *toys or teddy bears*.

There are also more acceptable compounds like schema:CivicStructure which is "a public structure such as a town hall or concert hall". This is of course a perfectly acceptable compound, which

[3] the search facility suggests schema:AnimalShelter

happens not to be in WordNet. In every case that an acceptable WordNet compound could not be found, we decided to make the *schema.org* concept a subclass of one or more WordNet synsets that captured part of the compound. For the above example of schema:CivicStructure, the obvious superclass is wn:structure#n#1.

Sometimes the compound nature of the *schema* terms is hidden. For example the terms that are subclasses of schema:LocalBusiness are a mixed group of explicit compounds (e.g., schema:MovingCompany, schema:IceCreamShop) and implicit compounds (e.g., schema:Electrician, schema:Locksmith, schema:HousePainter). That is, schema:Electrician is really meant to be something like "ElectricianBusiness" and not just "Electrician". The compound schema:HousePainter is even more complicated because it has an exact match in wn:house_painter#n#1, but in fact schema:HousePainter is really meant to be a HousePainterBusiness, so the exact match is illusory. The important modelling decision is whether or not to reintroduce the hidden compound in mapping to WordNet. That is, should schema:Electrician be regarded in its ordinary word sense as "a person who is an electrician", or should it be modelled as an "electrician business"? In other words, these concepts could simply be declared as subclasses of wn:place_of_business to maintain the intended interpretation in the *schema*. The most flexible solution was to declare an equivalent class relation between schema:Electrician and the person interpretation in WordNet, wn:electrician#n#1. This choice captures the notion that electricians are people. However it is also possible to infer that wn:electrician is a wn:place_of_business, as shown in Figure 5.

There is a small set of *schema.org* types for which we did not establish mappings. One group involved technical compounds describing the structure of web pages with terms like schema:AboutPage and schema:CheckoutPage. These are all subtypes of schema:WebPage, for which we did define a mapping. The second group was the primitive data types, schema:DataType which are not part of the main taxonomy subsumed by schema:Thing.

4 Using the WordNet Mappings

The practical motivation for mapping the *schema* to WordNet was to enrich the metadata that can be assigned to concepts in a web page. We have already seen this in examples such as *beach*. A secondary motivation was to make it easier for web masters to find the *schema* types without knowing anything about its structure. We have already developed a prototype of a tool in which the user can highlight any word in text, nominate its corresponding synset, and the application will attempt to guess the correct *schema* type. Consider the following example scenario.

There is a geological landmark called the Jenolan Caves in the Blue Mountains, Australia. Suppose a web master wanted to mark up the web site for Jenolan Caves. A quick search will reveal that there is no matching type in the *schema* for caves. Using the WordNet mappings it is possible for the designer to find the most appropriate types, without any knowledge of the *schema*. The synset wn:cave is a wn:geological_formation, which in turn maps to schema:Landform. However, the mapping ontology can also suggest additional useful classifications. The coordinate terms of wn:cave contain some terms which **are** defined in the *schema*, including our old friend *beach*. Recall that wn:beach is mapped to schema:CivicStructure through schema:Beach (see Figure 6). Thus Jenolan Caves could be marked with both *schema* types, and the properties of the facilities at the premises could be specified. Of course the annotation effort does not have to stop there. Since the WordNet synset is available, it can also be included in the markup, which in turn enables the markup to be used with a huge number of mappings to other resources[4].

While this process is currently being performed through our prototype tool where users specify the disambiguated sense (Veres and Elseth, 2013), this does not necessarily have to be performed manually. With sufficiently accurate disambiguation methods, any web page could be automatically annotated with *schema* and WordNet metadata. This would be useful for any downstream task including the construction of knowledge graphs, as previously mentioned.

The Jenolan Caves example requires the ability to declare multiple types. The original syntax for

[4] https://wordnet.princeton.edu/related-projects

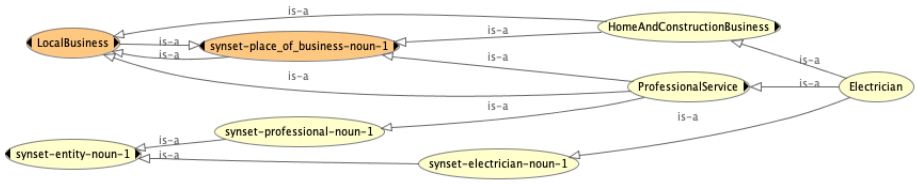


Figure 5: Electrician as both a person (wn:electrician#n#1) and place of business (wn:place_of_business#n#1).

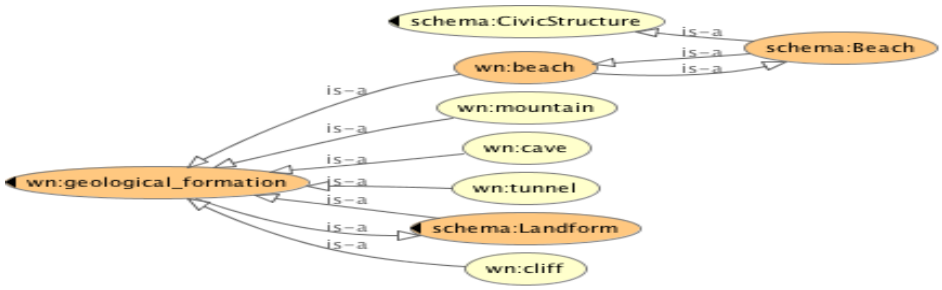


Figure 6: Mapping "cave" to schema.org types

the *schema*, *microdata* is not able to express multiple types. The recommendation therefore is to use *rdf-a*[5] or *json-ld*[6] which are inherently built to express multiple types from any vocabulary.

5 Conclusion

We proposed a method for evaluating the conceptual bias of *schema.org* by comparing the type terms against their usage in everyday language as stipulated in WordNet. The observation is that *schema.org* favours the markup of web sites promoting goods, services, and locations fulfilling some human centred need. This then results in the observed data that the majority of web sites which contain *schema.org*, are about products and goods and services. If search rankings favour sites with markup, and if most markup is about goods and services, then search results will come to favour goods and services. Anecdotally, this could be one factor for why it is sometimes easier to find where to buy something rather than information about the thing itself. The bias diminishes the potential for providing a rich source of general semantic metadata on the web, for use in diverse use cases.

We argued that the *schema* needs types that describe a more neutral view of the world, for example artefacts, to describe *things* independently of the *roles* they can play. A metadata specification should be able to annotate a *chicken* as a kind of *bird* as well as a kind of *food*.

Our suggestion to include WordNet mappings into the markup effort is one way to sneak more general markup into the annotation process. The requirement is that multiple types must be a standard feature of the annotation, with different types describing different aspects of the item. A *car* is an artefact designed for locomotion, but can also acquire its role as a *product* if it is put up for sale. This addition would not compromise people who want to advertise their products. In fact, it would give them more freedom to express physical properties of their products like size, construction material, origin, and so on.

In summary, we used WordNet as a standard representation of everyday word use, to provide clarity to the types proposed in *schema.org*. We proposed a method to help people mark up Web sites that do not fit neatly into the service oriented world view, by enabling them to annotate their contribution to world knowledge as broadly as possible. This is clearly of benefit to all users who see the web as a vehicle for disseminating informative structured data as freely as possible.

[5] https://rdfa.info/
[6] https://json-ld.org/

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003a. Sweetening wordnet with dolce. *AI Magazine*, 24:13–24.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003b. Sweetening wordnet with dolce. *AI Mag.*, 24(3):13–24, September.

RV Guha, Dan Brickley, and Steve Macbeth. 2016. Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59:44–51.

Harry Halpin, Patrick J. Hayes, James P. McCusker, Deborah L. McGuinness, and Henry S. Thompson. 2010. When owl:sameas isn't the same: An analysis of identity in linked data. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web – ISWC 2010*, pages 305–320, Berlin, Heidelberg. Springer Berlin Heidelberg.

Graeme Hirst, 2004. *Ontology and the Lexicon*, pages 209–229. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rubén Izquierdo, Armando Suárez, German Rigau, and Ixa Nlp. 2006. Exploring the automatic selection of basic level concepts. In *International Conference Recent advance in Natural Language Processing*.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA. ACM.

James Pustejovsky. 1991. The generative lexicon. *Comput. Linguist.*, 17(4):409–441, December.

Emilia Stoica and Marti A. Hearst. 2004. Nearly-automated metadata hierarchy creation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 117–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Csaba Veres and Eivind Elseth. 2013. Schema.org for the semantic web with madame. In *Proceedings of the I-SEMANTICS 2013 Posters & Demonstrations Track, Graz, Austria, September 4-6, 2013*, pages 11–15.

Csaba Veres, Kristian Johansen, and Andreas Opdahl. 2013. Synsettagger: A tool for generating ontologies from semantic tags. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 16:1–16:10, New York, NY, USA. ACM.

Anna Wierzbicka. 1984. Apples are not a "kind of fruit": the semantics of human categorization. *American Ethnologist*, 11(2):313–328.

Leaving No Stone Unturned When Identifying and Classifying Verbal Multiword Expressions in the Romanian Wordnet

Verginica Barbu Mititelu
RACAI
Bucharest, Romania
vergi@racai.ro

Maria Mitrofan
RACAI
Bucharest, Romania
maria@racai.ro

Abstract

We present here the enhancement of the Romanian wordnet with a new type of information, very useful in language processing, namely types of verbal multiword expressions. All verb literals made of two or more words are attached a label specific to the type of verbal multiword expression they correspond to. These labels were created in the PARSEME Cost Action and were used in the version 1.1 of the shared task they organized. The results of this annotation are compared to those obtained in the annotation of a Romanian news corpus with the same labels. Given the alignment of the Romanian wordnet to the Princeton WordNet, this type of annotation can be further used for drawing comparisons between equivalent verbal literals in various languages, provided that such information is annotated in the wordnets of the respective languages and their wordnets are aligned to Princeton WordNet, and thus to the Romanian wordnet.

1 Introduction

The Romanian wordnet (RoWN) is a rich lexical and semantic resource. Its development followed the expand method (Vossen, 2002) and started within the BalkaNet project (Tufiș et al., 2004). Alignment with Princeton WordNet (PWN) (Miller, 1995; Fellbaum, 1998) was a consequence of this working method and has always been one of the objectives whenever new synsets were developed for enlarging the RoWN. Consequently, alignment with all the other wordnets aligned with PWN is obtained, which is a great asset for both interlingual lexical comparison or for applications working in a multilingual environment.

The expand model in wordnets development implies importing the structure of the PWN (that is, its semantic relations) and translating the source synsets (from PWN), so that the meaning encoded by the English synset is rendered in the target language (Romanian, here). As a consequence, a Romanian synset may have one of the following structures: (i) list of words; (ii) list of free word combinations; (iii) empty list. (i) A list of words is a list of simple words (ex. *zâmbi* ("smile")) and/or expressions (ex. *casă de bani* (house of money "strong box")). These expressions are what in lexicographic terms is called idioms, terms, etc. (ii) Whenever no word or expression could be found in Romanian for rendering the meaning of the English synset, a free word combination, when possible, was used for implementing the respective synset: ex.: *pune jos* is a literal in the Romanian synset equivalent to the PWN 3.1 {ground:10} (gloss: place or put on the ground). These are examples of Recurrent Free Phrases, as Bentivogli and Pianta (2004) call them. (iii) In case not even such a combination could be found, the synset was left empty and a special tag is used for keeping track of them (they are marked as NL, i.e. non lexicalized): ex.: the English synset {change state:1, turn:4} (gloss: undergo a transformation or a change of position or action) has a non-lexicalized corresponding synset in RoWN. However, as already pointed out (Vincze et al., 2012; Bentivogli and Pianta, 2004; Agirre et al., 2005), these lexical gaps should be reduced as much as possible when use of wordnets is envisaged for tasks in a multilingual environment (see machine translation), but also for word sense disambiguation (Bentivogli and Pianta, 2004).

As far as this structure of its synsets is concerned, RoWN looks as rendered in Table 1. One should bear in mind the fact that it is impossible to distinguish automatically between expressions and free word combinations. That is why, on rows

4 and 5 in Table 1 both types of literals, expressions and free combinations of words, are counted together. As one can see, almost 70% of all Romanian synsets are made up of only simple literals. Those made up of only multiword literals represent 21.2% of all synsets. Less than 5% of the Romanian synsets are made up of both simple and multiword literals, having almost the same distribution as non-lexicalized synsets.

Types of synsets	*Number*	*Percent*
all synsets	59,348	-
synsets containing only simple literals	41,188	69.5%
synsets containing simple literals, expressions and free word combinations	2,813	4.7%
synsets containing expressions and/or free word combination	12,590	21.2%
non-lexicalized synsets	2,757	4.6%

Table 1: Distribution of different types of synsets in RoWN.

As far as the distribution of simple literals and expressions in RoWN is concerned, Table 2 shows that, at the literal level, the situation is somehow different: almost 65% of the whole number of unique literals are simple ones, whereas 35% are multiword ones. When considering their all occurrences, we notice that the simple ones are more frequent (76.5%), given their polysemy which is bigger than that of multiword units (see also (Bentivogli and Pianta, 2004)), which account for only 23.5% of the number of all literals in RoWN.

At present, we are carrying out a bilateral (Romanian-Bulgarian) project of annotating the different types of multiword expressions in the Romanian wordnet. The first step is annotating the verbal multiword expressions (VMWEs). This follows naturally from our participation in the PARSEME Cost Action[1] and in the creation and annotation of the corpora used in the PARSEME

[1] https://typo.uni-konstanz.de/parseme/

Types of synsets	*Number*	*Percent*
all literals	85,277	-
simple literals	65,246	76.5%
expressions and/or free word combination	20,031	23.5%
unique literals	50,480	-
unique simple literals	32,664	64.7%
unique expressions and/or free word combination	17,816	35.3%

Table 2: Distribution of different types of literals in RoWN.

shared tasks 1.0 (Savary et al., 2017) and 1.1 (Ramisch et al., 2018). This paper focuses on the annotation of Romanian wordnet data. We present the PARSEME typology of VMWEs and the types applicable to Romanian (section 2), the process of annotating the verbal literals in RoWN with these types of VMWEs (section 3) and we discuss the obtained results, as well as a comparison with those from the annotation of a Romanian news corpus with the same types of VMWEs (section 4), before concluding the paper.

2 Typology of verbal multiword expressions

For the organization of a shared task on the automatic identification and classification of VMWEs, the existence of an annotated corpus was one of the prerequisites. The interest in this initiative manifested by representatives of quite a large number of languages lead to fruitful discussions and the creation of an annotation manual defining the scope of the task, the types of VMWEs to be annotated and their characteristics. The annotation guidelines capture the idiosyncrasies of all the languages involved.

According to the last version of these guidelines[2], VMWEs fall into universal, quasi-universal

[2] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/index.php?

and language specific categories, the first two having some subcategories, as follows:

- universal categories are types of VMWEs that exist in all natural languages (at least in those participating in the PARSEME corpus annotation action). Their subcategories are:
  - *light verb constructions* (LVC) - they are made up of a verb and a predicative noun (directly following the verb or being introduced by a preposition), the latter having semantic arguments. Depending on the semantics of the verb, two subtypes are identified:
    - LVC.full - these are expressions in which the verb's contribution to the expression's semantics is (almost) null (we call the verb "light"): example: *pay a visit*;
    - LVC.cause - in these expressions the verb has a causative meaning, i.e. it identifies the subject as the cause or source of the event or state expressed by the noun in the expression: example: *give a headache*;
  - verbal idioms (VID) - they are made up of a verb and at least one of its arguments and have a totally non-compositional meaning (Vincze et al., 2012): example: *kick the bucket* (die);
- quasi-universal categories exist only in some of the languages under study. They are:
  - *inherently reflexive verbs* (IRV) - these are verbs that are accompanied by a clitic pronoun with a reflexive meaning: example: *help oneself*;
  - *verb-particle construction* (VPC) - these are verbs accompanied by a particle which totally or partially changes the meaning of the verb: example: *put off*;
  - *multi-verb constructions* (MVC) - they are sequences of two adjacent verbs functioning together as a single predicate with the same subject; this type does not exist in English.

Romanian displays only the following types of VMWEs from the PARSEME classification: LVC.full: *lua o decizie* (make a decision), LVC.cause: *da bătăi de cap* (give headaches), VID: *trage pe sfoară* (pull on rope "cheat") and IRV: *se preface* ("pretend"). These labels were used for the annotation of the Romanian corpus used in the shared task version 1.1 (as in version 1.0 the VMWEs types were slightly different). No language specific categories were necessary in the corpus annotation.

page=home

3 Annotation of the Types of VMWEs in RoWN

The task of annotating the VMWEs in a wordnet is different in some respects from their annotation in a corpus. First, all components are present as one literal in the synset, whereas in a corpus they need to be identified, according to the specifications available for all languages (e.g., auxiliaries, clitics or negation are not annotated as parts of the expression). Second, whenever at least one element of the VMWE inflects for number, gender, etc., it has a unique form in the wordnet, the one considered lemma, while in the corpus all inflected forms may be found and need to be recognized. Third, no voice alternation is to be found in the wordnet, while this can be spotted in a corpus. Fourth, when the decision on whether a word combination is a VMWEs depends on the meaning of that combination, the gloss attached to the synset is useful for this and the decision is based on it.

The annotation of VMWEs in RoWN was done by one linguist, with experience in annotating VMWEs in a corpus, following the PARSEME guidelines. Thus, we cannot discuss here the difficulty of this annotation or any controversial cases. The data are stored in a standoff file[3]. The file contains the literals in each synset, their VMWE label and the unique identifier of each synset, which is taken from PWN 3.0.

All VMWEs in RoWN were identified, extracted and were assigned to one of the types of VMWEs applicable to Romanian (LVC.full, LVC.cause, IRV and VID). However, these types proved not enough for this task. The free word combinations with a verb as head could not be annotated with any of these labels, as expected, in fact. Consequently, we marked them with a new label, NONE: they have a literal, compositional meaning, they do not display the characteristics of the VMWE classes: such an example is *culege nuci* (pick nuts).

[3] http://www.racai.ro/en/tools/text/

This type of annotation is done at the literal, not at the synset level (see also the discussion about the distribution of different types of VMWEs within a synset, in the next section).

Although the vast majority of VMWEs belong to only one type, there are literals which are annotated differently when belonging to different synsets, i.e. when having different meanings. Out of only a handful of such cases, here is one example: the expression *scoate fum* (give out smoke) is annotated as NONE when being in the synset corresponding to the English {fume:4; smoke:4} (gloss: emit a cloud of fine particles) and it is annotated as VID when belonging to the synsets corresponding to the English {steam:3} (gloss: get very angry).

4 Annotation Results

The distribution of the types of VMWEs in the RoWN is presented in Table 3. As one can see, there is a great number (1,211) of artificial verbal expressions (the label NONE). The most frequent type of expressions is IRV (989), followed by VID (614). The numbers of LVC.full and LVC.cause are quite low: 102 and 42, respectively.

Type	***No.***	***%***	***%*** *ignoring NONE*
LVC.full	102	3.4	5.8
LVC.cause	42	1.4	2.4
VID	614	20.9	35
IRV	989	33.3	56.5
NONE	1,211	40.8	
double ann.	5	0.2	0.3
TOTAL	2,963		

Table 3: The distribution of VMWEs types in the RoWN.

As far as the correlation of these figures with those found in the corpus annotated in PARSEME (see Table 4) is concerned, we notice that the frequency distribution is roughly the same, with IRV the most frequent type, followed by VID, while the subtypes of LVC are both rare.

We can conclude that the IRV type is the most frequent both at the lexicographic level and in language use for Romanian.

Figure 1 shows the presence of VMWEs in synsets of different lengths. We notice their greatest presence in shorter synsets (especially of lengths 1 or 2).

Type	***No.***	***Freq.***	***Rel. freq.***
LVC.full	39	312	5.31
LVC.cause	8	181	3.08
VID	171	1,602	27.28
IRV	268	3,777	64.32
TOTAL	486	5,872	-

Table 4: The distribution of VMWEs types in a Romanian news corpus.

		No. of VMWEs per synset					
		1	**2**	**3**	**4**	**5**	**6**
No. of literals per synset	**1**	867					
	2	246	220				
	3	79	54	41			
	4	30	18	15	12		
	5	11	7	6	3	4	
	6	3	0	2	0	0	1
	7	1	0	0	0	0	0
	8	1	1	0	0	0	0

Figure 1: The distribution of VMWEs in synsets of different lengths.

Figure 2 shows the distribution of RoWN synsets made up only of VMWEs by the number of literals in the synset. This is relevant for the productivity of the synonymy relation between VMWEs. As one can see, most of these expressions (867) do not have synonyms. It is noteworthy that this is the case mainly with those annotated as NONE, which is further proof of their artificial nature. There are 220 literals in which there are pairs of synonymous VMWEs. Synonymy among three VMWEs is displayed by 41 synsets, among four VMWEs by 12 synsets, among five VMWEs by 4 synsets, among six VMWEs by 1 synset, and among twelve VMWEs by 1 synset. This very rich synset is {fi de gardă, fi de pază, fi de strajă, fi de santinelă, face de gardă, face de strajă, face de pază, face de santinelă, sta de pază}, which is the equivalent of the PWN synset {stand guard:1, stand watch:1, keep guard:1, stand sentinel:1} (gloss: watch over so as to protect). This Romanian synset is based, on the one hand, on the synonymy among the nouns in the VMWEs

structure (*gardă, pază, strajă, santinelă*) and, on the other hand, on their collocation with three different verbs (*fi, sta, face*) for rendering the same meaning.

We analyzed the (277) synsets in which all literals are VMWEs in order to identify the synsets for which all types of MWEs occurring in the respective synsets are the same. After excluding those synsets containing only strings annotated as NONE (129), we counted 37 synsets in which the literals are all VID, 3 in which they are all LVC.full, 2 in which they are all LVC.cause and other 2 in which they are all IRV.

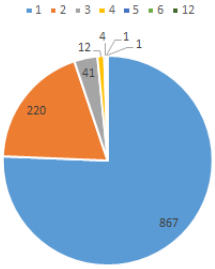


Figure 2: Distribution of synsets containing only VMWEs.

5 Conclusions

We have presented here the enhancement of the RoWN with a new type of syntagmatic information, namely labels for VMWEs. The importance of and, at the same time, the challenges raised by these lexical units for processing natural languages have been previously discussed (see, among many others, (Sag et al., 2002), (Baldwin and Kim, 2010)). Moreover, the impact of MWEs resources on the MWEs recognition in texts was proven by RiedlBiemann, : "In the case that high quality MWE resources exist, these should be used. If not, it is possible to replace them with unsupervised extraction methods". Savary et al. (2019) are also in favour of the creation of language resources containing MWEs, as many and diverse as possible; their presence in resources available for training systems for MWE identification being more important than their frequency (in annotated corpora). The results obtained in the annotation of the VMWEs in the RoWN are presented, as well as a a comparison with those obtained by annotating a news corpus with these types of VMWEs is drawn, showing that the distribution of types and their frequencies at the lexicon level are different from those at the corpus level. As further work, we envisage adding information about prepositional restrictions of the verbs in RoWN. This was another type of VMWEs in PARSEME, but annotating it was optional and we neglected it. The data annotated as presented here have been compared and discussed with the Bulgarian data, as the wordnets for both these languages have been annotated with VMWEs (Barbu Mititelu et al., 2019).

6 Acknowledgements

Part of the work reported here has been carried out within the Multilingual Resources for CEF.AT in the legal domain – MARCELL Action (http://marcell-project.eu/). Another part has been undertaken under the bilateral project *Enhancing Multilingual Language Resources with Derivationally Linked Multiword Expressions* (2018–2020) between the Institute for Bulgarian Language at the Bulgarian Academy of Sciences and the Research Institute for Artificial Intelligence at the Romanian Academy. The authors are grateful to the three anonymous reviewers for their valuable comments meant to improve the quality of the initially submitted form of this paper.

References

Eneko Agirre, Izaskun Aldezabal and Eli Pociello. 2005. *Lexicalization and Multiword Expressions in the Basque WordNet*. In Proceedings of the 3rd Global WordNet Conference, Jeju Island.

Timothy Baldwin and Su Nam Kim. 2010. *Multiword expressions.* Handbook of Natural Language Processing, Second Edition, 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Verginica Barbu Mititelu, Ivelina Stoyanova, Svetlozara Leseva, Maria Mitrofan, Tsvetana Dimitrova and Maria Todorova. 2019. *Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth.* Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), ACL, 2–12.

Luisa Bentivogli and Emanuele Pianta. 2004. *Extending WordNet with Syntagmatic Information*. In Proceedings of the 2nd Global Wordnet Conference (GWC 04), Czech Republic, 47-–53.

Christiane Fellbaum. 1998, ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

George A. Miller. 1995. *WordNet: A Lexical Database for English.* Communications of the ACM Vol. 38, No. 11: 39–41.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya and Abigail Walsh. 2018. *Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.* The Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), 222–240.

Martin Riedl and Chris Biemann. 2016. *Impact of MWE Resources on Multiword Recognition.* Proceedings of the 12th Workshop on Multiword Expressions (MWE 2016), ACL, 107–111

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. *Multiword Expressions: A Pain in the Neck for NLP.* In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), 1–15, Mexico City, Mexico.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova and Antoine Doucet. 2017. *The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.* Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), ACL, 31-47.

Agata Savary, Silvio Cordeiro and Carlos Ramisch. 2019. *Without lexicons, multiword expression identication will never fly: A position statement.* Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), ACL, 79–91.

Dan Tufiș, Dan Cristea and Sofia Stamou. 2004. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.* Journal on Information Science and Technology, Special Issue on BalkaNet, Romanian Academy, 7 (1-2), 7–41.

Veronika Vincze, Attila Almási and Janos Csirik. 2012. *Multiword verbs InWordNets*. In Proceedings of the 6th International Global Wordnet Conference, 337–381.

Piek Vossen. 2002. *EuroWordNet general document version 3.* Report, University of Amsterdam.

Thesaurus Verification Based on Distributional Similarities

Natalia Loukachevitch
Lomonosov Moscow State University
Moscow, Russia
Louk_nat@mail.ru

Ekaterina Parkhomenko
Lomonosov Moscow State University
Moscow, Russia
parkat13@yandex.ru

Abstract

In this paper we consider an approach to verification of large lexical-semantic resources as WordNet. The method of verification procedure is based on the analysis of discrepancies of corpus-based and thesaurus-based word similarities. We calculated such word similarities on the basis of a Russian news collection and Russian wordnet (RuWordNet). We applied the procedure to more than 30 thousand words and found some serious errors in word sense description, including incorrect or absent relations or missed main senses of ambiguous words.

1 Introduction

Large lexical-semantic resources such as Princeton WordNet (Fellbaum, 1998) and wordnets created for other languages (Bond and Foster, 2013) are important instruments for natural language processing. Developing and maintaining such resources requires special efforts, because it is difficult to find errors or gaps in structures consisting of thousands lexical units and relations between them.

In previous works, various methods on lexical enrichment of thesauri have been studied (Snow et al., 2006; Navigli and Ponzetto, 2012). But another issue was not practically discussed: how to find mistakes in existing thesaurus descriptions: incorrect relations or missed significant senses of ambiguous words, which were not included accidentally or appeared recently.

In fact, it is much more difficult to reveal missed and novel senses or wrong relations, if compared to novel words (Frermann and Lapata, 2016; Lau et al., 2014). So it is known that such missed senses are often found during semantic annotation of a corpus and this is an additional problem for such annotation (Snyder, Palmer, 2004; Bond, Wang, 2014).

In this paper, we consider an approach how to use embedding models to reveal problems in a thesaurus. Previously, distributional and embedding methods were evaluated in comparison with manual data (Baroni and Lenci, 2011; Panchenko et al., 2016). But we can use them in the opposite way: to utilize embedding-based similarities and try to detect some problems in a thesaurus.

We study such similarities for more than 30 thousand words presented in Russian wordnet RuWordNet (Loukachevitch et al., 2018[1]).

The structure of the paper is as follows. Section 2 is devoted to related work. In Section 3 we briefly present RuWordNet. Section 4 describes the procedure of calculating two types of word similarities based on thesaurus and a corpus. In Section 5 we analyze discrepancies between thesaurus-based and corpus-based word similarities, which can appear because of different reasons. In Section 6 we study groupings of distributionally similar words to an initial word using the thesaurus.

2 Related Work

In (Lau et al. 2014), the task of finding unattested senses in a dictionary is studied. At first, they apply the method of word sense induction based on LDA topic modeling. Each extracted sense is represented to top-N words in the constructed topics. To compute the similarity between a sense and a topic, the words in the definition are converted into the probability distribution. Then two probability distributions (gloss-based and topic-based) are compared using the Jensen-Shannon divergence. It was found that the proposed novelty measure could identify target lemmas with high- and medium-frequency novel senses. But the authors evaluated their method using word sense definitions in the Macmillan

[1] http://ruwordnet.ru/en/

dictionary and did not check the quality of relations presented in a thesaurus.

A series of works was devoted to studies of semantic changes in word senses (Gulordava and Baroni, 2011; Mitra et al., 2015; Frermann, Lapata, 2016) . Gulordava and Baroni (2011) study semantic change of words using Google n-gram corpus. They compared frequencies and distributional models based on word bigrams in 60s and 90s. They found that significant growth in frequency often reveals the appearance of a novel sense. Also it was found that sometimes the senses of words do not change but the context of their use changed significantly. For example, the context of word *parent* considerably change in 90s because of the most frequent collocation *single parent family*.

In (Mitra et al., 2015), the authors study the detection of word sense changes by analyzing digitized books archives. They constructed networks based on a distributional thesaurus over eight different time windows, clustered these networks and compared these clusters to identify the emergence of novel senses. The performance of the method has been evaluated manually as well as by comparison with WordNet and a list of slang words. But Mitra et al. did not check if WordNet misses some senses.

The task of revising and verifying of resources is important for developers of WordNet-like resources. Some ontological tools have been proposed to check consistency of relations in WordNet (Guarino and Welty, 2004; Alvez et al., 2018).

Some authors report about revision of mistakes and inconsistencies in their wordnets in the process of linking the wordnet and English WordNet (Cristea et al., 2004; Rudnicka et al., 2012). Rambousek et al. (2018) consider a crowdsourcing tool allowing a user of Czech wordnet to report errors. Users may propose an update of any data value. These suggestions can be approved or rejected by editors. Also visualization tools can help to find problems in wordnets (Piasecki et al. 2013; Johannsen et al., 2011).

Loukachevitch (2019) proposed to use embedding-based word similarities to find possible mistakes or inconsistencies in a WordNet-like thesaurus. In the current paper we provide some additional details for the (Loukachevitch, 2019) study.

3 RuWordNet

RuWordNet was created on the basis of another Russian thesaurus RuThes in 2016, which was developed as a tool for natural language processing during more than 20 years (Loukachevitch and Dobrov, 2002). Currently, the published version of RuWordNet includes 110 thousand Russian words and expressions.

The important feature of RuWordNet (and its source RuThes), which is essential for this study, is that a current news collection is used as a reference collection for maintenance of RuWordNet. Periodically, a new corpus (of last year news articles) is collected, single words and phrases absent in the current version of the thesaurus are extracted and analyzed for inclusion to the thesaurus (Loukachevitch, Parkhomenko, 2018). The monitoring of news flow is important because news articles concern many topics discussed in the current society, mention new terms and phenomena recently appeared.

The current version of RuWordNet comprises the following types of relations: hyponym-hypernym, antonyms, domain relations for all parts of speech (nouns, verbs, and adjectives); part-whole relations for nouns; cause and entailment relations for verbs. Synsets of different parts of speech are connected with relations of POS-synonymy. For single words with the same roots, derivational relations are described. For phrases included in RuWordNet, relations to component synsets are given.

4 Comparison of Distributional and Thesaurus Similarities

To compare distributional and thesaurus similarities for Russian according to RuWordNet, we used a collection of 1 million news articles as a reference collection. The collection was lemmatized. For our study, we took thesaurus words with frequency more than 100 in the corpus. We obtained 32,596 words (nouns, adjectives, and verbs).

Now we should determine what thesaurus relations or paths are taken to determine semantically similar entries. In the current study, we consider the following entries as semantically related to the initial thesaurus entry:

- its synonyms,
- all the entries located in the 3-relation paths, consisting of hyponym-hypernyms

relations or/and part-whole relations between synsets from the initial entry;

- all the entries linked with other direct relations to the initial entry;
- for ambiguous words, all sense-related paths were considered and thesaurus entries along these paths were collected together.

In such a way, for each word, we collected the thesaurus-based "bag" of similar words (TBag).

Then we calculated embeddings according to word2vec model with the context window of 3 words, planning to study paradigmatic relations (synonyms, hypernyms, hyponyms, co-hyponyms). Using this model, we extracted twenty the most similar words w_i to the initial word w_0. Each w_i should also be from the thesaurus. In such a way, we obtained the distributional (word2vec) "bag" of similar words for w_0 (DBag).

Now we can calculate the intersection between TBag and DBag and sum up the similarities in the intersection. Figure 1 shows the distribution of words according to the similarity score of the TBag-DBag intersection. The axis X denotes the total similarity in the TBag-DBag intersection: it can achieve more than 17 for some words, denoting high correspondence between corpus-based and thesaurus-based similarities.

Relative adjectives corresponding to geographical names have the highest similarity values in the TBag-DBag intersection, for example, *samarskii* (related to Samara city), *vologodskii* (related to Vologda city), etc. Also nouns denoting cities, citizens, nationalities, nations have very high similarity value in the TBag-DBag intersection.

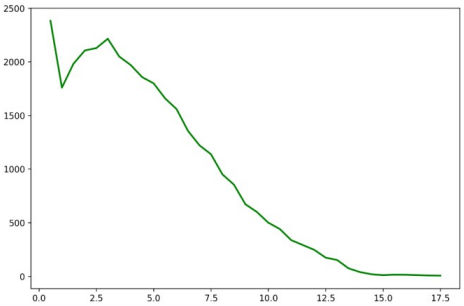


Figure 1. Distribution of numbers of thesaurus words according to total similarity in TBag-DBaf intersection

Among verbs, verbs of thinking, movement (*to drive - to fly*), informing (*to say - to inform - to warn - to assert*), value changing (*to decrease - to increase*), belonging to large semantic fields, have the highest similarity values (more than 13).

For example, according to the word2vec model, word *сказать* (*to say*) is most similar to such words as: *подчеркнуть* (*to stress*) 0.815, *заявить* (*to announce*) 0. 81, *добавить* (*to add*) 0.80, *заметить* (*to notice*) 0.79 .. And all these words are in TBag of this word in RuWordNet

On the other hand, the rise of the curve in low similarity values demonstrates the segment of problematic words.

5 Analyzing Discrepancies between Distributional and Thesaurus Similarities

We are interested in cases when the TBag-DBag intersection is absent or contains only 1 word with small word2vec similarity (less than the threshold (0.5)). We consider such a difference in the similarity bags as a problem, which should be explained.

For example, *троянец* (*troyanets*) is described in the thesaurus as a citizen of ancient Troya with the corresponding relations. But in the current texts, this word means a kind of malicious software (*troyan horse program*), this sense of the word was absent in the thesaurus. We can see that Dbag of word *троянец* contains:

вредоносный (*malicious)* 0.76, *программа* (*program*) 0.73, *троянский* (*trojan*) 0.71, ...*вирус* (*virus*) 0.61,...

This means that the DBag and TBag are completely different, Dbag of word *троянец* does not contain anything related to computers and software.

We obtained 2343 such problematic "words". Table 1 shows the distribution of these words according to the part of speech.

It can be seen that verbs have a very low share in this group of problematic words. It can be explained that in Russian, most verbs have two aspect forms (Perfective and Imperfective) and also frequently have sense-related reflexive verbs. All these verb variants (perfective, imperfective, reflexive) are presented as different entries in RuWordNet.

Therefore, in most cases altogether they should easily overcome the established threshold of discrepancies. In the same time, if some verbs are

found in the list of problematic words, they have real problems of their description in the thesaurus.

Part of speech	Number
Nouns	1240
Adjectives	877
Verbs	226
Total	2343

Table 1. Distribution of parts of speech among problematic words

To classify the causes of discrepancies, we ordered the list of problematic words in decreasing similarity of their first most similar word from the thesaurus, that is in the beginning words with the most discrepancies are gathered (further, Problem List). In the subsections, we consider specific reasons, which can explain discrepancies between thesaurus and corpus-based similarities.

5.1 Morphological Ambiguity and Misprints

The most evident source of the discrepancies is morphological ambiguity when two different words w_1 and w_2 have the same wordform and words from DBag of w_1 in fact are semantically related to w_2 (usually w_2 has larger frequency). For example, in Russian there are two words *bank* (financial organization) and *banka* (a kind of container). All similar words from Dbag to *banka* are from the financial domain: *gosbank* (*state bank*), *sberbank* (*saving bank*), *bankir* (*banker*), etc. The analyzed list of problematic words includes about 90 such words.

The technical reason of some discrepancies are frequent misprints. For example, frequent Russian word *заявить* (*zayavit* - to proclaim) is often erroneously written as *завить* (*zavit* - to curl). Therefore the DBag of word *zavit* includes many words similar to *zayavit* such as *сообщить* (*to inform*), or *otmetit* (*to remark*). Another example are words *statistka (showgirl)* and *statistika (statistics)*.

5.2 Named Entities and Multiword Expressions

The natural reason of discrepancies are named entities, which names coincide with ordinary words, they are not described in the thesaurus, and are frequent in the corpus under analysis. For example, *мистраль* (*mistral*) is described in RuWordNet as a specific wind, but in the current corpus French helicopter carrier Mistral is actively discussed.

Frequent examples of such named entities are names of football, hockey and other teams popular in Russia coinciding with ordinary Russian words or geographical names (*Zenith, Dynamo*, etc.). Some teams can have nicknames, which are written with lowercase letters in Russian and cannot be revealed as named entities, for example Russian word *ириска* (*iriska*) means a kind of candy. In the same time, it is nickname of Everton Football Club (*The Toffees*).

Word	The most frequent phrase	Phrase Freq. (Total freq.	Most similar word according to the corpus with frequency
Топленый (adj) (toplenyi – rendered)	Топленое масло (toplenoe maslo - rendered butter)	78 (112)	Миндальный (adj) (mindalnyi – adjective from миндаль (almond)) 180 Mindalnoe maslo (almond oil) 57
Размочить (verb) (razmochit' – to open (the score))	Размочить счет (razmochit' schet – to open the score)	183 (336)	Сравнять (verb) (sravnyat' – equalize) 6678 Сравнять счет (to equalize the score) 5294
Капитальный (adj) (kapitalnyi – capital)	Капитальный ремонт (kapitalnii remont – major repair)	12015 (17985)	Капремонт (noun) (kapremont – abbreviation from kapitalnii remont – major repair) 3504
Заварной (adj) (zavarnoi – boiled)	Заварной крем (zavarnoi krem – custard)	37 (126)	Тыквенный (adj) (tykvennyi – adjective from тыква (pumpkin) 175 Тыквенные семечки (pumpkin seeds) 15
Порывистый (adj) (poryvistii)-	Порывистый ветер (poryvistii veter – rough wind)	1176 (1512)	Метель (noun) (metel' – blizzard) 7479

Table. 3 Impact of multiword expressions on discrepancies between the thesaurus and corpus-based data

Some discrepancies can be based on frequent multiword expressions, which can be present or absent in the thesaurus. A component w_1 of multiword expression w_2 can be distributionally similar to other words frequently met with w_2 or it can be similar to words related to the whole phrase $w_1 w_2$.

It can be noted that if a word w_1 occurs in a phrase w_1w_2 more than half times (the order of components can be different), it can become distributionally similar to w_2 or w_3, which also often met in phrase w_3w_2, *even if* w_1 and w_3 are not similar in sense. Table 3 shows examples of similarity discrepancies, which seems to be explained with frequent co-occurrence in a specific phrase.

For example, word *топленый* (*toplenyi – rendered*) occurs in the phrase топленое масло (*toplenoe maslo* – rendered butter) 78 times of 112 of its total frequency. Because of this, this word is the most similar to word *миндальный* (*mindalnyi* – adjective to almond), which is met in the phrase *миндальное масло* (*mindalnoe maslo* – almond oil) 57 of 180 times. But two words *топленый* и *миндальный* cannot be considered as sense-related words.

5.3 Thesaurus Relations

In some cases, the idea of distributional similarity is clear, but the revision cannot be made the thesaurus. We found two types of such cases. First, such epithet as *гигант* (*giant*) in the current corpus is applied mainly to large companies (*IT-giant, cosmetics giant, technological giant,* etc.). But it can be strange to provide the relations between words *giant* and *company* in a thesaurus.

The second case can be seen on the similarity row to word *массажистка* (*women massager*), comprising such words as hairdresser, housekeeper, etc. This is a kind of specialists in specific personal services but it seems that an appropriate word does not exist in Russian to create a more detailed classification of such specialists.

Another interesting example of a similarity grouping is the group of "flaws in the appearance": word *целлюлит* (*cellulite*)[2] is most similar to words: *морщина* (*crease of the skin*), *перхоть* (*dandruff*), *кариес* (*dental caries*), *облысение* (*balding*), *веснушки* (*freckles*). It can be noted that a bald head or freckles are not necessary flaws of a specific person, but on average they are considered as flaws. On the other hand, such phrases as *недостатки внешности, недостатки внешнего вида* (*flaws in the appearance*) are quite frequent in Internet pages according to global search engines, therefore maybe it could be useful to introduce the corresponding concept for correct describing the conceptual system of the modern personality.

But also real problems of thesaurus descriptions were found. They included word relations, which could be presented more accurate. For example, word *тамада* (*tamada - toastmaster*) was linked to more general word, not to *ведущий* (*veduschii - master of ceremonies*).

5.4 Senses Unattested in Thesaurus

Also significant missed senses including serious errors for verbs were found. As it was mentioned before, in Russian there are groups of related verbs: perfective, imperfective, and reflexive. These verbs usually have a set of related senses, and also can have their own separate senses. In the comparison of discrepancies between TBag and Dbag of verbs, it was found that at least for 25 verbs some of senses were unattested in the current version of the thesaurus, which can be considered as evident mistakes. For example, the imperfective sense of verb *otpravlyatsya* (*depart*) was not presented in the thesaurus.

Several dozens of novel senses, which are the most frequent senses in the current collection, were identified. Most such senses are jargon (sports or journalism) senses, i.e. *дерби* (derby as a game between main regional teams) or *навес* as a type of a pass in football (*high-cross pass*). Also several novel senses that belong to information technologies were detected: *прошивка* (*proshivka - firmware*), *соцсеть* (abbreviation from *социальная сеть* (*social network*).

The modern news discourse allows using words and expressions of the colloquial register (Patrona, 2011; Busa, 2013). In our analysis, several colloquial (but well-known) word senses absent in RuWordNet were found. For example, verb *обжечься* (*obzech'sya*) in the main sense means 'burn oneself'. In Dbag the colloquial sense 'make a mistake' is clearly seen.

For word *корректор* (*corrector*), two most frequent unattested senses were found: cosmetic corrector and correction fluid. The Dbag of this word looks as a mixture of cosmetics and stationary terms: *гуашь* (*gouache*), *кисточка* (*tassel*),

[2] https://en.wikipedia.org/wiki/Cellulite

тональный (*tonal*), *чернила* (*ink*), *типографский* (*typographic*), etc.

Currently, about 90 evident missed senses (different from named entities), which are most frequent senses of words in the collection, are identified from the analysis of the differences in two similarity lists .

5.5 Other cases

In some cases, paths longer than 3 should be used to provide better correspondence between thesaurus-based and corpus-based similar words.

Besides, the collected news corpus contains some number of Ukrainian texts, which are also written in the Cyrillic alphabet. Some Russian words coincide with Ukrainian words but have different senses and contexts in texts. Therefore, distributional similarities of such words are very different from the Russian thesaurus similarities.

6 Searching for regularities in Dbags

We supposed that we can group words in the corpus-based set of similar words (DBag) of problematic words using synonyms and part-of-speech synonyms of RuWordNet.

In such a way we can find more clear indications to some missed relations or novel senses. We have gathered synonyms, summed up their similarity scores to the target word, and again reordered list according to the descending order of the maximum similarity in DBag. For example, we obtained for word *рассекать (to cut* in the thesaurus sense) the maximum similarity score 3.58 with the following group of words: *мчаться, промчаться, пронестись, нестись, носиться* (*rush, race, hasten*). And this is the clear indication of the novel sense of this word absent in the thesaurus.

At the same time we obtained for word *длинноногий* (*long-legged*) the following most similar group *белокурый светловолосый блондинистый* (*blond, blonde, light-haired*). There is no semantic similarity between words *длинноногий* (*long-legged*) and *светловолосый* (*light-haired*) but there frequent co-occurrence and occurrence with the same nouns (*девушка, красавица, красотка - girl, beauty*) generate such similarity values.

It is also evident, that word *кроссворд* (*crossword*) is distributionally similar to group *разгадывание, разгадывать, отгадывание* (*guess, guessing, solve*) (score 1.51) only because of their frequent co-occurrence.

From this experiment, we can conclude that trying to extract some novel senses or missed relations on the basis of corpus-based embeddings, it is important to account for the diversity of contexts and co-occurrence of words predicted to be related. Low diversity of frequent contexts and significant co-occurrence can lead to erroneous conclusion on word semantic similarity.

7 Conclusion

In this paper we discuss the usefulness of applying a checking procedure to existing thesauri. The procedure is based on the analysis of discrepancies between corpus-based and thesaurus-based word similarities. We applied the procedure to more than 30 thousand words of Russian wordnet RuWordNet, classified sources of differences between word similarities and found several dozens of serious errors in word sense description including too general relations, missed relations or unattested main senses of ambiguous words. It is impossible to find such diverse problems in short time without automatic support.

We highly recommend to use this procedure for checking wordnets - it is possible to find a lot of unexpected knowledge about the language and the thesaurus.

In future, we plan to develop an automatic procedure of finding thesaurus regularities in DBag of problematic words, which can make more evident what kind of relations or senses are missed in the thesaurus.

Acknowledgments

The reported study was funded by RFBR according to the research project N 18-00-01226 (18-00-01240).

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations* Association for Computational Linguistics: 7-12.

Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. Cross-checking WordNet and SUMO using meronymy. 2018. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In

Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, Edinburgh, Scotland, pages 1–11.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages. 1352-1362.

Francis Bond and Shan Wang. 2014. Issues in building English-Chinese parallel corpora with WordNets. In *Proceedings of the Seventh Global Wordnet Conference:* 391-399.

M.Grazia Busa. Introducing the language of the news: a student's guide. – Routledge, 2013.

Paul Cook and Graeme Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*.

Cristea, D., Mihaila, C., Forascu, C., Trandabat, D., Husarciuc, M., Haja, G., & Postolache, O. (2004). Mapping princeton WordNet synsets onto Romanian WordNet synsets. *Romanian Journal of Information Science and Technology*, *7*(1-2), 125-145.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Lea Frermann and Mirella Lapata. 2016. Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistic*s. V. 4. pages 31-45.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics: 67-71.

Nicola Guarino, and Christopher A. Welty. 2004. An overview of OntoClean. *Handbook on ontologies*. Springer, Berlin, Heidelberg: 151-171.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers):* 259-270.

Anders Johannsen, and Bolette Sandford Pedersen. “Andre ord”–a wordnet browser for the Danish wordnet, DanNet. *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*. 2011.

Natalia Loukachevitch and Boris Dobrov. 2002. Development and Use of Thesaurus of Russian Language RuThes. In *Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation*.(LREC 2002): 65-70.

Natalia Loukachevitch, German Lashevich and Boris Dobrov, Boris. 2018. Comparing Two Thesaurus Representations for Russian. In *Proceedings of Global WordNet Conference GWC-2018*, pages 35-44.

Natalia Loukachevitch and Ekaterina Parkhomenko. 2018. Recognition of Multiword Expressions Using Word Embeddings." *Russian Conference on Artificial Intelligence*. Springer, Cham, pages 112-124.

Natalia Loukachevitch. 2019. Corpus-based Check-up for Thesaurus. In Proceedings of ACL-2019: 5773-5779.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, *21*(5), 773-798.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*. V. 41, №. 2, pages 10.

Roberto Navigli and Simone Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pages 217-250.

Alexander Panchenko, Anastasiya Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 2018. RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In *Proceedings of Intern. conference Dialogue-2018*, pages 547--564.

Marianna Patrona. 2011. When journalists set new rules in political news discourse. Talking politics in broadcast media: *Cross-cultural perspectives on political interviewing, journalism and accountability*, 42, 157.

Maciej Piasecki, Michal Marcińczuk, Radoslaw and Marek Maziarz. 2013. WordNetLoom: a WordNet development system integrating form-based and graph-based perspectives. *International Journal of Data Mining, Modelling and Management*, *5*(3): 210-232.

Adam Rambousek, Ales Horak, and Karel Pala. 2018. Sustainable long-term WordNet development and maintenance: Case study of the Czech WordNet. *Cognitive Studies,* 18.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz 2012. A strategy of Mapping Polish WordNet onto Princeton WordNet. In Proceedings of COLING 2012, pages 1039-1048.

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Association for Computational Linguistics: 801-808.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. *In Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

Including Swiss Standard German in GermaNet

Eva Huber, Erhard Hinrichs
University of Tübingen
eva.huber@student.uni-tuebingen.de
erhard.hinrichs@uni-tuebingen.de

Abstract

GermaNet (Henrich and Hinrichs, 2010; Hamp and Feldweg, 1997) is a comprehensive wordnet of Standard German spoken in the Federal Republic of Germany. The GermaNet team aims at modelling the basic vocabulary of the language. German is an official language or a minority language in many countries. It is an official language in Austria, Germany and Switzerland, each with its own codified standard variety (Auer, 2014, p. 21), and also in Belgium, Liechtenstein, and Luxemburg. German is recognized as a minority language in thirteen additional countries, including Brasil, Italy, Poland, and Russia. However, the different standard varieties of German are currently not represented in GermaNet. With this project, we make a start on changing this by including one variety, namely Swiss Standard German, into GermaNet. This shall give a more inclusive perspective on the German language. We will argue that Swiss Standard German words, Helvetisms, are best included into the already existing wordnet GermaNet, rather than creating them as a separate wordnet.

1 Introduction

GermaNet is a comprehensive wordnet of Standard German spoken in the Federal Republic of Germany. German is an official language or a minority language in many countries. It is an official language in Austria, Germany and Switzerland, each with its own codified standard variety (Auer, 2014, p. 21), and also in Belgium, Liechtenstein, and Luxemburg. German is recognized as a minority language in thirteen additional countries, including Brasil, Italy, Poland, and Russia. However, the different standard varieties of German are currently not represented in GermaNet. More generally, among wordnets, there seems to be a lack of accounting for different standards of the same language. To the best of our knowledge, the Princeton WordNet (Fellbaum, 1998) is the only wordnet so far which accounts for standard varities by marking specifically American or specifically British words. Moreover, a colloquial wordnet of English has recently been created (McCrae et al., 2017). Therefore, it seems worthwhile integrating other German varieties into GermaNet. The central question to this paper, therefore, is how we can successfully model standard varieties. The present study focuses on Swiss Standard German (Swiss StdG). Swiss StdG differs from German on all linguistic levels (Dürscheid and Sutter, 2014, p.37). An orthographic difference pertains to the Eszett ß ("sharp S"), which is in all cases replaced by ss in Swiss StdG (Dürscheid and Sutter, 2014). There are also remarkable phonological differences, such as the primary stress of the initial syllable in, for instance, *Büffet* (Clyne, 1984, p.16). Grammar differences are also found in word order, gender differences, and word derivation patterns. However, lexical differences are by far the most frequent (Dürscheid and Sutter, 2014). At a train station, Swiss people buy a *Billet* (German variant: *Fahrschein*; "ticket") which they then show to a *Kondukteur* (*Schaffner*, "conductor") in the *Erstklasswagen* (*Wagen der ersten Klasse*; "first class carriage"). Since wordnets consist of lexemes, we are concerned with the lexical differences. As is common in the literature, we will refer to words which are idiosyncratic for Swiss StdG as *Helvetisms* and to those idiosyncratic for German StdG as *Teutonisms*.

Our approach shall attain a broader representation of German in wordnets and offer a framework for other languages, of which different standard varieties exist, such as Portuguese, Swedish

or French. The paper is structured as follows. First, we will give an overview of GermaNet (Section 2). In Section 3, we will demonstrate how words of Swiss StdG can be collected from lexicographic sources (Section 3.1) and by corpus-based methods (Section 3.2). Section 3.3 presents characteristic examples of Swiss StdG words that have been harvested from lexicographic and corpus-based sources. Section 4 suggests a framework of how to integrate Swiss Standard German . We conclude by discussing possible future work with regard to German varieties (Section 5).

2 GermaNet

GermaNet is a lexical semantic network that is modelled after the Princeton WordNet for English. The resource has been under development for more than twenty years and is still being extended on a continuous basis. The GermaNet team aims at constructing a lexical resource in digital form that models the basic vocabulary of the language. GermaNet covers the most frequently used German adjectives, nouns, and verbs. The coverage of GermaNet is determined by frequency lists compiled from very large digital text corpora of contemporary German. The current data release 13.0 of GermaNet contains 128,100 synsets, 164,814 lexical units, and 148,929 literals. In addition to the inventory of lexical and conceptual relations used in the Princeton WordNet, GermaNet contains a set of lexical relations for nominal compounds. These relations indicate the semantic relations that hold between the constituent parts of a compound. Compounds are also morphologically decomposed into their constituent parts. Release 13.0 contains a total of 82,309 compounds that have been decomposed in this way (Hinrichs et al., 2013).

The coverage of GermaNet is by and large restricted to Standard German. Regional variants and colloquial terms are included only to the extent that they occur frequently in large text corpora and are widely understood. The concept ”bread roll” is expressed in Standard German by the lemma *Brötchen* and has many regional variants. One such variant is the term *Wecken*, which is included in GermaNet. *Wecken* belongs to Southern dialects of Germany, but its meaning is widely known, and it occurs with considerable frequency in German corpus data. Therefore, it is reasonable to include such a variant in GermaNet. Compared to regional variants, colloquial words are included in GermaNet to a higher degree as long as their usage is stable over an extended period of time and as long as they are not offensive.

GermaNet is also linked to the Interlingual Index (ILI; Vossen 1998) that is used to link wordnets for different languages. The synsets for current release of the GermaNet records can be linked to the ILI via 28,566 ILI records. The lexical units in GermaNet can also be linked to a total of 29,550 Wiktionary sense descriptions (Henrich et al., 2014).

3 Detecting and Describing Helvetisms

Switzerland distinguishes itself from Austria and Germany in the sense that Swiss StdG is in a diglossic relationship with the Swiss dialects. While Swiss German dialects, so called *Mundarten*, are used in everyday communication, Swiss StdG occurs in written texts and in news media (Clyne, 1992, p. 119). The Swiss German dialects align themselves with canton boundaries and are acquired as children's first language. Swiss StdG is acquired only once children enter grade school. It is also worth noting that the German Alemannic dialects form a continuum that straddles the German and Swiss border. While it would be worthwhile to include regional varieties of both Germany and Switzerland, this project limits itself to the standard varieties only. In this section, we will discuss how relevant Swiss StdG words can be acquired by lexicographic resources and by data-driven methods.

3.1 Lexicographic Resources

The dictionary “Duden” is the common reference book for the German language, aiming at a full representation of the language [1](Duden, 2017). The “Schweizerhochdeutsche Duden” (Swiss High German Duden), however, merely lists specific Swiss StdG terms (Bickel and Landolt, 2018). Additionally, the German Duden marks typically *schweizerisch* (”Swiss”) or *österreichisch* (”Austrian”) words, while Teutonisms, such as *Tesafilm* (“sellotape”) are not marked. The German Duden allows for a detection of words which are present in Switzerland as well as in Southern Germany. For instance, the usage of *Nastuch* (”handkerchief”) is entered as *süddeutsch, schweizerisch*. Furthermore, the

[1]https://www.duden.de/

Swiss High German Duden specifies *mundartnahe* words, i.e., words derived from Swiss dialects. Thus, both of these reference works make the gradual characteristics of Swiss StdG to the Mundarten and to German StdG, to a certain degree, explicit. Lexicographic resources offer a valuable data set of words to include in a wordnet. However, some words listed in lexicographic resources are no longer widely used or are used only in certain regions. We, therefore, also consult data-driven methods, which will be described in the following section.

3.2 Data-Driven Methods

Word lists were obtained from two different data sources: The German and the Swiss section of the Leipziger Wortschatz Corpus Collection and news crawls for German and Swiss online materials. The Leipziger Wortschatz Corpus was data-mined by Schneider (Schneider, 2018) using a document classification technique. This method yielded a word list of 21,788 lemmas of all parts-of-speech for which the corpus was tagged. Each lemma was accompanied by a score that indicated the degree to which a word belongs to one standard variety or the other, or whether the word is likely to occur in both varieties. Since the document classification technique does not control for frequency, we also used a frequency-based approach that was facilitated by the frequency lists for the Swiss and German section that are made available along with the Leipziger Wortschatz data. Both frequency lists were truncated to obey a frequency threshold of 50 occurrences. In order to obtain candidate lemmas for Helvetisms, all lemmas from the German frequency list were eliminated from the Swiss frequency list. The same frequency-based method was also applied to filter frequency lists for the news crawls for German and Swiss online domains.

The word lists obtained by the document classification method and by the frequency-based method need to be manually inspected in order to acquire reliable lexical material for Helvetisms relevant for inclusion in a wordnet. Amongst other things, this also means that the candidate lemmas need to be restricted to the three word classes of nouns, verbs and adjectives. Filtering out the other word classes, we obtained 3,712 lemmas of Helvetisms from the Leipziger Corpus and 3,139 from the crawl. The Duden includes approximately 3,500 lemmas. In order to estimate how many of the words are Helvetisms, we analysed samples including 10% of each data set. Based on the analysis of the samples, 57.14% of the Duden, 9.19% of the list of the Leipziger Corpus, and 5.48% of the crawl list are expected to be Helvetisms. Thus, our data set includes approximatly 2,500 Helvetisms, without considering potential overlap between the data set. An analysis of the overlap between the samples of the Leipziger Corpus and the Duden and the crawl list and the Duden respectively shows that the overlap is relatively small. The overlap between the samples from the Leipziger Corpus and the samples from the Duden is 48.6% while the overlap between the samples from the crawl list and the samples from the Duden is merely 11.8%.

3.3 Swiss StdG Words

The Helvetisms that can be harvested from lexicographic resources or from digital corpora fall into different categories (see Lingg 2006; Clyne 1984): words that are derived from the Mundart, loanwords, particularly from French, and culture-specific words pertaining to domains such as politics or sports. The noun *Beiz* ("pub") is one example of a word that is derived from Mundart. It is used interchangeably with the word *Kneipe*, which belongs to the standard varieties spoken in Germany and Switzerland. French loanwords include lemmas such as *Jupe* ("skirt"), which corresponds to German StdG *Rock*. Additionally, Swiss StdG *Papeterie* ("stationary shop") is synonymous to the German StdG *Schreibwarengeschäft*. A further category includes words which are related to Switzerland's culture and tradition, administration and education, and government and political system. Switzerland has special sports, such as *Schwingen*, a kind of wrestling, and *Hornussen*, which obtains its name from a puck called *Hornuss*. Due to the different political systems in Germany and Switzerland, words related to politics are usually specific to its variety. The Swiss political system enables people to propose laws in the form of an *Initiative* ("popular initiative"). Furthermore, *Gegenvorschlag* ("counterproposal") is not as in the German variety merely a "counter proposal", but it is usually used to refer to a suggested alternative to a popular initiative. With regard to Switzerland's education system, we find words, such as *Sportferien* ("winter break") and

Maturitätsprüfung ("final exam").

One phenomenon that cuts across the various categories of Helvetisms is the word formation process of compounding that is as productive in the Swiss StdG variety as it is in other German varieties. Compounds in Swiss StdG can either be composed of two words which are not associated with any particular variety, or they can include one or more Helvetisms. The constituent words of the nominal compounds *Süssgetränke* ("soda"), *Todesschein* ("death certificate") and *Gratiseintritt* ("free admission") are all words that are used in both Swiss and German StdG. Yet, all three compounds are characteristic of Swiss StdG, and have as their German StdG counterparts *Erfrischungsgetränke*, *Totenschein* and *freier Eintritt* respectively. Compounds of Swiss StdG also include loanwords from French, such as *Veloschloss* and *Retourbillet*. In *Veloschloss* the modifier is taken from French, whereas in *Retourbillet* both the head and the modifier are French loanwords.

4 Introducing Swiss StdG into the World of Wordnets

Representing Swiss StdG in a wordnet can be approached in two different ways. In this section, we discuss the two options and illustrate the approach we adopted by specific examples that show how to model Swiss StdG words in a wordnet.

4.1 Two Possible Approaches

The first option is to build a separate wordnet for Swiss StdG and map this new wordnet to the existing GermaNet via the Inter-Lingual-Index (ILI; Vossen 1998). This would generalise the approach taken in EuroWordnet, where several European languages are connected via the ILI. This approach provides a means for systematically linking synonymous and hyponymic words between the two varieties. However, please note that this approach treats Swiss and German StdG as separate languages in the same way as is done in EuroWordnet for, among others, French and German. Such a solution has the following major drawback: it disregards the fact that the vocabulary of Swiss and German StdG is largely overlapping, so that the construction of a separate Swiss wordnet would, to a considerable extent, be redundant with the existing GermaNet in both structure and lexical coverage. Recall that our current estimates for Helvetisms amount to approximately 2,500 lemmas (see 3.2), which is only around 10% of the words present in GermaNet.

The second option is to integrate Swiss StdG words directly into GermaNet. This approach follows the strategy adopted in the Princeton WordNet, where words particular to American and British varieties of English are explicitly marked by means of so-called domain region pointers. These pointers link the lexical units to geographical places. For instance, the word *boot*, which is the British expression for the American *trunk*, is marked with the domain region marker relating the word to the synset [United Kingdom, UK, U.K., Britain, United Kingdom of Great Britain and Northern Ireland, Great Britain]. The introduction of domain region pointers into GermaNet allows the modelling of Helvetisms and Teutonisms by linking them to the synsets of [Helvetien, Schweiz] and [BRD, Bundesrepublik Deutschland, Deutschland] respectively. In this approach, words that are used in both varieties are not linked to either of the two synsets. Note also that such an approach is easily generalisable to additional standard varieties of German, whose variety-specific vocabulary would have to be linked to the appropriate synset of the region in which it is spoken.

4.2 Specific Examples

The Swiss StdG words will be integrated into GermaNet so that they are consistent with the overall structure of GermaNet. The same relations will be used, and the only new addition will be the added regional marker to [Helvetien, Schweiz] or [BRD, Bundesrepublik Deutschland, Deutschland] in order to include the three word categories (nouns, verbs and adjectives) [2].

For the integration of Helvetisms into GermaNet, five different cases need to be observed, which are summarised in table 1. They involve lemmas that are different in both varieties for the same concept (case 1), lemmas that are particular to Swiss StdG or German StdG in addition to synonymous lemmas occurring in both varieties (case 2), and, lastly, lemmas for concepts only used in Swiss or German StdG (case 3). The three different cases are exemplified in tables 2 to 4, and involve in each case different parts-of-speech. The cases in which different lemmas are used for the

[2] As opposed to the Princeton WordNet, GermaNet does not contain adverbs

case	description
case 1	different lemmas for the same concept
case 2	additional lemma in Swiss StdG
	additional lemma in German StdG
case 3	lemma and concept used in Swiss StdG only
	lemma and concept used in German StdG only

Table 1: Case distinction for Swiss and German StdG words

same concept, e.g. "breakfast" (see table 2), are treated as co-hyponyms in GermaNet, and each lexical unit is tagged by the regional markers linking it to Switzerland, e.g. *Morgenessen*, and to Germany, e.g. *Frühstück*. The treatment of case 2 in GermaNet is also straightforward: words that are particular to Swiss StdG, e.g. *Estrich*, and to German StdG, e.g. *Kraftfahrzeug* (see table 3), are introduced as additional lexical units into the relevant synset, e.g. the synset for "car" or "attic", and are tagged by the appropriate regional domain pointer. The other members in the synset, which belong to both varieties, e.g. *Dachboden* and *Auto*, remain untagged. The lemmas that belong to case 3 denote concepts only used in Swiss or German StdG, e.g. *Sechseläuten* (a Swiss spring holiday) and *Mettwurst* (a German sausage) (see table 4). Thus, the synsets which include lemmas of case 3 contain (a) lexical unit(s) that are all tagged by a regional domain pointer.

If one merges the two standard varieties of German spoken in Switzerland and Germany in the way just outlined, which steps does a lexicographer have to follow to enter all words that appear in a list consisting of Swiss StdG words into GermaNet? Such a word list may have been compiled from a lexicographic resource, such as the Swiss StdG Duden, or from a corpus of Swiss StdG texts, such as the data from the Leipziger Corpus. Given the assumption that the new word should be incorporated into the existing structure of GermaNet, lexicographers need to follow a sequence of steps summarized as the flow chart in Figure 1. The first step is to ensure that the word is not already included in GermaNet. If this is the case, the lexicographer determines whether the word is a true Helvetism or not. To make this decision, we rely on native speaker intuition, and also additional sources of information, such as Swiss High German corpora and German High German Corpora, are consulted. If the word, however, is not used in Swiss StdG only, the lexical unit is inserted as a new synset and tagged by the regional pointer to [BRD, Bundesrepublik Deutschland, Deutschland] if it is a Teutonism, else it is left unmarked. If the word is, indeed, a Helvetism, there are two possible next steps: either there is already a synset to which the Helvetism can be added (case 1 or case 2), or a new synset has to be created (case 3). In both cases, the lexical unit is marked with the regional domain pointer, linking it to [Helvetien, Schweiz]. If the Helvetism is inserted into an already existing synset, the other members of the synset have to be checked with respect to whether they are Teutonisms and have to be tagged by the regional domain pointer (case 1), or whether they are used in both varieties and are thus left unmarked (case 2).
Already existing words in GermaNet must be reexamined as to whether they are Helvetisms, Teutonisms or used in both varieties. This does not only concern words on the Swiss word list which are already included in GermaNet, but it applies to all words present in GermaNet.

5 Discussion and future work

In this paper, we have shown how to include Swiss StdG into GermaNet by following the approach taken in the Princeton WordNet for linking words from different standard varieties to regional domain pointers. We have emphasised the need for distinguishing between Swiss Mundarten and Swiss StdG and have limited our modelling to the latter. As data sources, we have consulted both lexicographic sources and corpus material and have shown the relative merits of these two sources. It would be worthwhile to broaden the empirical base for identifying Helvetisms by using other data sources, such as informant studies, a traditional method for collecting data on language varieties, and crowdsourcing, which has already been applied to collect colloquial words in a wordnet context by McCrae et al. (2017).

Once the integration of Helvetisms into GermaNet has reached a stable state, the additional data will be released with the yearly updates of the GermaNet resource. GermaNet can be licensed

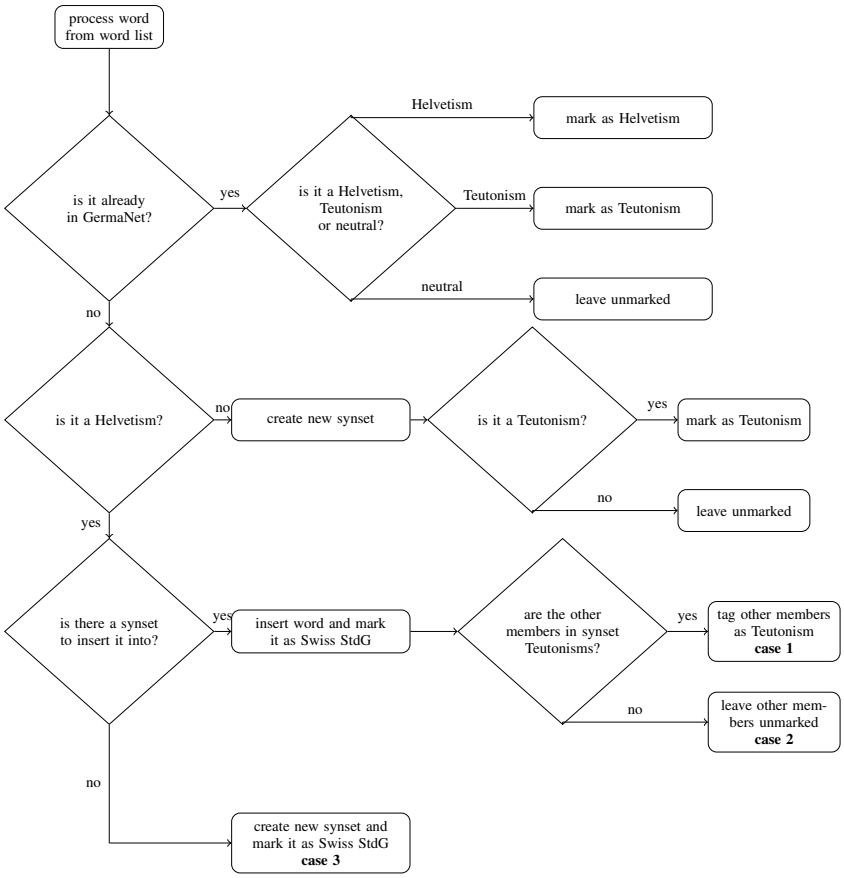


Figure 1: Workflow for lexicographers to include lexemes from the Swiss word list

	example	variety	meaning	part-of-speech
1.	Morgenessen Frühstück	Swiss StdG German StdG	breakfast	noun
2.	parkieren parken	Swiss StdG German StdG	park	verb
3.	Abdankung Trauerfeier	Swiss StdG German StdG	funeral service	noun
4.	Aktion Sonderangebot	Swiss StdG German StdG	bargain offer	noun
rule				

Table 2: Different lemmas in Swiss and German StdG for the same concept (case 1)

	example	variety	meaning	part-of-speech
1.	Beiz Kneipe	Swiss StdG Swiss StdG and German StdG	breakfast	noun
2.	Estrich Dachboden	Swiss StdG Swiss StdG and German StdG	attic	noun
3.	gehäuselt kariert	Swiss StdG Swiss StdG and German StdG	chequered	adjective
4.	überrissen übertrieben	Swiss StdG Swiss StdG and German StdG	excessive	adjective
5.	Kraftfahrzeug Auto	German StdG Swiss StdG and German StdG	car	noun
6.	artig brav	German StdG Swiss StdG and German StdG	well-behaved	adjective
7.	lauschen hinhören	German StdG Swiss StdG and German StdG	eavesdrop	verb
8.	schmuck dekorativ	German StdG Swiss StdG and German StdG	decorative	adjective

Table 3: Additional lemma in Swiss StdG (1-4) and German StdG (5-8) (case 2)

by academic institutions for research purposes free of charge. Non-academic institutions can license GermaNet for the purpose of internal research and development or for the development of commercial products or services.

A natural next step would be to extend the current approach to other standard varieties, such as the standard varieties spoken in Lichtenstein and Austria. These two countries are of particular interest since both border with Switzerland, and Austria also borders with Germany. Another variety of German worthwhile studying is the German spoken in Luxembourg, a country with Letzeburgisch, German and French as the three official languages. Letzeburgisch has been officially recognised as an independent language, but historically has been influenced by Dutch, French and German.

Another issue that we have only touched upon briefly in this paper is the modelling of regional varieties, such as the Swiss Mundarten or regional varieties spoken in Germany. It would be interesting to explore to what extent the approach taken in the Princeton WordNet and also in this paper to the treatment of standard varieties could be generalised to the treatment of regional varieties as well. Here, we can only give some examples from different regional varieties of Switzerland in or-

	example	variety	meaning	part-of-speech
1.	Ausgang -	Swiss StdG German StdG	nightlife	noun
2.	Gegenvorschlag -	Swiss StdG German StdG	counterproposal (in the context of a referendum)	noun
3.	strahlen -	Swiss StdG German StdG	to look for mountain crystals	verb
4.	Sechseläuten -	Swiss StdG German StdG	traditional spring holiday	noun
5.	- Mettwurst	Swiss StdG German StdG	German sausage	noun
6.	- Autohaus	Swiss StdG German StdG	car dealer	noun
7.	- Jahresurlaub	Swiss StdG German StdG	annual holiday	noun
8.	- dufte	Swiss StdG German StdG	smashing	adjective

Table 4: Lemma and concept used in Swiss StdG only (1-4) or in German StdG only (5-8) (case 3)

der to sketch what such an extension would look like. In Swiss Mundarten, the German and Swiss StdG verb *weinen* ("to cry") has the two variants *briäggä* and *brüele* in the dialect spoken in the canton of Zurich and *gränna* is the variant used in the canton of Berne. Similarly, the noun *Brötchen* ("bread roll") has the Mundarten variants *Weggli* used in the canton of Zurich, *Mütschli* in the canton of Berne and *Schwööbli* in the canton of Basel. Modelling such variants in GermaNet would mean to include the variants, e.g. *Weggli*, *Mütschli* and *Schwööbli* or *gränna*, *briäggä* and *brüele*, in one synset that also contains the lexical unit *Brötchen* used in Standard German. The regional variants are then linked to the appropriate domain pointers for the Swiss cantons, while the lexeme *Brötchen* remains unmarked.

Acknowledgments

We thank Reinhild Barkey, Çağrı Çöltekin and Christiane Fellbaum for providing insight and expertise from which this project has greatly benefitted. Furthermore, we gratefully acknowledge the financial support of our research by the German Ministry for Education and Research (BMBF) as part of the CLARIN-D research infrastructure grant given to the University of Tübingen.

References

Peter Auer. 2014. Enregistering pluricentric German. In Augusto Soares da Silva, editor, *Pluricentricity: Language Variation and Sociocognitive Dimensions*, chapter 1, pages 19–48. de Gruyter, Berlin/Boston.

Hans Bickel and Christoph Landolt. 2018. *Schweizerhochdeutsch*. Duden, Schweizerischer Verein für die deutsche Sprache.

Michael Clyne. 1992. German as a pluricentric language. In Michael Clyne, editor, *Pluricentric Languages: Differing Norms in Different Nations*, chapter 2, pages 117–148. Walter de Gruyter.

Michael G Clyne. 1984. *Language and society in the German-speaking countries*. Cambridge University Press Cambridge.

Duden. 2017. *Duden - Die deutsche Rechtschreibung*, 27th edition. Dudenredaktion.

Christa Dürscheid and Patrizia Sutter. 2014. Grammatische Helvetismen im Wörterbuch. *Zeitschrift für angewandte Linguistik*, 60(1):37–65.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Birgit Hamp and Helmut Feldweg. 1997. Germanet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid. Association for Computational Linguistics.

Verena Henrich and Erhard Hinrichs. 2010. Gernedit - the Germanet editing tool. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24, Uppsala, Sweden. Association for Computational Linguistics.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2014. Aligning Germanet senses with wiktionary sense definitions. In *Human Language Technology: Challenges for Computer Science and Linguistics*, volume 8387 of *Lecture Notes in Computer Science*, pages 329–342. Springer.

Erhard Hinrichs, Verena Henrich, and Reinhild Barkey. 2013. Using part-whole relations for automatic deduction of compound-internal relations in germanet. In *Language Resources and Evaluation, special issue on "Wordnets and Relations"*, volume 47:4. Springer.

Anna-Julia Lingg. 2006. Kriterien zur Unterscheidung von Austriazismen, Helvetismen und Teutonismen. In Christa Dürscheid and Martin Businger, editors, *Schweizer Standarddeutsch: Beiträge zur Varietätenlinguistik*, chapter 1, pages 23–49. Gunter Narr Verlag.

John P McCrae, Ian Wood, and Amanda Hicks. 2017. The colloquial Wordnet: Extending Princeton Wordnet with Neologisms. In *International Conference on Language, Data and Knowledge*, pages 194–202. Springer.

Gerold Schneider. 2018. Differences between Swiss High German and German German via data-driven methods. In *SwissText 2018: 3rd Swiss Text Analytics Conference, Winterthur, 12 Juni 2018*.

Piek Vossen. 1998. Introduction to EuroWordnet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer.

Danish in Wikidata lexemes

Finn Årup Nielsen
Cognitive Systems, DTU Compute, Technical University of Denmark
Kongens Lyngby, Denmark

Abstract

Wikidata introduced support for lexicographic data in 2018. Here we describe the lexicographic part of Wikidata as well as experiences with setting up lexemes for the Danish language. We note various possible annotations for lexemes as well as discuss various choices made.

1 Introduction

Wikipedia's structured sister Wikidata (Vrandečić and Krötzsch, 2014) at https://www.wikidata.org/ supports interlinking different language versions of Wikipedia as well as several other Wikimedia sites, such as Wikibooks and Wikimedia Commons. One wiki that has been missing from the list is Wiktionary, — the dictionary wiki. Wiktionary has a structure different from the other wikis as multiple different words and concepts might be described on the same page, only connected through the same orthographic representation.

In 2018, Wikidata enabled support for lexicographic data via special lexeme wiki pages. Compared to Wiktionary, Wikidata lexemes offer a solution with directly machine-readable data: it is not necessary to write parsers to obtain the lexeme data in a structured format. Wikidata lexemes also reduce the amount of redundant input: In Wiktionary, each language edition sets up its own dictionary, and a word described in one Wiktionary is not directly available in another Wiktionary. Further issues with Wiktionary are the linkage to the Wikidata concept ontology and the linkage to external resources such as WordNet (Miller, 1995). Neither of these links is non-trivial to set up, though matching lexical entries between Wiktionary and WordNet may be done with good accuracy (McCrae et al., 2012).

Below we will describe how lexemes are supported on Wikidata,[1] and list some of the Danish resources relevant for Wikidata lexemes. Then we will detail how Danish lexemes have been annotated and discuss some of the choices made.

[1]There is an introduction to Wikidata lexemes on Wikidata itself at https://www.wikidata.org/wiki/Wikidata_talk:Lexicographical_data.

2 Wikidata lexemes

Wikidata stores lexeme data on a new type of pages prefixed with the letter 'L' and further identified with an integer, e.g., the Danish lexeme *gentagelse* (repetition) has the identifier "L117" and available for view and edit at https://www.wikidata.org/wiki/Lexeme:L117. On the same page, multiple senses and forms for the lexeme may be defined. They are identified by suffixes to the lexeme identifier, e.g., the plural indefinite form *gentagelser* would be identified as "L117-F3", while the first sense—if it was defined—would have been identified as "L117-S1". Forms and senses are defined separately, so it is currently difficult to define a specific sense for a specific form. Lexemes, forms and senses may be associated with properties, and these properties are identified with integer prefixed with the letter 'P'.

The Wikidata lexeme data maps to an RDF representation,[2] and the RDF data is queryable via the *Wikidata Query Service* SPARQL endpoint at https://query.wikidata.org/. The mapping uses part of the *Lexicon Model for Ontologies* (LEMON) ontology (Cimiano et al., 2016; McCrae et al., 2012). The central OWL concepts for the lexeme data are ontolex:LexicalEntry, ontolex:Form and ontolex:LexicalSense for lexeme, form and sense respectively with the prefix http://www.w3.org/ns/lemon/ontolex#. Each of these three OWL concepts has associated basic data in Wikidata. Apart from the identifier, the lexeme has the lemma, language and lexical category, the form has its orthographic representation and grammatical features while the sense may have multiple glosses. This basic data cannot be associated with qualifiers and references like normal Wikidata properties.

Links from Wikidata lexemes (L-pages) to Q-items (i.e., the ordinary Wikidata items) are of two kinds: Either for the description of the lexical and grammatical "metadata" for the lexeme or for the description of the meaning of a sense. In the latter case, the Q-items function as the wordnet notion

[2]https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/RDF_mapping

Danish	Total	Description	SPARQL query fragment
1268	43816	Number of lexemes	`[] a ontolex:LexicalEntry`
4826	118742	Number of forms	`[] a ontolex:Form`
617	11194	Number of sense	`[] a ontolex:LexicalSense`
8594	218803	Number of grammatical feature links	`[] wikibase:grammaticalFeature []`

Table 1: Statistics for lexeme data in Wikidata. See also the statistics displayed on the Ordia website at https://tools.wmflabs.org/ordia/statistics/.

of *synsets*. Wikidata has specific properties to link Q-items to synsets in external lexical resources, including BabelNet (Navigli and Ponzetto, 2010) (P2581) and the Collaborative Interlingual Index (P5063) (Bond et al., 2016). Alternatively, the more generic property for Linked Open data URIs (P2888) can be used. Wikidata has linked some WordNet synsets used in ImageNet. The correspondence between the resources is not necessarily straightforward to establish (Nielsen, 2018).

Some statistics for the lexeme data in Wikidata are displayed in Table 1. It displays, e.g., that the number of forms is close to 120'000. In comparison, the English Wiktionary has currently around 5.9 million content pages, while the Danish Wiktionary has around 38 thousand.[3] The numbers are not directly comparable as multiple forms may be listed on one Wiktionary content page. A count on the distinct number of (monolingual) form representations in Wikidata gives 89'728 on 27 March 2019 based on the following SPARQL query:

```
SELECT
  (COUNT(DISTINCT(?representation))
  AS ?count)
{ [] ontolex:representation
  ?representation . }
```

3 Danish resources

There are some Danish resources relevant for Wikidata lexemes, e.g., corpora for language usage examples. As Wikidata is distributed under the Creative Commons Zero (CC0) license, the resources incorporated into Wikidata need to be compatible with that license.

Old out-of-copyright Danish works are typically with an antiquated spelling, e.g., where the first letter of nouns has a capital letter. Wikipedia and Wiktionary may not be used because their content is under an attribution and share-alike license, not compatible with the CC0 license. Modern Danish sentences can be retrieved from, e.g., Danish law texts at https://www.retsinformation.dk/, Danish translations of international treaties and conventions, such as the Treaty of Lisbon, and the Danish part of the Europarl corpus (Koehn, 2005). Fairy tales by Hans Christian Andersen can be found with modern spelling.

Of the lexicographical resources, the standard Danish dictionary, *Retskrivningsordbogen*, has a restrictive license. Another large Danish dictionary with over 300'000 entries and used, e.g., with the computer program *aspell*, is under the GNU General Public License and is not compatible with Wikidata's CC0. DanNet (Pedersen et al., 2009) has a WordNet-derived open license and a Wikidata property (P6140) for the DanNet words — corresponding to Wikidata lexemes—has been created in November 2018. DanNet is distributed as OWL, so should fit well with Wikidata lexemes.

NST Lexical database for Danish[4] has Speech Assessment Methods Phonetic Alphabet (SAMPA) pronunciation specification for over 235'000 Danish words and stated to have the CC0 license.

As of June 2019, we have used 160 sentences from the Danish part of the Europarl corpus,[5] and linked to 1258 DanNet 2.2 word identifiers,[6] while the NST phonetic data has hardly been used.

4 Annotating lexemes, forms and senses

Wikidata has a continuously growing number of properties that can be used to annotate lexemes, forms and senses. General properties—that are relevant for lexemes of most word classes—are *usage example* (P5831), *word stem* (P5187) and *derived from* (P5191), where the latter may indicate etymological origin or origin of derivations. Compound parts may be linked with a property (P5238) and the order of the parts may be specified with a property used as a qualifier (P1545). The Wikidata property for DanNet words (P6140) are linked to version 2.2 of the resource. As of March 2019, 844 lexemes with associated information about DanNet words are linked.[7] The data model of Wikidata al-

[3] https://en.wiktionary.org/wiki/Special:Statistics and https://da.wiktionary.org/wiki/Speciel:Statistik

[4] https://www.nb.no/sprakbanken/show?serial=sbr-26

[5] See Ordia's statistics at https://tools.wmflabs.org/ordia/reference/Q5412081

[6] The SPARQL query `SELECT ?dannet { ?lexeme wdt:P6140 ?dannet }` on the Wikidata Query Service.

[7] https://www.wikidata.org/wiki/Property_talk:P6140 displays the DanNet property statistics.

lows for the specification of "no value", thus it is possible to specify that a lexeme cannot be found in the DanNet 2.2 resource. For instance, adverbs and rare nouns, such as *lommevogn* (L40687), are not in DanNet and indicated as such. The *usage example* property (P5831) can store a short free-form text and the qualifier *stated in* (P248) can point to a Q-item with metadata about a work where the text appears. A related property is *attested in* (P5323) which also can point to a Q-item.

Lexemes may also be associated with classes via the *instance of* property (P31). Properties relevant for lexemes across word classes in Danish are, e.g., whether they loan words and/or compound words.

Forms may be associated with hyphenation and pronunciation specification. Wikidata has properties for X-SAMPA, IPA transcription and Kirshenbaum code. These pronunciation properties have been used on Wikidata's Q-items, but so far not (or very limited) for Danish lexemes.

Senses can be associated with language style (P6191) and perhaps most importantly with *item for this sense* property (P5137) which links the senses of lexemes to the Q-items and thus with the rest of the Wikidata knowledge graph. Synonyms, antonyms, hypernyms and hyponyms may be inferred from the information in that part of the Wikidata knowledge graph.

Links between lexemes in different languages can currently be made with a specific translation property (P5972) applicable for senses, or the connection between lexemes can be made through their senses and the P5137 property linking to Q-item that then binds lexemes from separate languages together.

4.1 Verbs

Verbs can be associated with conjugation class through the P5186 property. We have followed the scheme of (Allan et al., 1995) where there are four main Danish conjugation categories. The auxiliary verb(s) for a verb can be specified with P5401. Some verbs can be assigned to a class, e.g., motion verbs, auxiliary verb, transitive/intransitive verb or deponent verb. The valence (P5526) can also be specified.

4.2 Nouns

Danish nouns may be characterized by grammatical gender and class. Classes of commons nouns may be countable or mass noun, singulare tantum, plurale tantum, collective noun, 'nexual' or 'innexual' noun or nomen agentis. The distinction between nexual and innexual is based on (Hansen and Heltoft, 2019) where the former may refer to "actions and processes, activities and states," and the later "objects or compounds".

4.3 Images and audio

Senses may be associated with images by referencing filenames in the free media archive *Wikimedia Commons*. The link may help language learners and possibly be a resource for training natural language processing machine learning models in the same way that ImageNet has used WordNet, see (Nielsen, 2018; Nielsen and Hansen, 2018) for applications of the use of Wikidata. Typically the senses of nouns may be associated with images, while it may be difficult to identify good images to be associated with, e.g., adverbs. A few Danish verbs have been associated with images, e.g., *gå* (walk) and "visual" adjectives, such as *rød* (red), are also associated with images.

Lexemes can be associated with images. Photos of written signs may exemplify how words are used in the environment, e.g., a photo of a street sign reading "Cyklist vig for gående" is used to illustrate the usages of the lexeme *cyklist* (L43527, cyclist).

Audio files can be associated with the lexicographic data. For forms, the P443 property can link to one of the currently around 130 pronunciation audio files for Danish words, while senses can link to sound files with the P51 property, e.g., the sense for *bi* (L37259, bee) links to a sound recording of bees buzzing and the sense for *bil* (L36385, car) is associated with an audio file of a starting and driving car.

5 Discussion

Wikidata is entirely field-based and especially for lexemes there are very few means to enter free-form information. While exceptions can be noted in standard dictionaries such as Wiktionary, almost every piece of information added for a lexeme in Wikidata must be associated with a property. The explicitness of Wikidata complicates the modeling of language. Below we discuss a few of the issues that have appeared for the Danish language.

5.1 Lexeme splitting

The English lexeme *they* (L371) incorporates the forms *they*, *them*, *their*, *theirs*, *themselves* and *themself*, while French *vous* (plural *you*) and *votre* are separate lexemes (L9289 and L9289). In the Danish online dictionary *Den Danske Ordbog*, the corresponding forms for *they* are split into several dictionary entries, while the German Wikidata lexemes *ich* (L7877, the personal pronoun *I*) has currently no other form than *ich*. The issue was the subject of an inconclusive discussion on Wikidata.[8] As noted by one of the discussants, if the pronouns

[8] https://www.wikidata.org/wiki/Wikidata_talk:Lexicographical_data/Archive/2018/11#How_to_split_or_merge_stedord_(Q36224).

are split, a question is how to link between such different lexemes. A related issue for Danish appears for some adverbs, which could easily be regarded as separate lexemes, such as *hjem*, *hjemad* and *hjemme* (home, homeward, at home). Here the words are distinguished by telicity and a dynamic/static feature, e.g., *hjemad* is atelic and dynamic (Hansen and Heltoft, 2019, p. 216). Possibly new specific properties could describe the relationships.

Wikidata's choice of separating form and sense complicates modeling of some words. *vand* (*water*), *øl* (*beer*) and *tøj* (*cloths*) are examples of words that each are regarded as one lexeme but where the specific forms are associated with specific semantics: The common gender version of *øl* relates to a countable noun as in "one beer", while the neuter version relates to a mass noun. Here we could split the lexeme into two Wikidata lexemes, e.g., *vand* with common gender and *vand* with neuter, but that would complicate their relations to other lexemes, e.g., in terms of compounding and etymology. A related issue occurs for deponent verbs. For *finde*/*findes* (active/passive; find/exist) the lexemes have been separated (L39637 and L44601) following the convention of DanNet.

In case of, e.g., the lexemes *mor* and *moder* (mother) their singular forms are different but they have the same meaning and their plural form, *mødre*, is the same. There is no way of merging the separate plural forms when *mor* and *moder* are regarded as separate lexemes as the forms are tied to separate lexeme pages. The creation of a dedicated property could link such forms together.

5.2 Compound splitting

Danish is a language rich in compounds. The compounds and affixes of a lexeme can be specified with the P5238 property where other lexemes can be linked. The currently longest Danish lexeme in Wikidata, *ejendomsadministrationsvirksomhed* (building administration business), could be split as *ejendom-s-administration-s-virksomhed* with two *s*-interfixes and three words with a good semantic relation to the complete lexeme. With a more granular level, the word could be split into *ejen-dom-s-ad-ministr-ation-s-virk-somhed*, where affixes have been split from the roots. Here, *dom* and *virk* has little semantic relationship to the compound lexeme. With the current setup of the P5238 property and the structure of Wikidata, it is difficult to see how the two splits can coexist with the same lexeme. Currently, we typically split on the highest level, e.g., *ejendomsadministration-s-virksomhed*. The lexeme pages for *ejendomsadministration* and *virksomhed* can further split the compounds and derived words.

5.3 Linking compounds to parts

When orthographically similar words with the same etymology are split across multiple lexemes it may be unclear which lexeme a compound derives from. For instance, the compound *vaskemaskine* (L42991, washing machine) could be analyzed as consisting of: 1) a verb stem (*vask*), an interfix (*-e-*) and a noun (*maskine*), or 2) a verb in its infinitive form (*vaske*) and a noun, or 3) a noun (*vask*), an interfix and a noun. During data entry one would need to make an explicit choice. The same choice may appear for affixations, such as *for-be-handle*. While *be-* is arguably a prefix (L44579), *for* may be a prefix or an adverb, — or possibly an preposition.

5.4 Genitive

Danish genitive, where an *-s* suffix is added, has traditionally been regarded as a case, but newer words for Danish grammar challenge that notion and argues that it is a clitic and a derivation making a nominal to non-nominal (Hansen and Heltoft, 2019, p. 255). Originally, we began adding the genitive *-s* forms for the Danish nouns, but has discontinued it after becoming aware of the issue. The Swedish part of Wikidata lexeme continues to add the genitive *-s* forms for nouns. If we were to add the genitive for Danish nouns, then one could argue that genitive versions of other word classes should also be added, — as words from other word classes can be used as nouns, e.g., *de rødes valgsejr* (literally, *the reds' election victory*) where the adjective *røde* has the added genitive *-s*. The advantage of have the *-s* forms is that lookup, e.g., for spellchecking may be more convenient. Other Danish digital dictionaries record the *-s* form.

5.5 Data quality

The structured format of the data and the query tools associated with Wikidata enable us to perform some completeness and internal consistency checks. For instance, we may formulate a SPARQL query that returns Danish lexemes without any usage examples. We have used the Shape Expressions (ShEx) language (Baker and Prud'hommeaux, 2017) to formalize such checks, and these ShEx definitions are available on separate pages on Wikidata (Nielsen et al., 2019). As an example, the ShEx definition E65[9] checks Danish numerals regarding data about language, lemma, word stem, word class, DanNet, usage example, sense, form and hyphenation.

[9] https://www.wikidata.org/wiki/EntitySchema:E65

5.6 Applications

What can Wikidata lexemes be used for? Wikidata itself has a dedicated page for application ideas.[10] For spellchecking the current number of lexemes in Wikidata can hardly compete with already established larger word lists, but in the long run using the lexeme forms for spellchecking might be of interest. The advantage is that it is collaboratively extensible, likely able to quickly catch up on neologisms and evolving jargon in comparison to standard dictionaries. It is less clear if Wikidata lexemes can be used for more advanced natural language processing, such as part-of-speech tagging and grammar checking. The ability of the current Wikidata lexeme system has limited means for specifying grammar.

6 Related Research

Among related research, there are several studies reporting the extraction of data from Wiktionary and using the structured data for linguistic tasks or building a resource (Zesch et al., 2008; McCrae et al., 2012; Sérasset, 2014; Pantaleo et al., 2017). For instance, the Java- and database-based system by Zesch et al. (Zesch et al., 2008) for reading, storing and querying lexical semantic knowledge from Wikipedia and Wiktionary enables a user, e.g., to programmatically query for hyponyms of senses. The parser of the described system needs to be adjusted for each language edition of Wiktionary as each edition may use different markup for the lexical semantic information.

The lexicographic part of Wikidata is still comparably small, but contrary to many other online dictionaries with rich semantics, Wikidata users can add and edit the lexicographic information and more or less immediately see it becoming available in the powerful query facility of the SPARQL-based Wikidata Query Service. Our Ordia Web application at http://tools.wmflabs.org/ordia takes advantage of this service (Nielsen, 2019).

7 Acknowledgment

We thank Bolette Sandford Pedersen, Sanni Nimb, Sabine Kirchmeier, Nicolai Hartvig Sørensen and Lars Kai Hansen for discussions and answering questions, and the reviewers for suggestions for improvement of the manuscript. This work is funded by the Innovation Fund Denmark through the projects DAnish Center for Big Data Analytics driven Innovation (DABAI) and Teaching platform for developing and automatically tracking early stage literacy skills (ATEL).

[10] https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_tools.

References

[Allan et al.1995] Robin Allan, Philip Holmes, and Tom Lundskær-Nielsen. 1995. *Danish.* Routledge.

[Baker and Prud'hommeaux2017] Thomas Baker and Eric Prud'hommeaux. 2017. Shape Expressions (ShEx) Primer. July.

[Bond et al.2016] Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. *Proceedings of the Eighth Global WordNet Conference*, pages 50–57, January.

[Cimiano et al.2016] Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report, 10 May 2016. May.

[Hansen and Heltoft2019] Erik Hansen and Lars Heltoft. 2019. *Grammatik over det Danske Sprog.* University Press of Southern Denmark, February.

[Koehn2005] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *The Tenth Machine Translation Summit: Proceedings of Conference*, pages 79–86.

[McCrae et al.2012] John P. McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. 2012. Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics.*

[Miller1995] George Armitage Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38:39–41, November.

[Navigli and Ponzetto2010] Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: building a very large multilingual semantic network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, July.

[Nielsen and Hansen2018] Finn Årup Nielsen and Lars Kai Hansen. 2018. Inferring visual semantic similarity with deep learning and Wikidata: Introducing imagesim-353. *Proceedings of the First Workshop on Deep Learning for Knowledge Graphs and Semantic Technologies*, pages 56–61, April.

[Nielsen et al.2019] Finn Årup Nielsen, Katherine Thornton, and José Emilio Labra Gayo. 2019. Validating Danish Wikidata lexemes. June. Submitted to SEMANTiCS 2019.

[Nielsen2018] Finn Årup Nielsen. 2018. Linking ImageNet WordNet Synsets with Wikidata. *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France*, pages 1809–1814, April.

[Nielsen2019] Finn Årup Nielsen. 2019. Ordia: A Web application for Wikidata lexemes. May. From ESWC 2019.

[Pantaleo et al.2017] Ester Pantaleo, Vito Walter Anelli, Tommaso Di Noia, and Gilles Sérasset. 2017. Etytree: A Graphical and Interactive Etymology Dictionary based on Wiktionary. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pages 1635–1640.

[Pedersen et al.2009] Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299, August.

[Sérasset2014] Gilles Sérasset. 2014. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web: interoperability, usability, applicability.*

[Vrandečić and Krötzsch2014] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57:78–85, October.

[Zesch et al.2008] Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1646–1652.

Using Thesaurus Data to Improve Coreference Resolution for Russian

Ilya Azerkovich
Higher School of Economics
Moscow, Russia
ilazerkovich@edu.hse.ru

Abstract

Semantic information about entities, specifically, how close in meaning two mentions are to each other, can become very useful for the task of coreference resolution. One of the most well-researched and widely used forms of presenting this information are measures of semantic similarity and semantic relatedness. These metrics are often computed, relying upon the structure of a thesaurus, but it is also possible to use alternative resources. One such source is Wikipedia, which possesses the category structure similar to that of a thesaurus. In this work we describe an attempt to use semantic relatedness measures, calculated on thesaurus and Wikipedia data, to improve the quality of a coreference resolution system for Russian language. The results show that this is a viable solution and that combining the two sources yields the most gain in quality.

1 Introduction

Coreference resolution is a very important part of many natural language processing tasks, and for solving it generally information from several language layers is required. Among those, the importance of semantic information, as opposed to more shallow features, e.g. string-based, morphologic or syntactic ones, is sometimes debated (see e.g. Durrett and Klein (2013)), but it is nevertheless seen as useful for overcoming the potential plateau of quality, as V. Ng (2017) noted.

As far as English language is concerned, various thesauruses are usually used as sources of semantic information, the most popular of them being the WordNet (Harabagiu et al., 2001; Ponzetto and Strube, 2006 among others). Another such resource is Wikipedia that, while not a thesaurus by itself, is sometimes considered as such due to its structure of categories, connected to each other by the relation of inclusion (Ponzetto and Strube, 2006).

For Russian language the room for improvement of coreference resolution systems still exists, as has been demonstrated by results of the Ru-Eval-2014 competition for Russian coreference resolvers (Toldova et al., 2014). The usage of semantic information is also not as widespread, partly due to lesser volume of resources available: fewer thesauruses exist for Russian than there are for English, the most prominent of them being the RuThes (Loukachevitch et al., 2014), consisting of appr. 70 000 synsets, and the Russian segment of Wikipedia is also smaller. Consequently, fewer attempts at using semantic information have been made.

Nevertheless, the results of Toldova et al. (2014) mentioned above clearly show that semantic information needs to be explored to properly resolve cases such as (1) below.

(1) People who survived the wreck of **the ship** told that the main reason for the tragedy was the **oil-burner** being very old.

Additional information that can be obtained from a thesaurus is required to correctly join *oil-burner* to *the ship*. On the other hand, while thesauruses seldom contain information about named entities, such as people, additional resources would be required to obtain information of this kind. Data that can only be obtained from an encyclopedia such as Wikipedia is required for examples like (2):

(2) Victor Vekselberg would like to engage **Grigori Perelman** to work in the "Silicon Valley". The fortune has smiled upon **the mathematician**…

To deal with cases similar to the ones described above, a system would require to look-up the related content in a resource and properly infer the relation between the mentions.

This paper presents an attempt at using information, obtained from RuThes and Russian Wikipedia, to improve the quality of coreference resolution for the Russian language. More precisely, we explore the efficiency of using measures of semantic similarity and semantic relatedness, as quantified representations of how close the meanings of two concepts are. In our research we employ the measures, extracted from the aforementioned resources, as features used in machine learning solutions.

The achieved results suggest that integrating features based on semantic information does indeed improve the system performance, with the highest increase in quality being gained by combining the data from both resources.

2 Related Work

Thesauruses, in particular WordNet, have been widely used for purposes of coreference resolution in a variety of ways. Some of these include extracting hypernym chains or semantic classes, derived from high-level nodes (Poesio and Vieira, 2000; Soon et al., 2001) or calculating special confidence measures of different paths between concepts (Harabagiu et al., 2001). Semantic similarity has also been frequently employed in automated coreference resolution, either calculated from thesaurus data or unannotated corpora (Ponzetto and Strube, 2006; Versley, 2007), or based on word embeddings (Clark and Manning, 2016). A large spectrum of different semantic similarity values that can be calculated based on thesaurus structure has been suggested by various researchers. Overview of the most influential ones are given, e.g., in (Budanitsky and Hirst, 2006).

For Russian the research of coreference resolution using thesaurus data has been smaller in scale with the only participant system of RuEval-2014 that used semantic information relying on a proprietary ontology (Bogdanov et al., 2014). Recently, Toldova and Ionov (2017) have introduced a coreference resolution system, supplemented with semantic information from hypernym chains extracted from RuThes, achieving certain improvements in quality. Our research differs in approach with employing semantic similarity measures instead.

The Wikipedia data is also often used in systems of coreference resolution, including the Stanford parser (Raghunathan et al., 2010). Generally, the text content of the page is considered for analysis, with its category structure being used in a similar way to a thesaurus in (Ponzetto and Strube, 2006). The text information and categories of a page from Russian Wikipedia have been used by Azerkovich (2018) with a positive result, but the category tree as a whole was not considered.

3 Calculating Semantic Relatedness

3.1 Resources Used

Two main sources of semantic information were used in this research: RuThes thesaurus and the Russian segment of Wikipedia. RuThes is a thesaurus, created by a team of linguists, with its freely available part, RuThes-Lite, including 55 000 entities that correspond to 158 000 lexical entries. The structure of RuThes is similar to that of WordNet, with concepts in the thesaurus linked to each other by the set of labeled relations that includes IS-A, PART-WHOLE and a number of associative relations.

The Russian segment of Wikipedia with ~1.5 mln articles, while being smaller than the English one (over 5 mln articles), is still one of its largest, making it an important knowledge source. The feature of Wikipedia that allowed to include its information in our analysis is its category structure: each article can be placed within one or several categories, which, in its own turn, can be categorized further. Because one article can belong to several categories, and one category can be included in several parent categories, the structure of Wikipedia categories is not a tree in a strict sense, but a more general graph.

For both resources the following set of measures of semantic similarity was calculated: the path-based measures of Rada et al. (1989), Wu and Palmer (1994) and Leacock and Chodorow (1998); information content-based measure of Resnik (1995). Because the relations between parent and child categories in Wikipedia do not strictly correspond to IS-A relations, it would be more correct to consider the scores for this source as measures of semantic relatedness rather than semantic similarity.

For Wikipedia pages the measure of gloss overlap by Banerjee and Pedersen (2003) was also computed. This was not done for RuThes data, because not all synsets there are provided with a gloss, which is required to apply this measure.

3.2 Mining Semantic Information

In the case of RuThes, values of semantic similarity measures for two referential expressions

were obtained by calculating the scores for head lemmas of the groups in question. In case of heads of any or both groups being ambiguous, measures for all possible combinations of meanings were obtained, and after obtaining the values, the following two features were created: the maximum value of the similarity score, and the average value of the similarity score. If one or both mentions were absent from the thesaurus, the measure scores were considered to be zero.

In the case of Wikipedia, the problem of ambiguity had to be addressed slightly differently. To calculate the semantic relatedness measures, firstly, the pages corresponding to the referring expressions in question had to be obtained. For that purpose, the groups were queried to Wikipedia search engine. In case a disambiguation page was encountered, all hyperlinks from the page were analyzed. If a link led to the page, containing the other queried group, it was used as the hit. If no such links were found, the first hyperlink on the page was used. After resolving the referring expressions to their Wikipedia pages, the gloss overlap measure of the pages' texts was calculated.

The rest of the set of metrics was calculated in the same way as for RuThes, using the graph of categories to which the obtained pages belong. Following the observations of Ponzetto and Strube (2006), the possible depth of nodes was limited to 4 to assure less noisy results, due to higher levels of the category structure being too strongly connected. The values of path-based and information content-based measures were obtained for all combinations of categories for both pages, after which the same two features as for thesaurus data was calculated: the maximum value of the similarity score, and the average value of the similarity score. As with the RuThes data, if any of the mentions was not mapped to a corresponding Wikipedia page, the measures were considered zero.

3.3 Correlation with Human Judgement

As an additional step in preparing to use the values of measures, described above, as features for a coreference resolution algorithm, it was tested to what extent these measures correlate with human judgement on coreference.

To achieve that, the chosen set of measures was calculated for a set of referring expressions with pre-existing coreference annotation. As the source of annotation, the Russian coreference corpus RuCor was used. It is the corpus, created for the purposes of the task of automated anaphora and coreference resolution for RU-EVAL-2014 (Toldova et al., 2014). For 200-pair sets of coreferent and non-coreferent pairs semantic relatedness was calculated, and then the Pearson correlation coefficient with the annotation was calculated. To enable the calculations, the pairs from the evaluation set were assigned the maximum measure value if they were annotated as coreferent, and the minimum value if marked as not coreferent.

The results of evaluation are presented in Table 1. As can be seen from the tables, the values of measures generally do correlate with human judgement, justifying their usage as features for analysis, except from the gloss overlay, which was not used in further experiments. Different measures also correlate differently with coreference annotation: while the measures, obtained on the data from RuThes display higher correlation in general, the data from Wikipedia correlates relatively well with annotation for named entities. This leads to conclude that combining data from both resources can give the most coverage and, potentially, a larger improvement to quality of the analysis.

Source	***Rada***	***Wu***	***Leacock***	***Resnik***	***Gloss***
RuThes (non-empty)	0.56	0.59	0.51	0.30	n/a
Wikipedia (NEs)	0.7	0.6	0.1	0.2	0.2

Table 1: Correlation with coreference annotation

4 Using Semantic Relatedness for Machine Learning Feature Creation

4.1 Corpus Data Used

The research was conducted on the data of the aforementioned RuCor corpus (Toldova et al., 2014), as the largest available corpus of Russian with coreference annotation. It consists of 180 texts of a variety of genres that in total contain 3838 coreferential chains with 16557 referential expressions. For the Ru-Eval-2014 task it was split in the training and test sets (70% and 30% of the corpus volume, respectively), which were retained for our experiments. All texts in the corpus have been preprocessed and morphologically tagged using the set of instruments developed by Sharoff and Nivre (2011). The annotation, provided by the corpus creators, was used as the

golden standard, against which the systems were evaluated.

4.2 Learning Algorithm

For our research we used a machine learning algorithm based on a decision tree classifier, which has been tested in application to coreference resolution for Russian in (Toldova and Ionov, 2017). It is based on the work of (Soon et al., 2001), and uses a similar set of baseline features that we supplemented with described above features, derived from thesaurus data.

The system is based on a pairwise approach, according to which the classifier, being given a pair of referring expressions, decides whether they corefer or not, based on the feature values. The candidate pairs for analysis were created the following way: from each pair of coreferent expressions a positive instance is created, and then every NP between the anaphor and the antecedent is paired to the anaphor to create a negative instance. In our research we relied upon the NP boundaries, obtained from the corpus markup instead of automatically generated ones, in order to maximize the influence of the features we introduce in addition to the baseline set.

4.3 Baseline Features

The baseline system was based on the set of features, derived from the original set, suggested by Soon et al. (2001). It included features of various types: string-based, distance, morphological, syntactic and semantic. But, as it was originally created for the English language, several features, such as definiteness, were meaningless in the case of Russian, due to linguistic differences. Because of that, they were removed and, in some cases, replaced with alternative ones. The resulting feature set is given in Table 2.

Feature type	Features
String features	• Mention strings match • One of mentions is an identifier of the other • One of mentions is an abbreviation of the other
Distance features	• Number of sentences between mentions • Number of sentences is greater than 3
Morphological features	• Mentions match in gender • Mentions match in number • Both mentions are proper • Anaphor is a demonstrative pronoun • One of mentions is a pronoun
Syntactic features	• The potential anaphor is an appositive of the antecedent • Mentions are subject and object of the same sentence • Both mentions are subjects • Both mentions are first words in a sentence
Semantic features	• Both mentions are animate

Table 2: Baseline feature set

All features were represented by their numeric value if applicable, or indicator functions, equal to 1 in case the feature is true, and 0 in case it is false.

The performance of the system, using only the baseline set, was compared to performance of its version, using the set enhanced with features derived from thesaurus data of RuThes and Wikipedia: maximum and average values of the semantic relatedness measures.

4.4 Performance Evaluation

The performance of systems was evaluated, based upon a number of metrics: MUC (Vilain et al., 1995), B^3 (Baldwin and Bagga, 1998) and CEAF (Luo, 2005). The following versions of the baseline system were included in the comparison: enhanced with the RuThes-based features; enhanced with Wikipedia-based features; enhanced with features from both resources.

The Table 3 below contains the results of the comparison by metric, with maximum improvements over the baseline highlighted in bold. The improvements, achieved in the aforementioned work of Ponzetto and Strube (2006) by adding Wikipedia-based and Wordnet-based features are also given for comparison.

4.5 Discussion

The results of the evaluation show that features based on semantic relatedness measures do increase the system performance compared to the baseline to a certain degree. While the increase is similar in scale to the numbers demonstrated in earlier work of Ponzetto and Strube (2006), it may still be not large enough for statistical importance This prevents us from labelling it a decisive improvement and calls for further development of the method.

	MUC			B^3			CEAF
	P	R	F	P	R	F	
Baseline	72.76	59.49	65.46	71.01	44.50	54.71	49.02
Baseline + Wikipedia	70.28	59.71	64.56	66.50	44.63	53.41	46.36
Baseline + RuThes	72.72	59.43	65.41	71.15	44.44	54.71	48.91
Baseline + RuThes + Wikipedia	73.57	60.01	**66.10**	71.77	44.93	**55.26**	**49.66**
(Ponzetto and Strube, 2006), Wikipedia	+1.3%	-0.5%	+0.8%				
(Ponzetto and Strube, 2006), Wordnet	+2.2%	-0.9%	+1.3%				

Table 3: Evaluation metrics

Still, the resulting increase in quality is larger compared to that of similar work by Toldova and Ionov (2017): 0.54% of MUC score and 0.55% of B^3 score, compared to 0.26% and 0.19% correspondingly. As in our research we used semantic information in the form of semantic relatedness measures, compared to hypernym chains in (Toldova and Ionov, 2017), we can assume that more precise preprocessing of information and usage of features beyond Boolean ones can lead to more improvements in systems' performance.

Study of the results reveals that the largest increase in quality is observed when combining the features from both sources, with the improvement seen across all evaluation metrics. This corresponds to the assessment of correlation with human judgement described above.

The results also allow to conclude that information from both used sources serves to improve the quality of the analysis in different ways. While the data from RuThes can be used to improve the system's precision, the data from Wikipedia helps to increase the recall of the performance. This can be contributed to the difference in content between the sources: while RuThes, as a thesaurus created by a team of linguists, is less in size, but better structured than Wikipedia, the latter possesses a more contrived and not necessarily transparent category system, but contains more information about wider range of phenomena.

5 Conclusions and Future Work

In this paper we described an attempt to improve the quality of coreference resolution for Russian by introducing features, based on semantic information, obtained from thesaurus data. For that end, we used the thesaurus of Russian RuThes and the Russian segment of Wikipedia to compute several semantic relatedness measures to be used as features in a coreference resolution system.

While the results of evaluation of the system cannot yet be called final, they suggest that the quality of coreference resolution for Russian can be improved by using features based on semantic information. It is important to remark that the maximum profit was achieved by combining the features from both sources, with Wikipedia also being useful despite its open-source nature and being open to free editions by any user. While recent research relying on neural networks for coreference resolution achieve better results for Russian (e.g. (Le et al., 2019)), the gains of using semantic information observed by us and other researchers allow to assume that such algorithms could benefit from implementing it, as well.

Future work, inspired by this research, lies in exploring other coreference resolution algorithms and improving the quality of semantic features extraction. The former involves exploring more productive techniques of coreference resolution, in particular, assessing the potential of integrating semantic level information in neural networks. The latter involves employing a wider range of semantic relatedness measures, as well as increasing the efficiency of using Wikipedia-based information. As an alternative to the online encyclopedia, DBpedia can be used. It possesses clearer structure and labeled relations, which could simplify computing semantic relatedness from its data.

References

Ilya Azerkovich. 2018. Employing wikipedia data for coreference resolution in Russian. In *Artificial Intelligence and Natural Language*, volume 789, pages 107–112. Springer, Cham.

Breck Baldwin and Amit Bagga. 1998. Algorithms for Scoring Coreference Chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Satanjeev Banerjee and Ted Pedersen. 2003. Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805--810.

A. V. Bogdanov, S. S. Dzhumaev, D. A. Skorinkin, and A. S. Starostin. 2014. Anaphora analysis based on ABBYY Compreno linguistic technologies. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014)*, volume 13, pages 89–102.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, March.

Kevin Clark and Christopher D. Manning. 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653. June.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982.

Sanda M. Harabagiu, Razvan C. Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *2nd Meeting of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 55–62. The Association for Computational Linguistics.

T. A. Le, M. A. Petrov, Y. M. Kurato, and M. S. Burtsev. 2019. Sentence Level Representation and Language Models in The Task of Coreference Resolution for Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2019)*, pages 341–350.

Claudia Leacock and Martin Chodorow. 1998. Combining local context with wordnet similarity for word sense identification. In *WordNet: an electronic lexical database*, pages 265–283. MIT Press.

Natalia V. Loukachevitch, Boris Dobrov, and Ilia Chetviorkin. 2014. RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014)*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

Vincent Ng. 2017. Machine Learning for Entity Coreference Resolution : A Retrospective Look at Two Decades of Research. In *Proceedings of the 31th Conference on Artificial Intelligence (AAAI 2017)*, volume 6, pages 4877–4884.

Massimo Poesio and Renata Vieira. 2000. An Empirically Based System for Processing Definite Descriptions. *Computational Linguistics*, 26(4):539–593, December.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, volume 33, pages 192–199. Association for Computational Linguistics.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *IJCAI*:448–453.

S. Sharoff and J. Nivre. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2011)*, volume 10, pages 591–605.

Wee Meng Soon, Daniel Chung Yong Lim, and Hwee Tou Ng. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, December.

Svetlana Toldova and M. Ionov. 2017. Coreference Resolution for Russian: The Impact of Semantic Features. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2017)*, volume 16, pages 339–348.

Svetlana Toldova, Anna Roitberg, Alina Ladygina, M. D. Vasilyeva, Ilya Azerkovich, M. Kurzukov, Galina Sim, D. V. Gorshkov, A. Ivanova, Anna Nedoluzhko, and Yulia Grishina. 2014. RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2014)*, volume 13, pages 681–694.

Yannick Versley. 2007. Antecedent Selection

Techniques for High-Recall Coreference Resolution. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52.

Zhibiao Wu and Martha Palmer. 1994. Verb Semantics and Lexical Selection. *ACL*:133–138.

The Extended Arabic WordNet: a Case Study and an Evaluation Using a Word Sense Disambiguation System

Mohamed Ali Batita
Research Laboratory in Algebra,
Number Theory and
Nonlinear Analysis,
Faculty of Science,
Monastir, Tunisia
BatitaMohamedAli@gmail.com

Mounir Zrigui
Research Laboratory in Algebra,
Number Theory and
Nonlinear Analysis,
Faculty of Science,
Monastir, Tunisia
Mounir.Zrigui@fsm.rnu.tn

Abstract

Arabic WordNet (AWN) represents one of the best-known lexical resources for the Arabic language. However, it contains various issues that affect its use in different Natural Language Processing (NLP) applications. Due to resources deficiency, the update of Arabic WordNet requires much effort. There have only been only two updates it was first published in 2006. The most significant of those being in 2013, which represented a significant development in the usability and coverage of Arabic WordNet. This paper provides a study case on the updates of the Arabic WordNet and the development of its contents. More precisely, we present the new content in terms of relations that have been added to the extended version of Arabic WordNet. We also validate and evaluate its contents at different levels. We use its different versions in a Word Sense Disambiguation system. Finally, we compare the results and evaluate them. Results show that newly added semantic relations can improve the performance of a Word Sense Disambiguation system.

1 Introduction

Natural language processing (NLP) is part of computer linguistics, which is also part of artificial intelligence. There are many disciplines in NLP. Information extraction is one of them. It can be text mining, information retrieval, named entity recognition... All these disciplines require lexical and semantic resources to proceed and generate satisfactory results. The more inclusive the resource, the more accurate the results will be. Lack of resources, especially for less-resourced language such as Arabic, has always been a persistent problem. One of the reliable resources for the Arabic language is Arabic WordNet (AWN) (Black et al., 2006).

Princeton WordNet (PWN) (Miller, 1995; Miller, 1998), English WordNet or simply WordNet is the original and most developed of all wordnets. From its first publication, it proved its reliability with various NLP tasks. Many researchers were inspired by its usability and made a wordnet for their own languages. Now we have more than 77 wordnet[1], which AWN is one. Researches now are aiming either to create new wordnets for other languages (or dialects) or improve existing ones. Creating new wordnets can be done by gathering an exhaustive repository of meanings and senses, e.g. dictionary or corpora, and assigning all words for each sense. This approach is called the merge approach (Vossen, 1998). More common is the 'expansion' approach. It consists of translating the core of PWN[2] and extending it through more concepts related to the language. This is called the top-down approach. AWN has followed this approach.

Generally speaking, a wordnet is a group of *synsets* interconnected with different relations. A synset is a set of synonyms. In other words, it is a group of words that share the same meaning. Relations can be synonymy, antonymy, hyponymy, meronymy... The enrichment of a wordnet can follow the axe of synsets or relations. Besides, the coverage in terms of synsets with diverse relations can be very useful in many NLP applications, especially Question Answering (QA) and Word Sense Disambiguation (WSD). Numerous approaches present themselves to construct and extend wordnets, from statistics to word embedding-based approaches (Neale, 2018).

Even without enrichment, AWN showed great results with several NLP applications like infor-

[1] http://globalwordnet.org/resources/wordnets-in-the-world

[2] It contains the most frequently used words in any language and it has about 5,000 words.

mation retrieval (Abbache et al., 2016; Bouhriz et al., 2015) and query expansion (Abbache et al., 2018) even for e-learning applications (Karkar et al., 2015). But, AWN has seen many attempts to enrich its content with different approaches, either by adding new synsets or new entities or even new specificity of the Arabic language like broken plurals[3] (Abouenour et al., 2013; Saif et al., 2017; Ameur et al., 2017; Batita and Zrigui, 2017; Batita and Zrigui, 2018). Despite these efforts, AWN remains inadequate to the needs of complex modern systems. There remains a huge gap between the contents of AWN and the Arabic language itself, and also between AWN and other wordnets like PWN. This paper cites several significant programmes that have been undertaken to improve the contents of AWN. This paper also seeks to shine a light on the semantic relations of AWN and their importance for improving the performance of NLP applications. Finally, the paper provides an overview of tests we have undertaken with three versions of AWN in a concurrent NLP application.

The paper is structured as follows. The next section is an overview of the various updates and extensions of the AWN along a detailed discussion about its content. Section 3 summarises most of the significant research undertaken to enrich the semantic relations in AWN. Section 4 discusses the procedures that we follow to validate the newly added relations. Section 5 presents the conducted tests to show much the enriched AWN can affect a WSD system. Finally, section 6 will be our conclusion with some future works.

2 Versions of Arabic WordNet

The AWN project started in 2006. The goal was to build a freely open source lexical database for the Modern Standard Arabic available for the NLP community (Abbache et al., 2018). By that time, it has 9,698 synsets, corresponding to 21,813 words. Synsets were linked by 6 different types of semantic relations (hyponymy, meronymy, etc.), in a total of 143,715 relations (Cavalli-Sforza et al., 2013). Entities are distinguished by their part of speech POS: noun, verb, adverb, or adjective. Synsets are linked to their counterpart in PWN and the Suggested Upper Merged Ontology (SUMO) via the so-called Interlingual Index (ILI) (Black et al., 2006).

[3]It is non-regular plural that involves internal changing in the structure of an Arabic word.

In 2010, a second version has been published by Rodriguez et al. (Rodríguez et al., 2008). It has 11,269 synsets corresponding to 23,481 words with 22 types of semantic relationships in a total of 161,705 relations. This version has a browser written with JAVA that has an update and search functions (Rodríguez et al., 2008). This version is rich with more specific concepts related to the Arabic cultures like named entities and the Arabic language like broken plurals (Batita and Zrigui, 2018). Several researchers have taken advantage of this version in most of their work in different areas of NLP to improve the performance of their systems.

Recently, an extended version has been published in 2015 by Regragui el al. Regragui et al. (Regragui et al., 2016). This version is seen as an improvement of the coverage and usability of the previous version of AWN (Abouenour et al., 2013). It includes 8,550 synsets which correspond with 60,157 words, among which we find 37,342 lemmas, 2,650 broken plurals, and 14,683 verbal roots. Regragui et al. (Regragui et al., 2016) changed the structure of the database to the Lexical markup framework (LMF) (Francopoulo et al., 2006), the ISO standard for NLP abd machine-readable dictionary (MRD) lexicons. They made it publicly available and ready to use from the Open Multilingual Wordnet[4].

Table 1 below summarizes the statistics of entities, synsets, and relations of PWN and the three previous versions of AWN.

	PWN	**V1**	**V2**	**Ex.V**
Entities	206,978	21,813	23,481	60,157
Synsets	117,659	9,698	11,269	8,550
Relations	283,600 (22 types)	143,715 (6 types)	161,705 (22 types)	41,136 (5 types)

Table 1: Statistics of PWN with 3 versions of AWN.

First of all, we notice that the number of entities and synsets in PWN is very high compared to all the versions of AWN. In versions 1 and 2 (V1 and V2), we find that the number of entities is proportional to the number of synsets which is approximately two to three times the number of entities, which is not the case in the extended version (Ex.V). On the one hand, V2 contains more

[4]http://compling.hss.ntu.edu.sg/omw/

synsets and fewer entities than the Ex.V. On the other hand, V2 has 11,269 synsets connected with 161,705 relations and Ex.V has only 8,550 synsets connected with only 41,136 relations. By comparing the number of relations in PWN with V2, we note that V2 is nearly rich in terms of connections between synsets. As a result, we can say that Ex.V is more affluent than the other versions of AWN in terms of synsets but impoverished in terms of relations. Abouenour et Al. (Abouenour et al., 2013) put a focus on the entities, in this paper, we focus on the relations between them.

3 Related Works

Until now, there are several attempts to enrich the AWN using different methods and approaches. Most of the works focused on the improvement of the number of entities and synsets (Rodríguez et al., 2008; Alkhalifa and Rodríguez, 2009; Abouenour et al., 2010; Abouenour et al., 2013; Regragui et al., 2016; Ameur et al., 2017; Saif et al., 2017; Lachichi et al., 2018). The main reason behind those works is the richness of the Arabic Language. One study on both Arabic and English Gigaword corpus has shown that to deal with the same linguistic content of 100,000 words in English, it takes approximately 175,600 words in Arabic (Alotaiby et al., 2014). In other words, one English word can be processed with approximately two Arabic words. Thus, resource-based applications expect more coverage of the Arabic language.

In contrast, the work on the relations of AWN is much less. Boudabous et Al. (Boudabous et al., 2013) proposed a linguistic method based on two phases. The first one defines morpho-lexical patterns using a corpus developed from Arabic Wikipedia. The second one uses the patterns to extract new semantic relations from the entities in AWN. A linguistic expert has validated the obtained relations. While some of the new relations were good others were not - for various reasons, including the size of the corpus and the patterns applications.

In our first work on the AWN (Batita and Zrigui, 2017) we focused on the enrichment of antonym relations. As many studies have shown that the antonym relation is universal, but, it has been noted that there are different perspectives towards this lexical relation in different cultures (Hsu, 2015). Antonyms detection, in general, is a tough task for the NLP community. After a deep study, we have found that the extended version of AWN has only four types of relations. One of them is the antonym relations with only 14 pairs. This work has been concentrated on the extended version of AWN because it has been proved by Abouenour et al. (Abouenour et al., 2013) that it has given excellent results when testing in a Q/A system. We proposed a pattern-based approach to extract new antonym relations from the entities of AWN. For that, they extract patterns from an Arabic corpus and used a corpus analysis tool to recognize automatically the antonym pairs from other pairs. The analysis tool is the Sketch Engine (Kilgarriff et al., 2004). It has many useful metrics like the LogDice which gives a higher score to most likely related pairs. The results were filtered using the LogDice and the validation was manual.

After that our next step was the derivational relations in AWN (Batita and Zrigui, 2018). By that, we tackled another matter of the Arabic language which is the morphological aspect. The derivational and morphological problem has been a subject in different wordnet from other languages (Koeva et al., 2008; Mititelu, 2012; Šojat et al., 2012). Generally speaking, and when it comes to studying a language aspect, rule-based approaches seem the more promoting one because they rely on linguistic rules verified by an expert or by a native speaker. Based on that, wz relied on that kind of approach to add new derivational relations between entities in AWN. We studied the derivational aspect of the Arabic language to make a set of transformation rules. Those rules are based on the POS switch, for example between the verb كَتَبَ *kataba*[5] (write) and the noun كَاتِبٌ *kaātibun* (writer) there is a *HasDerivedVerb* relation. Rules are made by an expert and validated carefully to guaranty the precision of the results. For more information on the transformation rules see (Batita and Zrigui, 2018). In the end, we got 8 different relations with different frequencies. The validation of the rules and the finale results has been made by a lexicographer.

The knowledge-based systems in general and wordnet-based systems specifically shown good results when they used a rich wordnet with as many relations as possible (Fragos et al., 2003; Seo et al., 2004; Alkhatlan et al., 2018). Yet, the use of a wordnet, in general, has shown a great result in different areas of NLP such as humor detection

[5]We used the transliteration system of LaTeX.

(Barbieri and Saggion, 2014) and human feelings (Siddharthan et al., 2018) even in the cybercrime investigation (Iqbal et al., 2019). Given a sufficiently large database with many words and connections between them, many applications are quite capable of performing sophisticated semantic tasks. That is why work on the relations in AWN has to increase because richer resource can achieve significant results in a real-world NLP application. Evaluation and validation of the relations need to be considered as essential and continuous steps to guaranty the credibility of a resource. Basically, validation can be done either manually by verifying each relations individually or automatically using different approaches. In the next, we will describe how we validated the newly added relations in the previous updates.

4 Validation of the New Relations in Arabic WordNet

The previously cited works on the enrichment of the relations in AWN confronted different parts of the Arabic language, in general, using different methods and approaches. Table 2 summaries all the relations (new and pre-existing) of the extended version of AWN along with their frequency.

Relation	Frequency
Hyponym	21,851
Hypernym	21,851
NearSynonym	673
HasInstance	1,295
IsInstance	1,295
Antonym	800
HasDerivedVerb	2,005
ActiveParticiple	1,347
PassiveParticiple	1,004
Location	985
Time	752
Instrument	184
HasDerivedNoun	1,784
Relatedness	804
Total	**56,630**

Table 2: Relations of the extended version of Arabic WordNet with their frequencies.

We will focus on the extended version published by Regragui et al. (Regragui et al., 2016) and the new relations that we already added (Batita and Zrigui, 2017; Batita and Zrigui, 2018). Since many relations need to be validated (12), we initially used an automatic approach, which we developed. While the majority of the new relations are specific to the Arabic language (8 derivational relations), with the developed approach we will be working only on the three general relations: hyponyms, hypernyms, hasInstance, isInstance, and synonyms. We were inspired by the aspect of the dictionary and the construction of wordnets since they are based on the synonyms and the *is-a* relations (hyponym/hypernym).

Our automatic approach says that '*if a word w has a dictionary definition and belongs to a synset s with other words $w_1, \ldots, w_k$ then there is a strong probability that w mentions one or more of w_k in her definition and/or other words (w_k) from the synonym/hypernymy/instance of s*'. An example will simplify the point of the view:

- W: **تلف** *tlf* (dammage)
- $S = ta|kala\_\_v1AR$: صدأ ، **تلف**، تآكل *tākl, tlf, ṣdʿa* (corrosion, damage, rust)
- $Hyponym = AinohaAra\_v1AR$: إنْهار، تدهور، **فسَد** *āinhār, tdhwr, fsad* (collapsed, deteriorated, ruined)
- $Definitionofw$: عطِب ، **فسَد**، تلف الزّرع *tlf ālzrʿ, fsad, ṭib* (The implant is damaged, corrupted, damaged)

As we can see, $W \in S$ and its definition have a word (فسد *fsd*) that refers to the $hyponym$ of s. If so, then the relation is validated, otherwise it should be reviewed. We collect all the definition of the words that have one of the three relations from different dictionary[6]. All definitions are stored in one file. The file is structered as a table and each line contains one definition per word. Stop words are eliminated and remaining words have been lemmatized[7]. Finally, we applied our idea and we got the results of each relation as described in table 3. The high accuracy of the synonyms due to their limited number (we have only 412 relations). False relations are due to one of the following reasons (i) either a problem with the lemmatization or (ii) the granulate of the definition or (iii) the diacritization and/or correct written form of the word.

As a start-up, the first approach yields to promoting accuracy. To guaranty efficiency and high confidentiality, a second validation is done manually by native speakers and a linguistic expert. The

[6]For that we used the website of AlMaany `https://www.almaany.com/`.

[7]We used the Farasa toolkit (Abdelali et al., 2016).

Relations	Accuracy (%)
HasInstance/IsInstance	89,1
Hypernym/Hyponym	86,2
Synonym	96,7

Table 3: Accuracy of the automatic validation according to each relation.

remaining relations (derivational and wrongly validated by the first method) have been reviewed one by one. Native speakers made suggestions for some relations that may or may not hold between words. As an example, the two words وطن *wṭn* (prepare to do) and نظم *nẓm* (organize) are connected by the *hyponym* relation. Native speakers suggested that it should be eliminated but the expert said otherwise. So, the expert takes the final decisions. If a relation is obvious and does not exist, the expert can add it, as well as he can eliminate it otherwise. Besides his knowledge, the decisions of the expert are based on the following conditions:

- The suggestions of the native speakers.
- A clear definition of the words in the Arabic dictionary لسان العرب *lsān ālʿrb* (Lisan al-Arab).
- The existence of the relation between the words in question in AWN (some words do not have any relation at all).
- The correctness of two words that hold the relation.
- The existence of a relatedness between the words in the Arabic dictionary.

In the end, we got 81% correct relations, 5% wrong relations, 12% partially wrong relations (one of the pair of the words is wrong), and 2% of the words with no relations at all. Most of the wrong relations were found in the relations that are specifics to the Arabic language, like *Instrument* and *Relatedness* because they are based on transformation rules. Sometimes, words (irregular ones) that share this kind of relations do not follow any transformation rules. Some changes have been made by the linguistic expert regarding the 12% of the relations that are partially wrong by either changing one of the two words or replacing if the word does not exist in AWN. Finally, we could not do anything for the 2% of the words that have no relations at all.

5 Evaluation with a Word Sense Disambiguation System

In literature, we find different approaches to evaluate any lexical resources and the choice between them depending mainly on the kind of the resource itself and for what purpose (Brank et al., 2005). Since AWN is a lexical database in the first place, then its evaluation should follow one of the following strategy:

- Comparing it to a golden standard wordnet (in most cases, PWN).
- Using it in real-life NLP application and evaluating the obtained results.

As for the first approach of evaluation, many researchers have faced difficulties with it. Abouenour et al. (Abouenour et al., 2013) compared the content of AWN with the content of PWN and the Spanish WordNet. They found that the number of synsets in AWN is around 8% (too low) of those of PWN, while the Spanish wordnet represents 49%. Taghizadeh et al. (Taghizadeh and Faili, 2016), also, compared their newly constructed Persian WordNet with FarsNet and they found a precision of 19%, which is too low to consider their resource as a reliable one.

Basically, one can tell if a wordnet is a reliable resource or not by how far it can help a system to achieve better results. This kind of evaluation seems to be a better way to test the extended AWN. As mentioned above (section 2), many researchers used the AWN in their applications and it helped achieve great results. As we are concentrated on the relations of AWN, we looked into some NLP applications to see how the relations between the entities in AWN can affect the precision of an NLP application.

Word Sense Disambiguation WSD seemed the most successful system to show the effectiveness of the relations between the words. The choice of the WSD system was made following a study of different systems that profit from the relations in AWN. The aspect of the disambiguation is based on the similarity between words, which is exactly what the relations in AWN are made for in the first place. Besides, many WSD systems have been based on the relationship between words (Fragos et al., 2003; McCarthy, 2006; Kolte and Bhirud, 2009; Zouaghi et al., 2011; Zouaghi et al., 2012; Dhungana et al., 2015) and other applications, like information

retrieval and Q/A system, rely more on the words themselves rather than the relations between them. All of this gives the WSD the advantage to be our best candidate.

Since our aim is to evaluate the impact of the relations in AWN on a WSD system, the choice of the WSD algorithm is not the main task. We implement the very simple algorithm of Galley et al. (Michel and Kathleen R., 2003) with a slight difference. The algorithm proceeds as follows:

1. Build a representation of all possible combination of the text.
2. Disambiguate all words in the text.
3. Build a lexical chains.

The algorithm takes a text as an input and proceeds all of the possible combinations between the current word and all the previous words. After that, a weighted edge takes the place if one of the senses of the current word has a semantic relation with any senses of the previous words. At the end of the text, a *disambiguation graph* is built with the nodes represent the senses of each word of the text and the edges representing the semantic relations between the senses of the words since AWN links the senses and not the words. Finally, the weights of each edge are summed up to represent a final score to each sense for each word in the text. The correct sense of the target word have the highest score. One thing to mention here is that this algorithm works with only 4 semantic relations (synonym, hypernym/hyponym, and sibling) and the weight of each edge is assigned according to the type of relations and the distance between the two words.

We use the Khaleej-2004 corpus (Abbas et al., 2011). It contains 5690 documents divided to 4 categories; international and local news, economy, and sports. It has a total of nearly 3 millions words. We did not work on optimizing the weight nor the distance between the words. The only difference that we made is the number of relations. We implemented this algorithm to work with more relations. All relations in the extended version of AWN are token into consideration. We tested the algorithm with three versions of AWN; the version 2, the extended version with and without the new relations. Table 4 shows the obtained results.

As we can see from table 4, the enriched AWN with the semantic relations yields a significant improvement with a 78,6% of precision. We remark

Tested versions of AWN	**Precision (%)**	**Recall (%)**	**F1 score**
V2	69,2	57,6	72
Ex.V *without* new relations	72,7	66,9	69,6
Ex.V *with* new relations	**78,6**	**71,1**	**74,6**

Table 4: Precision, recall, and f1 score with different versions of AWN.

that the precision of V2 and the Ex.V without the new relations are very close. That is due to the diversity of the first one in terms of relations (22 types) and the richness of the second one in terms of hyponym/hypernym relations (19,806 relations). Despite the fact that V2 has more relations than Ex.V (161,705 and 50,787), the difference between their precisions is that V2 does not have much of specific relations related to the Arabic language. As an example, عزف *zf* is a polysemous verb. Two of his senses are completely different. One could be 'playing music' and the other 'strike.' In the extended AWN and without the enrichment of the relations, it has only two relations, *hyponym* with the verb شغّل *šġl* (fill) and *hypernym* with the verb أَخْرج *ʾaḫrǧ* (get it out). When we run the test in the WSD system, we could get the appropriate sense. After the test with the new relations, we got the *Instrument* relation a with a higher score.

The obtained results with the enriched AWN showed the importance of the resource and the relations between its words, even in a simple knowledge-based WSD algorithm like the one we used.

6 Conclusion

In this paper, we presented the different versions of the AWN along with a study case on the newly added relations to its extended version. Next, we described the content of different versions of AWN with some remarkable works done to enrich its relations. Then, we cited many evaluation approaches in general and how we evaluated AWN specifically. We provided an automatic method to validate some of the relations in AWN. In the end, we found the most reliable approach is the human evalua-

tion, despite the fact that it does not take advantage of computer programs and relies heavily on time-consuming work. To make the new content more accurate, we tested different versions of AWN with a real-life NLP application (WSD system). We attended interesting and promising results with the extended version of AWN. Before making it online and ready for the NLP community, we are still working on improving and refining the semantic relations in AWN to get more accuracy and we are running some test in different NLP applications.

References

Ahmed Abbache, Fatiha Barigou, Fatma Zohra Belkredim, and Ghalem Belalem. 2016. The use of arabic wordnet in arabic information retrieval. In *Business Intelligence: Concepts, Methodologies, Tools, and Applications*, pages 773–783. IGI Global.

Ahmed Abbache, Farid Meziane, Ghalem Belalem, and Fatma Zohra Belkredim. 2018. Arabic query expansion using wordnet and association rules. In *Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications*, pages 1239–1254. IGI Global.

Mourad Abbas, Kamel Smaïli, and Daoud Berkani. 2011. Evaluation of topic identification methods on arabic corpora. *JDIM*, 9(5):185–192.

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–16.

Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2010. Using the yago ontology as a resource for the enrichment of named entities in arabic wordnet. In *Proceedings of The Seventh International Conference on Language Resources and Evaluation (LREC 2010) Workshop on Language Resources and Human Language Technology for Semitic Languages*, pages 27–31.

Lahsen Abouenour, Karim Bouzoubaa, and Paolo Rosso. 2013. On the evaluation and improvement of arabic wordnet coverage and usability. *Language resources and evaluation*, 47(3):891–917.

Musa Alkhalifa and Horacio Rodríguez. 2009. Automatically extending ne coverage of arabic wordnet using wikipedia. In *Proc. Of the 3rd International Conference on Arabic Language Processing CITALA2009, Rabat, Morocco*.

Ali Alkhatlan, Jugal Kalita, and Ahmed Alhaddad. 2018. Word sense disambiguation for arabic exploiting arabic wordnet and word embedding. *Procedia computer science*, 142:50–60.

Fahad Alotaiby, Salah Foda, and Ibrahim Alkharashi. 2014. Arabic vs. english: Comparative statistical study. *Arabian Journal for Science and Engineering*, 39(2):809–820.

Mohamed Seghir Hadj Ameur, Ahlem Chérifa Khadir, and Ahmed Guessoum. 2017. An automatic approach for wordnet enrichment applied to arabic wordnet. In *International Conference on Arabic Language Processing*, pages 3–18. Springer.

Francesco Barbieri and Horacio Saggion. 2014. Automatic detection of irony and humour in twitter. In *ICCC*, pages 155–162.

Mohamed Ali Batita and Mounir Zrigui. 2017. The enrichment of arabic wordnet antonym relations. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 342–353. Springer.

Mohamed Ali Batita and Mounir Zrigui. 2018. Derivational relations in arabic wordnet. In *The 9th Global WordNet Conference GWC*, pages 137–144.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the arabic wordnet project. In *Proceedings of the third international WordNet conference*, pages 295–300. Citeseer.

Mohamed Mahdi Boudabous, Nouha Chaâben Kammoun, Nacef Khedher, Lamia Hadrich Belguith, and Fatiha Sadat. 2013. Arabic wordnet semantic relations enrichment through morpho-lexical patterns. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6. IEEE.

Nadia Bouhriz, Faouzia Benabbou, and Habib Benlahmer. 2015. Text conceptsextraction based on arabic wordnet and formal concept analysis. *International Journal of Computer Applications*, 111(16):30–34.

Janez Brank, Marko Grobelnik, and Dunja Mladenic. 2005. A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170. Citeseer Ljubljana, Slovenia.

Violetta Cavalli-Sforza, Hind Saddiki, Karim Bouzoubaa, Lahsen Abouenour, Mohamed Maamouri, and Emily Goshey. 2013. Bootstrapping a wordnet for an arabic dialect from other wordnets and dictionary resources. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Udaya Raj Dhungana, Subarna Shakya, Kabita Baral, and Bharat Sharma. 2015. Word sense disambiguation using wsd specific wordnet of polysemy words. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 148–152. IEEE.

Kostas Fragos, Yannis Maistros, and Christos Skourlas. 2003. Word sense disambiguation using wordnet relations. In *First Balkan Conference in Informatics, Thessaloniki*.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (lmf). In *International Conference on Language Resources and Evaluation-LREC 2006*.

Chan-Chia Hsu. 2015. A syntagmatic analysis of antonym co-occurrences in chinese: contrastive constructions and co-occurrence sequences. *Corpora*, 10(1):47–82.

Farkhund Iqbal, Benjamin CM Fung, Mourad Debbabi, Rabia Batool, and Andrew Marrington. 2019. Wordnet-based criminal networks mining for cybercrime investigation. *IEEE Access*.

Abdelghani Karkar, Jihad Mohamad Alja'am, Mohamad Eid, and Andrei Sleptchenko. 2015. E-learning mobile application for arabic learners. *Journal of Educational & Instructional Studies in the World*, 5(2).

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. Itri-04-08 the sketch engine. *Information Technology*, 105:116.

Svetla Koeva, Cvetana Krstev, and Duško Vitas. 2008. Morpho-semantic relations in wordnet–a case study for two slavic languages. In *Global wordnet conference*, pages 239–253. University of Szeged, Department of Informatics.

SG Kolte and SG Bhirud. 2009. Wordnet: a knowledge source for word sense disambiguation. *International Journal of Recent Trends in Engineering*, 2(4).

Cilia Lachichi, Chahrazad Bendiaf, Lamia Berkani, and Ahmed Guessoum. 2018. An arabic wordnet enrichment approach using machine translation and external linguistic resources. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.

Diana McCarthy. 2006. Relating wordnet senses for word sense disambiguation. In *Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*.

Galley Michel and McKeown Kathleen R. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 1486–1488.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Verginica Barbu Mititelu. 2012. Adding morpho-semantic relations to the romanian wordnet. In *LREC*, pages 2596–2601.

Steven Neale. 2018. A survey on automatically-constructed wordnets and their evaluation: Lexical and word embedding-based approaches. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Yasser Regragui, Lahsen Abouenour, Fettoum Krieche, Karim Bouzoubaa, and Paolo Rosso. 2016. Arabic wordnet: New content and new applications. In *Proceedings of the Eighth Global WordNet Conference*, pages 330–338.

Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, Musa Alkhalifa, M Antonia Martí, William Black, Sabri Elkateb, James Kirk, Adam Pease, et al. 2008. Arabic wordnet: Current state and future extensions. In *Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary*, pages 387–405.

Abdulgabbar Saif, Mohd Juzaiddin Ab Aziz, and Nazlia Omar. 2017. Mapping arabic wordnet synsets to wikipedia articles using monolingual and bilingual features. *Natural Language Engineering*, 23(1):53–91.

Hee-Cheol Seo, Hoojung Chung, Hae-Chang Rim, Sung Hyon Myaeng, and Soo-Hong Kim. 2004. Unsupervised word sense disambiguation using wordnet relatives. *Computer Speech & Language*, 18(3):253–273.

Advaith Siddharthan, Nicolas Cherbuin, Paul J Eslinger, Kasia Kozlowska, Nora A Murphy, and Leroy Lowe. 2018. Wordnet-feelings: A linguistic categorisation of human feelings. *arXiv preprint arXiv:1811.02435*.

Krešimir Šojat, Matea Srebačić, and Marko Tadić. 2012. Derivational and semantic relations of croatian verbs. *Journal of Language Modelling*, pages 111–142.

Nasrin Taghizadeh and Hesham Faili. 2016. Automatic wordnet development for low-resource languages using cross-lingual wsd. *Journal of Artificial Intelligence Research*, 56:61–87.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

A Zouaghi, L Merhbene, and M Zrigui. 2011. Word sense disambiguation for arabic language using the variants of the lesk algorithm. *WORLDCOMP*, 11:561–567.

Anis Zouaghi, Laroussi Merhbene, and Mounir Zrigui. 2012. Combination of information retrieval methods with lesk algorithm for arabic word sense disambiguation. *Artificial Intelligence Review*, 38(4):257–269.

On Hidden Semantic Relations between Nouns in WordNet

Tsvetana Dimitrova, Valentina Stefanova
Institute for Bulgarian Language "Prof. Lyubomir Andreychin",
Bulgarian Academy of Sciences
52 Shipchenski Prohod Blvd., bldg. 17
cvetana,valentina@dcl.bas.bg

Abstract

The paper presents an effort on transferability of noun – verb and noun – adjective derivative and semantic relations to noun – noun relations. The approach relies on information from semantic classes and existing inter-POS derivative and (morpho)semantic relations between noun and verb, and noun and adjective synsets. We have added semantic relations between nouns in WordNet that are indirectly linked via verbs and adjectives. Observations on the combination between the relations and semantic classes of nouns they link, may facilitate further efforts in assigning semantic properties to nouns pointing to their abilities to participate in predicate-argument structures.

1 Introduction[1]

The present work[2] aims at revealing hidden (indirect) semantic relations between nouns in WordNet by using information that is already available from the inter-POS derivative and (morpho)semantic relations between noun – verb, and noun – adjective synsets, and the semantic class of lexical concepts expressed by the members of a noun–noun pair.

The main relation among words in WordNet is synonymy (or near-synonymy; synonyms are defined as words which denote the same concept and are interchangeable in many (but not all) contexts). The synonyms (called 'literals') are grouped into unordered sets (synsets) which are linked via the so-called 'conceptual relations'. Most relations between synsets connect words of the same part-of-speech (POS). Noun synsets are linked via hypernymy / hyponymy (superordinate) relation, and meronymy (part-whole) relation. Verb synsets are arranged into hierarchies via hypernymy / hyponymy relation. Adjectives are organised in terms of antonymy and similarity, and relational adjectives (pertainyms) are linked to the nouns they are derived from. Adverbs are linked to each other via similarity and antonymy relations.

Thus, WordNet consists of four sub-nets, with few cross-POS relations – the so-called '(morpho)semantic' relations between semantically similar words that share a stem with the same meaning (e.g., *writer* is an **Agent** of *write*, see (Fellbaum et al., 2009)); pertainym relations: noun – adjective (e.g., *pope – papal*); adjective – adverb (e.g., *bad – badly*); derivative relations: noun – verb (e.g., *write – writer*); adjective – verb (e.g., *writing – write*); noun – adjective (e.g., *pope – papal*).

Lexical concepts expressed by the synsets are further semantically classified by assigning the so-called 'semantic primitives' (or 'semantic primes' or 'semantic classes') to each synset ((Fellbaum et al., 2009); (Miller et al., 1993)). Noun and verb synsets are subjected to elaborate semantic classifications – nouns are organised into 25 semantic classes such as *noun.person*, *noun.animal*, *noun.plant*, *noun.process*, *noun.act*, *noun.location*, etc., and verbs – into 15 classes – *verb.stative*, *verb.communication*, *verb.cognition*, *verb.perception*, etc. Only three labels are applied to the adjective synsets – *adj.all* (mainly) for descriptive adjectives, *adj.pert* for pertainyms, and *adj.ppl* for adjectival participles, but there are efforts on more detailed classifications of adjectives in wordnets for other languages (the

[1] For the requirements of the academic system, Tsvetana Dimitrova takes responsibility for sections 2 and 3, and Valentina Stefanova – for 1 and 4.

[2] We would like to thank three anonymous reviewers as well as the participants at the 10th Global WordNet Conference for their valuable comments and suggestions. Special thanks also go to Ivelina Stoyanova from the Institute for Bulgarian Language (BAS), for the help with data extraction.

WordNet for German (GermaNet), see (Hamp and Feldeg, 1997); WordNet for Russian (RussNet), see (Azarova and Sinopalnikova, 2004); the Polish WordNet (plWordNet), see (Maziarz et al., 1997); and the Bulgarian wordnet (BulNet), see (Stefanova and Dimitrova, 2017), (Dimitrova and Stefanova, 2018).

2 Nouns in WordNet

Nouns in WordNet are organised within the superordinate / subordinate (hypernymy / hyponymy) hierarchy. The hierarchical semantic organisation is limited in depth, and distinguishing features are added to create lexical inheritance system where each word inherits the distinguishing features (attributes (modification), parts (meronymy), functions (predication) from its superordinates ((Miller 1990, 1990)). An example would be {diarist:1} [10011486-n][3], which, as a hyponym of {writer:2} [10801291-n], is classified as *noun.person* and could be an **Agent** of the verb synsets {write:1} [00993014-v], {write:3} [01007027-v], *write:4* [01031966-v], and {write:5} [01691057-v] just like its hypernym.

Nouns are further related to verb synsets via derivative and/or (morpho)semantic relations – (morpho)semantic relations are applied to derivationally related noun – verb pairs, but not vice versa – not every derivationally related pair is (morpho)semantically linked, and to adjectives – via derivative and pertainym relations (pertainym relations are usually applied to *adj.pert* adjectives, and nouns and adjectives are derivationally linked but not every derivationally linked pair noun – adjective is in pertainym relation).

Some nouns linked via a verb have an explicit link through hypernym/hyponym relation: (1) they can be two hyponyms of the same hypernym, e.g., the nouns {exhibition:1} [eng-30-00522145-n] and {exposure:3} [eng-30-00522537-n] are derivationally linked via the verb {expose:9; exhibit:3} [eng-30-02140033-v], and are co-hyponyms of the noun synset {presentation:1; demonstration:1} [eng-30-00521562-n]; (2) One can be a hyponym of the other, as with {relish:2; flavour:2} [eng-30-05715864-n] which is a hyponym of {taste:9; taste sensation:1; taste perception:1} [eng-30-05715283-n], and the two are derivationally and morphosemantically (as **Event**) related to {taste:6; savor:4; savour:4} [eng-30-02194286-v]).

In the next section 2., we will discuss the relations between these nouns by taking into account the semantic class of the nouns and the 'linking verb', and the (morpho)semantic relations between the two (if available).

3 Nouns linked via verb synsets

In WordNet, verb and noun synsets are related via derivative and (morpho)semantic relations that link semantically similar verbs and nouns that share a stem with the same meaning. Verbs impose selectional restrictions on the entities selected for their argument positions, particularly on characteristics of the nouns taking specific semantic roles. For example, the **Agent** of cognitive verbs is expected to be animate and human (but not animal) while that of consumption verbs is animate but can be both human and animal. Selectional restrictions also apply to complements – for example, motion verbs may have as their **Instrument** nouns referring to vehicles and artifacts while their **Location** or **Direction** complement can be location, object or artifact.

Previous studies have further differentiated nouns which are linked via (morpho)semantic relations to different verb classes. (Paiva et al., 2014) and (Real and Rademaker, 2015) offer extension of the classification of deverbal nominals in Portuguese drawing upon work on Portuguese nominalisations (Real, 2014) where eight possible classes of eventive nominalisation have been proposed: action of, result of, physical result of, iteration of the act of, resulting state from, abstract result of, locative, collectivisation of.

In previous work on the Bulgarian wordnet, (morpho)semantic relations **Agent** and **Undergoer** were subdivided by taking into account the information about: verb and noun semantic classes, sentence frames encoding predicate-argument structure of the simple sentences that verbs can form, and noun suffixes, to formulate additional (morpho)semantic relations, such as **Experiencer**, **Actor**, **Recipient** ((Dimitrova, 2018)). (Leseva et al., 2018) have proposed subcategorisation

[3] Throughout the paper, the numbers of the literals follow those applied in the database used by the viewer Hydra available at: http://dcl.bas.bg/bulnet/. We do not give all literals and definitions due to space limitation but only ids of synsets acc. to PWN 3.0 – in square brackets, with POS marked at the end. There may be changes to semantic classes and (morpho)semantic relations between the PWN and the version on http://dcl.bas.bg/bulnet/, for detail see (Leseva et al., 2015).

of nouns by taking into account information from WordNet, VerbNet, and FrameNet, which resulted in formulating subcategories such as: **Agent_communicator**, **Agent_effector**, **Agent_experiencer**, **Agent_undergoer**, **Artifact_undergoer**, etc.

Our proposal on introducing noun – noun semantic relations is based on the assumption that selectional restrictions are imposed not only by verbs but also by nouns derived from verbs such as nominalisations (e.g., *writing*), agentive nouns (e.g., *writer*), resultative nouns (e.g., *written*), etc. They are related to the source verb (e.g., *write*) not only via (morpho)semantic but also via derivative relations. We additionally take into account the relations between the semantic classes of the nouns linked through derivative relations via verb synsets.

Some – but not all – derivationally linked nouns are linked also via (morpho)semantic relations, as in (1) where {writing:2} and {writer:1} are **Event** and **Agent**, respectively, of {write:7}. Other derivationally related nouns, however, such as {pen:3} below, are only derivationally (but not (morpho)semantically) linked:

Ex.

{write:7; compose:3; pen:1} [01698271-v] *verb.creation* 'produce a literary work'
has_Event: {writing:2; authorship:2; penning:1} [00929718-n] *noun.act*
has_Agent: {writer:1; author:3} [10794014-n] *noun.person*
derivative: {pen:3} [03906997-n] *noun.artifact*

We assume that in many cases, the (morpho)semantic relations between the nouns may reflect the (morpho)semantic relations between the respective nouns and the verb, i.e., {writing:2} is an event nominal which has an **Agent** {writer:1}. This assumption, however sketchy, can be tentatively extended to other derivationally related nouns; thus, we can add a semantic relation **Instrument** to {pen:3}, which can be additionally related as an **Instrument** for {writing:2; penning:1} and an **Instrument** of {writer:1}:

Ex.:

{writing:2; authorship:2; penning:1} *noun.act*
has_Agent: {writer:1; author:3} *noun.person*
has_Instrument: {pen:3} *noun.artifact*

{writer:1; author:3} *noun.person*
has_Instrument: {pen:3} *noun.artifact*

Some noun synsets have been already linked via hypernym/hyponym relations, f.ex. {squandering:1} **is_hyponym_of** {waste:5; wastefulness:1}, and {wastrel:1; waster:2} *is_hyponym_of* {prodigal:2; profligate:3; squanderer:1}, and all of them are linked to the verb {consume:4; squander:1; waste:6}. Thus, the relation between them is overtly exposed though it can be categorised further.

In the following section, we propose a set of semantic relations that can be applied to the noun – noun pairs[4]

3.1 Noun – noun relations through verbs

As already stated, noun synsets that are derivationally related to a verb synset, can be linked through semantic relations that mirror (or are inherited from) the (morpho)semantic relations between noun and verb synsets on the basis of the assumption that a deverbal noun may inherit the argument structure of the source verb. Some noun – verb relations in WordNet are derivative only, but (morpho)semantic ones can be additionally formulated (see (Stoyanova et al., 2013).

Nouns of all semantic classes can be derivationally related to verbs, as in: cook: cooking (*noun.act*) is done by using a cooker (*noun.artifact*) as an **Instrument** by a cook (*noun.person*) as an **Agent**; toast: toasting (*noun.act*) is done by using a toaster (*noun.artifact*) as an **Instrument** to produce a toast (*noun.food*) as a **Result**. Further, a cook (*noun.person*) uses a cooker (*noun.artifact*) as an **Instrument** for cooking (*noun.act*); a toaster (*noun.artifact*) produces a toast (*noun.food*) as a **Result** when toasting (*noun.act*); etc. We have formulated a number of noun – noun relations, some of which such as Agent, Instrument, Result, Property, Location, mirror or are inherited from noun – verb (morpho)semantic relations; in some cases the type of relation was changed (Event can become Result) or additionally specified as with Resulting_State. There are also newly formulated relations such as Actor, Causator, Patient, Possessor, Experiencer, Cause, Time, etc. Relations are inverse, asymmetric and intransitive, e.g., **is_Agent_of / has_Agent**; **is_Subevent_of / has_Subevent**, etc.

The new relations assigned to nouns, may allow us to further assign semantic subclasses (re-

[4]The set is to be extended further but for now we cover only the main relations.

flecting their properties) to the nouns at hand. Thus, if a noun classified as *noun.person* is related via **Experiencer** relation, we may assume that it lacks properties like agentivity and control. Moreover, these properties would restrict the noun's properties that enable its participation in certain predicate-argument structures (if a noun is classified as *noun.object* or noun.artifact and is linked to other noun(s) via a **Location** relation, we may assume that it may also participate in **Location** relations with other verbs selecting a **Location** relation.

3.1.1 Noun – noun relations: an overview

We have manually assigned[5] the semantic relations to 2,303 noun – noun pairs.

Persons

A noun labeled as *noun.person* can express a variety of relations to verbs and deverbal nouns such as Agent, Causator, Experiencer, Recipient, etc. Other semantic classes here are *noun.group* and *noun.animal*.

The **Agent** relation (513)[6] is inherited from noun – verb relations and links nouns mostly classified as *noun.person* related via verbs of semantic classes such as *verb.creation*, *verb.motion*, *verb.change*, *verb.competition*. Nouns classified as *noun.person* have conscious and active referents, while the other noun in the pair refers to explicitly active predicates such as *noun.act*, *noun.event*, *noun.process*, *noun.communication*.

Ex.: {etcher:1} [10064977-n] **is_Agent_of** {etching:1} [00938791-n].

The **Actor** relation (174) links a noun which cannot be considered an active participant in the situation but refers to an entity who has abilities to perform the action referred to by the other noun (*noun.animal*s linked to verbs via **Agent** relation are marked as **Actors**). Ex.: {inhabitant:1} [09620078-n] **is_Actor_of** {inhabitation:1} [01054545-n].

In the **Causator** relation (34), the other noun refers to a resultative phenomenon such as *noun.event*, *noun.phenomenon*, *noun.motive*, etc.

Ex.: {bell ringer:3; ringer:4} [10714851-n] **is_Causator_of** {ring:12; ringing:3} [07391863-n].

Three relations are labeled according to a semantic role differentiated on the basis of the verb class, (morpho)semantic relations and the class of the other noun in a pair. The **Experiencer** relation (98) holds between a *noun.person* and a noun classified mostly as **noun.feeling** or *noun.state* via *verb.emotion*, *verb.perception*, *verb.body*.

Ex.: {lover:1} [09622302-n] **is_Experiencer_of** {love:8} [07543288-n]

Nouns that are linked via **Patient** relation (85) are related to the verb via an **Undergoer** relation and can be *noun.person* or *noun.animal*, and the other noun in the pair is *noun.feeling*, *noun.possession*, *noun.cognition*, etc.

Ex.: {beloved:2; love:9} [09849598-n] **is_Patient_of** {love:8} [07543288-n].

The **Recipient** relation (17) holds between a noun related to the verb via an **Agent** relation, and a noun labeled as *noun.food*, *noun.competition*, *noun.possession*, *noun.communication*, *noun.artifact*, etc., as in: {luncher:1} [10277132-n] **is_Recipient_of** {lunch:3; luncheon:1} [07575076-n].

The **Possessor** relation (17) involves a noun labeled *noun.attribute*, and more rarely a *noun.possession*, as in: {economiser:1} [10044470-n] **is_Possessor_of** {economy:2} [05644727-n].

In a previous effort ((Dimitrova, 2018)), (morpho)semantic relations **Agent** and **Undergoer** were subdivided to formulate additional (morpho)semantic relations between nouns and verbs such as **Experiencer**, **Actor**, **Recipient** to be applied to the Bulgarian wordnet. Inthere, the relation **Experiencer** surpasses the relation **Agent** with two verb classes – *verb.perception* and *verb.emotion*. However, observations on the data about noun – noun relations show that if a *noun.person* is related to *noun.feeling* and *noun.state*, it is most likely to be **Experiencer** (53) or **Causator** (21) especially if linked via *verb.emotion* and *verb.body*. If a *noun.person* is linked to *noun.state*, it can be also **Patient**, **Possessor**, and **Actor** (e.g., {suspect:6} [10681383-n] **is_Patient_of** {suspicion:4} [13982839-n].

The **Agent** relation, however, still holds between *noun.person* and *noun.act* disregarding the class of the verb: a *noun.person* which is linked to a *noun.act* via *verb.cognition* is most likely to be **Agent** as referring to a person in professional function.

[5] For the resource, see: https://dcl.bas.bg/semantichni-mrezhi/ , with any further additions and changes.

[6] Due to space limitation, only the total number of relations added is given in brackets here.

A noun labeled as *noun.person* is most likely a **Possessor** or a **Recipient** in relation to *noun.possession* (esp. when linked via *verb.possession*).
Thus, one may assume that if a *noun.person* is related to other nouns of classes such as *noun.feeling* and *noun.state* via Experiencer relation, it may lack properties such as agentivity and control (a sleeper may snore (just like a snorer) but cannot read or drive a car).
In addition, there are nouns classified as *noun.group* which are linked via **Agent** or **Patient** relation, as in: {mover:1; moving company:1} [08478482-n] **is_Agent_of** {move:16} [01850315-v]. Here, we may assume that the group and/or its members have properties characteristic of a person.

Artifacts

A *noun.artifact* refers to non-animate nouns and is linked with **Instrument** (166) relation to nouns of all other classes but mostly predicative ones, as in:
Ex.: {printer:2} [-04004767-n] **is_Instrument_of** {printing:4; printing process:1} [06677302-n]
{machinist:1; mechanic:3} [10279018-n] **has_Instrument** {machine:4} [03699975-n]
The *noun.artifact* is usually linked to the verb synset via *Instrument* or *Means* (morpho)semantic relations.
noun.artifact can be also Result of a *noun.act*, as in:
excavation:3 [03302121-n] **is_Result_of** *excavation:2; digging:1* [00941974-n]
Another relation that can link a *noun.artifact* and a *noun.act* is **Theme** (306) as in:
{piece:9} [03932203-n] **is_Theme_of** {patching:1} [00267349-n]
The **Theme** relation often links non-animate nouns related to the verb via an **Undergoer** relation (and (*Uses*) which was subdivided into **Theme** and **Patient** depending on the characteristics of the noun's referent (a non-animate noun such as *noun.food*, *noun.plant*, etc. would be *Theme*, while animate and human nouns would be *Patient*), as in:
Ex.: {draft:12; tipple:2} [07883980-n] **is_Theme_of** { tippler:1; social drinker:1} [10712690-n]
{plant:1; flora:1} [00017222-n] **is_Theme_of** {planting:1} [00919513-n]
Most noun – noun pairs linked via **Instrument** relation contain a noun classified as *noun.artifact* – these nouns are related to verbs via **Instrument** and **Vehicle** (morpho)semantic relations. Nouns classified as *noun.substance* are linked to verbs via **Material** and **Uses** relations. In these cases, a *noun.substance* refers to a man-made entity.
If a noun is classified as *noun.object* and is linked to *noun.act*, *noun.event* or *noun.state*, it may be **Theme** (21) and **Result** (25) but also **Location** (11) and **Uses** (9); if it is linked to *noun.act* and *noun.state* via the same verb, it is **Result** of *noun.act* and **Theme** of **noun.state**. One may also assume that *noun.artifact* can be argument of various predicates (a cooker can be an **Instrument** of cooking (but also, indirectly, of frying or boiling) but also a **Location** of putting, or a **Theme** of repair, or a **Result** of producing, etc.).

Events

A noun – noun relation that is mostly inherited from the noun – verb relation is **Result** (219) which holds between a noun labeled as *noun.artifact*, *noun.food*, *noun.object*, etc. (linked to the verb synset via the (morpho)semantic relation **Result**) and a *noun.act*.
Ex.: {toast:3} [07686873-n] **is_Result_of** {toasting:1} 00246552-n
The subcategorised relation **Resulting_state** (89) holds between a noun classified as *noun.state* or *noun.feeling* and nouns of various classes such as *noun.state*, *noun.feeling*, *noun.event* via *verb.perception*, *verb.emotion*, *verb.change*, *verb.body* classes.
Ex.: {disturbance:7; upset:17} [14403282-n] **is_Resulting_state_of** {upset:4} [00554850-n]
The type of the relation can be changed, as in: {snap:23} [07394236-n] **is_Result_of** {snap:4} [00344699-n] (the noun – verb relation was **Event**).
A new relation that encodes the relation between two predicative nouns is **Subevent** (144) – it mostly holds between a noun referring to the act as such and a noun which may refer to the beginning, the end or any moment in-between the starting and ending point. This relation often holds between *noun.act* and *noun.event*, with the former referring to an event within the act, and between *noun.process* and *noun.act* assuming that a process consists of a series of acts. An example here is: {start:20} [07325190-n] **is_Subevent_of** {beginning:1; start:1} [00235435-n]. The as-

sumption that the lexical inheritance condition is valid here, would mean that any Subevent may have **Agent** or **Instrument** of the main event, e.g., if {barrage:2; bombardment:3} [00987863-n] **has_Agent** *blaster:1; chargeman:1* [09859557-n], and **has_Instrument** {shell:12} [04190464-n], its Subevent *blast:15* [07408171-n] would inherit these relations, and any of the verbal predicates related to the verb {blast:6; shell:4} [01135922-v] such as its hyponym {crump:2} [01136393-v] and its hypernym {bombard:3; bomb:1} [01131902-v], may select for arguments the nouns at hand (i.e., the person *blaster* as an **Agent**, the artifact *shell* as an **Instrument**, and the event *blast* as a **Subevent**).

Others

The relation **Location** (121) links nouns classified as *noun.location*, *noun.object*, and *noun.artifact* with *noun.process*, *noun.act*, *noun.state* via *verb.stative*, *verb.motion*, *verb.body* through **Location** and **Event** (morpho)semantic relations: {hatchery:1} [08581299-n] **is_Location_for** {hatch:8; hatching:2} [13491464-n]

Nouns labeled *noun.object* or noun.artifact can be linked not only to verbs but to other noun(s) via **Location** relation prompting an assumption that the noun classified as *noun.artifact* may also participate in **Location** relations with other verbs selecting a **Location** relation (a person can be hospitalised in a hospital as a **Location** but can also live or dance (however unusual it may seem) in a hospital as a **Location**).

The relation **Uses** (176) holds between nouns that refer to all non-human and non-predicative referents such as *noun.substance*, *noun.artifact*, including *noun.animal*, as in: {hawker:1} [10076604-n] **Uses** {hawk:3} [01605630-n])

The relation **Cause** (63) holds between a *noun.phenomenon* or *noun.motive* and a *noun.act*, *noun.process*, *noun.event*, etc., as in: {soaker:2} [11502102-n] **Causes** {drenching:1; soaking:2} [00277811-n]

The relation **Property** (52) links a noun classified as *noun.attribute* to a noun of any other class, as in: {invalid:5; shut-in:3} [10214230-n] **has_Property** {disability:1; disablement:1} [14548343-n], and this property may be characteristic of many other nouns of the same class (a chief executive can has a disability).

The relation **Time** (29) holds between a *noun.time* and a *noun.act*, *noun.process*, etc., as in: {period of play:1; play:52} [15256915-n] **is_Time_for** {playing:1} [00041188-n].

3.2 Case study

Here, we offer some observations on co-occurrence between the classes of nouns in a pair. We have manually assigned relations on noun – noun pairs linked via *verb.perception*, *verb.competition*, and *verb.consumption*. In Table 1, we give figures on *noun.person*s.

Noun.person are often Agents with *noun.act*, and Experiencers with *noun.feeling* and *noun.state*, and they Uses (incl. as Instruments) *noun.artifact*s. Further, with *verb.perception* and *verb.competition*, *noun.event* is Subevent and Result of *noun.act*, while *noun.act* is Subevent of *noun.process*. With *verb.consumption*, *noun.event*s (4) are much rarer.

Nouns labeled *noun.food* and *noun.artifact* are often Themes of *noun.act* when the two are linked via *verb.consumption*.

verb.perception

noun.class	noun.class	Rel [No]
person	act	Agent [45]
person	event	Causator [4]
person	communication	Agent [3], Actor [2]
person	feeling	Agent [3]
person	state	Experiencer [4]
person	cognition	Agent [4], Experiencer [5]

verb.consumption

person	act	Agent [29], Actor [7], Experiencer [1]
person	quantity	Agent [1]
person	cognition	Experiencer [1]
person	state	Experiencer [2], Actor [2]
person	feeling	Experiencer [2]

verb.competition

person	act	Agent [55], Actor [20], Recipient [2], Causator [1]
person	animal	Theme [4], Uses[2]
person	artifact	Uses [10], Theme [2], Instrument [4]

Table 1: *Noun.person* linked via *verb.perception*, *verb.consumption*, and *verb.competition*.

The Location relation links nouns classified as *noun.location* and *noun.artifact* with *noun.act*. (The (morpho)semantic relation Location is rarely found with the three verb classes.)
The observations on noun – noun relations may help us formulate some principles behind combinations between a semantic relation, a verb synset of a particular semantic class, and a set of noun synsets from other classes that are indirectly linked though a verb via derivative and morphosemantic relations. If we assume that the nouns linked to verbs are arguments to a predicate, the features associated with a particular concept in argument position, can be inferred also by observing other nouns linked to the same verb.

4 Nouns linked via adjective synsets

An adjective denotes a property that is permanently inherent for an entity it modifies or refers to and is attributed to it in its entirety. Therefore, an adjective can be defined as part-of-speech whose denotative function is realised through its connection to the noun. Adjectives and nouns in WordNet are linked to each other mostly via derivative relations. Descriptive adjectives (*adj.all*) are organised into clusters based on similarity of meaning (synonymy) and binary opposition (antonymy). Relational adjectives (*adj.pert*) are (derivationally) related and linked to the synset which contains their source noun (as a literal). Adjectival participles (*adj.ppl*) are related via participle relation to verbs they are derived from. Thus, adjectives are organized via a set of relations that encode their properties of attribution, antonymy, similarity, derivation; fuzzynymy and thematic category (in the EuroWordNet (Vossen, 2002).
However, from a derivational point of view, the distinction between descriptive and relational adjectives can be somewhat fuzzy, as descriptive adjectives can be also derived from nouns and refer to an attribute property of the defined entity (expressed by the noun). The property qualifies and characterises the entity expressed by the noun from which they are derived (e.g., *pitiful* - *pity*, etc.). Hence, an adjective may express one-sided relationship with the entity denoted by the motivating noun, though adjectives, which are derived from a noun, are motivated by it. In WordNet, an explicit noun – adjective relation with relational adjective (*adj.pert*) is pertainymy – an antisymmetric (derivative) relation between a relative adjective and the noun from which it is derived. The basic meaning of the relational adjective is determined by the noun from which it is derived, and these adjectives may inherit relations from the noun (Koeva, 2014). Some descriptive adjectives in WordNet may not be linked via pertainymy relation but can be derivationally related to a source noun.
We have extracted noun synsets which are indirectly linked via adjectives – a noun is derivationally related to an adjective which, in its turn, is related via similarity relation to another adjective which is related to another noun. We applied the following scheme of extracted nouns:
Noun *derivative* **Adjective** *similar_to* **Adjective** *derivative* **Noun**.
An example is given below where a noun – noun relation is assumed between {north wind:1; northerly:4; norther:1} and {north:3}.
Ex.:
{north wind:1; norther:1} [11487950-n] *noun.phenomenon*
derivative: {northerly:2; northern:1} [01601069-a]
similar_to: {north:2}
has_attribute: {north:3} 08561081-n *noun.location*
{north wind:1; norther:1} **is_Related_to** {north:3}

Some of these noun – noun pairs contain literals that are derivationally related (literals have the same root of at least one of the literals in the synset) though the synsets are not explicitly related via derivative relation; with others, only the adjectives are derivationally linked. We have identified only 31 noun – noun pairs that have at least one literal that is derivationally related, as in the example below.
Ex.
{salinity:1} [04993604-n] *noun.attribute*
derivative: {saline:1} [01074458-a]
similar_to: {salty:1} [01073822-a]
derivative: {salt:7; table salt:1} [07813107-n] *noun.food*

We have attempted to explore the dependence between the semantic classes of the nouns that are indirectly related via adjectives linked via similarity relation, to formulate noun – noun relations which were experimentally applied.

4.1 Noun – noun relations through adjectives

The majority of noun – noun pairs here contain literals that are not derivationally related – 1,193 pairs – but noun synsets are otherwise related through derivationally related adjectives, as exemplified below.
Ex.:
{ceremony:1} [01026897-n] *noun.act*
derivative: {ceremonial:1} [01042491-a]
similar_to: {formal:2} [01041916-a]
has_attribute: {formality:2; formalness:1} [04911420-n] *noun.attribute*
We have formulated four noun – noun semantic relations mostly drawing upon classes and definitions of the nouns. Here, we exemplify the co-occurrence of noun semantic classes that are most often found in our data. For a cleaner representation of dependencies between semantic classes of nouns we will present them in separate groups acc. to the formulated relations.

Result is a relation referring to a consequence of performing any action, process, event. Here, nouns classified as *noun.act* can express Result of *noun.artifact*[3][7], *noun.attribute* [33], *noun.cognition* [4], *noun.feeling* [4], etc. For example, {empiricism:2} [00635699-n] *noun.act*, which is **derivative** of : {empirical:1; empiric:1} [00858917-a] – **similar_to**: {experiential:1; existential:1} [00859632-a], has non-explicit relation with {experience:6} [05758059-n] *noun.cognition*. Hence, we can link {empiricism:2} with the relation **is_Result_of** to {experience:6} and formulate dependence of the type: *act* **Result** *cognition*, which means that an action can be a Result or can lead to a certain result of knowledge.
Nouns labeled as *noun.event* can be Result of *noun.attribute*s [11]. For example {discharge:17; outpouring:3; run:49} [07407777-n] *noun.event* **is_Result_of** {fluidity:2; fluidness:2; runniness:1} [04937043-n] *noun.attribute*.

Property is a relation that links nouns referring to concepts that are considered to be characteristic of another noun mostly classified as *noun.attribute* (but also *noun.state*, *noun.feeling*).
Nouns labeled as *noun.animal* are characterised by properties classified as *noun.attribute* [9] which are not obligatorily associated with the animal (body part). For example, {scale:5}[8] [01902877-n] *noun.animal* **has_Property** {roughness:3} or *animal* **has_Property** of some *attribute*.
Nouns classified as *noun.attribute* are Properties of *noun.act* [13], *noun.artifact* [8], *noun.cognition* [31], *noun.communication* [7], *noun.person* [11], *noun.state* [33], *noun.feeling* [21]. For example, {neurotic:3} [10354898-n] *noun.person* **has_Property** {obsessiveness:1} [04626062-n] *noun.attribute*.
Nouns classified as *noun.body* has property of nouns labeled as *noun.attribute* [12], *noun.state* [3]. So {fuzz:1} [05261894-n] *noun.body* **has_Property** {hairiness:1} [04683453-n] *noun.attribute*
Noun.state is property of nouns classified as *noun.feeling* [3] and *noun.person* [14]. For example, {subservience:2; subservientness:1} [13952466-n] *noun.state* **is_Property_of** { slave:2} [10609325-n] *noun.person*.
Nouns labeled as *noun.plant* [7], *noun.quantity* [4], *noun.shape* [11] have properties marked as *noun.attribute* like in the case of the example {thorn:3; prickle:4} [13089631-n] *noun.plant* **has_Property** {sharpness:3; keenness:1} [04705324-n] *noun.attribute*
Nouns classified as *noun.person* is characterised by *noun.attribute* [37], *noun.cognition* [4] or *noun.state* [6], e.g.: {teenager:1} [09772029-n] *noun.person* **has_Property** {youngness:1} [04928416-n] *noun.attribute*.

Part_of is a relation which links nouns referring to concepts as constituent elements of other concepts. This is a relation linking a noun referring to an event or entity which are associated with another event or entity. In this case **Part_of** is more often related to abstract nouns such as event and entity than to nouns having separate components as in the examples: 'the finger is part of the hand'; 'this piece is part of the pie', where the meronymy relation is to be applied.
Nouns labeled as *noun.communication* can be Part of *noun.cognition* [4] or *noun.attribute* [30], as in: {irony:3} [07106246-n] *noun.communication* **is_Part_of** {incongruity:1; incongruousness:1} [04714847-n] *noun.attribute*.

[7]The number in brackets shows the occurrences of the noun pairs.

[8]Here, we give only noun – noun pairs due to limitation of space.

Related is a general relationship that shows that there is connectivity between different objects, phenomena, dimensions but it is more of a free association relation that has not been properly defined yet.
Nouns of semantic class *noun.cognition* are related to *noun.attribute* [47], *noun.person* [4], *noun.state* [5]. For example {insightfulness:1} [05621808-n] *noun.cognition* **is_Related_to** {perceptiveness:1} [04843875-n] *noun.attribute*.
Nouns labeled *noun.feeling* are related to nouns of classified as *noun.attribute* [8] or *noun.state* [9], as in: {uneasiness:3} [07507329-n] *noun.feeling* **is_Related_to** {discomfort:2} [14446652-n] *noun.state*.
Nouns classified as *noun.food* are related to nouns classified as *noun.attribute* [14], *noun.substance* [3], as in: {fizz:2} [07919310-n] *noun.food* **is_Related_to** {bubbliness:1; frothiness:1} [04733347-n] *noun.attribute*.
Nouns labeled as *noun.object* are related to concepts classified as *noun.attribute* [10]: {reef:5} [09406793-n] *noun.object* **is_related_to** {shallowness:2} [05135725-n] *noun.attribute*
Noun.substance and *noun.time* are related to *noun.attribute* [25, 12] or *noun.state* [3, 2]: {vapor:2} [15055633-n] *noun.substance* **is_Related_to** {cloudiness:3} [14524198-n] *noun.state*

Considering the observed results, some dependencies have been formulated, which for the moment copy the information from the semantic classes of the related nouns:
act_Result_attribute [31];
attribute_Property_state [33];
attribute_Property_cognition [31];
attribute_Property_act [13];
attribute_Property_feeling [21];
body_Property_attribute [12];
state_Property_person [14];
shape_Property_attribute [11];
person_Property_attribute [37];
cognition_Related_attribute [47];
substance_Related_attribute [25];
time_Related_state [12].

To sum up, nouns, which refer to an attribute may be a result of a certain act, as well as a property of or related to a particular shape, person, physical body, cognition or substance. Further, they may have certain properties of state, cognition, act or feeling. Nouns for state are properties of a person, while nouns that indicate time may be related to a particular state. Some of these relations such as Property and Result can be traced back to noun – noun pairs linked via verbs, hence they may further deepen the lexical-semantic inter-relatedness.

5 Conclusion

The paper offers an approach to identification of semantic relations between nouns in WordNet that are indirectly linked via derivative relations through verbs and adjectives. In many cases, the derivationally related nouns preserve the semantics of the verb and the adjective, though there are some restrictions. We have formulated a basic set of semantic relations which mostly repeat the knowledge encoded on different levels of the network. Noun – noun relations also reflect certain restrictions on nouns that are related to verbs of certain classes. The new relations assigned to nouns, will not only increase the inter-relatedness and density of WordNet relations but would allow us to assign new semantic properties to nouns. The work will continue with extending both the number of related noun – noun pairs and the set of the semantic relations formulated.

Acknowledgments

The work is funded under the project "Towards a Semantic Network Enriched with a Variety of Relations" (DN 10-3 / 14.12.2016), financed by the Bulgarian National Science Fund (BNSF).

References

Irina Azarova and Anna Sinopalnikova. 2004. Adjectives in RussNet Proceedings of the Global WordNet Conference'2004, 251 – 258.

Tsvetana Dimitrova. 2018. Morfosemantichni relatsii i agentivni sashtestvitelni v Balgarskiya uardnet. Balgarski ezik, 65(2): 41 – 58.

Tsvetana Dimitrova, and Valentina Stefanova. 2018. Semantic Classification of Adjectives in the Bulgarian Wordnet: Toward a Multiclass Approach Etudes Cognitives, 18: 1 – 17.

Christiane Fellbaum, Anna Osherson, and Peter E. Clark. 2009. Putting semantics into WordNets morphosemantic links *Proceedings of the Third Language and Technology Conference*, Volume 5603, 350 – 358.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a lexical-semantic net for German. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9 – 15.

Svetla Koeva. 2014. WordNet i BulNet. Ezikovi resursi i tehnologii za balgarski ezik, 154 – 173.

Svetla Koeva, Svetlozara Leseva, Ivelina Stoyanova, Tsvetana Dimitrova, and Maria Todorova 2016. Automatic prediction of morphosemantic relations *Proceedings of the Eighth Global Wordnet Conference*, 168 – 176.

Svetlozara Leseva, Ivelina Stoyanova, Maria Todorova, Tsvetana Dimitrova, Borislav Rizov, and Svetla Koeva. 2015. Automatic classification of WordNet morphosemantic relations. The 5th Workshop on Balto-Slavic Natural Language Processing, 59 – 64.

Svetlozara Leseva, Ivelina Stoyanova, Hristina Kukova, and Maria Todorova. 2018. Integrirane na subkategorizatsionna informatsia v relatsionnata struktura na UardNet. Balgarski ezik, 65(2): 11 – 40.

Marek Maziarz, Stanisaw Szpakowicz and Maciej Piasecki. 2012. Semantic relations among adjectives in Polish WordNet 2.0: A new relation set, discussion and evaluation *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Cognitive Studies / tudes Cognitives*, Volume 12, 149 – 179.

George A. Miller. 1990. Nouns in WordNet: a lexical inheritance system. International journal of Lexicography, 3.4: 245 – 264.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to WordNet: an On-line Lexical Database. Five Papers on WordNet Princeton, NJ: Princeton University.

Valeria de Paiva, Livy Real, Alexandre Rademaker, and Gerard de Melo, Gerard. 2014. NomLex-PT: A Lexicon of Portuguese Nominalizations. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), 2851 – 2858.

Livy Real. 2014. Nominalizacoes. Ph.D. thesis, Universidade Federal do Parana, Curitiba, Brazil.

Livy Real and Alexandre Rademaker. 2015. An overview on Portuguese nominalisation. Workshop on Type Theory and Lexical Semantics, 119 – 128.

Valentina Stefanova and Tsvetana Dimitrova 2017. Classification of adjectives in BulNet: Notes on an effort *Proceedings of the Challenges for Wordnets Workshop within the First International Conference, LDK*, Volume 1899, 188 – 196.

Ivelina Stoyanova, Svetla Koeva, and Svetlozara Leseva. 2013. Wordnet-based Cross-language Identification of Semantic Relations. Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, 119 – 128.

Piek Vossen. 2002. EuroWordNet General Document. Version 3. Final. July 1, 2002 University of Amsterdam.

Linking Russian Wordnet RuWordNet to WordNet

Natalia Loukachevitch
Lomonosov Moscow State University
Moscow, Russia
Louk_nat@mail.ru

Anastasia Gerasimova
Lomonosov Moscow State University
Moscow, Russia
anastasiagerasimova432@gmail.com

Abstract

In this paper we consider the linking procedure of Russian wordnet (RuWordNet) to Wordnet. The specificity of the procedure in our case is based on the fact that a lot of bilingual (Russian and English) lexical data have been gathered in another Russian thesaurus RuThes, which has a different structure than WordNet. Previously, RuThes has been semi-automatically transformed into RuWordNet, having the WordNet-like structure. Now, the RuThes English data are utilized to establish matching from the RuWordNet synsets to the WordNet synsets.

1 Introduction

The Princeton WordNet thesaurus (Fellbaum, 1998, Miller, 1998) created for the English language is one of the most popular linguistic resources used in natural language processing. In many countries their own projects on creating WordNet-like resources (wordnets) for national languages have been initiated (Vossen, 1998).

The Open Multilingual WordNet project is currently being developed (Bond and Paik, 2012; Bond and Foster, 2013; Rudnicka et al., 2017). The goal of the project is to link together the existing wordnets created for different languages with an open license[1]. To connect a new language to the project, it is necessary to associate synsets of this language with WordNet synsets and present the data in the required format.

Sources of links of a specific wordnet to English synsets of Princeton WordNet can be different (Vossen, 1998; Piante et al., 2002). Some wordnets have been developed with semi-automatic translation of Princeton WordNet synsets, and therefore these links exist from the beginning. The creators of the Finnish wordnet (FiWN) translated Princeton WordNet manually, using the work of professional translators. As a result, the Finnish wordnet was created on the basis of translation of more than 200 thousand word senses of Princeton WordNet words within 100 days (Lindén and Niemi, 2014). Other wordnets are developed from scratch using own-language text corpora and dictionaries (Rudnicka et al., 2017). In such cases, their linking to WordNet synsets should be organized as a special procedure based on bilingual dictionaries and expert verification.

In the current study, we describe another way of aligning the Russian wordnet (RuWordNet) and WordNet synsets. RuWordNet was semi-automatically generated from another Russian thesaurus RuThes, which is being developed for more than 20 years (Loukachevtich et al., 2018; Kirillovich et al., 2017). For bilingual text processing, the RuThes concepts also have English representation. This English part of the RuThes thesaurus has been collected from various sources, including several text collections (news articles, European Community documents, etc.), English and Russian-English dictionaries, and others. Currently, the RuThes concepts have more than 140 thousand English text entries. In the paper we describe the process of linking RuWordNet with WordNet, which exploits the previously gathered bilingual data in RuThes.

The paper is structured as follows. In Section 2 we consider related work. Section 3 describes RuWordNet thesaurus and its source - RuThes thesaurus, including representation of bilingual Russian-English lexical units and phrases. Also the general scheme of links. In Section 4 we consider the general scheme of linking RuWordNet and WordNet using RuThes bilingual data. Section 5 presents two main steps of linking RuWordNet and WordNet: automated linking through RuThes bilingual information and manual linking of WordNet core concepts.

[1] http: // compling.hss.ntu.edu.sg/omw/

2 Related Work

For the first time, the idea of linking wordnets was proclaimed in EuroWordNet project (Vossen, 1998). In order to establish communication between different languages, the synsets of each wordnet should refer to the so-called interlingual index (ILI), for which the Princeton WordNet synsets were used. The index is an unordered list of synsets with glosses. To accurately describe the correspondence of specific synsets of each language and overcoming lexical gaps that may arise in a particular language, several different equivalence relations from synsets of a specific language to the ILI index were proposed: synonym, near-synonym, hyperonym, hyponym.

Christea et al. (2004) list the main problems of linking English-language WordNet and another wordnet using Romanian wordnet (Tufiș et al., 2013) as an example. The first type of difficulties is related to the fact that potential matches in WordNet correspond to several synsets denoting similar senses, and the explanations of synsets are very similar. Additional analysis is needed to choose the most appropriate synset.

The second type of problems is associated with the absence of lexicalized means of naming a concept denoted by the English synset. In such cases, an additional synset is introduced into the Romanian wordnet, which contains a non-lexicalized expression. The next type of problems stems from the fact that the word sense system in the English WordNet is more fractional than in the Romanian wordnet. In such cases, new senses were entered into the Romanian wordnet.

Linking between Polish wordnet (plWordNet) and WordNet was performed in 2012 (Rudnicka et al., 2012). To establish links, the following set of interlingual (I) relationships was used: I-synonymy, I-hyponymy, I-hyperonymy, I-meronymy, I-holonymy, I-quasi-synonymy (near synonymy), I- inter-register synonymy. The latter relation is established when the synsets in Polish and English have the same meaning, but refer to different language registers. The matching between the Polish and English synsets was performed manually. In the process of searching for equivalents, inaccurate descriptions of Polish word senses could be corrected.

Maziarz et al. (2013) provide quantitative characteristics of the established relations: the I-hyponymy relation was the most frequent link between synsets of WordNet and plWordNet. This can be explained by the existence of a large number of lexical and cultural lacunae, greater lexicalization of the category of gender in the Polish language (for example, for the names of roles, posts of people), the use of diminutive names in Polish, etc.

3 RuWordNet Thesaurus

The Russian wordnet RuWordNet (Loukachevitch et al., 2016; Loukachevitch et al., 2018) has been created on the basis of another Russian thesaurus RuThes in 2016 (Loukachevitch, Dobrov, 2002).

Main units of RuThes are concepts, each concept has a monosemous and clear name and the set of text entries that convey the corresponding concept in texts. The text entries of a concept can include single words of different parts of speech, multiword expressions and also compositional phrases, with the same meaning. To represent bilingual data, the RuThes concept has the English name of concept and the set of English text entries with the same variety of text entries.

To create RuWordNet, the RuThes data were transformed: the concepts were subdivided to part-of-speech-related synsets and traditional WordNet-like relations were established between the synsets. Table 1 presents the quantitative characteristics of synsets and language units in RuWordNet.

Further we consider the organization of English part in the RuThes because we use these data for linking RuWordNet and WordNet.

Part of speech	Number of synsets	Number of unique Russian entries	Number of senses
Noun	29,296	68,695	77,153
Verb	7,634	26,356	35,067
Adj.	12,864	15,191	18,195

Table 1. Quantitative characteristics of the synsets and Russian entries in RuWordNet

3.1 RuThes as a Bilingual Resource

RuThes is a linguistic ontology presented as a hierarchy of concepts. Each concept has a unique name in Russian and in English (if existing). A concept is associated with a set of Russian text entries and English text entries.

Text entries of the same concepts in both languages can include single words of different parts of speech, multiword expressions, and compositional phrases that can express this con-

cept. Current volume of RuThes is more than 60 thousand concepts, 200 thousand Russian text entries and 146 thousand English text entries.

The English text entries were collected for many years from several sources, including bilingual dictionaries, analysis of English documents in various projects, such as knowledge-based text categorization.

During last years, each new concept introduced into RuThes is provided with the English name and English text entries, if they exist. These English translations are specially searched in bilingual resources or translated with online-translation services. Then all English variants are verified on Internet-pages to check if they really exist and express the intended senses, because any found translations can be incorrect.

Besides direct translations, also cross-category synonyms are added as text entries, for example, adjective or verb derivations expressing the same concept. Additionally, multiword phrases expressing the same concept are searched for and introduced, because for various applications it is important to match a thesaurus concept in texts using its variant forms.

For example, for concept *ПРОМЫШЛЕННОСТЬ (promyshlennost')/ INDUSTRY* the following English text entries have been introduced: *industry, industrial, industrial sphere, sphere of industry.* From this example, the importance of adding such multiword variants can be seen: they are unambiguous, but their components have several senses.

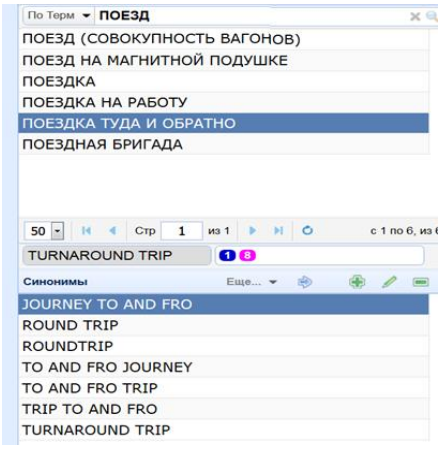


Figure 1. English text entries for the RuThes concept *ПОЕЗДКА ТУДА И ОБРАТНО* (*TURNAROUND TRIP*)

Figure 1 shows English variants collected for the RuThes concept *ПОЕЗДКА ТУДА И ОБРАТНО* (*TURNAROUND TRIP*). It could be noted that corresponding synset in WordNet contains only the *round trip* lexical entry.

Figure 2 demonstrates English text entries for the RuThes concept *ПОЕЗДКА НА РАБОТУ* (*COMMUTE TO WORK*). In WordNet word *commute* has 1 noun sense and 5 verb senses, which means that this word can be quite difficult for word sense disambiguation. But when we introduce unambiguous variant phrases *commute for work* and *commute to work*, we provide reliable way to detect this concept in texts because these phrases are quite frequent according to Google (*commute for work* – 143 thousand pages, *commute to work* – 12 mln. pages).

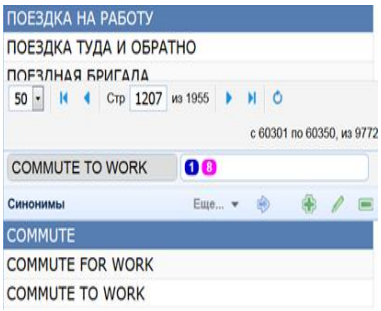


Figure 2. English text entries for the RuThes concept *ПОЕЗДКА НА РАБОТУ* (*COMMUTE TO WORK*)

RuThes is a Russian-oriented resource. In such cases when a single Russian word corresponding to an English word sense is absent, the following solutions can be made:

- If the sense can be expressed with an existing Russian phrase (multiword expression or a compositional phrase) then an additional concept can be introduced,
- in other cases, such English word can be attached to the closest RuThes concept. For example, English word *watch* (portable timepiece) is linked to the RuThes concept *ЧАСЫ* (*TIMEPIECE*) (Figure 3)

On Figure 3 the upper left form contains a list of concepts with "часы" substring. The lower left form shows text entries for the highlighted concept. In the middle between these forms, the English concept name (*TIMEPIECE*) can be seen. The right upper form presents the relations of the highlighted concept.

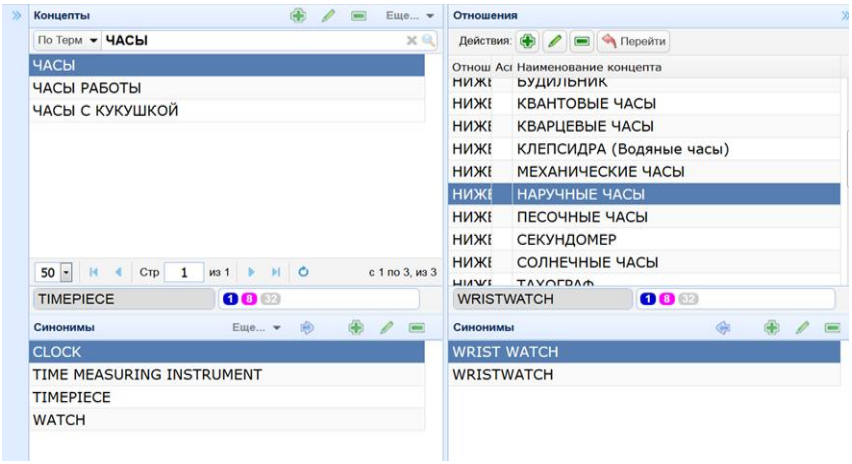


Figure 3. The differences in conceptualization of timepieces in Russian and English: there is no Russian word for English *watch*, as a portable timepiece

The low right form of Fig. 3 describes text entries of the highlighted concept *НАРУЧНЫЕ ЧАСЫ* (*WRIST WATCH*).

4 General Scheme of Linking RuWordNet to WordNet

The synsets of RuWordNet contain reference links to RuThes concepts from which these synsets were generated. Therefore English text entries collected in the English part of RuThes now can be used for matching RuWordNet and WordNet synsets.

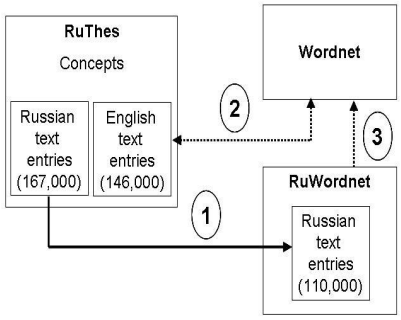


Figure 4. The scheme of linking RuWordNet to WordNet through the RuThes concepts with English text entries

Figure 1 shows the connections between the resources. Initially, thesaurus RuThes has been created. Most concepts of RuThes have Russian and English names and Russian and English text entries. Then the Russian part of RuThes was semi-automatically transformed to the WordNet-like thesaurus RuWordNet (**link 1**). Currently, we are semi-automatically creating links between the English part of RuThes and the WordNet synsets (**link 2**). From these two procedures, we obtain links from the RuWordNet synsets to the WordNet synsets (**link 3**).

5 Linking Procedure

The process of linking of WordNet and RuWordNet synsets includes two parts:

- Automatic matching the RuThes English entries with the WordNet units with further validation by experts and the transfer of the Russian established link from RuThes to RuWordNet, which has direct correspondence with RuThes,
- Analysis of the core wordnet synsets (Boyd-Graber et al., 2006), which are considered to be frequent and most salient. The task of the analysis is to check if the English-Russian links were established, or some corrections are needed, or the link cannot be established because of the absence of proper lexicalization in Russian.

Currently, I-S (inter-language synonym) and I-NS (inter-language quasi-synonym) are established between WordNet and RuWordNet synsets (through RuThes concepts). The relationship of interlanguage synonymy is established if the synset and concept have very close sets of denotations, but there are some features of the

word meanings that are different in the two languages

In subsections we consider these two procedures and their results.

5.1 Linking translated RuThes Concepts

English text entries of RuThes were automatically matched with WordNet entries. Table 2 shows the main types of situations that occurred as the result of the performed matching for nouns. Let us consider some examples for each type of linking of the RuThes concepts and WordNet synsets.

Type 1.1. (one-to-one) links are usually represented by the concepts of certain domains, for example, chemistry (*hydrogen, helium*), finance (*credit system, central bank*), politics (*communist party, iron curtain*), medicine (*thrombophlebitis, bronchial asthma*), geographical names (*Minsk, White sea*), names of animals and plants, etc.

Types of matching between RuThes concepts and WordNet noun synsets	Number of RuThes concepts
1. RuThes concept has only single English text entry, among them:	**9,629**
1.1. One-to-one matching with WordNet synset	1,373
1.2. One-to-many matching with WordNet synsets	4 ,935
1.3. No matching with WordNet synsets	3,803
2. RuThes concept has several English text entries, among them:	**19,715**
2.1. Only one English text entry has single matching with a WordNet synset	4,343
2.2. Several English text entries correspond to monosemous WordNet units	3,344
2.2.1. Several English text entries mainly match with one of the WordNet synsets	1,611
2.3. Several text entries and all their matches with WordNet are ambiguous	4,425
2.4. Several English text entries but none of them matches with WordNet units	5,589

Table 2. The quantitative results of automatic matching English text entries in RuThes and the WordNet synsets

As an example of the **1.2 type of links,** the word *energy* can be considered, which is the only option in RuThes for the concept *ENERGY* as a physical characteristic, and also corresponds to the concept *HUMAN ENERGY* in the group of synonyms (*energy, human energy, life energy, vigor, vigor*).

In WordNet, the word *energy* is included into 7 synsets of nouns, one of which obviously corresponds to the physical meaning of the word energy (as in RuThes). One of the senses in WordNet corresponds to energy as a specific state of mind, enthusiasm. This sense clearly exists in Russian, but is absent in RuThes, and should be added.

Therewith, the word *energy* is attributed by the authors of WordNet to the synset: *Department of Energy, DOE (Department of Energy, United States; created in 1977).* In RuThes, there is a similar entity, called *Министерство топлива и энергетики* (*Ministry of Fuel and Energy*) with the translations: *Department of Energy, Energy department,* etc, but the text entry *energy* is absent. In this case, the RuThes concept and the WordNet synset will be matched by other text entries (**type of comparison 2.3.).**

Some of the RuThes concepts and WordNet synsets cannot be matched, when a WordNet synset includes only single words, but in RuThes the related concept is linked only with phrases as text entries. For example, for the RuThes concept *ЗОЛОТОЙ ЦВЕТ* (*golden color*) there is a direct analogue in WordNet, namely synset: *(n) amber, gold (a deep yellow color).* However, RuThes contains only English noun phrases as text entries: *golden color, gold color, golden colour, gold colour*.

The above-mentioned example of the synset *amber, gold* also demonstrates another problem, which arises from the comparison of two thesauri for different languages, namely the differences in conceptualization, i.e. what exactly is considered in each resource to be the same concepts, and what is considered to be different. Conceptualization may be erroneous in one of the resources. In some cases it may be not clear enough how it is better to divide words into synsets (attributed to concepts).

The unified synset *amber, gold* in WordNet means that the concepts of golden and amber colors are united in WordNet, while in RuThes they have different concepts. Description and comparison of different colors and their shades is a difficult task. However, the existing systems for presenting colors on the html pages of the Internet, for example, distinguish between amber and gold colors, matching code FFD700 to the gold color, and code FFBF00 to the amber color, that is, the RuThes presentation is more correct.

It is possible to find examples of another kind, when two synsets of WordNet correspond to a single RuThes concept. For example, in RuThes there is the concept *АТОМНАЯ ЭНЕРГИЯ* (*atomic energy*), the text entries for which in Russian are the phrases *атомная энергия* (*atomic energy*) and *ядерная энергия (nuclear energy*), and in English the name of this concept is formulated as *NUCLEAR ENERGY*, and the following phrases are listed as text entries*: atomic energy, atomic power, nuclear energy, nuclear powe*r.

In WordNet, two synsets correspond to this single RuThes concept: 1) atomic energy, nuclear energy (energy released by a nuclear reaction); 2) atomic power, for civilian use. In the second synset, *atomic power* is considered as a function of the atomic energy from the first synset, namely the use in power engineering. However, it seems that the same treatment of this sense cannot be reproduced in Russian.

Another example of the differences in conceptualization is related to the concept of *clock*. There are three basic concepts in WordNet: *time piece, timekeeper, horologe* and its two hyponyms: *clock* (a timepiece that shows the time of day) and *watch, ticker* (a small portable timepiece), including wrist or pocket watches.

Wikipedia shows a different type of conceptualization of these concepts for the English language, when *clock* and *timepiece* are united into one article, and the watch has another article. In RuThes, there is one concept of *ЧАСЫ* (*Timepiece),* with English-language translations: *clock, watch, timepiece,* and various subspecies of clocks, since in Russian there is no more general concept corresponding to the dimension of time than *часы* (clock), nor individual words that correspond to small, "portable" clocks.

Thus, it can be seen that the comparison between semantic systems of different resources reveals flaws (repetition of sense, lack of senses) in one of the descriptions or different conceptualizations. Therefore, it is hardly worth setting the task of complete linking of all concepts (synsets).

It can be seen from the Table 2 that the published version of RuThes contains about 9 thousand concepts (of 31 thousand concepts), which have English text entries but no matching with WordNet noun synsets (**Types 1.3** and **2.4**). These concepts include:

- Russian and near-to Russia geographic names (about 1300 concepts),
- concepts having only verbs or adjectives as text entries,
- Russia-specific cultural and social concepts: *gzhel* (Russian style of blue and white ceramics), *sopka* (specific hills in Siberia), *kalach* (Eastern European bread), *kissel* (viscous fruit dish), *gorodki* (ancient Russian folk sport),etc.,
- concepts based on multiword expressions, which are absent in WordNet.

The direct matching of RuThes concepts and WordNet synsets, utilizing unambiguous and the most frequent correspondences (with post-editing), gave the following numbers of the established links between RuWordNet and WordNet synsets:

- 8,608 from 29,296 noun synsets,
- 996 from 7,634 verb synsets,
- 2,100 from 12,864 adjective synsets.

5.2 Translating Core Concepts

Additionally to the above-described matching to WordNet based on the RuThes English text entries, the independent examination of the WordNet core synsets is necessary because some English words can be absent in the English counterpart of the RuThes thesaurus. In this case, a professional linguist searches for each WordNet core synset direct link to a RuWordNet synset using both English text entries from RuThes and also any additional resources.

Currently, we have 90% of synonym and near-synonym links for the WordNet core concepts with the RuWordNet synsets, and it seems a very high level for the resources, which have been developed independently. About 400 new RuWordNet synsets have been proposed to introduction.

Table 3 shows statistics on established relations between RuWordNet and WordNet synsets for core synsets.

Part of Speech	Number of core concepts	Percent of established links (%)
Nouns	3300	90.3
Adjectives	698	85.0
Verbs	999	94.0
Total	4997	90.0

Table 3. Statistics on established relations between the RuWordNet and WordNet synsets for the core synsets

Some examples of core WordNet noun synsets for which the correspondece in RuWordNet are metonymic transfer of source senses:

- (n) village, small town, settlement (a community of people smaller than a town)
- (n) university (the body of faculty and students at a university)
- (n) manner of speaking, speech, delivery (your characteristic style or manner of expressing yourself orally)

Other examples of absent noun links are quite diverse:

- (n) style (editorial directions to be followed in spelling and punctuation and capitalization and typographical display)
- (n) **s**urvivor (one who outlives another) *"he left his farm to his survivors"*
- (n) search (an investigation seeking answers) *"a thorough search of the ledgers revealed nothing*

For adjectives, the most frequent problems of linking between two resources is the absence of an adjective form for a specific concept, which can be expressed with a particle (that is a verb form) in Russian. For example, the following "core" adjectives senses are absent in Russian:

- absent – *отсутствующий* (otsutstvuyushchiy),
- *afraid – испуганный* (*ispugannyy*),
- *asleep – спящий* (*spyashchiy*).

The main reason of absense of verbal links is due that such senses are expressed only with *ligh*t *verb+noun* constructions in Russian:

- [cast]: select for a play or movie,
- [cater] supply food ready to eat,
- [demonstrate] march, march in protest.

6 Conclusion

In this paper we have considered the procedure for linking Russian wordnet (RuWordNet) to WordNet. The specificity of the procedure is based on the fact that a lot of bilingual (Russian and English) lexical data have been gathered in another Russian thesaurus RuThes, which has the structure different from WordNet. At first, Russian wordnet was semi-automatically generated from RuThes. Now, the RuThes English data are utilized to establish matching from the RuWordNet synsets to the WordNet synsets (through RuThes concepts).

Additionally, the WordNet core concepts are manually looked through to establish direct relations between RuWordNet and WordNet. Currently, 90% of the core Wordnet synsets are provided with links to RuWordNet, which is quite a large percentage for the independently developed resources.

Acknowledgments

The reported study was funded by RFBR according to the research project N 18-00-01226 (18-00-01240).

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: 1352-1362

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index, in: *Proceedings of the 8th Global WordNet Conference 2016* (*GWC2016*): 27-30.

Jordan Boyd-Graber, Christiane, Fellbaum, D. Osherson, and R. Schapire. 2006. Adding dense, weighted connections to WordNet.' In: *Proceedings of the Third Global WordNet Meeting, GWC-2006*.

Dan Cristea, Catalin Mihaila, Corina Forascu, Diana Trandabat, Maria Husarciuc, Gabriela Haja, and Oana Postolache. 2004. Mapping Princeton WordNet synsets onto Romanian WordNet synsets. *Romanian Journal of Information Science and Technology*, *7*(1-2): 125-145.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Alexander Kirillovich, Olga Nevzorova, Emil Gimadiev, and Natalia Loukachevitch. RuThes Cloud: Towards a Multilevel Linguistic Linked Open Data Resource for Russian. In: P. Różewski and C. Lange (eds.) Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017). Communications in Computer and Information Science, vol. 786, pp. 38-52. Springer (2017)

Krister Lindén and Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? An evaluation. *Language resources and evaluation*, 48.2: 191-201.

Natalia Loukachevitch and Boris Dobrov. 2002. Development and Use of Thesaurus of Russian Lan-

guage RuThes. *Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation, LREC 2002*: 65-70.

Natalia Loukachevitch, German Lashevich, Anastasia Gerasimova, Vladimir Ivanov, and Boris Dobrov. 2016. Creating Russian WordNet by Conversion. In *Proceedings of Conference on Computational linguistics and Intellectual technologies Dialog-2016*: 405-415

Natalia Loukachevitch, German Lashevich, and Boris Dobrov. 2018. Comparing Two Thesaurus Representations for Russian. *Proceedings of Global WordNet Conference GWC-2018*: 35-44.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*: 443-452.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. *Proceedings of the First International Conference on Global WordNet*: 293-302.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2012. A strategy of Mapping Polish WordNet onto Princeton WordNet. *In Proceedings of COLING 2012*: 1039-1048.

Ewa Rudnicka, Maciej Piasecki, Piotrowski, T., L. Grabowski, and Francis Bond. 2017. Mapping wordnets from the perspective of inter-lingual equivalence. *Cognitive Studies| Études cognitives*, (17).

Dan Tufiş, Verginica Mititelu, Dan Ştefănescu, and Radu Ion. 2013. The Romanian wordnet in a nutshell. *Language resources and evaluation*, *47*(4), 1305-1314.

Piek Vossen. 1998. Introduction to EuroWordNet. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer: 1-17.

Fast developing of a Natural Language Interface for a Portuguese Wordnet: Leveraging on Sentence Embeddings

Hugo Gonçalo Oliveira
CISUC, Dept. Informatics Engineering
Universidade de Coimbra, Portugal
hroliv@dei.uc.pt

Alexandre Rademaker
IBM Research and EMAp/FGV
Rio de Janeiro, Brazil
arademaker@gmail.com

Abstract

We describe how a natural language interface can be developed for a wordnet with a small set of handcrafted templates, leveraging on sentence embeddings. The proposed approach does not use rules for parsing natural language queries but experiments showed that the embeddings model is tolerant enough for correctly predicting relation types that do not match known patterns exactly. It was tested with OpenWordNet-PT, for which this method may provide an alternative interface, with benefits also on the curation process.

1 Introduction

A natural way of interacting with computational systems or knowledge bases is to use the same language we use for interacting with other humans. However, due to all the complex phenomena of natural language, most systems rely on browsing, keyword-based search interfaces or their combination. This is simpler at the technical level and avoids having to deal with Natural Language Understanding issues. The previous phenomena include ambiguity and language variability and are the reason why matching natural language with formal queries is not a trivial task. To overcome this challenge, we investigate how a model of sentence similarity can be exploited by a natural language interface (NLI) for a wordnet. Our approach is tested in OpenWordNet-PT (de Paiva et al., 2012) (OWN-PT), probably the most active Portuguese wordnet (de Paiva et al., 2016b).

The development of this system, dubbed NELIO, requires only a small set of handcrafted templates for each query to be covered. Instantiating those templates with arguments from OWN-PT results in a large set of sentences, used for training a doc2vec (Le and Mikolov, 2014) model. The latter is a variation of the popular word2vec (Mikolov et al., 2013) but, besides learning dense vector representations of words, it learns a representation for documents (sentences, in our case), based on the words used and on a document label. Such a model can be used e.g., for predicting the most suitable label for an unseen document. In this work, we rely on the trained doc2vec model for predicting the relation type that a natural language query is asking for. We then use this information for querying OWN-PT and retrieving suitable answers. This process is fast enough and avoids writing a set of rules for parsing natural language queries. Besides providing a more natural way of interacting with OWN-PT, NELIO turns out to be an alternative way of exploring OWN-PT and reveal flaws that, otherwise, would not be easy to spot.

The remainder of this paper briefly overviews OWN-PT, describes the development of NELIO, reports on performed experiments, including a systematic evaluation of the model in this context, and, before concluding, overviews related work.

2 OpenWordNet-PT

OpenWordNet-PT (OWN-PT) is an ongoing project to build a wordnet for Portuguese. It is aligned with Princeton WordNet (Fellbaum, 1998) (PWN), but still has about half of its size. So far, only partial evaluations of its coverage were performed, namely of verbs (de Paiva et al., 2016a, 2014), nouns (Rademaker et al., 2014), and (gentilic) adjectives (Real et al., 2016).

OWN-PT is freely available in RDF/OWL. Its data can be retrieved via a SPARQL endpoint, but it can also be explored through its own web interface[1] or through the interface of the Open Multilingual WordNet (Bond and Foster, 2013). As previously suggested (Real et al., 2015), a visual interface helps to discover interesting issues to work on. The research presented here is related to lessons previously learned.

[1] http://openwordnet-pt.org

3 System Development

NELIO interprets questions, in Portuguese, that ask for concepts, lexicalised as y, which are related in some way to another concept lexicalised as x, mentioned in the question. This section describes the steps for developing its current version.

3.1 Question Templates

To enable the generation of prototypical questions, a small set of templates for each covered relation was handcrafted by the first author of this paper. Such templates generalise possible ways of asking the desired questions in Portuguese. All templates currently used (between 3 and 10 per relation) are revealed in table 1, grouped according to the target relation. Most semantic relations in OWN-PT are covered. Yet, due to their different scope, lexical relations were left out of this set.

3.2 Model Training

The generation of prototypical questions results from filling the templates, automatically, with real examples from OWN-PT. Those questions were used to train a doc2vec (Le and Mikolov, 2014) model, with the name of the target semantic relation set as their label. Examples of generated questions include:

(hyponymOf)	*que formas há de correr?*
(hypernymOf)	*qual é o hiperónimo de maçã?*
(memberHol..Of)	*quais os membros de Liga Árabe?*
(substanceHol..Of)	*de que é feito molho de soja?*
(partHolonymOf)	*que partes tem Portugal?*
(partMeronym)	*de que faz parte Breslávia?*
(antonymOf)	*qual é o contrário de líquido?*
(x_causes)	*qual é o efeito de ferir?*
(entails)	*o que implica migrar?*

The learned model can be exploited in a classification task. More precisely, given a fragment of text, it can be used for predicting the appropriate label. Once predicted, the label is used together with the relation argument that appears on the question (x) for generating a SPARQL query, which can be made to OWN-PT for retrieving the possible answers.

3.3 Fixed Argument Extraction

Besides classifying the question into a relation type, the fixed relation argument x must be extracted from the input text. In all handcrafted templates, this argument is the last term of the question. In fact, for the type of considered questions, there would not be many variations where this was not the case. Therefore, the extraction of x was simplified in such a way that it is always the last sequence of words in the question. More precisely, in order to cover multiword expressions, the system searches for the longest lexical form in OWN-PT starting with the $i^{th}, i \in (1, n]$, and ending in the last token of the question. For instance, given the question *que tipos há de intoxicação alimentar?* (what types are there of food poisoning?), the system checks, in the following order, whether OWN-PT covers the forms: *tipos há de intoxicação alimentar*, *há de intoxicação alimentar*, *de intoxicação alimentar*, *intoxicação alimentar*. It stops once it finds that the lexical form *intoxicação alimentar* (food poisoning) exists.

3.4 SPARQL Generation

With the label and the fixed argument, a SPARQL query can be generated to get all the valid lexical forms for y. Figure 1 shows the generated query for the question *que formas há de correr?*, with label [hyponymOf] and $x = correr$. It retrieves lexical forms (lf) in OWN-PT synsets ($s2$) for which the aligned PWN synset ($sen2$) is a hyponym of another PWN synset ($sen1$) that is aligned with an OWN-PT synset with the lexical form *correr*.

```
prefix wn30: <https://w3id.org/own-pt/wn30/schema/>
prefix owl: <http://www.w3.org/2002/07/owl#>

SELECT ?lf WHERE {
  ?spt1 wn30:containsWordSense ?ws1 .
  ?ws1  wn30:word ?word .
  ?word wn30:lexicalForm "correr"@pt .
  ?sen1 owl:sameAs  ?spt1 .
  ?sen2 owl:sameAs ?spt2 .
  ?sen2 wn30:hyponymOf ?sen1 .
  ?spt2 wn30:containsWordSense ?ws2 .
  ?ws2  wn30:word/wn30:lexicalForm ?lf .
}
```

Figure 1: SPARQL query for retrieving the hyponyms of *correr*. Query is available in OWN-PT's SPARQL endpoint at https://ibm.co/2OCptyv.

4 Experiments

NELIO was implemented in Java, using Apache Jena[2] for querying OWN-PT and DeepLearning4J[3] for training the doc2vec model, more specifically, the ParagraphVectors class. This sections illustrates NELIO's usage and reports on a simple evaluation made automatically.

[2] https://jena.apache.org/
[3] https://deeplearning4j.org/

X hyponymOf Y (8 templates)	
que (tipos|géneros|espécies|sub-classes especificações|formas) há de <Y>?	*what (types | genres | species | subclasses | specifications | forms) are there of ⟨Y⟩?*
que (hipónimos|subordinados) tem <Y>?	*what (hyponyms|subordinates) does ⟨Y⟩ have?*
X hypernymOf Y (4 templates)	
qual é a classe de <Y>?	*what is the class of ⟨Y⟩*
qual é o hiperónimo de <Y>?	*what is the hypernym of ⟨Y⟩?*
qual é o conceito superordenado de <Y>?	*what is the superordinate concept of ⟨Y⟩??*
o que é <Y>?	*what is ⟨Y⟩?*
X memberHolonymOf Y (10 templates)	
quais os (membros|constituintes|componentes) de <Y>?	*what are the (members|constituents|components) of ⟨Y⟩?*
que membros tem <Y>?	*what members does ⟨Y⟩ have?*
o que tem <Y>?	*what does ⟨Y⟩ have?*
de que é constituído <Y>?	*what is ⟨Y⟩ made of?*
o que inclui <Y>?	*what does ⟨Y⟩ include?*
o que está em <Y>?	*what is there in ⟨Y⟩?*
em que se (divide|decompõe) <Y>?	*in what can ⟨Y⟩ be divided|decomposed?*
X partHolonymOf Y (9 templates)	
quais as (partes|constituintes|componentes) de <Y>?	*what are the (parts|constituents|components) of ⟨Y⟩?*
que partes tem <Y>?	*what parts does ⟨Y⟩ have?*
o que tem <Y>?	*what does ⟨Y⟩ have?*
de que é constituído <Y>?	*what is ⟨Y⟩ made of?*
o que inclui <Y>?	*what does ⟨Y⟩ include?*
em que se (divide|decompõe) <Y>?	*in what can ⟨Y⟩ be (divided|decomposed)?*
X substanceHolonymOf Y (7 templates)	
quais as substâncias de <Y>?	*what are the substances of ⟨Y⟩?*
que substâncias tem <Y>?	*what substances does ⟨Y⟩ have?*
o que tem <Y>?	*what does ⟨Y⟩ have?*
de que é (constituído|feito) <Y>?	*what is ⟨Y⟩ made of?*
o que inclui <Y>?	*what does ⟨Y⟩ include?*
o que está em <Y>?	*what is there in ⟨Y⟩?*
X memberMeronymOf Y / X partMeronymOf Y (3 templates)	
de que faz parte <Y>?	*what is part of ⟨Y⟩?*
onde se inclui <Y>?	*where is ⟨Y⟩ included?*
a que pertence <Y>?	*what does ⟨Y⟩ belong to?*
X substanceMeronymOf Y (4 templates)	
de que faz parte <Y>?	*what is part of ⟨Y⟩?*
onde se inclui <Y>?	*where is ⟨Y⟩ included?*
onde encontramos <Y>?	*where can we find ⟨Y⟩?*
onde se encontra <Y>?	*where is ⟨Y⟩ found?*
X causes Y (7 templates)	
qual é o (efeito|resultado) de <X>?	*what is the (effect|result) of ⟨X⟩?*
qual é a consequência de <X>?	*what is the consequence of ⟨X⟩?*
o que (causa|faz|origina) <X>?	*what does ⟨X⟩ (cause|make|originate)?*
em que resulta <X>?	*what does ⟨X⟩ result in?*
X causes Y (2 templates)	
o que leva a <Y>?	*what leads to ⟨Y⟩?*
o que resulta em <Y>?	*what does ⟨Y⟩ result in?*
X entails Y (4 templates)	
o que (acarreta|implica) <Y>?	*what does ⟨Y⟩ (entail|implies)?*
o que se (infere|conclui) de <Y>?	*what may one (infer|conclude) of ⟨Y⟩?*
X antonymOf Y (5 templates)	
qual é o (antónimo|contrário|oposto|inverso) de <X>?	*what is the (antonym|contrary|opposite|inverse) of ⟨Y⟩?*
o que é diferente de <X>?	*what is different from ⟨Y⟩?*

Table 1: Handcrafted templates for querying the NLI interface. The vertical bar is an abbreviation, in the templates files, each word used constitute one template.

4.1 Examples

The resulting doc2vec model identifies the correct relation type in most situations. Besides being easy to train, a good thing about it is that no syntactic analysis is required and, still, the text of the questions does not have to match the original templates exactly. This provides an interesting level of tolerance while dealing with syntactic variability. Table 2 shows some of the questions answered correctly that, despite sharing some words, do not perfectly match any template.

On the other hand, since the model is based on the surface forms of words, it does not handle variation of vocabulary. For instance, the question "*qual é a matéria de chocolate?*" (what is the matter of chocolate?) is answered with hyponyms of 'chocolate', instead of its substances, because no substanceHolonymOf template includes the word *matéria* (matter). However, this situation can be easily fixed by adding a single template with this word for the previous relation, e.g., `matéria-prima de <Y>?`.

4.2 Generalisation

Combining all possible question templates with all the relation instances in OWN-PT results in 298,249 different questions. Those questions were generated, shuffled and used as a dataset for training and testing NELIO.

In a simple experiment for analysing how well

Question	Answer
tipos de peixe? types of fish?	*sardinha, arenque, enchova, Peixe Esportivo, atum, salmonete, manjuba, eperlano, Salmão ...* sardine, herring, anchovy, sport fish, tuna, surmullet, smelt, salmon ...
forma de andar? way of walking?	*andar a galope, retornar, afastar-se, serpear, correr terras, descarrilhar, perseguir, coxear ...* gallop, return, depart, creep, drift, derail, seek, limp, ...
o que é um cão? what is a dog?	*animal_doméstico, canídeo, cão* domestic animal, canid
membros de Passeridae? members of Passeridae?	*Pardal, pardal, Passer* sparrow, Passer
que tem na Polónia? what is there in Poland?	*Auschwitz-Birkenau, Prússia, Rio vístula, Bydgoszcz, Rio oder, Czestochowa, Varsóvia, ...* Auschwitz-Birkenau, Prussia, Vístula river, Bydgoszcz, Oder river, Czestochowa, Warsaw, ...
onde fica a Polónia? where is Poland?	*Europa* Europe
que resulta de extender? what results from stretching?	*esticar* to stretch
que implica olhar? what implies looking?	*olhar, ver, mirar, inspecionar, assistir, examinar, observar* to see, to eye, to inspect, to watch, to skim, to observe, to lay eyes on
contrário de alto? opposite of tall?	*baixo* short

Table 2: Questions correctly answered by NELIO.

the model generalises, it was tested with different proportions of training and testing data. Table 3 presents the accuracy, i.e., the proportion of questions correctly answered in this experiment.

This also showed that, some of the incorrect answers were in fact empty, due to misclassification of the relation type, which suggested a second experiment: similar to the previous but, when the given answer was empty, NELIO tried to get an answer with the second or third relation type predicted by doc2vec. As expected, this resulted in higher accuracies, also in table 3 (Top-3).

Training		Test	Accuracy	
Prop.	#Questions	Prop.	1st label	Top-3
90%	(268,424)	10%	93.3%	97.2%
75%	(223,687)	25%	92.9%	97.5%
50%	(149,125)	50%	93.2%	97.6%
25%	(74,562)	75%	91.9%	97.6%
20%	(59,650)	80%	92.1%	97.9%
15%	(44,737)	85%	89.6%	97.0%
10%	(29,825)	90%	80.3%	95.9%
5%	(14,912)	95%	79.2%	94.9%

Table 3: Accuracy when answering questions depending on proportion of training data.

When considering only the top label, training the model with 90% (≈268k), 50% (≈149k), or even 20% (≈59k) of the questions, results in accuracies above 90%. This happens mainly because, although there are only a few templates, they are instantiated many times. With lower training proportions, accuracy drops more considerably. Yet, with only 5% it is still close to 80%.

Accuracy is different for different relations. For instance, with 90% of training data, it ranges from 100%, for entails, antonymOf and hyponymOf, to 73%, for substanceMeronymOf. A closer look shows that, except for the meronym-holonym relations, all accuracies are higher than 94% (hypernymOf). The problem with the former is that they are very similar and, for this reason, share several templates among them, which confuses the model.

The aforementioned issue is significantly minimised when the top-3 labels are considered. In this case, accuracies are 97% or higher with 15% or more training data. Specifically, they are 98% or higher for all relation types, except for the meronym-holonym, which are still the most problematic. The lower accuracy in this scenario is for memberHolonymOf (87.9%).

5 Related Work

Traditional Automatic Question Answering (QA) follows an Information Retrieval perspective (Kolomiyets and Moens, 2011). Queries are typically natural language questions (NLQs) and answers are retrieved from a collection of written documents. But the development of natural language interfaces (NLIs) for databases has also been a research topic for a long time (Androutsopoulos et al., 1995). Here, the primary challenge involves translating NLQs to formal queries made to a database. Knowledge-based QA systems are a specific case of the previous.

Several NLIs for ontologies — e.g., Querix (Kaufmann et al., 2006), PANTO (Wang et al., 2007), FREyA (Damljanovic et al., 2010) — translate NLQs to SPARQL with a set of rules on the result of syntactically parsing NLQs, possibly using PWN for synonym expansion. A

similar approach (Unger et al., 2012) may be based on SPARQL templates, to be filled with entities and predicates identified in the NLQ.

Other systems rely on domain-independent semantic parsers that learn how to map NLQs to predicates in a large knowledge base, based on question-answer pairs. SEMPRE (Berant et al., 2013) maps words to predicates and then combines the predicates to the final logical form. Another possibility (Kwiatkowski et al., 2013) is to parse utterances for producing an underspecified logical form, before mapping lexical predicates to the target ontology predicates. The previous systems were assessed while resorting to Freebase for answering NLQs. Yet, as opposing to Freebase or DBPedia, wordnets have a much smaller number of predicates. So, it could be worth exploring how semantic parsers could be adapted for our work.

Once translated to SPARQL, generally to a subset of this language, expressiveness is limited. To avoid this, SQUALL (Ferré, 2014) is a controlled natural language for querying and updating RDF datasets. Nouns and intransitive verbs are used as classes; relation nouns and transitive verbs as properties; and proper nouns as resources. Syntactic and semantic analysis is implemented as a Montague grammar, an approach that would work for querying a wordnet, considering the simplicity of its RDF model. On the other hand, SQUALL requires that end-users comply with its controlled syntax, and know the RDF vocabulary.

An alternative approach (Bordes et al., 2014) learns low-dimensional embeddings of words and entities, respectively in questions and relation types of Freebase. This way, representations of questions and of their corresponding answers are close to each other in the joint embedding space. More recent works (Neelakantan et al., 2016; Zhong et al., 2017) rely on neural networks for translating NLQs to formal queries, thus avoiding domain-specific grammars or rules.

6 Conclusion

We have described how we can leverage on sentence embeddings in the development of a NLI for a wordnet. The proposed procedure was applied to OWN-PT with some success. When trained in a subset with at least 20% of the possible questions, generated with a small set of templates, and tested with the remaining questions, accuracies were higher than 91%, when using the first prediction, or 97%, when trying with the first three predictions, in case the previous did not return an answer. This simple experiment confirmed that the proposed approach works well with the doc2vec model for predicting the correct relation type. Despite the positive results, this experiment revealed that the system is confused by similar relations, for which the templates share vocabulary, namely the three types of meronymy. The problem can be minimised by considering the top-3 predictions, but others, such as merging the three relations, can be analysed in the future.

Still, this was a limited experiment, where known limitations of the system had a low impact. This includes questions with vocabulary not covered by the templates, or questions that do not end with the fixed word. The former can be minimised by adding alternative templates. The second is due to a simplification that works for many cases, but fails for some, as in the question '*quais frutas existem?*' (what fruits exist?), where the target word is *frutas*. The previous question has to be made like '*quais os tipos de fruta?*'. In the future, we will devise more general ways of extracting the target argument from the question, e.g., having in mind that, among the words/expressions in the question, it should be the least frequent in the dataset; or maybe training an automatic sequence labeller for identifying the target argument in the context of a question. In the latter case, training data should also include templates that do not end with the target argument.

Other possible directions for future work include: (i) Presenting the answers according to the senses they apply to, because context is not enough for disambiguation (currently, there is an option for considering only the first sense); (ii) Adding alternative types of question e.g., *what is the relation between* $\langle x \rangle$ *and* $\langle y \rangle$*?* or *is* $\langle y \rangle$ *related to* $\langle x \rangle$*?*, to be answered, respectively, with the name of a relation between x and y in OWN-PT, or yes/no, depending on the existence of such a relation; (iii) Exploring recent models for representing sentence meaning, learned from natural language inference data (Conneau et al., 2017), though available data in Portuguese (Fonseca et al., 2016; Real et al., 2018) may not be enough.

Despite its limitations, NELIO was already helpful for finding issues in OWN-PT that need to be fixed. It showed flaws such as inconsistencies

in the capitalization (e.g., *Salmão*, *Pardal*), presence of underscores instead of spaces (e.g., *animal_doméstico*), or plural instead of singular form (e.g., *epidemias*, *montanhas*), not to mention actual errors (e.g, *dançar* entails *andar*, in English, dancing entails walking).

A mid-term goal is to make NELIO available from a web interface. In the meantime, its source code is available online, at `https://github.com/hgoliv/nli_openwordnet-pt`. Although, so far, the proposed approach was only used as a NLI for a wordnet, in principle, a similar approach could be used in the development of a NLI for any knowledge base represented as *a*-relatedTo-*b* triples.

Acknowledgments

This work was partially funded by FCT's INCoDe 2030 initiative, in the scope of the demonstration project AIA, "Apoio Inteligente a empreendedores (chatbots)".

References

I. Androutsopoulos, G.D. Ritchie, and P. Thanisch. 1995. Natural language interfaces to databases — an introduction. *Natural Language Engineering*, 1(1):29–81.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544. ACL Press.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1352–1362. ACL Press.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620. ACL Press.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. ACL Press.

Danica Damljanovic, Milan Agatonovic, and Hamish Cunningham. 2010. Natural language interfaces to ontologies: Combining syntactic analysis and ontology-based lookup through the user interaction. In *Proceedings of the 7th International Conference on The Semantic Web: Research and Applications - Volume Part I*, ESWC'10, pages 106–120. Springer.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Sébastien Ferré. 2014. Squall: a controlled natural language as expressive as SPARQL 1.1. *Data Knowl. Eng.*, 94(PB):163–188.

Erick Rocha Fonseca, Leandro Borges dos Santos, Marcelo Criscuolo, and Sandra Maria Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.

Esther Kaufmann, Abraham Bernstein, and Renato Zumstein. 2006. Querix: A natural language interface to query ontologies based on clarification dialogs. In *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, pages 980–981.

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A Survey on Question Answering Technology from an Information Retrieval Perspective. *Information Sciences*, 181(24):5412–5434.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1556, Seattle, Washington, USA. ACL Press.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of the Workshop track of the Intl. Conf. on Learning Representations (ICLR)*, Scottsdale, Arizona.

Arvind Neelakantan, Quoc V. Le, Martín Abadi, Andrew McCallum, and Dario Amodei. 2016. Learning a natural language interface with neural programmer. *CoRR*, abs/1611.08945.

Valeria de Paiva, Fabricio Chalub, Livy Real, and Alexandre Rademaker. 2016a. Making virtue of necessity: a verb lexicon. In *Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR 2016)*, pages 271–282, Tomar, Portugal. Springer.

Valeria de Paiva, Cláudia Freitas, Livy Real, and Alexandre Rademaker. 2014. Improving the verb lexicon of OpenWordnet-PT. In *Proceedings of Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish (ToRPorEsp)*, São Carlos, Brazil. Biblioteca Digital Brasileira de Computação, UFMG, Brazil.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India. The COLING 2012 Organizing Committee.

Valeria de Paiva, Livy Real, Hugo Gonçalo Oliveira, Alexandre Rademaker, Cláudia Freitas, and Alberto Simões. 2016b. An overview of portuguese wordnets. In *Global Wordnet Conference 2016*, Bucharest, Romenia.

Alexandre Rademaker, Valeria de Paiva, Gerard de Melo, Livy Real, and Maira Gatti. 2014. OpenWordNet-PT: A project report. In *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia.

Livy Real, Fabricio Chalub, Valeria de Paiva, Claudia Freitas, and Alexandre Rademaker. 2015. Seeing is correcting: curating lexical resources using social interfaces. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 20–29, Beijing, China. ACL Press.

Livy Real, Valeria de Paiva, Fabricio Chalub, and Alexandre Rademaker. 2016. Gentle with gentilics. In *Joint Second Workshop on Language and Ontologies (LangOnto2) and Terminology and Knowledge Structures (TermiKS) (co-located with LREC 2016)*, Slovenia.

Livy Real, Ana Rodrigues, Andressa Vieira, Beatriz Albiero, Bruna Thalenberg, Bruno Guide, Cindy Silva, Guilherme de Oliveira Lima, Igor C. S. Câmara, Miloš Stanojević, Rodrigo Souza, and Valeria De Paiva. 2018. Sick-br: A portuguese corpus for inference. In *13th International Conference, PROPOR 2018*, pages 303–312, Canela, Brazil.

Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. 2012. Template-based question answering over rdf data. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 639–648, New York, NY, USA. ACM.

Chong Wang, Miao Xiong, Qi Zhou, and Yong Yu. 2007. Panto: A portable natural language interface to ontologies. In *Proceedings of the 4th European Conference on The Semantic Web: Research and Applications*, ESWC '07, pages 473–487. Springer.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Two experiments for embedding Wordnet hierarchy into vector spaces*

Jean-Philippe Bernardy and Aleksandre Maskharashvili
Gothenburg University, Department of philosophy, linguistics and theory of science,
Centre for linguistics and studies in probability

jean-philippe.bernardy,aleksandre.maskharashvili@gu.se

Abstract

In this paper, we investigate mapping of the WORDNET hyponymy relation to feature vectors. Our aim is to model lexical knowledge in such a way that it can be used as input in generic machine-learning models, such as phrase entailment predictors. We propose two models. The first one leverages an existing mapping of words to feature vectors (*fast*Text), and attempts to classify such vectors as within or outside of each class. The second model is fully supervised, using solely WORDNET as a ground truth. It maps each concept to an interval or a disjunction thereof. The first model approaches but not quite attain state of the art performance. The second model can achieve near-perfect accuracy.

1 Introduction

Distributional encoding of word meanings from large corpora (Mikolov et al., 2013; Mikolov et al., 2018; Pennington et al., 2014) have been found to be useful for a number of NLP tasks.

While the major goal of distributional approaches is to identify distributional patterns of words and word sequences, they have even found use in tasks that require modeling more fine-grained relations between words than co-occurrence in word sequences. But distributional word embeddings are not easy to map onto ontological relations or *vice-versa*. We consider in this paper the hyponymy relation, also called the *is-a* relation, which is one of the most fundamental ontological relations. We take as the source of truth for hyponymy WORDNET (Fellbaum, 1998), which has been designed to include various kinds of lexical relations between words, phrases, etc. However, WORDNET has a fundamentally symbolic representation, which cannot be readily used as input to neural NLP models.

Several authors have proposed to encode hyponymy relations in feature vectors (Vilnis and McCallum, 2014; Vendrov et al., 2015; Athiwaratkun and Wilson, 2018; Nickel and Kiela, 2017). However, there does not seem to be a common consensus on the underlying properties of such encodings. In this paper, we aim to fill this gap and clearly characterize the properties that such an embedding should have. We additionally propose two baseline models approaching these properties: a simple mapping of FASTTEXT embeddings to the WORDNET hyponymy relation, and a (fully supervised) encoding of this relation in feature vectors.

2 Goals

We want to model the hyponymy relation (ground truth) given by WORDNET — hereafter referred to as HYPONYMY. In this section we make this goal precise and formal. Hyponymy can in general relate common noun phrases, verb phrases or any predicative phrase, but hereafter we abstract from all this and simply write "word" for this underlying set. In this paper, we write $(\subseteq)$ for the reflexive transitive closure of the hyponymy relation (ground truth), and $(\subseteq_M)$ for relation predicted by a model M.[1] Ideally, we want the model to be sound and complete with respect to the ground truth. However, a machine-learned model will typically only approach those properties to a certain level, so the usual relaxations are made:

Property 1 *(Partial soundness) A model M is*

*Supported by Swedish Research Council, Grant number 2014-39.

[1] We note right away that, on its own, the popular metric of cosine similarity (or indeed any metric) is incapable of modeling HYPONYMY, because it is an asymmetric relation. That is to say, we may know that the embedding of "animal" is close to that of "bird", but from that property we have no idea if we should conclude that "a bird is an animal" or rather that "an animal is a bird".

partially sound with precision α iff., for a proportion α of the pairs of words w, w' such that $w \subseteq_M w'$ holds, $w \subseteq w'$ holds as well.

Property 2 *(Partial completeness) A model M is partially complete with recall α iff., for a proportion α of the pairs of words w, w' such that $w \subseteq w'$ holds, then $w \subseteq_M w'$ holds as well.*

These properties do not constrain the way the relation $(\subseteq_M)$ is generated from a feature space. However, a satisfying way to generate the inclusion relation is by associating a subset of the vector space to each predicate, and leverage the inclusion from the feature space. Concretely, the mapping of words to subsets is done by a function P such that, given a word w and a feature vector x, $P(w, x)$ indicates if the word w applies to a situation (state of the world, sentence meaning, sentory input, etc.) described by feature vector x. We will refer to P as a classifier. The inclusion model is then fully characterized by P, so we can denote it as such $(\subseteq_P)$.

Property 3 *(Space-inclusion compatibility) There exists $P : (Word \times \mathbb{R}^d) \to [0, 1]$ such that*

$$(w' \subseteq_P w) \iff (\forall x. P(w, x) \leq P(w', x))$$

Any model given by such a P yields a relation $(\subseteq_P)$ which is necessarily reflexive and transitive (because subset inclusion is such) — the model does not have to learn this. Again, the above property will apply only to ideal situations: it needs to be relaxed in some machine-learning contexts. To this effect, we can define the measure of the subset of situations which satisfies a predicate $p : \mathbb{R}^d \to [0, 1]$ as follows:

$$\mathsf{measure}(p) = \int_{\mathbb{R}^d} p(x)dx$$

(Note that this is well-defined only if p is a measurable function over the measurable space of feature vectors.) We leave implicit the density of the vector space in this definition. Following this definition, a predicate p is included in a predicate q iff.

$$\frac{\mathsf{measure}(p \wedge q)}{\mathsf{measure}(p)} = \frac{\int_{\mathbb{R}^d} p(x)q(x)dx}{\int_{\mathbb{R}^d} p(x)dx} = 1$$

Following this thread, we can define a relaxed inclusion relation, corresponding to a proportion of ρ of p included in q:

Property 4 *(Relaxed Space-inclusion compatibility) There exists $P : Word \to \mathbb{R}^d \to [0, 1]$ and $\rho \in [0, 1]$ such that*

$$(w' \subseteq_P w) \iff \frac{\int_{\mathbb{R}^d} P(w', x)P(w, x)dx}{\int_{\mathbb{R}^d} P(w, x)dx} \geq \rho$$

In the following, we call ρ the relaxation factor.

3 Mapping WordNet over *fast*Text

Our first model of HYPONYMY works by leveraging a general-purpose, unsupervised method of generating word vectors. We use *fast*Text (Mikolov et al., 2018) as a modern representative of word-vector embeddings. Precisely, we use pre-trained word embeddings available on the *fast*Text webpage, trained on Wikipedia 2017 and the UMBC webbase corpus and the statmt.org news dataset (16B tokens). We call FTDom the set of words in these pre-trained embeddings.

A stepping stone towards modeling the inclusion relation correctly is modeling correctly each predicate individually. That is, we want to learn a separation between *fast*Text embeddings of words that belong to a given class (according to WordNet) from the words that do not. We let each word w in *fast*Text represent a situation corresponding to its word embedding $f(w)$. Formally, we aim to find P such that

Property 5 $P(w, f(w')) = 1 \iff w' \subseteq w$

for every word w and w' found both in WordNet and in the pre-trained embeddings. If the above property is always satisfied, the model is sound and complete, and satisfies Property 3.

Because many classes have few representative elements relative to the number of dimensions of the *fast*Text embeddings, we limit ourselves to a linear model for P, to limit the possibility of overfitting. That is, for any word w, $P(w)$ is entirely determined by a bias $b(w)$ and a vector $\theta(w)$ (with 300 dimensions):

$$P(w, x) = \delta(\theta(w) \cdot x + b(w) > 0)$$

where $\delta(\text{true}) = 1$ and $\delta(\text{false}) = 0$.

We learn $\theta(w)$ and $b(w)$ by using logistic regression, independently for each WordNet word w. The set of all positive examples for w is $\{f(w') \mid w' \in \mathsf{FTDom}, w' \subseteq w\}$, while the set of negative examples is $\{f(w') \mid w' \in \mathsf{FTDom}, w' \not\subseteq w\}$. We train and test for all the predicates with at least 10 positive examples. We

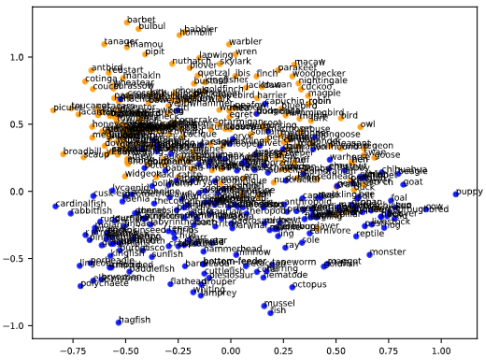


Figure 1: PCA representation of animals. Birds are highlighted in orange.

use 90% of the set of positive examples (w') for training (reserving 10% for testing) and we use the same number of negative examples.

We then test Property 5 on the 10% of positive examples reserved for testing, for each word. On average, we find that 89.4% of positives are identified correctly (std. dev. 14.6 points). On 1000 randomly selected negative examples, we find that on averale 89.7% are correctly classified (std dev. 5.9 points). The result for positives may look high, but because the number of true negative cases is typically much higher than that of true positives (often by a factor of 100), this means that the recall and precision are in fact very low for this task. That is, the classifier can often identify correctly a *random* situation, but this is a relatively easy task. Consider for example the predicate for "bird". If we test random negative entities ("democracy", "paper", "hour", etc.), then we may get more than 97% accuracy. However, if we pick our samples in a direct subclass, such as (non-bird) animals, we typically get only 75% accuracy. That is to say, 25% of animals are incorrectly classified as birds.

To get a better intuition for this result, we show a Principal Component Analysis (PCA) on animals, separating bird from non-birds. It shows mixing of the two classes. This mixture can be explained by the presence of many uncommon words in the database (e.g. types of birds that are only known to ornithologists). One might argue that we should not take such words into account. But this would severely limit the number of examples: there would be few classes where logistic regression would make sense.

We are not ready to admit defeat yet as we are ultimately not interested in Property 5, but rather in properties 1 and 2, which we address in the next section.

4 Inclusion of subsets

A strict interpretation of Property 3 would dictate to check if the subsets defined in the previous section are included in each other or not. However, there are several problems with this approach. To begin, hyperplanes defined by θ and b will (stochastically) always intersect therefore one must take into account the actual density of the *fast*Text embeddings. One possible approximation would be that they are within a ball of certain radius around the origin. However, this assumption is incorrect: modeling the density is a hard problem in itself. In fact, the density of word vectors is so low (due to the high dimensionality of the space) that the question may not make sense. Therefore, we refrain from making any conclusion on the inclusion relation of the subsets, and fall back to a more experimental approach.

Thus, we will test the suitability of the learned $P(w)$ by testing whether elements of its subclasses are contained in the superclass. That is, we define the following quantity $Q(w', w) =$

$$average\{P(w', x) \mid x \in \mathsf{FTDom}, P(w, f(x))\}$$

which is the proportion of elements of w' that are found in w. This value corresponds to the relaxation parameter ρ in Property 4.

If $w' \subseteq w$ holds, then we want $Q(w', w)$ to be close to 1, and close to 0 if w' is disjoint from w. We plot (figure 2) the distribution of $Q(w', w)$ for all pairs $w' \subseteq w$, and a random selection of pairs such that $w' \not\subseteq w$. The negative pairs are generated by taking all pairs (w', w) such that $w' \subseteq w$, and generate two pairs (w_1, w) and (w', w_2), by picking w_1 and w_2 at random, such that neither of the generated pairs is in the HYPONYMY relation. We see that most of the density is concentrated at the extrema. Thus, the exact choice of ρ has little influence on accuracy for the model. For $\rho = 0.5$, the recall is 88.8%. The ratio of false positives to the total number of negative test cases is 85.7%. However, we have a very large number of negatives cases (the square of the number of classes, about 7 billions). Because of this, we get about 1 billion false positives, and the precision is only 0.07%. Regardless, the results are comparable with state-of-the art models (section 6).

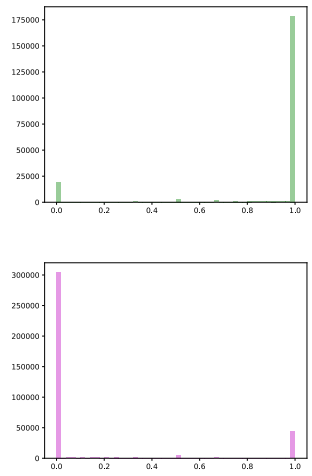


Figure 2: Results of inclusion tests. On the left-hand-side, we show the distribution of correctly identified inclusion relations in function of ρ. On the right-hand-side, we show the distribution of (incorrectly) identified inclusion relations in function of ρ.

5 WordNet predicates as disjunction of intervals

In this section we propose a baseline, fully supervised model model for HYPONYMY.

The key observation is that most of the HYPONYMY relation fits in a tree. Indeed, out of 82115 nouns, 7726 have no hypernym, 72967 have a single hypernym, and 1422 have two hypernyms or more. In fact, by removing only 1461 direct edges, we obtain a tree. The number of edges removed in the transitive closure of the relation varies, depending on which exact edges are removed, but a typical number is 10% of the edges. In other words, when removing edges in such a way, one lowers the recall to about 90%, but the precision remains 100%. Indeed, no pair is added to the HYPONYMY relation. This tree can then be mapped to one-dimensional intervals, by assigning a position to each of the nodes, according to their index in depth-first order ($ix(w)$ below). Then, each node is assigned an interval corresponding to the minimum and the maximum position assigned to their leaves. A possible directed acyclic graph (DAG) and a corresponding assignment of intervals is shown in Fig. 3. The corresponding definition of predicates is the following:

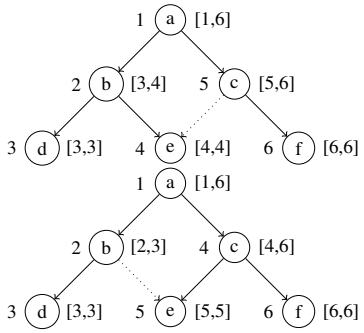


Figure 3: Two trees underlying the same dag. Nodes are labeled with their depth-first index on the left and their associated interval on the right. Removed edges are drawn as a dotted line.

$$\begin{aligned} P(w,x) &= x \geq lo(w) \wedge x \leq hi(w) \\ lo(w) &= min\{ix_T(w') \mid w' \subseteq_T w\} \\ hi(w) &= max\{ix_T(w') \mid w' \subseteq_T w\} \end{aligned}$$

where $(\subseteq_T)$ is the reflexive-transitive closure of the T tree relation (included in HYPONYMY). The pair of numbers $(lo(w), hi(w))$ fully characterizes $P(w)$. In other words, the above model is fully sound (precision=1), and has a recall of about 0.9. Additionally, Property 3 is verified.

Because it is fully sound, a model like the above can always be combined with another model to improve its recall with no impact on precision — including itself. Such a self-combination is useful if one does another choice of removed edges. Thus, each word is characterized by an n-dimensional co-product (disjoint sum) of intervals.

$$w \subseteq_M w' \triangleq \bigvee_i \Big(lo_i(w') \geq lo_i(w) \wedge hi_i(w') \leq hi_i(w) \Big)$$

$$lo_i(w) = min\{ix_{T_i}(w') \mid w' \subseteq_{T_i} w\}$$
$$hi_i(w) = max\{ix_{T_i}(w') \mid w' \subseteq_{T_i} w\}$$

By increasing n, one can increase the recall to obtain a near perfect model. Table 4b shows typical recall results for various values of n. However Property 3 is not verified: the co-product of intervals do not form subspaces in any measurable set.

6 Related Work: Precision and recall for hyponymy models

Many authors have considered modeling hyponymy. However, in many cases, this task was

not the main point of their work, and we feel that the evaluation of the task has often been partially lacking. Here, we review several of those and attempt to shed a new light on existing results, based on the properties presented in section 2.

Several authors (Athiwaratkun and Wilson, 2018; Vendrov et al., 2015; Vilnis et al., 2018) have proposed Feature-vector embeddings of WORDNET. Among them, several have tested their embedding on the following task: they feed their model with the transitive closure of HYPONYMY, but withhold 4000 edges. They then test how many of those edges can be recovered by their model. They also test how many of 4000 random negative edges are correctly classified. They report the average of those numbers. We reproduce here their results for this task in Table 4a. As we see it, there are two issues with this task. First, it mainly accounts for recall, mostly ignoring precision. As we have explained in section 4, this can be a significant problem for WORDNET, which is sparse. Second, because WORDNET is the only input, it is questionable if any edge should be withheld at all (beyond those in the transitive closure of generating edges). We believe that, in this case, the gold standard to achieve is precisely the transitive closure. Indeed, because the graph presentation of WORDNET is nearly a tree, most of the time, the effect of removing an edge will be to detach a subtree. But, without any other source of information, this subtree could in principle be reattached to any node and still be a reasonable ontology, from a purely formal perspective. Thus we did not withhold any edge when training our second model on this task (the first one uses no edge at all). In turn, the numbers reported in Table 4a should not be taken too strictly.

Authors	Result
(Vendrov et al., 2015)	90.6
(Athiwaratkun and Wilson, 2018)	92.3
(Vilnis et al., 2018)	92.3
us, *fast*Text with LR and $\rho = 0.5$	87.2
us, single interval (tree-model)	94.5
us, interval disjunctions, $n = 5$	99.6

(a) Authors, systems and respective results on the task of detection of HYPONYMY in WORDNET

n	recall
1	0.91766
2	0.96863
5	0.99288
10	0.99973

(b) Typical recalls for multi-dimensional interval model. (Precision is always 1.)

Figure 4: Tables

7 Future Work and Conclusion

We found that defining the problem of representing HYPONYMY in a feature vector is not easy. Difficulties include 1. the sparseness of data, 2. whether one wants to base inclusion on an underlying (possibly relaxed) inclusion in the space of vectors, and 3. determining what one should generalize.

Our investigation of WORDNET over *fast*Text demonstrates that WORDNET classes are not cleanly linearly separated in *fast*Text, but they are sufficiently well separated to give a useful recall for an approximate inclusion property. Despite this, and because the negative cases vastly outnumber the positive cases, the rate of false negatives is still too high to give any reasonable precision. One could try to use more complex models, but the sparsity of the data would make such models extremely sensitive to overfitting.

Our second model takes a wholly different approach: we construct intervals directly from the HYPONYMY relation. The main advantage of this method is its simplicity and high-accuracy. Even with a single dimension it rivals other models. A possible disadvantage is that the multi-dimensional version of this model requires disjunctions to be performed. Such operations are not necessarily available in models which need to make use of the HYPONYMY relation. At this stage, we make no attempt to match the size of intervals to the probability of a word. We aim to address this issue in future work.

Finally, one could see our study as a criticism for using WORDNET as a natural representative of HYPONYMY: because WORDNET is almost structured like a tree, one can suspect that it in fact misses many hyponymy relations. This would also explain why our simple *fast*Text-based model predicts more relations than present in WORDNET. One could think of using other resources, such as JEUXDEMOTS (Lafourcade and Joubert, 2008). Yet our preliminary investigations suggest that these suffer from similar flaws — we leave a complete analysis to further work.

References

[Athiwaratkun and Wilson2018] Ben Athiwaratkun and Andrew Gordon Wilson. 2018. On modeling hierarchical data via probabilistic order embeddings. In *International Conference on Learning Representations*.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

[Lafourcade and Joubert2008] Mathieu Lafourcade and Alain Joubert. 2008. JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, pages 657–666, France.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

[Mikolov et al.2018] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[Nickel and Kiela2017] Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

[Vendrov et al.2015] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *CoRR*, abs/1511.06361.

[Vilnis and McCallum2014] Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. *CoRR*, abs/1412.6623.

[Vilnis et al.2018] Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 263–272. Association for Computational Linguistics.

Towards Interpretable, Data-derived Distributional Semantic Representations for Reasoning: A Dataset of Properties and Concepts

Pia Sommerauer, Antske Fokkens and Piek Vossen
Computational Lexicology and Terminology Lab
Vrije Universiteit Amsterdam
De Boelelaan 1105 Amsterdam, The Netherlands
pia.sommerauer@vu.nl, antske.fokkens@vu.nl, piek.vossen@vu.nl

Abstract

This paper proposes a framework for investigating which types of semantic properties are represented by distributional data. The core of our framework consists of relations between concepts and properties. We provide hypotheses on which properties are reflected in distributional data or not based on the type of relation. We outline strategies for creating a dataset of positive and negative examples for various semantic properties, which cannot easily be separated on the basis of general similarity (e.g. **fly**: *seagull*, *penguin*). This way, a distributional model can only distinguish between positive and negative examples through evidence for a target property. Once completed, this dataset can be used to test our hypotheses and work towards data-derived interpretable representations.

1 Introduction

When it comes to representations of word meaning, we currently have to choose between relatively transparent, interpretable representations that are low in coverage and opaque embedding representations with high coverage. While the former lend themselves well to reasoning, the latter are hard to interpret and their reasoning potential remains limited. Ideally, we would have 'the best of both worlds': data-derived, high-coverage transparent representations we can reason over. Reasoning over such vectors would open new opportunities for the study of phenomena at the core of lexical semantics, such as similarity and ambiguity (one form - multiple meanings) and variation (one meaning - multiple forms).

In this paper, we present a framework for analyzing what type of semantic information is present in distributional data as a first step towards such semantic representations. We consider word meaning from the perspective of semantic properties, which enables us to explain semantic similarity and dissimilarity and reason over word meanings. We propose a methodology that can be used to create datasets representing concepts and their semantic properties, which can be used to test hypotheses about what type of information is present in distributional models.

When trying to model the type of semantic information represented by linguistic context, the following questions arise: Which aspects about the meaning of a word can be expected to be mentioned in (written) utterances? Do people talk about the yellowness of lemons? Or would they rather give accounts of what lemons are used for? We propose a number of hypotheses about which type of semantic knowledge is encoded in the linguistic context based on the semantic relation between a particular concept and property.

If distributional vectors contain information about a semantic property, it should be possible to distinguish positive examples of the property from negative examples purely on the basis of the distributional vector. As distributional semantic representations usually provide good indications for general relatedness or similarity, one major pitfall of our approach is that words can easily be separated into positive and negative examples because they happen to fall into rather distinct categories. Therefore, we specifically aim to collect challenging examples (e.g. **fly**: *seagull*, *penguin* rather than **fly**: *seagull*, *table*). We propose a framework for sampling and defining concept-property pairs that, in future work, will be annotated and used to test our hypothesis. To the best of our knowledge, this will be the first dataset specifically designed to analyze the ability of embeddings to encode property information.

Besides being a diagnostic tool, we hope that

the resulting resource will provide complementary information to traditional lexical semantic representations. A core notion in lexical semantics is semantic similarity. Different lexical resources reflect this notion in different ways. Whereas Princeton Wordnet (Fellbaum, 2010; Miller, 1995) structures semantic knowledge in terms of hierarchical categories, we approach similarity from the perspective of property overlap. Implicitly, knowledge about property overlap is also represented in hierarchically structured categories, as they capture information about shared and distinguishing properties. We expect that the final dataset will be a complementary resource to WordNet as it could yield insights into semantic categorization in terms of semantic properties. Currently, our setup only takes English data in consideration, but we think that valuable insights could be gained from extending it to more languages thus enabling cross-linguistic comparisons.

The remainder of this paper is structured as follows: Section 2 outlines insights on semantic properties from various research domains. Based on this, we present a framework of properties and concepts in Section 3, followed by our method for creating our dataset suitable for testing our hypotheses in Section 4. We conclude and discuss the implications of our framework in Section 5.

2 Theoretical Background

This section provides an overview of theories and observations about the type of knowledge encoded in linguistic contexts. In general, we assume that semantic information can either be encoded explicitly (e.g. by expressions such as *lemons are yellow*) or implicitly (e.g. *the lemon rolled off the table*, which indirectly indicates that lemons have a round shape). Both sources of evidence provide sufficient information for humans to infer these properties. It is an open question to what extent this is represented by embedding models. We start this investigation by raising the question of what type of information is likely to be mentioned (either implicitly or explicitly) in natural language.

Different theoretical and applied fields have addressed this question, namely, language generation, corpus linguistics and cognitive theories of word meaning. We draw from approaches about referential expressions (Section 2.1), typical properties and concepts revealed in similes (Section 2.2) and afforded actions and processes (Section 2.3). The remainder of this section provides an outline of these factors which form the basis of our proposed framework, introduced in Section 3.

2.1 Gricean Maxims

One major function of language is to 'point' towards things in the world. This is explicitly modeled in approaches to natural language generation, which include referring expression generation (REG) as a subtask (Gatt and Krahmer, 2018). Dale and Reiter (1995)'s seminal work proposes to model REG in terms of Gricean maxims (Grice, 1975). In essence, humans are expected to refer to objects by being maximally informative while not providing more information than necessary, resulting in the use of maximally discriminative attributes. When given a choice of objects with a range of different, but partly overlapping attributes and the task of singling out a particular one, humans are expected to use only the attribute(s) which is (are) most informative.

Experimental data show that people do tend to overspecify (in as much as 50% of cases (Koolen et al., 2011)) for several reasons: Arts et al. (2011) argue that overspecification in terms of highly salient attributes may facilitate identification of the referent. Rubio-Fernández (2016) claim that the overspecification of color attributes can facilitate object search as it is easier to find something based on multiple pieces of information. For instance, finding a blue cup is easier if you can look for something blue and for a cup, in particular when the target object is the only cup *and* the only blue object. A complementary observation was made by Koolen et al. (2011), who show that overspecification increases with the difficulty of the reference task. However, color attributes also tend to be overspecified for objects which are typically described in terms of color, such as clothes. This later phenomenon possibly is language-dependent, as it was observed for English speakers but not Spanish speakers. More generally, Sedivy (2003) found that color attributes tend to be used redundantly for objects that have a high color-variability (i.e. things that naturally come in several colors, such as t-shirts). Complementary, Koolen et al. (2011) observe that overspecification occurs for concepts whose instances can be described in terms of many different attributes.

These insights have been obtained from highly controlled lab settings with limited situational

context. An attempt to generalize to the information included in utterances 'in the wild' can be seen as somewhat of a leap. Nevertheless, we expect that in general, people tend to avoid mentioning information which is already available to their interlocutors through their (physical) experience of the world. For instance, we expect that people would hardly ever specify the color or taste of a lemon (unless it is a highly unusual one), since this information is already available to people who have had some sort of experience with lemons. In contrast, we expect that people are more likely to specify target objects in terms of attributes (e.g. color) in case of high variability of attributes or in case strong association between concepts and attributes (typicality). The former could either be due to (1) the reference task being actually harder because of the high variety of attributes or (2) the observed tendency to overspecify in cases of high attribute-variability. The next section discusses how typicality can result in contexts that explicitly reflect shared knowledge.

2.2 Stereotypicality

Veale (2013) explores the way different semantic properties of concepts (most of which can be seen as having 'multifaceted' meanings) can be extracted from text corpora. He proposes that " [...] words[1] are represented as bundles of the typical properties and behaviors they are commonly shown to exhibit in everyday language" (Veale, 2013, p.1) and presents an automatic system to extract and reason over the different affective contents associated with concepts via their most salient properties. For instance, the word *baby* can receive a positive interpretation when appearing in a context highlighting cuteness and peacefulness, but just as well be used in less flattering descriptions such as *cry like a baby*.

Veale's approach shows that information about stereotypical concepts of a property is mentioned in natural language, as it relies on pattern extraction from corpora. Specifically, stereotype information tends to be expressed in similes of forms like *as* ADJECTIVE *as a* NOUN (e.g. *as mindless as a zombie*) or in the case of activities VERBing *like a* NOUN (e.g. *drooling like a zombie*) (Veale and Hao, 2007) .

It seems that implied information about concepts tends to be mentioned explicitly if the concept can serve as a particularly good example to illustrate the (implied) property. While it is unlikely to find instances stating the obvious (e.g. *coal is black*), it is more likely to find utterances in which the stereotypical concept is used to illustrate a property of something else (e.g. *eyes as black as coal*).

2.3 Common Actions and Affordances

Based on accounts in cognitive psychology and cognitive linguistics, we expect (highly implied) knowledge relating to specific types of afforded actions (as introduced by Gibson (1954)) likely to be reflected by linguistic context. Glenberg (1997) argues that a central component of our memory is a set of actions that are available to an agent in a certain situation, which he calls 'mesh'.

Glenberg and Robertson (2000) explore this notion by comparing embodied to high-dimensional (i.e. distributional) theories of meaning. Their experimental results indicate that distributional models provide good indications about the kinds of actions and processes concepts are usually involved in. They are, however, unable to reflect possible (i.e. *afforded*) actions that are highly unusual.

We hypothesize that this is due to a tendency of people to describe and report on specific events in the world, which consist of combinations of actions and processes. Specific events, in contrast to general properties, are very unlikely to be implied knowledge and therefore have to be communicated (e.g. *dogs have four legs* versus *My dog ran towards the ball*). A large corpus is more likely to contain patterns that arise from specific activities and processes (e.g. dogs will often be involved in running events), while unusual activities will be too erratic to lead to meaningful regularities in the data that end up represented in distributional models.

2.4 Summary of Factors

When determining whether a specific semantic property is likely to be encoded by distributional information, we consider the following factors to be relevant:

> **Impliedness**: Which information is already known, Which information has to be made explicit?
> **Variability**: Do the instances of a concept vary with respect to the target property?

[1]This paper is on word meaning. The expression 'word' should be read as referring to word meaning.

Typicality: Is a concept likely to be used to illustrate a property?
Affordedness: Do certain properties afford activities that instances of a concept engage in? In other words: Are there properties of a concept which enable certain activities?

3 Contextually Encoded Properties

Based on the observation outlined in Section 2, we present predictions about whether a specific distributional vector representation of a concept is likely to encode information about a specific semantic property or not. To operationalize this, we translate the factors discussed in Section 2 to descriptions of relations between concepts and properties. We assume that knowledge about properties of concepts is generally implied and hence unlikely to be expressed explicitly (following the Gricean maxim of quantity). However, there are a number of factors which cause violations against this general tendency. We translate these competing forces to relations between properties and concepts. We outline them below and summarize them in Table 1, which also provides an overview of our hypotheses.

Typicality. Typical properties of instances of a concept are usually also highly implied (e.g. *rose* - **red**). While a high level of impliedness in combination with Gricean maxims would mean that the property is unlikely to be mentioned explicitly, typicality may have the opposite effect. Based on the observations by Veale (2013), we expect that typical examples of a property can often serve to illustrate the property in a another concept (e.g. *coal* serves to illustrate blackness in the phrase *eyes as black as coal*, *rose* may serve to illustrate redness, etc). In contrast, properties that immediately come to mind when thinking of a concept, but not vice-versa are unlikely to be represented, but can be seen as highly implied (e.g. **green** is a typical property of *broccoli*, but *broccoli* is usually not used to illustrate greenness).

Affordedness. In general, we propose that afforded and usually performed activities are represented, while afforded and not usually performed activities are not (e.g. bowling ball - roll v.s. candle - roll). Usually performed activities can be seen as highly implied knowledge about a concept. However, the fact that activities usually form part of specific events (which are not part of our implied knowledge) makes them much more likely to be mentioned in communication than other highly implied properties. In addition to being afforded properties themselves, activities can also provide indirect evidence for other properties. In particular, they provide indirect evidence for those properties which enable the activity. For instance, *bowling balls* are commonly involved in rolling-activities. The context is likely to provide direct evidence of the activity **rolling** (e.g. *The bowling ball rolled by 5-foot-10*).[2] The same evidence can also serve as an indirect indication for the property affording the rolling-activity, namely being **round**. Many properties of a concept are, however, not necessarily reflected in activities. Consider, for instance *candles*: even though they are often round (an affording property for the activity of rolling), rolling is not something they typically do. In the remainder of this paper, we use the following sub-types of properties: We distinguish activities from attributes. Activities can be afforded and usually performed or afforded and not usually performed (or not afforded at all). Attributes can fall under any of the relations outlined here. In addition, they can afford activities.

Variability. This factor refers to the degree of variation in instances of a concept. In general, we propose that variable properties are likely to be represented by linguistic contexts because they can be relevant for further distinctions and are not automatically implied. For instance, a color attribute can distinguish between different subcategories of bears or distinguish between peppers with different tastes, knives can be used for different cooking activities or processes, etc. These variable properties can have different degrees of discriminatory power. On one end of the spectrum, they distinguish between different conceptual categories (e.g. subcategories of bears). At the other end of the spectrum, they distinguish instances of the same category (e.g. t-shirts of different colors or dogs trained for different activities). While in this later case, there is a very high probability of properties to be mentioned explicitly, we do not expect the evidence to be enough to be captured by a distributional semantic model: due to the high degree of variance, individual properties will be mentioned sporadically at best. Properties that can only apply to instances of concepts in exceptional cases are not expected to be represented.

[2] https://www.latimes.com/archives/la-xpm-1991-05-30-sp-3586-story.html

factor	present	absent
typicality	concept is typical of the property	property is typical of the concept
afforded activities	usually performed	possible but not usually performed
affording attributes	affording usually performed activities	not relevant for usually performed activities
variability (options)	limited (also values on a scale or opposites)	wide selection
variability (categories)	subcategories	not relevant for subcategories

Table 1: Overview of relations between concepts and properties: **present** and **absent** indicate whether the concept-property relation is hypothesized to be apparent from distributional data.

Table 1 provide an overview of the relevant factors and related prediction. A single concept-property pair can be related to more than one factor. For instance, *sky - blue* can be described in the following terms:

Implied : **blue** is a highly implied property of *sky*

Typical (concept) : **blue** is a typical property of *sky*

Typical (property) : *sky* is a stereotypical example of something which is **blue**

Variable (limited) : *skies* can also be **grey** or **black**

If at least one description falls under *present* in Table 1, we expect the context to contain evidence for the property. Whether this evidence is sufficient for a distributional model to represent the property is an open question.

4 A Dataset of Concepts and Properties

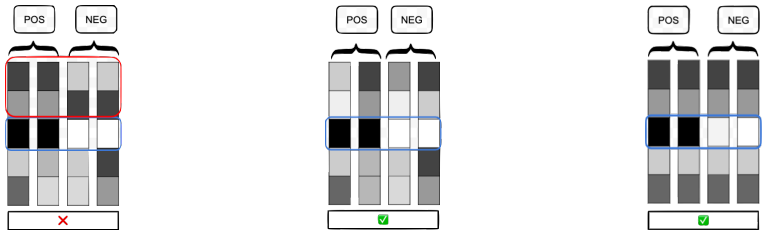
This section describes the design of our dataset. We first outline the experiments we envision, because they provide the motivation of some of the key properties of our dataset.

To conduct experiments on whether the predictions introduced in Section 3 hold, we plan to use approaches suggested in the field of investigating neural network representations, such as diagnostic classification (Belinkov et al., 2017; Hupkes et al., 2018; Derby et al., 2018). In particular, we plan to extend the experiments presented in Sommerauer and Fokkens (2018), which try to investigate whether dimensions of embedding representations can capture semantic properties. While this seems to be implied by the method of inferring the missing word in an analogy pair by means of vector subtraction and addition (Mikolov et al., 2013; Levy and Goldberg, 2014), analogy calculation methods have been heavily criticized, calling this notion into question (Linzen, 2016; Gladkova and Drozd, 2016; Gladkova et al., 2016). To shed light on this, we proposed an experimental set-up in which we tested whether a supervised machine learning system could successfully learn to distinguish vectors of words clearly associated with a property from vectors of words which are clearly not associated with the property.

Any supervised classification approach relies on finding regularities which are shared among all or most examples of a particular class and distinguish them from other classes. Therefore, the distribution of positive and negative examples of properties is crucial to ensure that the vector dimensions discovered by the classifier actually correspond to the semantic property under investigation rather than some other information which happens to correlate with it. To illustrate the importance of the similarity distribution of positive and negative examples, consider the following: Suppose our dataset for the property **red** consists of names of red fruits (positive examples) and green garden plants (negative examples). If we train and test a classifier on such a dataset, it is very likely that it can reach relatively high performance. But did it learn to identify the semantic property **red** in a distribution? In such a case, it would be impossible to draw a clear conclusion for the following reasons: The names of the red fruits most likely share more properties than being red, such as having a sweet taste, being used for similar things, or largely falling into the category of berries. Consequently, more information connects these examples than the property **red**. The same holds for the negative examples: they belong to a relatively coherent category and probably share many properties. Many of these properties will not be shared with the positive examples. This means that a classifier can rely on a multitude of indications, none of which are necessarily evidence of the target property **red**. Figure 1 illustrates different scenarios of shared and distinguishing features.

To address this challenge, our dataset has to ad-

(accessed 2019/09/30)



(a) Target property (present in black, absent in white) and other dimensions correlates with positive and negative classes.

(b) The only dimension correlating with the positive and negative classes is the target dimension of the target property.

(c) Highly similar positive and negative examples that can only be distinguished by the dimension of the target property

Figure 1: A schematic representation of vectors of positive and negative examples of a property. To ensure that shared and distinguishing patterns identified by a classifier are representative of the target property, positive and negative examples should only be separable based on the target property.

here to the following requirements:

1. For each property, there is a sufficient number of positive and negative examples.
2. The distinction between positive and negative examples cannot be made on the basis of general similarity alone (see Figure 1). The candidates should include (1) positive examples that differ with respect to most properties except the target property, i.e. that have low overall similarity (e.g. **fly**: *seagull*, *airplane*) and (2) negative examples that share a number of properties with positive examples, creating high similarity between positive and negative examples (e.g. **fly**: *seagull*, *penguin*)

Most existing feature norm sets (McRae et al., 2005; Devereux et al., 2014) do not contain information about negative examples, as they only list (salient) properties of concepts. One might consider to derive negative examples by viewing all concepts *not* labeled with a certain feature as negative examples of the feature. This approach, however, results in a number of wrongly labeled instances, as positive cases are not always labeled as such (for instance, 18 out of 36 concepts labeled as **is_a_bird** are not labeled as **is_an_animal** in the CSLB feature norms (Devereux et al., 2014)).

Our main objective is to collect fine-grained information for property-concept pairs to fill this gap. Through crowd annotations, we aim to divide these property-concept pairs into three categories: Properties which apply to **all or most**, **some** or **hardly any or no** instances of a concept. We draw the line in the middle of the 'some' category, which encompasses different degrees of variability: while we expect attributes with little variance to have enough evidence for a model, attributes with a high degree of variability are most likely not encoded.

The second requirement can be fulfilled by controlling (a) the selection of target properties (see Section 4.1), (b) the selection of candidate concepts from resources (Section 4.2) and (c) the selection of particularly challenging examples in the distributional semantic space (Section 4.3). Sections 4.4 and 4.5 provide further details on the setup of our crowd sourcing task.

4.1 Selecting Challenging Properties

We select semantic properties which apply to concepts that are spread across traditional, taxonomic categories. We consider the following types of properties: perceptual attributes (e.g. colors, shapes, temperature), part attributes (e.g. *having wheels*), complex attributes (high level semantic categories such *being dangerous*) and activities (e.g. *swim*, *fly*). We hand-selected specific properties (listed in Section B of the Appendix) for each type based on the criteria of them cutting across taxonomic categories and applying to a large number of concepts.

4.2 Selecting Challenging Concepts

We collect candidate concepts from existing computational and psycholinguistic resources, listed in Table 2, and from a distributional model. By exploiting the feature norm sets and the stereotype data, we get a limited set of candidates 'for free' by searching for the selected properties directly.

By searching for target properties directly (e.g.

concepts associated with **round** in ConceptNet via the relations *HasProperty* or *NotHasProperty*), we only receive limited sets of examples, in particular with respect to negative candidates. Therefore, we extend the search by including concepts of particular traditional, taxonomic categories whose members we expect to have or not have the target property. We explain the idea through the activity **fly**.

Concepts that are similar and only differ with respect to **fly** or categories which contain positive and negative examples are particularly useful. We exploit this in our sampling strategy: we know that while most birds can fly, some cannot. The category of insects also contains both cases. In addition, we could add vehicles. While the first two categories contain similar concepts that share a large number of properties, the later category introduces words that share almost no properties with the first two except the target property.

For this type of search, we exploit the hyponymy relations of WordNet as well as properties from the feature norm sets and the corpus data. For WordNet, we manually select the synset representative of a category (based on synset members and definitions) and collect all lemmas of its hyponym-synsets. In the feature norm data, we simply search for the target property. In addition, we use the positive and negative examples derived from the CSLB norms and annotated by the crowd as described by Sommerauer and Fokkens (2018).

This strategy is successful for some properties (e.g. 105 probably positive and 256 probably negative candidates for the **black**) but less for others (e.g. 6 probably positive and 63 probably negative candidates for **round**). While we try to select negative examples that are difficult to distinguish from positive ones through other properties than the target property, it is not entirely clear whether this is the case. To extend our examples and at the same time target particularly challenging negative examples, we use an existing distributional model as a source of additional examples.

4.3 Challenging Examples using Embeddings

Distributional semantic models provide relatively good indications of word similarity, reflecting the assumption that words with similar meanings tend to appear in similar linguistic contexts. However, they cannot give us precise information about what makes words similar. The main challenge of our approach is to select examples that could not be distinguished purely on the basis of this distributional similarity. Therefore, we specifically select examples from a distributional model which have a very high chance of being classified wrongly based on their similarity (e.g. *penguin* for **fly**, or *heroine* for **dangerous** while other positive examples are weapons or animals). If it can be classified correctly, we can interpret this as good evidence for the property to be encoded in the distributional vector representation.

type	**resources**
feature norm sets	McRae et al. (2005), CSLB norms (Devereux et al., 2014)
lexicon	WordNet (Fellbaum, 2010; Miller, 1995) ConceptNet (Speer and Havasi, 2012)
stereotype data	concepts representing stereotypes of properties (Veale, 2013)
feature norms negative extension	subset annotated on top of the CSLB norms (Sommerauer and Fokkens, 2018), quantified McRae norms (Herbelot and Vecchi, 2015)

Table 2: Overview of resources.

To operationalize this, we select positive 'seed' words and calculate a vector representation for them by taking the average of the seeds. This allows us to specifically select candidates with embedding representations that are overall similar to positive examples of a property (by taking the n nearest neighbors of the averaged representation). We select these positive 'seeds' by using positive examples of a property we are confident about (i.e. we do not include concepts returned by a search for a category containing 'mixed' examples).

This results in a selection of candidate concepts which are very difficult to separate into positive and negative examples based on general similarity. We collect the 200 nearest neighbors of this approximate property representation. We exclude negative examples further away from the centroid than the furthest positive example by manual inspection. The embedding model used in this step is the skip-gram model with negative sampling (using recommended settings according to Levy et al. (2015)), trained on the full Wikipedia corpus (dump from August 2018).

4.4 Sampling for the Crowd

The strategies outlined above result in rather large numbers of candidates not all of which are useful

(e.g. the distributional model returns non-standard spelling variants and words other than nouns). We reduce and clean the resulting sets (1) by means of preprocessing and (2) sampling based on characteristics with potential impact on how well distributional data can represent information. The characteristics we consider are (1) different types of ambiguity, (2) psycholinguistic factors such as concreteness and familiarity represented in the MRC database (Coltheart, 1981), word frequency (3) the distance to the centroid vector calculated over all positive examples of a property.

type	**n syns**	**wup sim**	**min wup sim**	**cos syns**	**abs-conc**
homonyms	8.28 (6.97)	0.32 (0.16)	0.20 (0.18)	0.20 (0.20)	0.60
metaphors	8.32 (7.68)	0.35 (0.19)	0.24 (0.21)	0.24 (0.23)	0.45
metonymy (approx.)	3.01 (2.72)	0.53 (0.32)	0.48 (0.35)	0.57 (0.38)	0.24
monosemy	1.97 (2.43)	0.78 (0.32)	0.76 (0.35)	0.80 (0.31)	0.09

Table 3: Averages on nouns only (standard deviation in parentheses).

We create bins for each characterization, distinguishing four types of polysemy and three histogram bins for each of the other characteristics. Except for cosine to centroid, we use the distribution of all nouns recorded in the LDOCE dictionary (Proctor, 1978) to divide candidates across bins. For each characterization, we randomly draw examples from each bin until we reach a certain predefined number of examples for probably positive, probably negative or undecided candidates. The resulting distributions are summarized in Table 4.

We aim to include different types of ambiguity, since ambiguity is of particular interest for our further research. We are not aware of a lexical resource providing fine-grained information about types of ambiguity. To approximate it, we exploit metaphor annotations in the MIPVU corpus (Steen, 2010) and the distinction between homonymy, polysemy and monosemy information in the LDOCE dictionary. The third group we distinguish consists of other forms of polymsemy (metonymy, specialization and generalization). While it is not feasible to verify this approximation manually, we tested a number of tendencies which should hold if our approximation strategies are appropriate: The similarity between senses of ambiguous words should correlate with the semantic phenomena involved in it: Senses of homonymous words should be least similar while mentonymous senses should be most similar. This can be measured in terms of WordNet similarity or embedding vector similarity with the monosemous synset members of the senses. The sense similarity/distance can also be analyzed in terms of very broad semantic areas that a sense can fall into. Homonymous senses accidentally share the same form and metaphorical words often express mappings between abstract and concrete domains. Therefore, we expect that the latter two tend to have senses in both the abstract and concrete part of the WordNet hierarchy, while this should not be the case for metonymous senses (which typically remain restricted to one part of the hierarchy).

property	pos	neg	pos/neg	total
warm	20	28	118	166
hot	19	20	108	147
red	46	59	69	174
square	6	23	90	119
green	57	58	60	175
cold	18	22	81	121
sweet	28	1	145	174
blue	22	60	61	143
yellow	45	65	64	174
round	37	2	101	140
black	60	58	34	152
juicy	20	6	148	174
swim	57	61	62	180
roll	4	1	115	120
lay_eggs	61	61	32	154
fly	58	61	61	180
dangerous	63	61	17	141
used_in_cooking	59	60	60	179
female	57	11	48	116
wheels	54	16	45	115
wings	58	60	29	147
made_of_wood	59	12	81	152

Table 4: Overview of dataset size after sampling.

As the results summarized in Table 3 indicate, the ambiguity bins seem to provide a decent representation homonyms, words with metaphorical and metonymous senses and (for the same of comparison) monosemous words. We therefore use them for sampling.

4.5 Framework for Collecting Judgments

The resulting candidate concepts should be annotated in terms of their relations to the target property. To do this in an efficient way, we present

relation	examples	T/F
unusual	In an unusual situation, *chocolate* could be **pink**.	True
	In an unusual situation, *chocolate* could be brown.	False
affording_activity	Having **ink** is necessary for things a *pen* usually does or for things we usually do with a pen.	True
	Being **grey** is necessary for things a *car* usually does or for things we usually do with a *car*.	False
typical_of_concept	Being **spicy** is typical of a *chili pepper*.	True
	Being **sweet** is a typical property of a *carrot*.	False
variability_open	A *t-shirt* can be **white** or of another property of the same category as **white** there is a very wide set of possible options.	True
	A *pepper* can be **white** or of another property of the same category as *white* there is a very wide set of possible options.	False

Table 5: Examples of concept-property relations for crowd annotation with most appropriate True/False-judgment.

crowd workers with statements about the relation between a concept and a property and ask them to indicate whether it is generally true or false. We opt for this set up rather than presenting workers with all options, as it is faster and will most likely seem more attractive.[3] Rather than presenting generic, abstract descriptions of a property-concept pair, we present sentences such as the examples presented in Table 5, which are supposed to be natural-sounding and easy to judge.

5 Discussion and Conclusion

In this paper, we have outlined a method to create a dataset of semantic properties of concepts which can be used to evaluate whether and to what extent distributional models reflects semantic properties. This work can be positioned in our larger research goals, which involve creating transparent, interpretable lexical semantic representation in terms of semantic properties which lend themselves well for reasoning over ambiguity and variation. The dataset will be made available upon completion.[4]

The main goal of this paper is to propose a design for a dataset that can be used to test the ability of word embeddings to represent semantic properties. A more precise understanding of what information word embeddings can provide is highly relevant for improving NLP systems relying on embeddings as lexical semantic representations. Moreover, it can help in deciding whether embeddings are an appropriate representation in computational models of cognitive processes (as for instance discussed by Utsumi (2011)). Eventually, we plan to move towards data-derived interpretable word representations in terms of semantic properties.

The dataset proposed here enables us to use methods suggested in the area of studying representations and learning processes in neural networks, specifically diagnostic classification to test whether embeddings represent properties. In particular, we can go beyond the approach presented by Derby et al. (2018), who use all concepts for which a property has not been elicited as negative examples of a property.

In addition to proposing a dataset design, we offer specific hypotheses based on a variety of observations from different fields about information that is likely or unlikely to be expressed in English natural language corpora. Rather than making claims based on entire categories of semantic properties, we base our predictions on underlying factors involved in the relations between concepts and properties. By testing these hypotheses, we hope to go beyond insights from experimental approaches comparing the information captured in embeddings to semantic feature norm sets (e.g. Fagarasan et al. (2015), Herbelot and Vecchi (2015), Tsvetkov et al. (2015), Derby et al. (2018), Sommerauer and Fokkens (2018)).

Finally, we hope that comparing the relations captured by our dataset to traditional, taxonomic categories represented in WordNet may yield insights about the relation between properties of concepts and categorization. This could be extended to other languages to enable cross-linguistic comparisons.

[3] At this point, the exact set-up of the task is still under development. The resulting dataset will be made available once data have been collected.

[4] https://github.com/cltl/semantic_property_dataset

Acknowledgments

This research is funded by the PhD in the Humanities Grant provided by the Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) PGW.17.041 awarded to Pia Sommerauer and NWO VENI grant 275-89-029 awarded to Antske Fokkens. We would like to thank Emily Bender and anonymous reviewers for feedback that helped improve this paper. All remaining errors are our own.

References

Anja Arts, Alfons Maes, Leo Noordman, and Carel Jansen. 2011. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 1–10.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. 2018. Representation of word meaning in the intermediate projection layer of a neural language model. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 362–364.

Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 52–57.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

James J Gibson. 1954. The visual perception of objective motion and subjective movement. *Psychological Review*, 61(5):304.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Arthur M Glenberg and David A Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401.

Arthur M Glenberg. 1997. What memory is for. *Behavioral and brain sciences*, 20(1):1–19.

HP Grice. 1975. Logic and conversation. *Foundations of Cognitive Psychology*, page 719.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

P Proctor. 1978. *Longman Dictionary of Contemporary English*. Longman Group, Essex, UK.

Paula Rubio-Fernández. 2016. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, 7:153.

Julie C Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, 32(1):3–23.

Pia Sommerauer and Antske Fokkens. 2018. Firearms and tigers are dangerous, kitchen knives and zebras are not: Testing whether word embeddings can tell. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2049–2054, Lisbon, Portugal.

Akira Utsumi. 2011. Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive science*, 35(2):251–296.

Tony Veale and Yanfen Hao. 2007. Learning to understand figurative language: from similes to metaphors to irony. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

Tony Veale. 2013. The agile cliché: using flexible stereotypes as building blocks in the construction of an affective lexicon. In *New Trends of Research in Ontologies and Lexical Resources*, pages 257–275. Springer.

A Framework of semantic relations between concepts and properties

factor	relation description	example	represented	instances
impliedness	(A/an) [concept] is part of a larger category of which all members are [attribute].	**animate** - *cat*	no	all/most
typicality (of the concept)	(A/n) [concept] is a typical examples of things which are [attribute].	**green** - *broccoli*	no	all/most
typicality (of the property)	[attribute] is a typical property of (a/an) [concept].	**blue** - *sky*	yes	all/most
afforded (attribute)	Being [attribute] is necessary for activities/processes (a/an) [concept] is usually involved in.	**has a point** - *dagger*	yes	most/all
variability (distinction)	[attribute] is an important factor to distinguish different subcategories of members of the category [concept].	**grey** - *bear*	yes	some
variability (limited)	(A/an) [concept] can be [attribute] or another attribute of same category as [attribute] - there is a limited set of possible options.	**red** - *pepper*	yes	some
	(A/an) [concept] can be [attribute] or a bit more [attribute] or the opposite of [attribute].	**warm** - *water*	yes	some
variability (open)	(A/n) [concept] can be [attribute] or another attribute of same category as [attribute] - there is a very wide set of options.	**pink** - *t-shirt*	no	some
Variability (unlikely)	(A/an) [concept] is [attribute] could only be true in a rather unusual situation.	**blue** - *horse*	no	few/none
Variability (creative)	(A/an) [concept] is [attribute] can only be true in a creative, figurative way of speaking.	**round** - *idea*	no	few/none
Impossible	It is impossible that (a/an) [concept] is [attribute].	**solid** - *steam*	no	none

Table 6: Overview of relations between attributes and concepts.

factor	relation description	example	represented	instances
impliedness	(A/an) [concept] is/are part of a larger category of which all members can do/are involved in [activity].	**breathe** - *cat*	no	most/all
typicality (of the concept)	'[activity]' is a typical activity or process of (a/an) [concept].	**fly** - *bird*	no	all/most
typicality (of the property)	(A/an) [concept] is/are typical example(s) of things which do/are involved in the activity or process '[activity]'.	**hunt** - *tiger*	yes	all/most
afforded (activity)	(A/an) [concept] usually does/is involved in the activity or process '[activity]'.	**run** - *horse*	yes	most/all
	(A/an) [concept] can do/be involved in the activity or process '[activity]' but this is not what it/they usually does/do.	**roll** - *pen*	no	most/all
variability (distinction)	Doing/being involved in the activity or process '[activity]' is an important factor for distinguishing different subcategories of members of the category [concept].	**cooking** - *knife*	yes	some
variability (open)	(A/an) [concept] can do/be involved in the activity or process '[activity]' or not, but this is not an important factor for distinguishing different subcategories of members of the category [concept].	**play** - *dog*	no	some
Variability (unlikely)	(A/an) [concept] does/is involved in the activity or process '[activity]' could only be true in a highly unusual situation.	**fly** - *car*	no	few/none
Variability (creative)	(A/an) [concept] does/is involved in the activity or process '[activity]' can be only true in a creative, figurative way of speaking.	**fly** - *idea*	no	few/none
Impossible	It is impossible that (A/an) [concept] does/is involved in the activity or process [activity].	**fly** - *horse*	no	none

Table 7: Overview of relations between activities and concepts.

B Overview of selected properties

property type	category	properties
attributes	perceptual	warm, hot, red, square, green, cold, sweet, blue, yellow, round, black, juicy
	parts	wheels, wings, made_of_wood
	complex	dangerous, found_in_seas, used_in_cooking, female
activities	swim, roll, lay_eggs, fly	

Table 8: Overview of properties currently included (open for expansion).

Connections between the semantic layer of *Walenty* valency dictionary and PLWORDNET

Elżbieta Hajnicz
Institute of Computer Science,
Polish Academy of Sciences
Warsaw, Poland
hajnicz@ipipan.waw.pl

Tomasz Bartosiak
Institute of Computer Science,
Polish Academy of Sciences
Warsaw, Poland
tomasz.bartosiak@gmail.com

Abstract

In this paper we discuss how *Walenty* is using PLWORDNET to represent semantic information. We decided to use PLWORDNET lexical units and synsets to describe both the predicate meaning and the semantic fields of its arguments. The original design decision required some further refinement caused by the structure of PLWORDNET and complex relations between arguments.

1 Introduction

Walenty, a comprehensive valency dictionary of Polish developed at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS), is created to a large degree as a part of CLARIN-PL (Przepiórkowski et al., 2014a; Przepiórkowski et al., 2014b).[1] It was meant to be used both by computer programs (e.g. it is employed by two parsers of Polish, POLFIE[2] (Patejuk and Przepiórkowski, 2012) and Świgra[3] (Woliński, 2004)) and by linguists.

The dictionary comprises above 18,000 entries (with over 101,000 schemata and 31,000 frames), including 13,000 verbs, 4,000 nouns, 950 adjectives and 200 adverbs. Therefore, nonverbal entries form 28% of the lexicon.

Walenty is composed of two main layers: syntactic and semantic. The syntactic layer was described in (Przepiórkowski et al., 2014c; Przepiórkowski et al., 2014a; Hajnicz et al., 2016b), whereas (Przepiórkowski et al., 2014b) focuses on its phraseological component. On the other hand, the semantic layer was sketched in (Hajnicz et al., 2016a).

The semantic layer of *Walenty* is strictly connected with PLWORDNET (Piasecki et al., 2009; Piasecki et al., 2016), one of two Polish wordnets.[4] PLWORDNET describes the meaning of a lexical unit by placing this unit in a network of relations (such as synonymy, hypernymy, meronymy, etc.).

In this paper we want to focus on how semantic layer of *Walenty* was influenced by PLWORDNET and its structure.

2 Related works

There exist valency dictionaries connecting syntactic and semantic information about predicates and their arguments. The most famous is FrameNet[5](Fillmore et al., 2003; Ruppenhofer et al., 2006) based on a theory called Frame Semantics (Fillmore, 1976; Fillmore and Baker, 2001). It is organised around the notion of a *semantic frame* representing a situation. A semantic frame is evoked by lexical units representing corresponding meanings of words (not only verbs). Frames are lists of semantic roles called *frame elements* (FEs).

FrameNet contains about 800 hierarchically organised frames evoked by 10 000 lexical units. Frames are organised in a hierarchy which relates lexical units evoking them. Apart from a hierarchy, frames are organised into scenarios. Nevertheless, FrameNet lexical units are not related to a wordnet (in particular, Princeton WordNet, (Fellbaum, 1998; Miller and Fellbaum, 2007)) and create independent structure[6].

Another important valency dictionary is VerbNet[7] (Kipper-Schuler, 2005) based on the classification of verbs by Levin (1993). Each verb class in VerbNet is completely described by semantic roles, selectional restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a temporal function. VerbNet describes about 5250 senses of 3800 verb lemmas. Each verbal sense in VerbNet may refer to a set of Wordnet senses that captured the meaning appropriate to the corresponding Levin's class

[1] http://www.clarin-pl.eu/en/

[2] http://zil.ipipan.waw.pl/LFG

[3] http://zil.ipipan.waw.pl/%C5%9Awigra

[4] The other one is *PolNet* (Vetulani et al., 2009; Vetulani, 2014; Vetulani and Kochanowski, 2014) developed at Adam Mickieiwcz University by Zygmunt Vetulani Group.

[5] https://framenet.icsi.berkeley.edu/fndrupal/

[6] There were several attempts to relate the resources, cf. (Cao et al., 2010).

[7] https://verbs.colorado.edu/~mpalmer/projects/verbnet.html

obléci$^{1}_{pf}$ / oblékat$^{1}_{impf}$ / obléknout$^{1}_{pf}$ / ustroit$^{1}_{pf}$ / stroit$^{1}_{impf}$

```
=canbepassive   yes
=class          dress-41.1.1
```

1 obléci:1 / oblékat:1 / obléknout:1

```
-frame:     AG<person:1>[obl kdo1]   VERB
            PAT<person:1>[obl komu3]
            ART<garment:1>[obl co4]
-synonym:   ustroit:1 / stroit:1
-use:       prim
-refl:      obj_dat
```

2 obléci:1 / oblékat:1 / obléknout:1 / ustroit:1 / stroit:1

```
-frame:     AG<person:1>[obl kdo1]   VERB
            PAT<person:1>[obl koho4]
            ART<garment:1>[obl do čeho2]
-synonym:
-use:       prim
-refl:      obj_ak
```

Figure 1: An exemplary entry of VerbaLex valency dictionary

(Dang et al., 1998; Kipper et al., 2000). Moreover, selectional restrictions are based on semantic categories labelling WordNet files. The syntactic valency information is represented by means of *LTAG* trees.

There exist several Czech valency dictionaries. Two of them, VALLEX (Lopatková et al., 2003; Žabokrtský and Lopatková, 2007) and PDT-VALLEX (Hajič et al., 2003; Urešová, 2009), are based on Functional Generative Description (Sgall et al., 1986). Despite common origins those dictionaries have been developed independently, following different approaches. While the first one tries to encompass all frames for a given lexeme, the latter is connected with Prague Dependency Treebank and has only those frames that were encountered in the corpus. In both dictionaries frames representing semantics are syntax driven, with multiple syntactic realisations of a single word meaning creating multiple (often different) frames. Nonetheless, frames are not connected to any wordnet.

A third one, VerbaLex (Hlaváčková and Horák, 2006) is connected with the Czech WordNet (Pala and Smrž, 2004; Rmbousek et al., 2017). Valency frames are connected with whole synsets, not particular lexical units. The semantic characteristic of arguments has two level representation and consists of a set of semantic roles including 40 elements from EuroWordNet top ontology (Vossen, 1998) and more precise semantic types including specific literals (lexical units) from the set of Princeton WordNet Base Concepts with relevant sense numbers. Semantic types correspond to selectional restrictions/preferences. On the other hand, the frames are connected to Levin's classes and hence with VerbNet.

Figure 1 presents an exemplary entry of VerbaLex. A frame corresponds to a synset containing five lexical units, but only three of them can be used in **1** as other two do not follow the same syntax.

There exist some Polish valency dictionaries as well. The most important are (Polański, 1980 1992; Świdziński, 1994). Only the first one includes semantic information, i.e. abstract selectional restrictions (cf. Figure 2, e.g. NP^1_A has to have 'Anim' property, while NP^2_A has to have 'Abstr' property). A corpus-based dictionary including some purely syntactic valency information is (Bańko, 2000).

LUBIĆ

$$NP_N \quad - \left\{ \begin{array}{l} NP^1_A \quad + \quad \left(\left\{ \begin{array}{l} za \cap NP^2_A \\ za \cap Ts_A, że \cap S \end{array} \right\} \right) \\ NP^3_A \\ żeby \cap S \\ IP \end{array} \right\}$$

$$NP_N \longrightarrow [+Hum]$$
$$NP^1_A \longrightarrow [+Anim] \qquad NP^3_A \longrightarrow \left[\begin{array}{l} -Abstr \\ -Anim \end{array} \right] \; [+Abstr]$$
$$NP^2_A \longrightarrow [+Abstr]$$

Figure 2: Exemplary entry for the verb LUBIĆ 'like' in Polański's valency dictionary

3 Basic information about the dictionary

The representation language of *Walenty* is in general universal w.r.t. parts of speech. Each lexical entry is identified by its lemma (e.g. verb GNIEWAĆ 'irritate' noun GNIEW 'anger', 'irritation' or adjective GNIEWNY 'angry', 'irritated').

On the syntactic level, each entry is divided into subentries according to its grammatical properties. Reflexive mark, aspect (both only for verbs), predicativity (only for adjectives and adverbs) and negativity are taken into account. For instance, the entry GNIEW has exactly one subentry **gniew (,,)**, whereas GNIEWAĆ has two subentries **gniewać (_,,imperf)** and **gniewać się (_,,imperf)**.

Each subentry may have any number of syntactic valency schemata[8] assigned, each being a set of syntactic positions. A syntactic position is a set of phrase types – if two morphosyntactically different phrases may occur coordinated, they are taken to be different realisations of the same position (Szupryczyńska, 1996). Labels are used to distinguish special argument positions – subject and object (if they occur). In *Walenty* we decided that

[8] We use a term *schema* for the syntactic level representation and a term *frame* for the semantic level representation.

subject and object syntactic positions[9] are marked only for verbs. However, there exist theories, e.g. generative ones, in which nouns, at least some of them (*derived nominals*), have (deep) subjects and objects (Chomsky, 1970). The required information can be inferred from dependencies between derivationally connected entries as both syntactic positions represent the same argument, cf. section 6. Additional label `head` was introduced in order to represent a non-local control dependency between the head of an adjective and its infinitival argument (e.g. *Szukają kompozytorów gotowych tworzyć z nimi nowoczesny teatr.* '[They] are looking for composers [who are] ready to create a modern theater with them.'). This matter, similarly as other issues specific for syntax of nonverbal predicates, goes beyond the scope of this article.

4 Semantic layer

The semantic layer is composed of semantic frames. Each frame is a set of semantic arguments represented as pairs ⟨semantic role, selectional preferences⟩. The set of semantic roles is presented in Figure 3 – they have colours assigned to them in a fixed way. More information about semantic roles in *Walenty* is included in (Hajnicz et al., 2016a). We assume that there cannot be two identical frames for a single entry, as otherwise there would be no way to distinguish between their meanings. This requirement does not concern frames identified by multi-word lemmas if they correspond to a different meaning.

	Initial Group	Accompanying Group	Ending Group
Main Roles	Initiator Stimulus	Theme Experiencer Factor Instrument	Recipient Result
Auxilary Roles	Condition	Attribute Manner Measure Location Path Time Duration	Purpose
Attributes	Source	Foreground Background	Goal

Figure 3: Table of *Walenty*'s roles

[9]Representation of subject and object in *Walenty* was described in (Przepiórkowski et al., 2014a).

4.1 Identification of the meaning

Each frame is connected to the meaning of a predicate. Those meanings are identified by PLWORDNET lexical units (LUs). We use PLWORDNET version 2.1, as it was the current version at the moment we started works on the semantic layer of *Walenty*.

Contrary to VerbaLex, *Walenty* frames are assigned to predicate lemmas, not to synsets. Therefore, synonyms are not related within the dictionary. This approach prevents us from overlooking some subtle differences between frames concerning selectional preferences or even presence of a particular argument (e.g. *Instrument*). The technical matter concerning potential side-effects of changes in PLWORDNET are also important.

Nevertheless, it is possible for multiple LUs to correspond to the same frame. There are three main reasons for that to happen:

1. Lexical units are derivationally connected. This includes:
   - reflexive and non-reflexive verbs, provided that they represent the same meaning (diathesis alternations, e.g. GNIEWAĆ 'to irritate' and GNIEWAĆ SIĘ 'to be angry'),
   - noun and adjective derivatives of verbs (e.g. DBAĆ 'to care', DBAŁOŚĆ 'a care', DBAŁY 'careful' and NIEDBAŁY 'careless').
2. A single word describing different aspects of situation (e.g. POŻYCZAĆ can mean either 'to borrow' or 'to lend' depending on syntactic structure being a convers of itself).
3. Despite having different hypernyms, a lexical unit cannot be distinguished by semantic frame only (e.g. KOMENTOWAĆ 'to comment' has two lexical units in PLWORDNET– the first with hypernym KRYTYKOWAĆ 'to criticise' and the other with hypernym INTERPRETOWAĆ 'to interpret'– both taking same types of arguments, but being used in different larger contexts).

On the other hand, some lexical units may be absent in PLWORDNET. In such cases new LUs are added, indicated by capital letters instead of numbers following the lemma of an LU (wordnet standard), in order to differentiate them from the original wordnet LUs. Such new LUs are provided with glosses[10] as well as potential location in PLWORDNET structure. For instance, *mleć-A* lit. 'mill' from Figure 5 should be a hyponym of *kręcić-4* 'rotate'. This will facilitate including them by PLWORDNET developers.

[10]Original PLWORDNET LUs may have glosses in *Walenty* as well.

4.2 Selectional preferences

Arguments, identified by semantic roles, are provided with selectional preferences (Katz and Fodor, 1964; Resnik, 1993). Unlike some other dictionaries, we do not use a fixed set of qualifiers, like *abstract/concrete, solid/liquid/gaseous* etc. We want to be much more precise, hence we use PLWORDNET synsets (represented by LUs) and relations to represent selectional preferences. Therefore, it is *dogs* that generally BARK, we tend to DRINK *beverages* (not all *liquids*), and we prefer to use *bandages* to BANDAGE (not every *cloth*).

The selectional preferences are represented as a list of elements of the following four types (elements of different types can cooccur in the same list):

1. a PLWORDNET synset,
2. a predefined set of synsets,
3. a PLWORDNET relation to another argument,
4. a PLWORDNET relation to another synset.

The most basic way to represent selectional preferences is a direct use of PLWORDNET synsets. For instance, the frame of the verb BANDAŻOWAĆ 'bandage' with a strictly constrained meaning is presented in Fig. 4: *istota ludzka-1* 'human being' bandages *część ciała-1* 'body part' of *stworzenie-5* 'creature' by means of *bandaż-1* 'bandage'. Contrary to VerbaLex, we use selectional preferences form a Polish wordnet, not an English one. As a consequence, no interlingual relations are required to check whether selectional preferences are satisfied in a particular sentence. However, the rich structure of PLWORDNET disallow us to use only hyponymy relation in this respect.

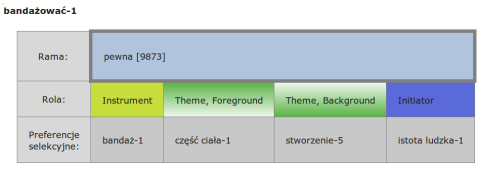


Figure 4: A frame for the verb BANDAŻOWAĆ with PLWORDNET selectional preferences only

In many situations, groups of PLWORDNET synsets commonly occur together in a single selectional preference. For example, both foods and drinks can be tasted or pasteurised. Similarly, both people and organisations/companies can buy, sell or store goods. What is more, people can speak about anything – objects, abstracts and situations. As such semantically connected concepts may be composed of many unrelated PLWORDNET synsets, we decided to add symbols representing such common combinations.

Table 1 lists all the predefined selectional preferences. The first column contains their labels, the second column contains their English meaning whereas the third column contains lists of corresponding PLWORDNET LUs. Such organisation of information simplifies the work of lexicographers elaborating *Walenty*, decreases its sensitivity to changes in PLWORDNET and increases the readability of the dictionary, the more so as such lists can be really long. What is most important, we can modify these lists without bothering of revising all corresponding entries. This feature has a positive impact on the cohesion of the resource.

Complicated structure of PLWORDNET (caused by specifics of Polish language) made us also introduce PLWORDNET relations to another synset as a way of representing selectional preferences. For instance, an *Instrument* for PISAĆ 'write' could be *a pen, a ballpen, a pencil* etc. However, in PLWORDNET their direct hypernym is *artykuł papierniczy-1* 'writing materials' which is evidently too wide (as it includes, e.g. 'notebook'). They are correctly joined by the *holonymy (collection)* relation to *przybory do pisania-1* 'writing implements', as this term is used in Polish only in plural. This representation is equivalent to listing directly all relevant synsets, bu less sensitive to changes in PLWORDNET.

For some predicates, arguments considered separately represent a wide class of entities, but actually they are closely related to each other. For instance, one meaning of MLEĆ 'mill' concerns objects moving their parts through some substance. For example, windmill can mill air with its sails, while water wheel can mill water with its blades (but not with sails as it has none). Classic selectional preferences tell us nothing about what can be used by those objects for milling, but we can clearly see that they have to have to be internal parts of original object. Therefore, we introduced selectional preferences determined by means of relations to another argument. Meronymy seems to be a appropriate relation here, cf. Figure 5.

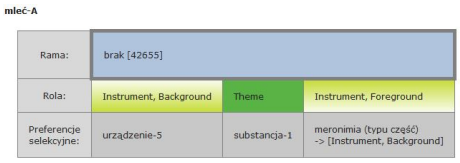


Figure 5: Selectional preferences based on relations between arguments for the verb MLEĆ

5 Connecting both layers

In *Walenty*, syntactic and semantic valency information are represented separately. Nevertheless,

they are closely connected, but this relation is a many-to-many one. On one hand, one semantic frame can be syntactically implemented by several schemata (diathesis alternation). On the other, one schema can be used in several frames. Relating a frame and a schema we directly link semantic arguments with corresponding syntactic positions. Let us consider the verb GNIEWAĆ SIĘ 'be angry' / GNIEWAĆ 'irritate'. The corresponding frame together with some schemata being its realisations are presented in Figure 6.

This is yet another difference between *Walenty* and VerbaLex. Two VerbaLex frames presented in Figure 1 differ only in the syntactic realisations of arguments. Nevertheless, the joint representation forces duplication of all information – syntactic and semantic. Moreover, lexical units involved in both syntactic realisations are connected with both frames, whereas in *Walenty* a lexical unit can label only one frame. For example, one *Walenty* frame in Figure 6 is connected to 9 verb schemata.

6 Common frames

Representation of verbs, nouns and adjectives does not differ on semantic level. What is important, derivationally connected entries of different PoSes are attached to the same frames. This is important for a correct interpretation of paraphrase. For historical reasons, this does not concern aspectual pairs.

It is worth noting that VerbNet and VerbaLex are focused solely on verbs, whereas FrameNet and PDT-VALLEX concern nouns and adjectives as well.

Let us consider the noun GNIEW 'anger' derivationally connected with the verb GNIEWAĆ SIĘ 'be angry', cf. Figure 7 (4 out of 15 schemata are visualised on the figure). Please note that the frame presented in Figures 6 and 7 is connected with the six PLWORDNET lexical units: *gniewny-1*, *gniewać-1*, *gniewać się-1*, *gniewać się-2*, *gniew-1* and *gniew-2*. This means that the frame is shared by three entries: GNIEWAĆ, GNIEW and GNIEWNY, and units representing the meaning of the current entry is written in bold.

7 Lexical units with multi-word lemmas

Walenty has a rich phraseological component (Przepiórkowski et al., 2014b). Hajnicz et al. (2016a) considers the simpler case when a lexicalised dependant does not change the meaning of a predicate and represents a fixed form of an argument (or a modifier). However, the more interesting case is when an idiomatic construction changes the meaning of the predicate, and its lexicalised dependant semantically is not an argument.

PLWORDNET contains lexical units having multi-word lemmas, and we decided to adapt this approach in *Walenty*. The semantic frame for the idiom *kraść całusa* 'steal a kiss' is presented in Figure 8. The fact that the frame is linked to an idiom is marked with a white rectangle with **Lemma** inside; a lexicalised dependant is marked white as well. Such phraseology appears for nonverbal entries as well[11]. We have chosen an idiom having both verbal and nominal realisation, which is not a typical case.

LUs identifying such idioms have multi-word lemmas composed of a lemma of the main predicate (here: the verb KRAŚĆ 'steel') and its syntactically dependant part (here: the noun CAŁUS 'kiss' in accusative) in a syntactically coherent way, see Figure 8. The structure of such a lemma could be more complicated, e.g. *płakać nad rozlanym mlekiem* 'cry over spilt milk', cf. 9. Similarly as in the general case, such lemma can be present in PLWORDNET or added in *Walenty*.

8 Conclusions and future works

This article describes the relations between two Polish language resources PLWORDNET and *Walenty* valency dictionary. The relations appear on two levels. First, PLWORDNET lexical units are connected to each semantic valency frame as their meaning identifiers. In particular, this concerns LUs with multi-word lemmas. Moreover, synsets (represented by LUs) are used to represent selection preferences of arguments.

Walenty is based on PLWORDNET version 2.1. Therefore, one of the main future tasks is to update the connection to the current version of PLWORDNET. This will be a very complicated task due to the fact that the changes in PLWORDNET are deep, which sometimes may cause a shift of the meaning of a particular LU. We plan to apply mappings between LUs from the source and the target PLWORDNET versions and estimate their reliability comparing their neighbourhood in the net. The special attention should be paid to to the LUs deleted from the PLWORDNET. On the other hand, we plan to automatically check, for all LUs added by Walenty developers, whether there exist relevant new PlWordNet units. The operation will be based on the synonymy/hypernymy relations. The whole procedure aims at maximal limitation of manual work.

In further future we want to connect semantically related frames of different entries in a hierarchical structure similar to hypernymy. This may involve unification of frames into a FrameNet-like hierarchy with inheritance. We are also interested in enriching the semantic layer with other semantic relations like presupposition or causation. The (morpho)syntactic level will not be influenced by these changes.

[11] However, most of nominal or adjectival idioms are fixed and do not open any valency positions. Such idioms are not considered in *Walenty*.

Table 1: List of predefined selectional preferences

ALL		
LUDZIE	PEOPLE	⟨osoba-1, grupa ludzi-1⟩
ISTOTY	CREATURES	⟨istota żywa-1, grupa istot-1⟩
PODMIOTY	FIRMS	⟨LUDZIE, podmiot-3, media-2⟩
WYTWÓR	ARTEFACT	⟨rzecz-4, wytwór-1, element-3, zbiór rzeczy-1⟩
JADŁO	FOOD	⟨pokarm-1, napój-1⟩
DOBRA	ESTATE	⟨JADŁO, mienie-1, przedmiot-1, wytwór-1, zbiór rzeczy-1⟩
KOMUNIKAT	COMMUNICATION	⟨informacja-1, wypowiedź-1⟩
KONCEPCJA	IDEAS	⟨informacja-1, wytwór umysłu-1, dzieło-2, dyscyplina-2, treść-1, zależność-3, model-1, rzecz-2, tematyka-1, struktura-2, wiedza-1, zwyczaj-1, prawo-3⟩
POŁOŻENIE	LOCATION	⟨miejsce-1, przestrzeń-1, obiekt-2⟩
MIEJSCE	PLACE	⟨lokal-1, budowla-1, rejon-1, obszar-1, państwo-1, jednostka administracyjna-1, woda-4⟩
OTOCZENIE	SURROUNDINGS	⟨powierzchnia-2, rzecz-4, wytwór-2, pomieszczenie-3, istota żywa-1⟩
CZAS	TIME	⟨chwila-1, czas-3, czas-8, godzina-3⟩
OBIEKTY	OBJECTS	⟨obiekt-2, element-3, zbiór-1⟩
CECHA	ATTRIBUTE	⟨cecha-1, zespół cech-1, atrybut-3⟩
CZYNNOŚĆ	ACT	⟨czynność-1, czyn-1⟩
SYTUACJA	SITUATION	⟨CZYNNOŚĆ, zdarzenie-2, stan-1, okoliczność-1, okoliczności-1, ciąg zdarzeń-1, działalność-1⟩
KIEDY	WHEN	⟨CZAS, SYTUACJA⟩
CZEMU	WHY	⟨CECHA, SYTUACJA, LUDZIE⟩
ILOŚĆ	AMOUNT	⟨ilość-1, rozmiar-1, rozmiar-2, jednostka-4, wielkość-6⟩

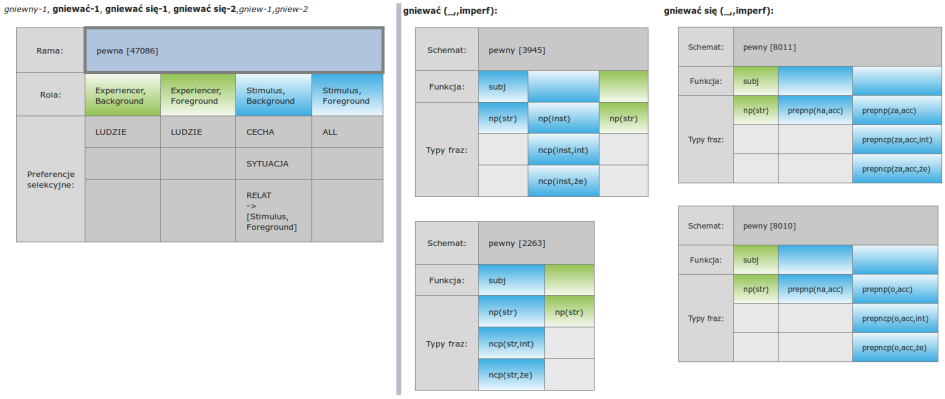


Figure 6: A screenshot with a semantic frame and schemata being its syntactic realisation

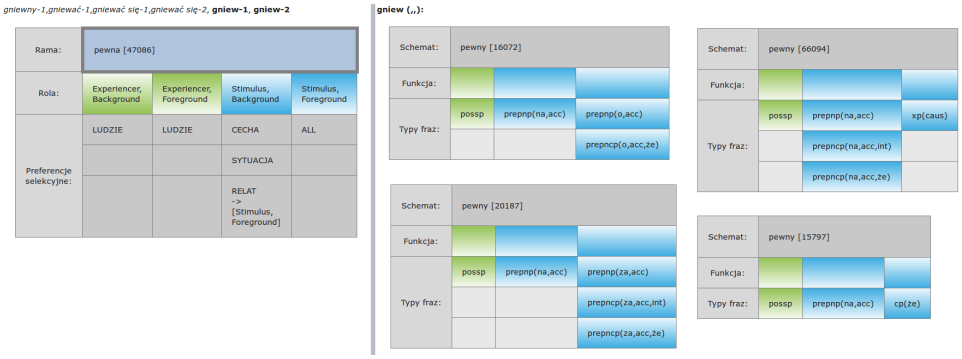


Figure 7: A screenshot with a semantic frame and schemata being its syntactic realisation form the noun perspective

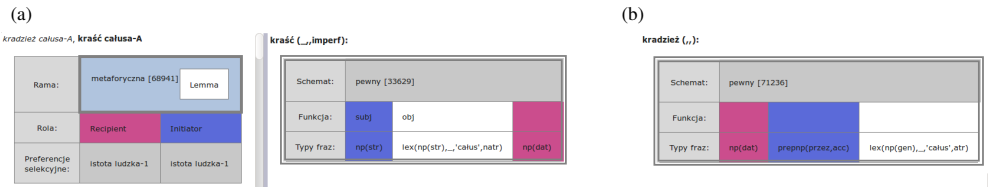


Figure 8: A frame representing idiom *kraść całusa*(a) from the verb perspective (b) schema of the noun

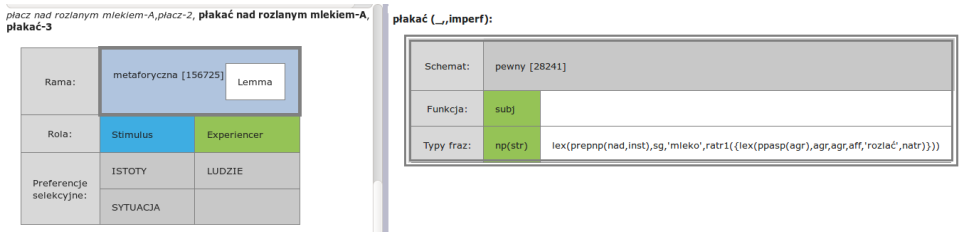


Figure 9: A frame representing idiom *płakać nad rozlanym mlekiem*

Acknowledgements Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education

References

Mirosław Bańko, editor. 2000. *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, Poland.

Diego De Cao, Danilo Croce, and Roberto Basili. 2010. Extensive evaluation of a framenet-wordnet mapping resource. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pages 2752–2757, Valetta, Malta. ELRA.

Noam Chomsky. 1970. Remarks on nominalization. In Roderic A. Jacobs and Peter S. Rosenbaum, editors, *Readings in English transformational grammar*, pages 184–221. Ginn and Company, Waltham, MA.

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational LinguisticsCOLING-ACL'98*, pages 293–299, Montreal, Canada.

Christiane Fellbaum, editor. 1998. *WordNet — An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Charles J. Fillmore and Colin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of the WordNet and Other Lexical Resources Workshop*, Pittsburgh. NAACL.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32. John Wiley.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. *Mathematical Modelling in Physics, Engineering and Cognitive Science*, 9:57–68.

Elżbieta Hajnicz, Anna Andrzejczuk, and Tomasz Bartosiak. 2016a. Semantic layer of the valence dictionary of Polish *Walenty*. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-2016)*, pages 2625–2632, Portorož, Slovenia. ELRA.

Elżbieta Hajnicz, Agnieszka Patejuk, Adam Przepiórkowski, and Marcin Woliński. 2016b. *Walenty*: słownik walencyjny języka polskiego z bogatym komponentem frazeologicznym. In Karolína Skwarska and Elżbieta Kaczmarska, editors, *Výzkum slovesné valence ve slovanských zemích*, pages 71–102. Slovanský ústav Akademie věd ČR, Prague, Czech Republic.

Dana Hlaváčková and Aleš Horák. 2006. VerbaLex — new comprehensive lexicon of verb valences for Czech. In *Proceedings of the Third International Seminar on Computer Treatment of Slavic and East European Languages*, pages 107–115, Bratislava, Slovakia.

J. J. Katz and J. A. Fodor. 1964. The structure of a semantic theory. In J. A. Fodor and J. J. Katz, editors, *The Structure of Language*, pages 479–518. Prentice Hall.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 691–696, Austin, TX. AAAI Press.

Karin Kipper-Schuler. 2005. *VerbNet: A broad coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Department, University of Pennsylvania.

Beth Levin. 1993. *English verb classes and alternation: a preliminary investigation*. University of Chicago Press, Chicago, IL.

Markéta Lopatková, Zdeňek Žabokrtský, Karolina Skwarska, and Václava Benešová. 2003. VALLEX 1.0 valency lexicon of Czech verbs. Technical Report TR-2003-18, ÚFAL/CKL MFF UK, Prague, Czech Republic.

George A. Miller and Christiane Fellbaum. 2007. WordNet then and now. *Language Resources and Evaluation*, 41:209–214.

Karel Pala and Pavel Smrž. 2004. Building the Czech WordNet. *Romanian Journal of Information Science and Technology*, 7(2–3):79–88.

Agnieszka Patejuk and Adam Przepiórkowski. 2012. Towards an LFG parser for Polish. an exercise in parasitic grammar development. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3849–3852, Istanbul, Turkey. ELRA.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, Poland.

Maciej Piasecki, Stan Szpakowicz, Marek Maziarz, and Ewa Rudnicka. 2016. PlWordNet 3.0 – Almost There. In *Proceedings of the 8th International WordNet Conference (GWC 2016)*, pages 290–299, Bucharest, Romania. Global Wordnet Association.

Kazimierz Polański, editor. 1980–1992. *Słownik syntaktyczno-generatywny czasowników polskich*, volume I–V. Zakład Narodowy imienia Ossolińskich, Wrocław · Warszawa · Kraków · Gdańsk, Poland.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Filip Skwarski, Marcin Woliński, and Marek Świdziński. 2014a. Walenty: Towards a comprehensive valence dictionary of Polish. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2785–2792, Reykjavík, Iceland. ELRA.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. 2014b. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland.

Adam Przepiórkowski, Filip Skwarski, Elżbieta Hajnicz, Agnieszka Patejuk, Marek Świdziński, and Marcin Woliński. 2014c. Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, XXXIII:159–178.

Philip Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, December.

Adam Rmbousek, Karel Pala, and Sandra Tukačová. 2017. Overview and future of Czech WordNnet. In *LDK Workshops: OntoLex, TIAD and Challenges for Wordnets*, pages 146–151, Galway, Ireland.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, Holland.

Maria Szupryczyńska. 1996. Problem pozycji składniowej. In Krystyna Kallas, editor, *Polonistyka Toruńska Uniwersytetowi w 50. Rocznicę Utworzenia UMK*, Językoznawstwo, pages 135–144. Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń, Poland.

Marek Świdziński. 1994. *Syntactic Dictionary of Polish Verbs*. Uniwersytet Warszawski / Universiteit van Amsterdam.

Zdeňka Urešová. 2009. Building the PDT-Vallex valency lexicon. In *Proceedings of the 5th Corpus Linguistics Conference*. University of Liverpool.

Zygmunt Vetulani and Bartłomiej Kochanowski. 2014. "PolNet – polish wordnet" project: PolNet 2.0 – a short description of the release. In *Proceedings of the 7th International WordNet Conference (GWC 2014)*, pages 400–404, Tartu, Estonia. University of Tartu.

Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Jacek Marciniak, Paweł Konieczka, and Przemysław Rzepecki. 2009. An algorithm for building lexical semantic network and its application to PolNet — Polish WordNet project. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human Language Technology. Chalenges of the Information Society. 3rd Language & Technology Conference*, volume 5603 of *LNAI*, pages 369–381. Springer-Verlag. Revised Selected Papers.

Zygmunt Vetulani. 2014. PolNet – Polish WordNet. In Zygmunt Vetulani and Joseph Mariani, editors, *Human Language Technology Chalenges for Computer Science and Linguistics. LTC 2011*, volume 8387 of *LNAI*, pages 408–416. Springer-Verlag. Revised Selected Papers.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic network*. Kluwer Academic Publishers, Dordrecht, Holland.

Marcin Woliński. 2004. *Komputerowa weryfikacja gramatyki Świdzińskiego*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Zdeňek Žabokrtský and Markéta Lopatková. 2007. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, 87:41–60.

Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation

Loïc Vial **Benjamin Lecouteux** **Didier Schwab**
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{loic.vial, benjamin.lecouteux, didier.schwab}@univ-grenoble-alpes.fr

Abstract

In this article, we tackle the issue of the limited quantity of manually sense annotated corpora for the task of word sense disambiguation, by exploiting the semantic relationships between senses such as synonymy, hypernymy and hyponymy, in order to compress the sense vocabulary of Princeton WordNet, and thus reduce the number of different sense tags that must be observed to disambiguate all words of the lexical database. We propose two different methods that greatly reduce the size of neural WSD models, with the benefit of improving their coverage without additional training data, and without impacting their precision. In addition to our methods, we present a WSD system which relies on pre-trained BERT word vectors in order to achieve results that significantly outperforms the state of the art on all WSD evaluation tasks.

1 Introduction

Word Sense Disambiguation (WSD) is a task which aims to clarify a text by assigning to each of its words the most suitable sense labels, given a predefined sense inventory.

Various approaches have been proposed to achieve WSD: Knowledge-based methods rely on dictionaries, lexical databases, thesauri or knowledge graphs as primary resources, and use algorithms such as lexical similarity measures (Lesk, 1986) or graph-based measures (Moro et al., 2014). Supervised methods, on the other hand, exploit sense annotated corpora as training instances for a classifier such as SVM (Chan et al., 2007; Zhong and Ng, 2010), or more recently by a neural network (Kågebäck and Salomonsson, 2016). Finally, unsupervised methods automatically identify the different senses of words from unannotated or parallel corpora (e.g. Ide et al. (2002)).

Supervised methods are by far the most predominant as they generally offer the best results in evaluation campaigns (for instance (Navigli et al., 2007)). State of the art classifiers used to combine specific features such as the parts of speech and the lemmas of surrounding words (Zhong and Ng, 2010), but they are now replaced by neural networks which learn their own representation of words (Raganato et al., 2017b; Le et al., 2018).

One major bottleneck of supervised systems is the restricted quantity of manually sense annotated corpora: In the annotated corpus SemCor (Miller et al., 1993), the largest manually sense annotated corpus available, words are annotated with 33 760 different sense keys, which corresponds to only approximately 16% of the sense inventory of WordNet (Miller, 1995), the lexical database of reference widely used in WSD. Many works try to leverage this problem by creating new sense annotated corpora, either automatically (Pasini and Navigli, 2017), semi-automatically (Taghipour and Ng, 2015), or through crowdsourcing (Yuan et al., 2016).

In this work, the idea is to solve this issue by taking advantage of the semantic relationships between senses included in WordNet, such as the hypernymy, the hyponymy, the meronymy, the antonymy, etc. Our method is based on the observation that a sense and its closest related senses (its hypernym or its hyponyms for instance) all share a common idea or concept, and so a word can sometimes be disambiguated using only related concepts. Consequently, we do not need to know every sense of WordNet to disambiguate all words of WordNet.

For instance, let us consider the word “mouse” and two of its senses which are the *computer* mouse and the *animal* mouse. We only need to know the notions of “animal” and “electronic de-

vice" to distinguish them, and all notions that are more specialized such as "rodent" or "mammal" are therefore superfluous. By grouping them, we can benefit from all other instances of electronic devices or animals in a training corpus, even if they do not mention the word "mouse".

Contributions: In this paper, we hypothesize that only a subset of WordNet senses could be considered to disambiguate all words of the lexical database. Therefore, we propose two different methods for building this subset and we call them sense vocabulary compression methods. By using these techniques, we are able to greatly improve the coverage of supervised WSD systems, nearly eliminating the need for a backoff strategy that is currently used in most systems when dealing with a word which has never been observed in the training data. We evaluate our method on a state of the art WSD neural network, based on pretrained contextualized word vector representations, and we present results that significantly outperform the state of the art on every standard WSD evaluation task. Finally, we provide a documented tool for training and evaluating neural WSD models, as well as our best pretrained model in a dedicated GitHub repository[1].

2 Related Work

In WSD, several recent advances have been made in the creation of new neural architectures for supervised models and the integration of knowledge into these systems. Multiple works also exploit the idea of grouping together related senses. In this section, we give an overview of these works.

2.1 WSD Based on a Language Model

In this type of approach, that has been initiated by Yuan et al. (2016) and reimplemented by Le et al. (2018), the central component is a neural language model able to predict a word with consideration for the words surrounding it, thanks to a recurrent neural network trained on a massive quantity of unannotated data.

Once the language model is trained, it is used to produce sense vectors that result from averaging the word vectors predicted by the language model at all positions of words annotated with the given sense.

At test time, the language model is used to predict a vector according to the surrounding context, and the sense closest to the predicted vector is assigned to each word.

These systems have the advantage of bypassing the problem of the lack of sense annotated data by concentrating the power of abstraction offered by recurrent neural networks on a good quality language model trained in an unsupervised manner. However, sense annotated corpora are still indispensable to contruct the sense vectors.

2.2 WSD Based on a Softmax Classifier

In these systems, the main neural network directly classifies and attributes a sense to each input word through a probability distribution computed by a softmax function. Sense annotations are simply seen as tags put on every word, like a POS-tagging task for instance.

We can distinguish two separate branches of these types of neural networks:

1. Those in which we have several distinct and token-specific neural networks (or classifiers) for every different word in the dictionary (Iacobacci et al., 2016; Kågebäck and Salomonsson, 2016), each of them being able to manage a particular word and its particular senses. For instance, one of the classifiers is specialized in choosing between the four possible senses of the noun "mouse". This type of approach is particularly fitted for the lexical sample tasks, where a small and finite set of very ambiguous words have to be sense annotated in several contexts, but it can also be used in all-words word sense disambiguation tasks.
2. Those in which we have a larger and general neural network that is able to manage all different words and assign a sense in the set of all existing sense in the dictionary used (Raganato et al., 2017b).

The advantage of the first branch of approaches is that in order to disambiguate a word, limiting our choice to one of its possible senses is computationally much easier than searching through all the senses of all words. To put things in perspective, the average number of senses of polysemous words in WordNet is approximately 3, whereas the total number of senses considering all words is 206 941.

The second approach, however, has an interesting property: all senses reside in the same vector space and hence share features in the hidden layers of the network. This allows the model to predict

[1] https://github.com/getalp/disambiguate

an identical sense for two different words (i.e. synonyms), but it also offers the possibility to predict a sense for a word not present in the dictionary (e.g. neologism, spelling mistake...).

Finally, in two recent articles, Luo et al. (2018a) and Luo et al. (2018b) have proposed an improvement of these type of architectures, by computing an attention between the context of a target word and the gloss of its different senses. Thus, their work is one of the first to incorporate knowledge from WordNet into a WSD neural network.

2.3 Sense Clustering Methods

Several works exploit the idea of grouping together mutiple WordNet sense tags in order to create a coarser sense inventory which can potentially be more useful in some NLP tasks.

In the works of Ciaramita and Altun (2006), the authors propose a supervised system that learns and predicts "Supersense" tags, which belong to the set of the broad semantic categories of senses, organizing the sense inventory of WordNet. This tagset consists, in their work, of 26 categories for nouns (such as "food", "person" or "object"), and 15 categories for verbs (such as "emotion" or "weather"). By predicting supersense tags instead of the usual fine-grained sense tags of WordNet, the output vocabulary of their system is shrinked to only 41 different classes, and this leads to a small and easy-to-train model able to perform partial WSD, which could be useful and sufficient for other NLP tasks where the fine-grained distinction is not necessary.

In Izquierdo et al. (2007), the authors propose several methods for creating "Basic Level Concepts" (BLC), groups of related senses with a generally smaller size than supersenses, and which can be controlled by a threshold variable. Their methods rely on the semantic relationships between senses of WordNet, and, in the same way as Ciaramita and Altun (2006), they evaluated their clusters on a modified WSD task, where supersenses or BLC have to be predicted instead of the original sense tags from WordNet.

The main difference between our work and these works is that our end goal is to improve fine-grained WSD systems. Even though our methods generate clusters of related senses, we guarantee that two different senses of a lemma reside in two different clusters, so at the end, even if our supervised system produces a cluster tag for a target word, we are still able to find back the true sense tag, by simply keeping track of which sense key of its lemma belongs to the predicted group.

3 Sense Vocabulary Compression

Current state of the art supervised WSD systems such as Yuan et al. (2016), Raganato et al. (2017b), Luo et al. (2018a) and Le et al. (2018) are all confronted to the following issues:

1. Due to the small number of manually sense annotated corpora available, a target word may never be observed during the training, and therefore the system is not able to annotate it.
2. For the same reason, a word may have been observed, but not all of its senses. In this case the system is able to annotate the word, but if the expected sense has never been observed, the output will be wrong, regardless of the architecture of the supervised system.
3. Training a neural network to predict a tag which belongs to the set of all WordNet senses can become extremely slow and requires a lot of parameters with a large output vocabulary. And this vocabulary goes up to 206 941 if we consider all word-senses of WordNet.

In order to overcome all these issues, we propose a method for grouping together multiple sense tags that refer in fact to the same concept. In consequence, the output vocabulary decreases, the ability of the trained system to generalize improves, as well as its coverage.

3.1 From Senses to Synsets: A Vocabulary Compression Based on Synonymy

In the lexical database WordNet, senses are organized in sets of synonyms called synsets. A synset is technically a group of one or more word-senses that have the same definition and consequently the same meaning. For instance, the first senses of "eye", "optic" and "oculus" all refer to a common synset which definition is "the organ of sight".

Illustrated in Figure 1, the word-sense to synset mapping is hence a way of compressing the output vocabulary, and it is already applied in many works (Yuan et al., 2016; Le et al., 2018), while not being always explicitly stated. This method clearly helps to improve the coverage of supervised systems however. Indeed, if the verb "help" is observed in the annotated data in its first sense, the context surrounding the target word can be used to later annotate the verb "assist" or "aid" with the same valid synset tag.

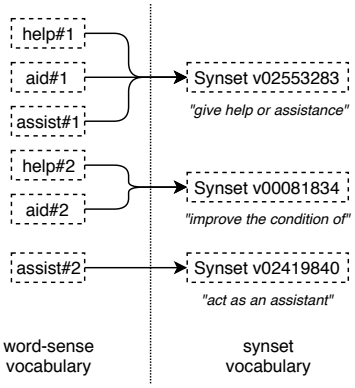


Figure 1: Word-sense to synset mapping (compression through synonymy) applied on the first two senses of the words "help", "aid" and "assist".

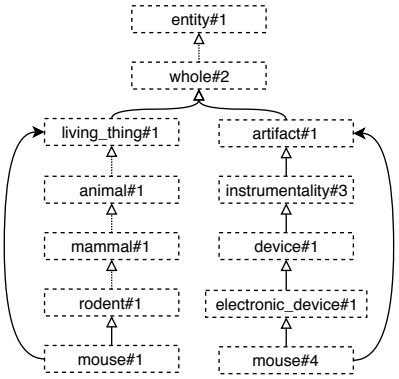


Figure 2: Sense vocabulary compression trough hypernymy hierarchy applied on the first and fourth sense of the word "mouse". Dashed arrows mean that some nodes are skipped for clarity.

Going further, other information from WordNet can help the system to generalize. Our first new method takes advantage of the hypernymy and hyponymy relationships to achieve the same idea.

3.2 Compression through Hypernymy and Hyponymy Relationships

According to Polguère (2003), hypernymy and hyponymy are two semantic relationships which correspond to a particular case of sense inclusion: the hyponym of a term is a specialization of this term, whereas its hypernym is a generalization. For instance, a "mouse" is a type of "rodent" which is in turn a type of "animal".

In WordNet, these relationships bind nearly every noun together in a tree structure[2] that goes from the generic root, the node "entity" to the most specific leaves, for instance the node "white-footed mouse". These relationships are also present on several verbs: for instance "add" is a way of "compute" which is a way of "reason".

For the sake of WSD, just like grouping together the senses of the same synset helps to better generalize, we hypothesize that grouping together the synsets of the same hypernymy relationship also helps in the same way. The general idea of our method is that the most specialized concepts in WordNet are often superfluous for WSD.

Indeed, considering a small subset of WordNet that only consists of the word "mouse", its first sense (the small rodent), its fourth sense (the electronic device), and all of their hypernyms. This is illustrated in Figure 2. We can see that every concept that is more specialized than the concepts "artifact" and "living_thing" could be removed. We could map every tag of "mouse#1" to the tag of "living_thing#1" and we could still be able to disambiguate this word, but with a benefit: all other "living things" and animals in the sense annotated data could be tagged with the same sense. They would give examples of what is an animal and then show how to differentiate the small rodent from the hand-operated electronic device.

Therefore, the goal of our method is to map every sense of WordNet to its highest ancestor in the hypernymy hierarchy, but with the following constraints: First, this ancestor must discriminate all the different senses of the target word. Second, we need to preserve the hypernyms that are indispensable to discriminate the senses of the other words in the dictionary. For instance, we cannot map "mouse#1" to "living_thing#1", because the more specific tag "animal#1" is essential to distinguish the two senses of the word "prey" (one sense describes a person, the other describes an animal).

Our method thus works in two steps:

1. We mark as "necessary" the children of the first common ancestor of every pair of senses of every word of WordNet.
2. We map every sense to its first ancestor in the hypernymy hierarchy that has been previously marked as "necessary".

[2] We computed that 41 607 on the 44 449 polysemous nouns of WordNet (94%) are part of this hierarchy.

As a result, the most specific synsets of the tree that are not indispensable for discriminating any word of the lexical inventory are automatically removed from the vocabulary. In other words, the set of synsets that is left in the vocabulary is the smallest subset of all synsets that are necessary to distinguish every sense of every word of WordNet, following the hypernym and hyponym links.

3.3 Compression through all semantic relationships

In addition to hypernymy and hyponymy, WordNet contains several other relationships between synsets, such as the instance relationship (e.g. "Albert Einstein" is an instance of "physicist'), the meronymy (X is part of Y, or X is a member of Y) and its counterpart the holonymy, the antonymy (X is the opposite of Y), etc.

We hence propose a second method for sense vocabulary compression, that considers all the semantic relationships offered by WordNet, in order to form clusters of related synsets.

For instance, using all semantic relationships, we could form a cluster containing "physicist", "physics" (domain category), "Albert Einstein" (instance of), "astronomer" (hyponym), but also further related senses such as "photon", because it is a meronym of "radiation", which is a hyponym of "energy", which belongs to the same domain category of "physics".

Our method works by constructing these clusters iteratively: First, we initialize the set of clusters C with one synset in each cluster.

$$C = \{c_0, c_1, ..., c_n\} \qquad S = \{s_0, s_1, ..., s_n\}$$
$$C = \{\{s_0\}, \{s_1\}, ..., \{s_n\}\}$$

Then at each step, we sort C by sizes of clusters, and we peek the smallest one c_x and the smallest related cluster to c_x, c_y. We define a cluster being related to another if they contain at least one synset that have a semantic link together. We merge c_x and c_y together, and we verify that the operation still allows to discriminate the different senses of all words in the lexical database. If it is not the case, we cancel the merge and we try another semantic link. If no link is possible, we try to create one with the next smallest cluster, and if no further link can be created, the algorithm stops.

In Figure 3, we show a possible set of clusters that could result from our method, focusing on two senses of the word "Weber" and only on a few relationships.

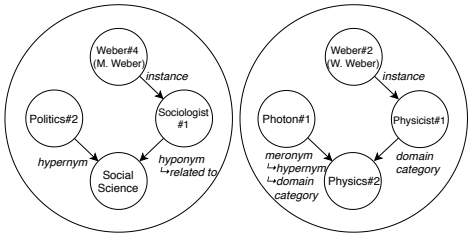


Figure 3: Example of clusters of sense that could result from our method, if we limit our view to two senses of the word "Weber" and only some relationship links.

This method produces clusters significantly larger than the method based on hypernyms. On average, a cluster has 5 senses with the hypernym method, whereas it has 17 senses with this method. This method, unlike the previous one, is also stochastic, because the formation of clusters depends on the underlying order of iteration when multiple clusters are the same size. However, because we always sort clusters by size before creating a link, we observed that the final vocabulary size (i.e. number of clusters) is always between 11 000 and 13 000. In the following, we consider a resulting mapping where the algorithm stopped after 105 774 steps.

Method	Vocabulary size	Compression rate	SemCor Coverage
No compression	206 941	0%	16%
Synonyms	117 659	43%	22%
Hypernyms	39 147	81%	32%
All relations	11 885	94%	39%

Table 1: Effects of the sense vocabulary compression on the vocabulary size and on the coverage of the SemCor.

In Table 1, we show the effect of the common compression through synonyms, our first proposed compression through hypernyms, and our second method of compression through all semantic relationships, on the size of the vocabulary of WordNet sense tags, and on the coverage of the SemCor corpus. As we can see, the sense vocabulary size is drastically decreased, and the coverage of the same corpus really improved.

4 Experiments

In order to evaluate our sense vocabulary compression methods, we applied them on a neural WSD system based on a softmax classifier capable of classifying a word in all possible synsets of WordNet (see subsection 2.2).

We implemented a system similar to Raganato et al. (2017b)'s BiLSTM but with some key differences. In particular, we used BERT contextualized word vectors (Devlin et al., 2018) in input of our network, Transformer encoder layers (Vaswani et al., 2017) instead of LSTM layers as hidden units, our output vocabulary only consists of sense tags seen during training (mapped according to the compression method used), and we ignore the network's predictions on words that are not annotated.

4.1 Implementation details

For BERT, we used the model named "bert-large-cased" of the PyTorch implementation[3], which consists of vectors of dimension 1024, trained on BooksCorpus and English Wikipedia.

Due to the fact that BERT's internal tokenizer sometimes split words in multiples tokens (i.e. ["rodent"] becomes ["rode", "##nt"]), we trained our system to predict a sense tag on the first token only of a splitted annotated word.

For the Transformer encoder layers, we used the same parameters as the "base" model of Vaswani et al. (2017), that is 6 layers with 8 attention heads, a hidden size of 2048, and a dropout of 0.1.

Finally, because BERT already encodes the position of the words inside their vectors, we did not add any positional encoding.

4.2 Training

We compared our sense vocabulary compression methods on two training sets: The SemCor, and the concatenation of the SemCor and the Princeton WordNet Gloss Corpus (WNGC). The latter is a corpus distributed as part of WordNet since its version 3.0, and it consists of the definitions (glosses) of every synset of WordNet, with words manually or semi-automatically sense annotated. We used the version of these corpora given as part of the UFSAC 2.1 resource[4] (Vial et al., 2018).

[3] https://github.com/huggingface/pytorch-pretrained-BERT

[4] https://github.com/getalp/UFSAC

We performed every training for 20 epochs. At the beginning of each epoch, we shuffled the training set. We evaluated our model at the end of every epoch on a development set, and we kept only the one which obtained the best F1 WSD score. The development set was composed of 4 000 random sentences taken from the Princeton WordNet Gloss Corpus for the models trained on the SemCor, and 4 000 random sentences extracted from the whole training set for the other models.
For each training set, we trained three systems:

1. A "baseline" system that predicts a tag belonging to all the synset tags seen during training, thus using the common vocabulary compression through synonyms method.
2. A "hypernyms" system which applies our vocabulary compression through hypernyms algorithm on the training corpus.
3. A "all relations" system which applies our second vocabulary compression through all relations on the training corpus.

We trained with mini-batches of 100 sentences, truncated to 80 words, and we used Adam (Kingma and Ba, 2015) with a learning rate of 0.0001 as the optimization method.

System	SemCor	SemCor+WNGC
baseline	77.15M	120.85M
hypernyms	63.44M	79.85M
all relations	55.16M	60.27M

Table 2: Number of parameters of neural models.

All models have been trained on one Nvidia's Titan X GPU. The number of parameters of individual models are displayed in Table 2. As we can see, our compression methods drastically reduce the number of parameters, by a factor of 1.2 to 2.

4.3 Evaluation

We evaluated our models on all evaluation corpora commonly used in WSD, that is the English all-words WSD tasks of the evaluation campaigns SensEval/SemEval. We used the fine-grained evaluation corpora from the evaluation framework of Raganato et al. (2017a), which consists of SensEval 2 (Edmonds and Cotton, 2001), SensEval 3 (Snyder and Palmer, 2004), SemEval 2007 task 17 (Pradhan et al., 2007), SemEval 2013 task 12 (Navigli et al., 2013) and SemEval 2015 task 13 (Moro and Navigli, 2015), as well as the "ALL" corpus consisting of the concatenation of all pre-

System	SE2	SE3	SE07 17	SE13	SE15	ALL (concat. of previous tasks) nouns	verbs	adj.	adv.	total	SE07 07
First sense baseline	65.6	66.0	54.5	63.8	67.1	67.7	49.8	73.1	80.5	65.5	78.9
HCAN (Luo et al., 2018a)	72.8	70.3	-	68.5	72.8	72.7	58.2	77.4	84.1	71.1	-
LSTMLP (Yuan et al., 2016)	73.8	71.8	63.5	69.5	72.6	†73.9	-	-	-	†71.5	83.6
SemCor, baseline	77.2	76.5	70.1	74.7	77.4	78.7	65.2	79.1	85.5	76.0	87.7
SemCor, hypernyms	77.5	77.4	69.5	76.0	78.3	79.6	65.9	79.5	85.5	76.7	87.6
SemCor, all relations	76.6	76.9	69.0	73.8	75.4	77.2	66.0	80.1	85.0	75.4	86.7
SemCor+WNGC, baseline	**79.7**	76.1	**74.1**	78.6	80.4	80.6	68.1	82.4	86.1	78.3	90.4
SemCor+WNGC, hypernyms	**79.7**	77.8	73.4	**78.7**	**82.6**	81.4	68.7	83.7	85.5	**79.0**	90.4
SemCor+WNGC, all relations	79.4	**78.1**	71.4	77.8	81.4	80.7	68.6	82.8	85.5	78.5	**90.6**

Table 3: F1 scores (%) on the English WSD tasks of the evaluation campaigns SensEval/SemEval. The task "ALL" is the concatenation of SE2, SE3, SE07 17, SE13 and SE15. The first sense is assigned on words for which none of its sense has been observed during the training. Results in **bold** are to our knowledge the best results obtained on the task. Scores prefixed by a dagger (†) are not provided by the authors but are deduced from their other scores.

vious ones. We also compared our result on the coarse-grained task 7 of SemEval 2007 (Navigli et al., 2007) which is not present in this framework.

For each evaluation, we trained 8 independent models, and we give the score obtained by an ensemble system that averages their predictions through a geometric mean.

System	No Backoff	Backoff on Monosemics
SemCor, baseline	93.23%	98.13%
SemCor, hypernyms	98.75%	99.68%
SemCor, all relations	99.67%	99.99%
SemCor+WNGC, baseline	98.26%	99.41%
SemCor+WNGC, hypernyms	99.83%	99.96%
SemCor+WNGC, all relations	99.99%	100%

Table 4: Coverage of our systems on the task "ALL". "Backoff on Monosemics" means that monosemic words are considered annotated.

In the results in Table 3, we first observe that our systems that use the sense vocabulary compression through hypernyms or through all relations obtain scores that are overall equivalent to the systems that do not use it.

Our methods greatly improves their coverage on the evaluation tasks however. As we can see in Table 4, on the total of 7 253 words to annotate for the corpus "ALL", the baseline system trained on the SemCor is not able to annotate 491 of them, while the vocabulary compression through hypernyms reduces this number to 91 and 24 for the compression through all relations.

When adding the Princeton WordNet Gloss Corpus to the training set, only one word (the monosemic adjective "cytotoxic") cannot be annotated with the system that uses the compression through all relations because its sense has not been observed during training.

If we exclude the monosemic words, the system based on our compression method through all relations miss only one word (the adverb "eloquently") when trained on the SemCor, and has a coverage to 100% when the WNGC is addded.

In comparison to the other works, thanks to the Princeton WordNet Gloss Corpus added to the training data and the use of BERT as input embeddings, we outperform systematically the state of the art on every task.

4.4 Ablation Study

In order to give a better understanding of the origin of our scores, we provide a study of the impact of our main parameters on the results. In addition to the training corpus and the vocabulary compression method, we chose two parameters that differentiate us from the state of the art: the pre-trained word embeddings model and the ensembling method, and we have made them vary.

For the word embeddings model, we experimented with BERT (Devlin et al., 2018) as in our main results, with ELMo (Peters et al., 2018), and with GloVe (Pennington et al., 2014), the same pre-trained word embeddings used by Luo et al. (2018a). For ELMo, we used the model trained on

Training Corpus	Input Embeddings	Ensemble	F1 Score on task "ALL" (%)					
			Baseline		Hypernyms		All relations	
			$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
SemCor+WNGC	BERT	Yes	78.27	-	79.00	-	78.48	-
SemCor+WNGC	BERT	No	76.97	±0.38	77.08	±0.17	76.52	±0.36
SemCor+WNGC	ELMo	Yes	75.16	-	74.65	-	70.58	-
SemCor+WNGC	ELMo	No	74.56	±0.27	74.36	±0.27	68.77	±0.30
SemCor+WNGC	GloVe	Yes	72.23	-	72.74	-	71.42	-
SemCor+WNGC	GloVe	No	71.93	±0.35	71.79	±0.29	69.60	±0.32
SemCor	BERT	Yes	76.02	-	76.73	-	75.40	-
SemCor	BERT	No	75.06	±0.26	75.59	±0.16	73.91	±0.33
SemCor	ELMo	Yes	72.55	-	73.09	-	69.43	-
SemCor	ELMo	No	72.21	±0.13	72.83	±0.24	68.74	±0.29
SemCor	GloVe	Yes	70.77	-	71.18	-	68.44	-
SemCor	GloVe	No	70.51	±0.16	70.77	±0.21	67.48	±0.55
HCAN (Luo et al., 2018a) (fully reproducible state of the art)								
SemCor+WordNet glosses	GloVe	No	71.1					
LSTMLP (Yuan et al., 2016) (state of the art scores but use private data)								
SemCor+1K (private)	private	No	71.5					

Table 5: Ablation study on the task "ALL" (i.e. the concatenation of all SensEval/SemEval tasks). For systems that do not use ensemble, we display the mean score ($\bar{x}$) of eight individually trained models along with its standard deviation (σ).

Wikipedia and the monolingual news crawl data from WMT 2008-2012.[5] For GloVe, we used the model trained on Wikipedia 2014 and Gigaword 5.[6] Due to the fact that GloVe embeddings do not encode the position of the words (a word has the same vector representation in any context), we used bidirectional LSTM cells of size 1 000 for each direction, instead of Transformer encoders for this set of experiments. In addition, because the vocabulary of GloVe is finite and all words are lowercased, we lowercased the inputs, and we assigned a vector filled with zeros to out-of-vocabulary words.

For the ensembling method, we either perform ensembling as in our main results, by averaging the prediction of 8 models trained separately or we give the mean and the standard deviation of the scores of the 8 models evaluated separately.

As we can see in Table 5, the additional training corpus (WNGC) and even more the use of BERT as input embeddings both have a major impact on our results and lead to scores above the state of the art. Using BERT instead of ELMo or GloVe improves respectively the score by approximately 3 and 5 points in every experiment, and adding the WNGC to the training data improves it by approximately 2 points. Finally, using ensembles adds roughly another 1 point to the final F1 score.

[5] https://allennlp.org/elmo
[6] https://nlp.stanford.edu/projects/glove/

Finally, through the scores obtained by invidual models (without ensemble), we can observe on the standard deviations that the vocabulary compression method through hypernyms never impact significantly the final score. However, the compression method through all relations seems to negatively impact the results in some cases (when using ELMo or GloVe especially).

5 Conclusion

In this paper, we presented two new methods that improve the coverage and the capacity of generalization of supervised WSD systems, by narrowing down the number of different sense in WordNet in order to keep only the senses that are essential for differentiating the meaning of all words of the lexical database. On the scale of the whole lexical database, we showed that these methods can shrink the total number of different sense tags in WordNet to only 6% of the original size, and that the coverage of an identical training corpus has more than doubled. We implemented a state of the art WSD neural network and we showed that these methods compress the size of the underlying models by a factor of 1.2 to 2, and greatly improve their coverage on the evaluation tasks. As a result, we reach a coverage of 99.99% of the evaluation tasks (1 word missing on 7 253) when training a system on the SemCor only, and 100% when adding the WNGC to the training data, on the pol-

ysemic words. Therefore, the need for a backoff strategy is nearly eliminated. Finally, our method combined with the recent advances in contextualized word embeddings and with a training corpus composed of sense annotated glosses, our system achieves scores that considerably outperform the state of the art on all WSD evaluation tasks.

References

[Chan et al.2007] Yee Seng Chan, Hwee Tou Ng, and Zhi Zhong. 2007. Nus-pt: Exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 253–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Ciaramita and Altun2006] Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 594–602, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

[Edmonds and Cotton2001] Philip Edmonds and Scott Cotton. 2001. Senseval-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SENSEVAL '01, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Iacobacci et al.2016] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany, August. Association for Computational Linguistics.

[Ide et al.2002] Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 61–66. Association for Computational Linguistics, July.

[Izquierdo et al.2007] Rubén Izquierdo, Armando Suárez, and German Rigau. 2007. Exploring the automatic selection of basic level concepts. In *Proceedings of RANLP*, volume 7. Citeseer.

[Kågebäck and Salomonsson2016] Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional lstm. In *5th Workshop on Cognitive Aspects of the Lexicon (CogALex)*. Association for Computational Linguistics.

[Kingma and Ba2015] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

[Le et al.2018] Minh Le, Marten Postma, Jacopo Urbani, and Piek Vossen. 2018. A deep dive into word sense disambiguation with lstm. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 354–365. Association for Computational Linguistics.

[Lesk1986] Michael Lesk. 1986. Automatic sense disambiguation using mrd: how to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC '86*, pages 24–26, New York, NY, USA. ACM.

[Luo et al.2018a] Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018a. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411. Association for Computational Linguistics.

[Luo et al.2018b] Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018b. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2473–2482. Association for Computational Linguistics.

[Miller et al.1993] George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Miller1995] George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Moro and Navigli2015] Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June. Association for Computational Linguistics.

[Moro et al.2014] Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *TACL*, 2:231–244.

[Navigli et al.2007] Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *SemEval-2007*, pages 30–35, Prague, Czech Republic, June.

[Navigli et al.2013] Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

[Pasini and Navigli2017] Tommaso Pasini and Roberto Navigli. 2017. Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88. Association for Computational Linguistics.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[Peters et al.2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

[Polguère2003] Alain Polguère. 2003. *Lexicologie et sémantique lexicale*. Les Presses de l'Université de Montréal.

[Pradhan et al.2007] Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Raganato et al.2017a] Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017a. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April. Association for Computational Linguistics.

[Raganato et al.2017b] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017b. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1178. Association for Computational Linguistics.

[Snyder and Palmer2004] Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

[Taghipour and Ng2015] Kaveh Taghipour and Hwee Tou Ng. 2015. One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Beijing, China, July. Association for Computational Linguistics.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

[Vial et al.2018] Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2018. UFSAC: Unification of Sense Annotated Corpora and Tools. In *Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, May.

[Yuan et al.2016] Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *COLING 2016*.

[Zhong and Ng2010] Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 78–83, Stroudsburg, PA, USA. Association for Computational Linguistics.

Estimating senses with sets of lexically related words for Polish word sense disambiguation

Szymon Rutkowski
Institute of Computer Science
Warsaw, Poland
szymon@szymonrutkowski.pl

Piotr Rychlik
Institute of Computer Science
Warsaw, Poland
rychlik@ipipan.waw.pl

Agnieszka Mykowiecka
Institute of Computer Science
Warsaw, Poland
agn@ipipan.waw.pl

Abstract

We propose a new algorithm for word sense disambiguation, exploiting data from a WordNet with many types of lexical relations, such as plWordNet for Polish. In this method, sense probabilities in context are approximated with a language model. To estimate the likelihood of a sense appearing amidst the word sequence, the token being disambiguated is substituted with words related lexically to the given sense or words appearing in its WordNet gloss. We test this approach on a set of sense-annotated Polish sentences with a number of neural language models. Our best setup achieves the accuracy score of 55.12% (72.02% when first senses are excluded), up from 51.77% of an existing PageRank-based method. While not exceeding the first (often meaning most frequent) sense baseline in the standard case, this encourages further research on combining WordNet data with neural models.

1 Introduction

Ambiguity is an inherent feature of natural languages. There is no one-to-one relation between the vocabulary of word units and the set of meanings which these words represent. Although there are more and more applications in which disambiguation step is not clearly distinguished, explicit identification in which sense a particular word is used in a given context remains important in many situations.

If we aim at selecting a specific sense from a given inventory like WordNet (A. Miller, 1995), this task is called Word Sense Disambiguation (WSD) and was commonly addressed in one of two ways. The first one treats the task as a standard word classification problem solved using any of the supervised learning techniques. The hard part of applying this approach is obtaining satisfactorily large annotated data sets for relatively big subset of senses, even if the annotation can be partially bootstrapped in a semi-supervised manner, for example using label propagation (Yuan et al., 2016). Manual labelling of data with word senses takes time, and agreement between annotators is usually not very high. Another problem is that a lot of text has to be processed to collect occurrences of several (or even more) senses of each word.

This is why the second approach to WSD seems to be more common. In this type of solutions, information included directly or indirectly in lexical databases, especially WordNet, is used either to generate additional features or as the only data source (in the algorithms based on analysis of knowledge graph structure). Recently, vector word representations and neural network architectures have started to be widely used. Our solution combines neural models trained on a large text corpus with information extracted from the plWordNet (Piasecki et al., 2009).

2 Related Work

The problem of resolving lexical ambiguity has a long and complicated history. This task is one of the oldest problems in computational linguistics and machine translation research, but its definition and role in natural language processing (NLP) community's efforts changed over time in many ways. Although solutions of one specific version of the problem – an explicit task of resolving fine-grained and coarse-grained ambiguity to a fixed inventory of senses – showed, at the Senseval-3 conference (Mihalcea et al., 2004a), consistent and respectable accuracy levels, Agirre and Edmonde (2006) observed that this success did not lead to better performance in real applications. They opined that WSD as a topic of study found it-

self "in a strange position", and seemed to diverge from research on NLP applications, "despite several efforts to investigate and demonstrate its utility".

The authors of the best solution at that time (Mihalcea et al., 2004b) reported an accuracy score of 0.65, which was at human levels according to inter-annotator agreement. Their method requires constructing a graph with all senses of words that are present in the text. A PageRank-like algorithm is applied to this graph for choosing the most salient senses, combined with the Lesk algorithm (Lesk, 1986) and most frequent senses heuristics.

Although this system achieved the best result, accuracy of 0.65 was not satisfactory for industrial NLP applications. It should be noted, however, that these results were obtained on Princeton WordNet, which distingishes fairly multiple senses for words, with granularity than can exceed the needs in many situations. With no direct enhancement in view, research on the WSD task was receiving waning interest, but did not cease entirely.

Many researchers explored different measures for graph connectivity which might be useful for the WSD task (Navigli and Lapata, 2007). In the SemEval-2013 Task 12, linked data for different languages were also used for this purpose (Navigli et al., 2013; Panchenko et al., 2017). With the increasing popularity of distributional semantic approach, many experiments exploiting word embeddings as an additional or the only source of information were performed (Iacobacci et al., 2016; O et al., 2018).

While the evidence from research on the WSD task for English appears contradictory, it should be instructive to see how approaches perform on data in different languages with their unique problems and qualities. For Polish, relatively little was investigated on this subject, but some results were published. Leaving out very early experiments which constrained themselves to a purpose-built set of senses for a group of selected words, we should mention (Kędzia et al., 2015) who employed the graph-based method proposed by (Mihalcea et al., 2004b) and (Agirre et al., 2014), utilizing data from plWordNet integrated by the authors with existing SUMO ontology.

Recently, (Wawer and Mykowiecka, 2017) proposed an approach where probability of senses in context is assessed by replacing the disambiguated word with unambiguous members of their synsets. This method, while obviously limited to cases where such unambiguous words can be found in the token's synsets, produced promising results when tested on data from (Hajnicz, 2014). The general idea of estimating context probability with replacements from a WordNet is similar to the one presented in this paper, but we argue that it can be exploited more fully using lexical relations.

3 Test Data Description

Our test data consists of a small sample of 1000 sentences selected from the manually annotated part of the NKJP (National Corpus of Polish) (Przepiórkowski et al., 2012). The sentences were chosen randomly, but we excluded transcribed speech and internet sources. We collected 24,535 tokens of 9,741 token types in total. All nouns, verbs, adjectives and adverbs were manually annotated with plWordNet 3.1 senses by appropriately trained linguists.

As the annotation process is very time consuming, only a part of the data was annotated by both of them and they agreed on 83% of tokens. This is comparable to the measures of inter-annotator agreement in Senseval competitions (Green et al., 2017). In Senseval-1, the 80% agreement was eventually achieved by allowing for discussion and revisions of ambiguities in lexical entries before final tagging. In Senseval-2, the agreement on verb annotation was initially 71%, but after grouping some senses into more coarse-grained ones it rose to 82%.

4 Method

Intuitively, when people have to disambiguate senses, they look at the context and choose the most fitting meaning – that is, the sense that would produce an interpretation of the sentence (and of the text) that the author would probably "have in mind". This presupposes knowledge of the inventory of senses, and some way of representing them for evaluation.

In computer contexts, we usually use a WordNet as an authority on senses. The vague concept of "fitting" may be expressed in terms of probabilities. As to representation, unless we devise some way of obtaining sense embeddings, we have to employ some tricks, like the one presented below.

Speaking a little more formally, for every ambiguous word (w), we would like to select the

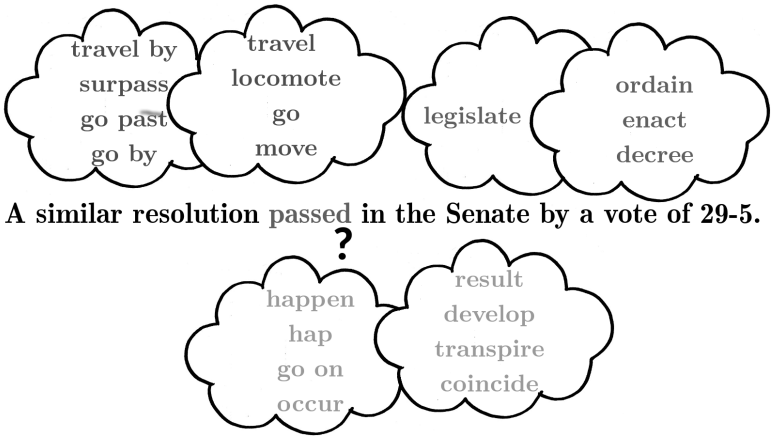


Figure 1: A visual example of representing three senses of the verb *pass* (here taken from Princeton WordNet, for English) with related words – other synset members to the left, and hypernyms to the right. (Both the senses and their associated words are a selection from larger sets.) These "neighbours" could be substituted for the original word to estimate the likelihood of the sense occuring here.

sense (s^*) with the highest probability given the form and context (c) of the word:

$$s^* = \arg\max_s P(s|w,c) \tag{1}$$

However, since there is no clear way to obtain $P(s)$ directly, we approximate it with some set R_s of word forms related to the sense in question. One way of combining the evidence from members of R_s is to average their probabilities in the context:

$$s^* = \arg\max_s \frac{\sum_{r \in R_s} P(r|w,c)}{|R_s|} \tag{2}$$

We also test the variant where the highest probability estimated for a related word is taken to represent the whole sense:

$$s^* = \arg\max_s \max_{r \in R_s} P(r|w,c) \tag{3}$$

Once r is an explicitly designated word form or lemma, a language model capable of predicting probability of word sequences can be used to predict $P(r|w,c)$.

Note that we only have to decide whether the word is likely to occur in the context or not; there is no need for a full distribution of words that could occur there otherwise. Thus, following word2vec's negative sampling method (Mikolov et al., 2013), we train our language model only to discriminate between true and "garbled" fragments of text. Specifically, we obtain negative samples for training from positive (real) ones by shuffling the order of words and replacing some of them with random entries from vocabulary.

We define the set of related words (neighbours) as follows, using relations between lexical units, i.e. senses, and synsets in plWordNet (compare Figure 1). For relations among lexical units, we include lemmas of the related units. For relations between synsets, we include lemmas of all lexical units belonging to the related synsets. Also words from the same synset as the lexical unit in question are taken into account. Finally, words from the lexical unit description (gloss) can also be treated as neighbours.

Intuitively, swapping the ambiguous word for related terms, such as hyponyms or hypernyms, is a method similar to heuristics that a human could use. To give an English example, to disambiguate the word *plants* in the phrase *People there liked to surround themselves with plants*, one might try to substitute some synonyms, and estimate how much sense they would make semantically in the context: *People there liked to surround themselves with factories*, *People there liked to surround themselves with flora*, *People there liked to surround themselves with contrivances*, etc. The ones that have the highest probability of occurring would tend to be those which are related to true sense of the original word.

Since it is possible for a sense to not yield any

neighbours, because of having no relevant relations, we use the probability of the original context (that is, the one containing the word being disambiguated) as the baseline probability for all senses. Only when a sense does have some other words related to it, the baseline is replaced with either the average or the maximum of their estimated probabilities.

Estimates for all senses, computed separately, in practice rarely sum to one. We normalise them before making the decision, although this does not influence the final verdict of the model. If many senses have the same, highest estimated probability, we choose from among them at random.

5 Experiments

In plWordNet, there are many types of relations – over 40 in the 3.1 version, not counting subtypes, which makes experimenting with them attractve. We selected some of relation types that seemed particularly useful for our task, and grouped them into three primary subsets.

The first subset contains synonymy (including belonging to the same synset), hypernymy and hyponymy, the second contains also antonyms, and the third one, apart from everything from the first subset, incorporates various types of meronymy and other relation types that seem to connect to words that would be adequate replacements for their neighbours in the sentence. For example, in plWordNet there is a number of relations connecting verbs that presuppose or imply each other, or adjectives that differ by magnitude of the quality that they describe.

We test [1] how accurate are predictions based on (1, 2, 3) those three subsets, (4) combination of all of them, (5) on words from glosses only, (6) on words in glosses and all words obtained from relation subsets.

The basic context probability estimator, serving as the core of our system, is an LSTM (Hochreiter and Schmidhuber, 1997) network, taking nine word vectors as its input, with the disambiguated word position in the middle. The hidden size of an LSTM cell is as little as 9 – we have tried bigger values, such as 64 and 128, but they performed worse.

The last output of the LSTM is squashed with sigmoid function and interpreted as probability. Previously published set of word embeddings (Mykowiecka et al., 2017) was used for vectorising sentences. We used 300-dimensional vectors from a word2vec model, trained using continuous bags-of-words and negative sampling on lemmatised corpus consisting of NKJP and the Polish Wikipedia. As an alternative, we also tested vectorising contexts with ELMo embeddings (Peters et al., 2018), using the ELMoForManyLangs package (Fares et al., 2017; Che et al., 2018). It provided a pretrained model for Polish and an appropriate interface. Both LSTMs were trained on the manually annotated, balanced portion of NKJP.

These setups were compared with an existing hierarchical softmax model that was trained on full, unbalanced version of NKJP and Polish Wikipedia corpus. It exists in Gensim (Řehůřek and Sojka, 2010) format, which allows for scoring probabilities (or more precisely, log likelihoods) of entire sentences, which can be also applied to sentence fragments. As explained in (Taddy, 2015), log likelihood of a sentence $S = [w_1, w_2, \ldots, w_n]$ is defined as the pairwise composite log likelihood:

$$\mathcal{L}(S) = \sum_{i=1}^{|S|} \sum_{j=1}^{|S|} \ell(w_i, w_j),$$

where

$$\ell(w_i, w_j) = \begin{cases} \log P(w_i|w_j) & \text{if} \quad 1 \leq |j-i| \leq b \\ 0 & \text{otherwise} \end{cases}$$

With the skipgram variant of word2vec model which was used here, $P(w_i|w_j)$ denotes the conditional probability of a context word w_i for a target word w_j. The number b is the window size used in model training. In our case, it is 5, so the whole window contains 11 words.

The Gensim implementation uses a shallower regular word2vec architecture instead of recurrent networks. It is also, in contrast to the RNN, not intrinsically aware of word order.

6 Results

Results in Table 1 show, for all models, a sharp improvement of quality when all types of relations are considered, as opposed to smaller subsets. It seems that regardless of whether neighbour words make sense as replacements for the word being disambiguated, their semantic relatedness to the

[1] The source code is available at `zil.ipipan.waw.pl/CoDeS`.

Neighbour subset	RNN/avg	Gensim/avg		ELMo/avg	RNN/max	Gensim/max		ELMo/max
Relations 1	42.94%	41.36%	40.28%	43.45%	43.90%	40.36%	39.53%	43.75%
Relations 1+2	44.70%	43.06%	42.53%	44.99%	43.89%	40.73%	39.68%	43.52%
Relations 1+3	45.58%	44.68%	44.77%	46.04%	44.37%	40.34%	40.62%	44.00%
Relations 1+2+3	53.93%	50.83%	54.00%	54.08%	54.92%	50.57%	54.97%	55.08%
Glosses	43.93%	43.37%	43.90%	44.18%	44.70%	42.85%	44.80%	42.85%
Glosses + Rels	53.88%	50.88%	54.09%	54.01%	**55.12%**	50.52%	54.89%	55.08%

Table 1: Prediction accuracy measured for all ambiguous cases in our corpus: 'RNN' – basic model, 'Gensim' – Gensim implementation of sequence likelihood (for nine word window and full sentence case), 'ELMo' – RNN with ELMo embeddings instead of word vectors; 'avg' – taking the average probability of all neighbours, 'max' – taking the maximal value.

Neighbour subset	RNN/avg	Gensim/avg		ELMo/avg	RNN/max	Gensim/max		ELMo/max
Relations 1	55.51%	54.16%	52.97%	55.69%	56.70%	53.26%	51.45%	55.75%
Relations 1+2	57.73%	56.23%	55.72%	57.95%	56.68%	53.65%	51.85%	55.64%
Relations 1+3	58.40%	59.63%	57.23%	58.94%	57.58%	53.27%	52.84%	56.40%
Relations 1+2+3	70.01%	66.94%	69.99%	70.22%	71.77%	65.35%	71.82%	72.02%
Glosses	56.58%	57.23%	56.33%	57.01%	56.93%	56.53%	57.85%	58.38%
Glosses + Rels	70.60%	66.97%	70.05%	70.12%	**72.02%**	65.29%	71.61%	71.74%

Table 2: Prediction accuracy measured for cases where the first sense was not the correct one.

context facilitates recognition of the correct sense. On the other hand, glosses appear to work relatively poorly as a source of neighbours for our solution. This may be partially explained by the lack of consistent formatting of glosses in Polish WordNet, where definitions, examples and other metadata are mixed in a couple of ways in one field of the database.

For almost all methods, the approach of taking the maximum probability instead of the average yielded better results. The only exceptions are some weaker versions of Gensim and ELMo approaches. We hypothesise that neighbours that seem the most likely in given context may indeed reflect the best whether the sense that they represent is appropriate. A possible counterargument would point towards negligible improvements caused by this change to the approach based entirely on words from glosses. Although one would think that ignoring junk words from metadata would markedly raise chances of the true sense, this appears not to be the case.

It should be noted that these results, unfortunately, are still lower than the baseline of 59.77% cases where the correct sense is the first variant in Polish WordNet (which often, but not always, happens to be the most frequent one in Polish language). It is a known issue in development of WSD solutions, and for our data this result is even higher than MFS (Most Frequent Sense) accuracy cited for English in (Agirre and Edmonds, 2006), i.e. 46.4%. However, most measurements exceed the lower baseline of assigning sense annotations at random (45.08% accuracy).

Among all the models of context probability evaluation, the basic word vector-based LSTM performed the best. Its superiority over ELMo seems to be linked to operating on lemmas, instead of forms, as the pretrained ELMo embedder. Due to rich morphology of Polish, information in a corpus is markedly easier to generalise if the inflections are abstracted away. Our preliminary tests with training a form-based LSTM operating on word vectors confirmed this hypothesis by degrading maximum accuracy, although it still fared better than ELMo on smaller relation subsets.

It is true that any RNN shows an improvement over the Gensim non-recursive solution, which is unaware of word order. We additionally ran more relaxed tests where this model was allowed to see whole sentences (as the Gensim package interface suggests to do), and even then it was not able to reach the level of RNNs.

Analysis of differences between sets of incorrectly classified words has shown the gains to be incremental. This is supported by our experiments with disambiguation by voting of various models, which yielded little improvement. This, along with moderate differences in accuracy, shows that the behaviors of individual variants appear, ultimately, similar. One should keep in mind, however, that our corpus size makes it difficult to draw conclusions concerning particular morphological features in Polish that might be the stronger points

of some models.

We also present results obtained on non-first variant cases only, in Table 2. It appears that our algorithm is capable of relatively precise treatment of less frequent senses, even though it has issues with separating them from the dominant ones. Here we still observe the superiority of LSTM based on word vectors with taking the maximum probability.

We compared our results with the only one other general purpose method for solving Polish WSD task described in (Kędzia et al., 2015). We carried out the test on our test set using two taggers: WCRFT2 (Radziszewski and Warzocha, 2014) and MorphoDiTa (Straka and Straková, 2014). In both cases, we have achieved accuracy of around 51% (more precisely, 51.05% for WCRFT2 and 51.77% for MorphoDiTa). All versions of our algorithm surpassed these scores, as long as they considered all the subsets of plWordNet relations.

7 Conclusions

We present a new method of disambiguating senses in Polish texts using lexical relations from the plWordNet database. We test various relation subsets and approaches to modeling probability of contexts.

The WSD problem for Polish is still far from being solved. No published results were able to exceed 70% accuracy, which would move them closer to matching those published for English. It is worth pointing out, however, that our accuracy for cases where the first WordNet sense was excluded does approach this level of performance. Perhaps finding a way to distinguish the most typical contexts, where one can expect these most frequent senses to occur, can greatly help the overall usefulness of the system.

Judging from our findings, there is little to be gained by enhancing language models within the same framework of estimating sense likelihoods. The results do show potential in combining modern machine learning with creative use of existing knowledge bases, and should encourage further research in this direction.

Acknowledgments

This work was supported by the Polish National Science Centre project 2014/15/B/ST6/05186.

References

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–, 11.

Eneco Agirre and Philip Edmonds, editors. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Comput. Linguist.*, 40(1):57–84, March.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October. Association for Computational Linguistics.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden, May. Association for Computational Linguistics.

Meredith Green, Orin Hargraves, Claire Bonial, Jinying Chen, Lindsay Clark, and Martha Palmer. 2017. Verb/ontonotes-based sense annotation. In *Handbook on Linguistic Annotation*. Springer.

Elżbieta Hajnicz. 2014. Lexico-semantic annotation of *składnica* treebank by means of PLWN lexical units. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 7th International WordNet Conference (GWC 2014)*, pages 23–31, Tartu, Estonia. University of Tartu.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 897–907.

Paweł Kędzia, Maciej Piasecki, and Marlena Orlińska. 2015. Word sense disambiguation based on large scale Polish CLARIN heterogeneous lexical resources. *Cognitive Studies| Études cognitives*, 15:269–292.

Michael E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference*.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004a. The Senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3 Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28. ACL.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004b. Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Piotr Rychlik. 2017. Testing word embeddings for Polish. *Cognitive Studies / Études Cognitives*, 17:1–19.

Roberto Navigli and Mirella Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the IJCAI*, pages 1683–1688.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

Dongsuk O, Sunjae Kwon, Kyungsun Kim, and Youngjoong Ko. 2018. Word sense disambiguation based on word similarity calculation using word vector representation from a knowledge-based graph. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

Alexander Panchenko, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. 2017. Using linked disambiguated distributional networks for word sense disambiguation. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 72–78. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Adam Radziszewski and Radosław Warzocha. 2014. WCRFT2. CLARIN-PL digital repository.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Milan Straka and Jana Straková. 2014. MorphoDiTa: Morphological dictionary and tagger. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Matt Taddy. 2015. Document classification by inversion of distributed language representations. *CoRR*, abs/1504.07295.

Aleksander Wawer and Agnieszka Mykowiecka. 2017. Supervised and unsupervised word sense disambiguation on word embedding vectors of unambigous synonyms. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 120–125. Association for Computational Linguistics.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models.

Merging DanNet with Princeton Wordnet

Bolette S. Pedersen[1], Sanni Nimb[2], Ida R. Olsen [3], Sussi Olsen [4]
University of Copenhagen [1,3,4] & The Danish Society for Language and Literature[2]
Njalsgade 136, DK-2300 Copenhagen S[1,3,4,] Christians Brygge 1, DK-1219[2]
bspedersen@hum.ku.dk, sn@dsl.dk, jms862@hum.ku.dk, saolsen@hum.ku.dk

Abstract

In this paper we describe the merge of the Danish wordnet, DanNet, with Princeton Wordnet applying a two-step approach. We first link from the English Princeton core to Danish (5,000 base concepts) and then proceed to linking the rest of the Danish vocabulary to English, thus going from Danish to English. Since the Danish wordnet is built bottom-up from Danish lexica and corpora, all taxonomies are monolingually based and thus not necessarily directly compatible with the coverage and structure of the Princeton WordNet. This fact proves to pose some challenges to the linking procedure since a considerable number of the links cannot be realised via the preferred cross-language **synonym** link which implies a more or less precise correlation between the two concepts. Instead, a subpart of the links are realised through near synonym or hyponymy links to compensate for the fact that no precise translation can be found in the target resource. The tool WordnetLoom is currently used for manual linking but procedures for a more automatic procedure in future is discussed. We conclude that the two resources actually differ from each other quite more than expected, both vocabulary- and structure-wise.

1 DanNet - a monolingually compiled wordnet

In contrast to the majority of wordnets following the Princeton standard, DanNet (Pedersen et al. 2009) is constructed using the so-called merge approach where the wordnet is built on monolingual grounds and thereafter merged with Princeton WordNet (PWN, cf. Fellbaum 1998).

DanNet is open source and currently contains 65,000 synsets available from www.wordnet.dk in owl/rdf and csv formats (Pedersen et al. 2009). It can be browsed online from www.andreord.dk or from wordties.cst.ku.dk.

The wordnet has been compiled as a collaboration between the University of Copenhagen and the Society for Danish Language and Literature and is based on Den Danske Ordbog (DDO, Hjorth et al. 2003-2005). In other words, our starting point was the corpus-based, at that time newly completed dictionary of Danish, accessible in a machine-readable version and with genus proximum information explicitly specified for each sense definition (DDO). The motivation for a monolingual approach seemed obvious since by taking this approach we were enabled to compile the wordnet in a rather efficient and semi-automatic fashion using the genus proximum of the dictionary as the driving factor. The result was a resource truly based on the Danish language and vocabulary and not biased by English.

The SIMPLE lexicons (cf. Lenci et al. 2000) and particularly the Danish version of it (Pedersen & Keson 1999, Pedersen & Paggio 2004) have also influenced the construction of DanNet in the sense that it includes qualia information[1] such as the telic (PURPOSE) and the agentive role (ORIGIN), roles which corresponded well with the content of the word definitions in DDO. Qualia roles are encoded in DanNet in terms of relations such as *used_for*, *made_by* and *concerns* as well as by means of features such as SEX and CONNOTATION. Apart from these additional features, DanNet follows wordnet standards wrt. relation types and synset structure, and all synsets are tagged with EuroWordNet Top Ontology types (Vossen et al 1999).

[1] We apply Qualia Structure and Qualia information as proposed by Pustejovsky 1995.

2 Linking procedure – manual or semi-automatic?

Not surprisingly, a major disadvantage of applying the monolingual strategy is that subsequent linking to PWN becomes really complex and cumbersome, which is also why it was not prioritized in the first phase of the Danish wordnet project. Over time, however, it has become more and more evident that a full linking of the resource is indispensable if we want to operate in all sorts of multilingual contexts and if our vision of applying language transfer where it is meaningful and does not involve too strong a bias, should be realistic. To this end, we have been much inspired by the work around the Polish wordnet, plWordNet (Maziarz et a. 2014), a resource which is compiled monolingually in a fashion comparable to that of DanNet and subsequently merged with PWN. Thus, much of the linking experiences resembled in i.e. Rudnicka et al. (2012) such as differences in taxonomies/structures have counterparts in our work even if the difficulties are not exactly the same. [2]

Driven by the METANET/METANORD initiatives (cf. www.meta-net.eu) where we wanted to validate wordnets across the Nordic countries (cf. Pedersen et al. 2013), we initiated the merge with PWN by focusing on Princeton Core wordnet (http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt) which is a subset 5,000 central concepts of English. Going from English to Danish, these concepts where linked semi-automatically to DanNet and missing concepts where established in the Danish resource. A bilingual dictionary was used as a first automatic look-up and link suggestion for the core concepts and from here on the encoder could accept or modify the proposed links applying a wizard-like routine in the encoding tool.

When embarking in 2018 the ELEXIS project (cf. elex.is, Krek et al. 2018), which is concerned with opening up linguistic and lexicographical data and language tools for European communities, we were finally prompted to start the full linking process of DanNet. This time the process is switched, going from Danish to PWN and thus taking point of departure in the Danish coverage and taxonomy. [3]

In this process, we also make use of a bilingual dictionary, but no semi-automatic linking to PWN is applied at the current stage. The reason for this is that it was not very evident which particular automatic procedure to pursue because of the many cases where no exact match can be found in PWN to a Danish synset, as also depicted in Figure 1.

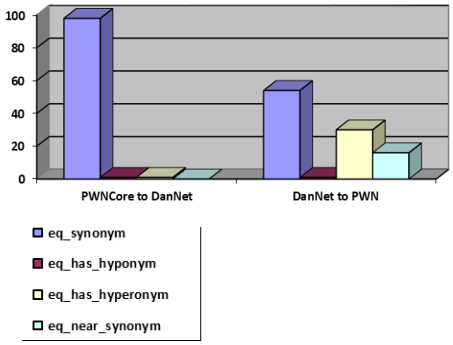


Figure 1. Percentage of different linking relations used when linking core concepts from English to Danish compared to linking general vocabulary from Danish to English.

Figure 1 illustrates how the use of linking relations differ quite radically when linking from PWNCore compared to when linking the other way around from DanNet to PWN. When going from PWNCore to DanNet, i.e. linking between core concepts in the two languages, almost all links are direct links in terms of eq synonym relations (for more details see Section 4). This means that the lexicographer has in almost all cases identified (through the semiautomatic procedure) what is considered to be an exact match between the English and the Danish resource.

The opposite proves to be the case when it comes to the linking of non-core concepts, now with the Danish resource as starting point for the linking process.[4] In the cases where no direct links are

[2] For instance, Rudnicka et al. (2012) show that since lexical units are the main building blocks in plWordNet (and not synsets as in PWN), linking to PWN is not straightforward.

[3] The linking is funded partly by ELEXIS, partly by The Carlsberg Foundation.

[4] Note however that DanNet contains less than one third of the number of senses in PWN. Nonetheless, the coverage differs quite substantially in particular when it comes to compounds, for more discussion see Section 4.

found, a rather complex cognitive procedure is initiated by i.e. looking up the Danish hypernym, finding the corresponding PWN synset, and looking for candidates among the related PWN hyponyms. Alternatively, by searching for a potential PWN hyponym to be linked to (for more details see Section 5).

To this end, we have at the current stage estimated that an automatic procedure for this process requires a rather precise cross-lingual hypernym or hyponym detection as a minimum. Nevertheless, some links can be established semi-automatically once a certain amount of relations have been established. Either vertically in cases where a Danish synset is synonym-linked to a PWN synset where it can be suggested that the hypernym of the PWN synset is also a hypernym of the Danish synset. Or horizontally, e.g. if two Danish synsets are near-synonymous, and only one is synonym-linked to PWN, then the second Danish concept can inherit that near-synonym link.

Another possibility is to apply an automatic prompt system as proposed by Kędzia et al. (2013) where the linguist/lexicographer is prompted in the process of manual mapping plWordNet on PWN. This system is based on the extended Relaxation Labelling algorithm, and suggests potential target synset candidates based on the synset positions in both wordnet structures, bilingual dictionaries and/or input from the linguist. Finally, the linguist verifies (or rejects) suggested links. It seems plausible to adjust this system to our mapping process and speed up the manual linking: it partially resembles the cognitive procedure described above, and also provides a possibility to determine the desired type of semantic relation.

At a later stage, when a more substantial part of the vocabulary has been linked, we will consider whether to follow for example Joshi et al. (2012) who generate lists of potential linking candidates with a heuristic based measure by pruning and ranking information from bilingual dictionaries. Better results are achieved with this measure when a number of links are already established. This approach could potentially be implemented when being able to utilize the high-quality established links to PWN already made by language experts. Arcan et al. (2016) use existing relations across wordnets and parallel corpora to identify contextual information for wordnet senses, and thereby expand the wordnets. Such an approach could also be adapted in our case and, again, build on the established links.

The approach of McCrae et al. (2017) for linking English-German knowledge graphs combines machine translation and cross-lingual ontology alignment. This approach, which makes use of the NAISC tool (McCrae et al. 2018), could be adapted for linking DanNet to PWN, and tested on the established links. It would require high-quality machine translation and sufficiently rich synset information, which additionally could be reinforced with contextual information as in Arcan et al. (2016).

Certainly, such automatic approaches would not achieve the precision of the manually created links, but they could be integrated as part of a semi-automatic procedure in order to speed up the process.

3 Linking complexities due to taxonomical differences

A major challenge when merging two wordnets concerns the often found discrepancies in taxonomical structure (Pedersen et al. 2013, Rudnicka 2012). Taxonomical discrepancies may have different origins, such as:

- different overall compilation approaches regarding how to organize the wordnet
- cultural differences in how to conceive a (group of) concept(s),
- idiosyncracies of the wordnet developers.

In our linking work, we encounter discrepancies of all three types. Where DanNet is compiled on the basis of a layman's dictionary of Danish, PWN is compiled without basis in any specific previous resource, but generally more true to expert knowledge in particular in relation to i.e. natural taxonomies. Consider the taxonomical complexity of the concept *plante* ('plant') in DanNet in Figure 2 compared to that of PWN in Figure 3. Even if the graphical interfaces differ, it proves quite evident that DanNet uses a layman's much simpler organization principles of plants than does PWN. Another overall discrepancy worth mentioning is different approaches taken wrt. the treatment of systematic polysemy. For instance, in DanNet all countries have a 'geographical' and a 'people' reading, a dichotomy which is not

equally found in PWN and which makes a one-to-one linking procedure impossible.

Figure 2: Taxonomical complexity of *plante* ('plant') in DanNet based on a layman's approach

- S: (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)
  - ***direct hyponym*** / *full hyponym*
    - S: (n) phytoplankton (photosynthetic or plant constituent of plankton; mainly unicellular algae)
    - S: (n) microflora (microscopic plants; bacteria are often considered to be microflora)
    - S: (n) crop (a cultivated plant that is grown commercially on a large scale)
    - S: (n) endemic (a plant that is native to a certain limited area) *"it is an endemic found only this island"*
    - S: (n) holophyte (an organism that produces its own food by photosynthesis)
    - S: (n) non-flowering plant (a plant that does not bear flowers)
    - S: (n) plantlet (a young plant or a small plant)
    - S: (n) wilding (a wild uncultivated plant (especially a wild apple or crabapple tree))
    - S: (n) ornamental (any plant grown for its beauty or ornamental value)
    - S: (n) pot plant (a plant suitable for growing in a flowerpot (especially indoors))
    - S: (n) acrogen (any flowerless plant such as a fern (pteridophyte) or moss (bryophyte) in which growth occurs only at the tip of the main stem)
    - S: (n) apomict (a plant that reproduces or is reproduced by apomixis)
    - S: (n) aquatic (a plant that lives in or on water)
    - S: (n) cryptogam (formerly recognized taxonomic group including all flowerless and seedless plants that reproduce by means of spores: ferns, mosses, algae, fungi)
    - S: (n) annual ((botany) a plant that completes its entire life cycle within the space of a year)
    - S: (n) biennial ((botany) a plant having a life cycle that normally takes two seasons from germination to death to complete; flowering biennials usually bloom and fruit in the second season)
    - S: (n) perennial ((botany) a plant lasting for three seasons or more)
    - S: (n) escape (a plant originally cultivated but now growing wild)
    - S: (n) hygrophyte (a plant that grows in a moist habitat)
    - S: (n) neophyte (a plant that is found in an area where it had not been recorded previously)
    - S: (n) embryo ((botany) a minute rudimentary plant contained within a seed or an archegonium)
    - S: (n) monocarp, monocarpic plant, monocarpous plant (a plant that bears fruit once and dies)
    - S: (n) sporophyte (the spore-producing individual or phase in the life cycle of a plant having alternation of generations)
    - S: (n) gametophyte (the gamete-bearing individual or phase in the life cycle of a plant having alternation of generations)
    - S: (n) houseplant (any of a variety of plants grown indoors for decorative purposes)
    - S: (n) garden plant (any of a variety of plants usually grown especially in a flower or herb garden)
    - S: (n) vascular plant, tracheophyte (green plant having a vascular system: ferns, gymnosperms, angiosperms)

Figure 3. *Plant* with hyponyms in PWN

Cultural differences regarding how for instance the educational or the juridical system is organized is also clearly reflected in the taxonomical structures. Finally, pure idiosyncracies are found all over the resources, maybe even to some extent also culturally based; for instance cheese has a taxonomical division of concepts in DanNet (Figure 4) based on whether the cheese is cut or spread on the bread (typically on open sandwiches of rye bread); a division which is not made in PWN.

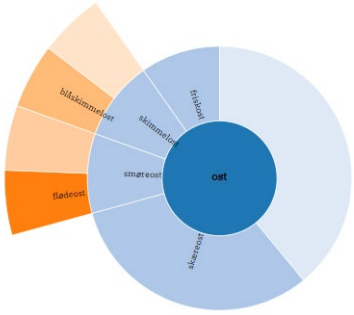


Figure 4. *ost* ('cheese') taxonomical complexity in DanNet.

4 Core concepts: Linking complexities and lexicographic characteristics

The core concepts of PWN have been selected based on two criteria: Importance of synsets measured by a) the number of relations with other synsets and b) a high position in the hierarchy.

Oflazer & Murat (2018) describes how the six Balkanet WordNets successfully used the latter criterion, a relatively high level of the English words in the PWN hierarchy, as a common starting point for the expand method, based on the assumption that language-specific information gets more important as one moves down the hierarchy. Also Green (2006) states that concepts at a basic level are more likely to be shared across classificatory systems than concepts at more general or more specific levels. In our case this is confirmed. As already described in Section 2, the linking process of the core concepts when going from PWN to DanNet results in many direct links, and equivalents were likely to be part of vocabulary covered by DanNet - only in a few cases new synsets had to be created.

The fact that DanNet is linked directly to a medium-sized corpus-based monolingual dictionary giving access to all types of lexical information now allows us to study the lexicographic characteristics of the core vocabulary in detail. We would in the case of Danish expect the core concepts to be simplex words rather than compounds and are now able to find out whether it is in fact the case. Simplex lemmas in DDO are opposite to

compound lemmas characterized by often being part of the manually selected ~65,000 lemmas that constituted the vocabulary of the first printed version of the dictionary, and thereby to carry information on etymology, phonetics and compounding to a much higher degree than the ~35,000 lemmas added in the later years, after the first published edition. As seen in Table 1, the DanNet core-concept lemmas do in fact have a far higher number of all these types of information than the non-core lemmas.

Information on: DanNet Lemma	**Core**	**Non-core**
Etymology	65 %	33 %
Compounding	61%	8%
Phonetics	87%	45%
Part of DDO priority selection	99,98%	69%

Table 1. Comparison of information types across core and non-core vocabulary, percentage per lemma.

We would also expect the core concepts to be much more polysemous than the non-core concepts. The linking challenges we encountered when mapping the core synsets of PWN to DanNet are well-known to all WordNet developers (see for example Rudnicka et al. 2012, Cristea et al. 2004), typically being caused by the differences in sense distinctions and sense granularity. Often the case would be that one English synset corresponds to two or more Danish synsets, or vice versa, or even more challenging, the distinction between senses has been drawn in a slightly different way in the two resources. When looking at the number of senses of the Danish core vocabulary, it becomes obvious why the mapping was not trivial. Even though the core concept lemmas in DDO constitute only 4.6 % of the total number of lemmas in the dictionary, they cover 21.6 % of the senses in the dictionary. And while 69 % of the core lemmas are polysemous, this is only the case for 28 % of the non-core lemmas. The polysemous core lemmas have 2.65 times as many senses as the non-core polysemous lemmas. When it comes to fixed expression, the 4.6% core lemmas cover 56% of the total number in the dictionary, and they are much more likely to be part of one: 37% of them have at least one. This is only the case for 6.5% of the non-core lemmas. The core lemmas have an average of 2.76 times as many fixed expressions as the non-core lemmas, cf. Table 2. The high degree of polysemy and the high number of fixed expressions is of course a complicating factor when core concepts are linked between PWN and DanNet.

DanNet vocabulary	**Core**	**Non-core**
Lemmas ≥ 2 senses	69%	28%
Sense per polysemous lemma (incl. fixed expressions)	6.55	2.47
Lemmas with fixed expression	37%	6,5%
Fixed expressions (of lemmas with fixed expression	4.41	1.6
% of definitions (total DDO = 98,944)	21,6% 21,407	78,4% 77,537

Table 2. DanNet - core and non-core vocabulary, polysemous lemmas and fixed expressions.

When it comes to the challenges caused by different sense granularities in the two lexical resources, the Danish lexicographers who mapped the core concepts often got the impression that the sense inventory of PWN was more fine-grained than the one of DanNet/DDO. This seems to be for a good reason. When studying 20 highly polysemous Danish nouns with their English equivalents (see Table 3), we calculated PWN to have an average of 10.3 % more senses. A similar comparison of highly polysemous verbs and adjectives would probably show an even bigger difference in the number of senses.

	Number of senses	
Lemma, Danish/ English	**DDO**	**PWN**
selskab / company	10	9
kontakt /contact	9	9
kort / card, map	10	11
Plads /room, space..	13	16
slag (stroke; blow; knock)	17	12 (*stroke*)
top /top	8	11
hul /hole	14	8
plade / plate; sheet	11	15 (*plate*)
lys / light	13	15
Model	8	9
skud / shot	12	17
kurs / course	3	9
hold / hold	12	9
ansigt / face	7	13
skade / damage; harm	4	5 (*damage*)
blik / look; gaze	5	4 (*look*)

stykke / piece; bit; part	18	13 (*piece*)
stand / condition; state	9	8 (*condition*)
støtte / support	5	11
vold_1 / violence; force	6	10 (*force*)
Total number	**194**	**214 = 10,3 % more**

Table 3. Number of senses for selected polysemous Danish nouns and their English equivalents.

5 Linking complexities of non-core concepts (going from DanNet to PWN)

When it comes to the vocabulary of the non-core concepts, the linking complexities are of a different nature. One might think that the task of mapping less polysemous words would be easier, but confirming the assumptions of Oflazer & Murat, (2018) mentioned in Section 4, it seems that language-specific peculiarities tend to evoke more translation difficulties as one moves down the hierarchy. In spite of their considerable size, the two lexical resources cover quite different vocabulary and it is often difficult to find exact equivalents. Although the two wordnets seem to have more or less the same level of specificity, it is not carried out in detail for exactly the same vocabulary. Sometimes PWN is more specific wrt. to hyponyms than DanNet, and sometimes DanNet covers the highest number of specific concept lemmas, typically in the form of compounds. As an example to this, the noun *forhandling* ('negotiation') has 13 hyponyms in DanNet, all compounds, e.g. *kontraktforhandling* ('contractual negotiations'), *skilsmisseforhandling* ('divorce proceedings'). The English equivalents are not included in OED, nor in PWN. And the English equivalent to the hypernym *forhandling* ('negotiation') has no hypernyms in PWN.

Even when it comes to mapping the hyponyms of concrete core concepts which are already mapped, and where we find roughly the same number of hyponyms in the two wordnets, we might still not find many equivalents among these hyponyms. Compare for example the types of carpets in DanNet, the hyponyms of *tæppe* (*axminstertæppe*, *bedetæppe*, *kludetæppe*, *kokostæppe*, *løber*, *perser*, *rya*, *måtte*, *filttæppe*, *forligger*, *sengeforligger*, *tæppebelægning*) in Figure 5 with the types of carpets in PWN, the hyponyms of *rug* in PWN in Figure 6. Among the 14 English hyponyms, only *prayer carpet*, *runner*, *scatter rug* and *shag rug* have Danish equivalents among the 12 hyponyms of *tæppe*.

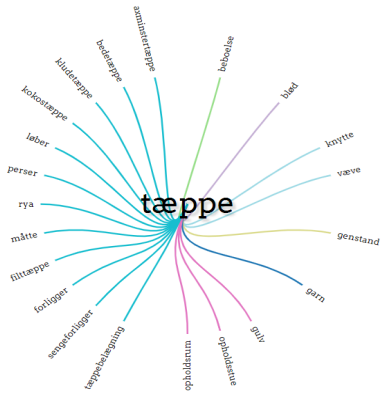


Figure 5. Hyponyms of *tæppe* ('rug') in DanNet

- S: (n) **rug**, carpet, carpeting (floor covering consisting of a piece of thick heavy fabric (usually with nap or pile))
  - ***direct hyponym*** / *full hyponym*
    - S: (n) broadloom (a carpet woven on a wide loom to obviate the need for seams)
    - S: (n) Brussels carpet (a carpet with a strong linen warp and a heavy pile of colored woolen yarns drawn up in uncut loops to form a pattern)
    - S: (n) drugget (a rug made of a coarse fabric having a cotton warp and a wool filling)
    - S: (n) flying carpet ((Asian folktale) an imaginary carpet that will fly people anywhere they wish to go)
    - S: (n) hearthrug (a rug spread out in front of a fireplace)
    - S: (n) Kurdistan (an oriental rug woven by Kurds that is noted for fine colors and durability)
    - S: (n) numdah, numdah rug, nammad (an embroidered rug made from a coarse Indian felt)
    - S: (n) prayer rug, prayer mat (a small rug used by Muslims during their devotions)
    - S: (n) red carpet (a strip of red carpeting laid down for dignitaries to walk on)
    - S: (n) runner (a long narrow carpet)
    - S: (n) scatter rug, throw rug (a small rug; several can be used in a room)
    - S: (n) shag rug (a rug with long pile)
    - S: (n) stair-carpet (a strip of carpet for laying on stairs)
    - S: (n) Wilton, Wilton carpet (a carpet woven on a Jacquard loom with loops like a Brussels carpet but having the loops cut to form a close velvety pile)

Figure 6. Hyponyms of *rug* in PWN

Also culture-specific differences as discussed in section 3 result in many lexical gaps. This is a problem that wordnet developers encounter even when applying the expand model in the first place. In BalkaNet for example, once a core wordnet was developed by translating from PWN, the 6 language partners had to add a number of language-specific synsets which were afterwards linked to PWN via hypernymy relations (Oflazer & Murat 2018, p. 328). In our case such synsets are already included in DanNet and have Danish hypernyms, and they are now supplied with an English hypernym as well, also in cases where an English translation equivalent does in fact exist but is not (yet) part of PWN. One example is the vocabulary of handball, a common sport in Denmark, however

less important in the Anglo Saxon community and therefore not (yet) included in PWN.

Finally it should be mentioned that some linking complexities are caused by differences in word formation in Danish and English. Where noun-noun compounding is indeed very productive in Danish, English in many cases construct similar content by using an attributive and a noun. For example, compounds with *andels-* (co-op, cooperative) e.g. *andelssamfund* and *andelsbutik* translate into English by using an attributive and a noun as in 'cooperative society', 'cooperative store'. There seems to be a tendency that such terms are not lexicalized in English to the same degree and thus not present in PWN.

6 The linking tool

For the linking from DanNet to PWN (which is currently ongoing) we apply the wordnet editing system WordnetLoom 2.0 (Naskręt et al. 2017). WordnetLoom is a graph-based system where several users can access and edit the nodes (lexical units) edges (semantic relations), and synsets as well as view glosses and usage examples. The complex ontological types of the synsets (following The EuroWordNet top-ontology (Vossen 1999)) are also visible in the accustomed version suitable for browsing DanNet, developed by Tomasz Naskręt[5] and adapted by Mitchell J. Seaton[6].

An advantage of the system is that users can view and directly edit the relations in the interface, avoiding problems on manual editing of a wordnet representation file. As seen at the top of Figure 7, multiple bars of slices of the wordnet graph can be open at the same time, and are found by a given search query to the left. The results can, in the DanNet adjusted version, be filtered by part-of-speech, synsets, supersenses, lexical units, and lexicons. The presentation of results includes relations and nodes from both DanNet and PWN.

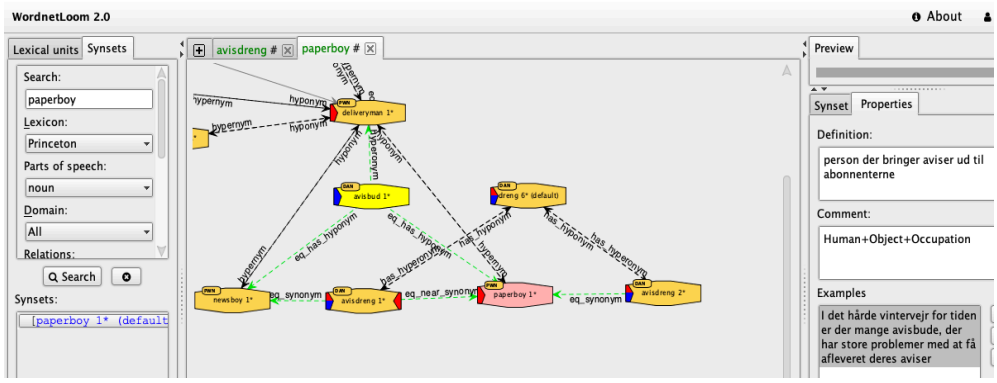


Figure 7: Linking synsets in WordnetLoom

Figure 7 shows an example where *avisbud 1* ('paper deliveryman') is placed between 'deliveryman 1' as a hypernym, and 'newsboy 1' as a hyponym. *avisdreng 1* is synonymous with 'newsboy 1', which is nearly the same as 'paperboy 1'. Every relation can be established, edited or deleted. The synonym, near-synonym, hypernym and hyponym relations (see the green lines) are prioritized (in that order) when linking. The relation is chosen from a drop-down menu as seen in Figure 8.

[5]G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Science and Technology, Wrocław, Poland

[6] Centre for Language Technology, Department of Nordic Studies and Linguistics, Copenhagen University

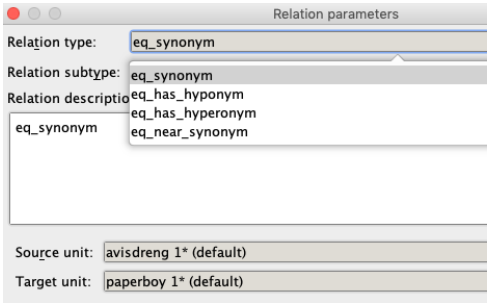


Figure 8: WordnetLoom drop-down menu of relation types.

7 Concluding remarks

The merging of DanNet with PWN is still ongoing and proves both cumbersome and complex as we have exemplified in the previous sections. To speed up the process, we hope to be able to introduce more semi-automatic procedures at a later stage when a substantial number of links have already been established, even if it has become evident that manual inspection and correction will always be a considerable part of the job. Within the ELEXIS project the NAISC tool (McCrae 2018) will soon be available and we hope to examine to which degree a semi-automatic linking with this tool involving interaction between lexicographers and developers can be useful.

It has generally been a surprise to us to acknowledge to which extent the two resources actually differ, both vocabulary- and structure-wise. A fact which has made us realize that a merge of the resources will really only be approximate. Nevertheless, it is our conviction that even such an approximate merge will be useful for several future NLP tasks where Danish is involved. Further, in line with the goals of the ELEXIS project, we hope that it will help interconnect existing resources in the lexicographical milieus in Europe. As such, the merge will provide the interlingual access to a substantial part of the lexical resources available for Danish.

References

Arcan, M., McCrae, J.P., & Buitelaar, P. (2016). Expanding wordnets to new languages with multilingual sense disambiguation. In *Proceedings of {COLING} 2016, the 26th International Conference on Computational Linguistics: Technical Papers, p. 97,* Osaka.

Cristea, D.; Mihaila, C., Forascu, C., Trandabat, D., Husarciuc, M., Haja, M., Postostolache, O. (2004): Mapping Princeton WordNet Synsets onto Romanian Wordnet Synsets. In *Romanian Journal of Information Science and Technology, Vol. 7*, Numbers 1–2, p. 125–145.

Fellbaum, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT press.

Green, R. J. (2006). *Vocabulary alignment via basic level concepts*. OCLC/ ALISE research grant report published electronically by OCLC Research. http://www.oclc.org/research/grants/reports/green/rg2005.pdf.

Hjorth, E. & Kristensen, K. red. (2003-2005). *Den Danske Ordbog, bind 1-6*, DSL / Gyldendal, Online: ordnet.dk/ddo

Joshi, S., Chatterjee, A., Karra, A. K., and Bhattacharyya, P. U. (2012a). Eating your own cooking: automatically linking wordnet synsets of two languages. In *Proceedings of COLING 2012: Demonstration Papers*, p.239—246, Mumbai.

Kędzia P., Piasecki M., Rudnicka E., Przybycień K. (2013). Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies* 13: 123–141.

Krek, S., Kosem, I., McCrae, J., Navigli, R., Pedersen, B. S., Tiberius, C., Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress, Lexicography in Global Contexts*, pp 881-892.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., et al. (2000). SIMPLE—A general framework for the development of multilingual lexicons. *International Journal of Lexicography,* 13(4), 249–263.

Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S. (2014). plWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources. In *Proceedings of the Seventh Global Wordnet Conference*, 2014.

McCrae, J. P. & Arčan, M. & Buitelaar, P. (2017). Linking Knowledge Graphs across Languages with Semantic Similarity and Machine Translation. *The First Workshop on Multi-Language Processing in a Globalising World* (MLP 2017).

McCrae, J. P. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies. 18*, p.109-123. 10.2478/cait-2018-0010.

Naskręt, T., Dziob, A., Piasecki, M., Saedi, C., & Branco, A. (2018). WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation. In *Proceedings of the 9th Global WordNet Conference (GWC2018)*, Singapore.

Oflazer, K., Saraçlar, M. (eds.) (2018). *Turkish Natural Language Processing.* Springer International Publishing AG, Switzerland.

Pedersen; B.S. & Keson, B. (1999). 'SIMPLE - Semantic Information for Multifunctional Plurilingual Lexica: Some Danish Examples on Concrete Nouns'. In: *SIGLEX99: Standardizing Lexical Resources.* Association of Computational Linguistics, ACL99 Workshop, Maryland.

Pedersen, B. S., Paggio, P. (2004). The Danish SIMPLE Lexicon and its Application in Content-based Querying, in *Nordic Journal of Linguistics Vol 27*:1 p.97-127.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L. and Lorentzen, H. (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources & Evaluation* 43:269–299.

Pedersen, B. S., Lindén, K., Vider, K., Forsberg, M., Kahusk, N., Niemi, J., Nygaard, L., Seaton, M., Orav, H., Borin, L., Voionmaa, K., Nisbeth, N. and Rögnvaldsson, E. (2013). Nordic and Baltic wordnets aligned and compared through "WordTies". In *Proceedings from the 19th Nordic Conference on Computational Linguistics (NODALIDA)*. Linköping Electronic Conference.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA

Rudnicka, E., Maziarz, M., Piasecki, M., Szpakowicz, S. (2012). A strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proceedings of COLING 2012*, Posters, pp. 1039–1048, Mumbai.

Vossen, P (ed). (1999). *EuroWordNet, A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, The Netherlands.

.

Development of Assamese Rule based Stemmer using WordNet

Anup Kumar Barman
Dept. of IT
Central Institute of Technology
Kokrajhar, India
ak.barman@cit.ac.in

Jumi Sarmah
Dept. of IT
Gauhati University
Guwahati, India
jumis884@gmail.com

Shikhar Kr. Sarma
Dept. of IT
Gauhati University
Guwahati, India
sks001@gmail.com

Abstract

Stemming is a technique that reduces any inflected word to its root form. Assamese is a morphologically rich, scheduled Indian language. There are various forms of suffixes applied to a word in various contexts. Such inflected words if normalized will help improve the performance of various Natural Language Processing applications. This paper basically tries to develop a Look-up and rule-based suffix stripping approach for the Assamese language using WordNet. The authors prepare the dictionary with the root words extracted from Assamese WordNet and Named Entities. Appropriate stemming rules for the inflected nouns, verbs have been set to the rule engine and later tested the stemmed output with the morphological root words of Assamese WordNet and Named Entities by computing hamming distance. This developed stemmer for the Assamese language achieves accuracy of 85%. Also, the authors reported the IR system's performance on applying the Assamese stemmer and proved its efficiency by retrieving sense oriented results based on the fired query. Thus, Morphological Analyzer will embark the research wing for developing various Assamese NLP applications.

1 Introduction

Computationally, stemming is the process to automatically extract the base form of a given inflected word. The stemmed word is not required to be identical with the morphological root of the word. Most Indian languages are highly inflectional and many words in a document appear in many morphological forms. Indexing is the important subtask of an IR system. Indexing all words in a document appearing in various morphological forms is highly tedious and time-consuming. Thus, it is necessary to stem the words to reduce them to their original base form. Reducing to their original base form will help the indexer in IR to detect the important terms in a document, detect Named entities, multi-word expression and extract stopwords. Looking deeply into the matter, we found that two parts-of-speech Nouns and Verbs have a wide list of inflections for the Assamese language. The main objective of this paper is to perform stemming task on a group of inflected words to retrieve root words with an acceptable accuracy.

Many approaches to stemming have been identified. They are classified into three categories- Rule-based, Statistical and Hybrid approaches.

Rule-based approach- Such approaches apply a set of morphotactic rules of a language to an inflected word. Such rules may derive the base form by emitting the suffix or the prefix.

Statistical approach- One of the drawbacks of rule- based approach is that it is language dependent and it is dependent on the database. Statistical approach overcomes both the problems by calculating probabilistic distributions of the terms.

Hybrid approach- Combination of both rule-based and statistical approaches.

In this paper, the authors have researched and implemented a rule-based stemmer for Assamese language embedding the Look-up based approach. The quick Look-up approach is made on the dictionary prepared from Assamese WordNet and Named Entities. Assamese WordNet is a large lexical knowledge database developed by the team (Sarma et al., 2010). It contains four major components-

- ID: an unique identification number
- CAT: the Parts-Of-Speech category
- Synsets: the main building block of WordNet. A number of 30K synsets are present in

Assamese WordNet

- Gloss: The concept or meaning of the given synset

Named entities are a collection of terms that has a unique concept. They are mainly the names of people, organization, places, festivals etc.

Assamese is the official language of the North-eastern state- Assam of India. It is spoken by nearly 15 million people. Assam shares an international border with Bhutan and Bangladesh. It is a computationally less aware language which belongs to the Indo-Aryan language family. But, recently some development is done for this language from Natural language processing perspective. Development of Assamese WordNet, Corpus, IR system is some of them.

This research paper aims to implement a rule-based Morphological Analyser for the Assamese language to be embedded as a plug-in to Assamese IR system. No such work implementing 22 morphotactic rules for Assamese language is defined before in previous works. We believe this would mark a great contribution to Assamese NLP area.

The road-map of the paper is as follows- Section 2 discusses some related work to stemming implemented in Indian languages, Section 3 describes the rule based stemmer for the Assamese language with the system architecture. Section 4 discusses the performance of the stemmer computing the hamming distance. The IR system performance is evaluated on performing stemming to the inflected terms and the results are reported in section 5 of this paper. The paper is summarized in Section 6.

2 Background work

This section gives us an overview of stemmers developed in Indian Languages. For the English language, the most commonly used stemming algorithm is the Porter stemming algorithm (Willett, 2006) which followed a rule-based approach. The Indian language (Ramanathan and Rao, 2003; Aswani and Gaizauskas, 2010; Mahmud et al., 2014; Kumar and Rana, 2010; Majgaonker and Siddiqui, 2010; Prajitha et al., 2013; Thangarasu and Manavalan, 2013; Kumar et al., 2011) in which stemmer is developed along with the approaches used and accuracies derived is mentioned in Table 1

Table 1: Indian language stemmer

Language	Approache	Correctness
Hindi	Rule-based	Accuracy 88%
Gujarati	Dictionary and Rule-based	Precision 83%
Bengali	Rule-based	Accuracy 88%
Punjabi	Brute-force	Accuracy 81.27%
Marathi	Hybrid (Rule-based + suffix stripping +statistical)	Precision 82.50%
Malayalam	Finite state machines	Accuracy 94.76%
Tamil	Light Stemmer (preserves word meaning)	Accuracy 83.28%
Telugu	Unsupervised approach	Accuracy 85.40%

3 Development of Assamese stemmer

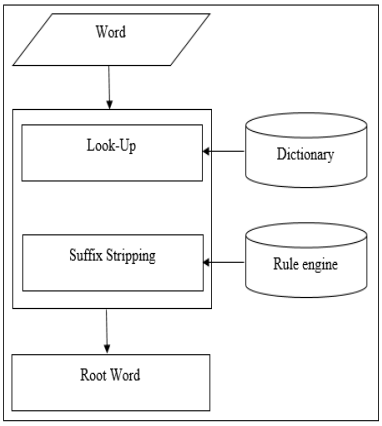


Figure 1: Assamese stemmer system diagram

Assamese words in a text take a series of suffixes in a sequential manner. For developing a rule-based stemmer, our first intention is to determine the sequence of various suffixes a word can occur in a text. Some of them were collected by consulting with the Linguistic scholars of GU NLP team. They may be divided into eight possible suffix categories such as:

- Plural- "সকল", "মখা", "বিলাক", "সোপা"
- Case markers- "ক", "ত", "লৈ", "এ", "ৰ"
- Pleonastic suffix- "হে", "চোন", "নে", “গৈ”
- Definitive- "জন", "জনী", "জনা", "খন", "টো"
- In-definitive- "কেইজন", "কেইজনী”
- Verbal- "াই", "াইছ", "াইছা", "াইছিল"
- Kinship noun- “য়েক”
- Extra- “দৰে”,"য়ে", "নো", "কৈ"

Step1: Dictionary Lookup

Assamese dictionary of size about 2 lakh root words is prepared by our Linguists from Assamese WordNet and Named Entities. Our module first looks at the dictionary table to determine if the words are already in the root form. If true then, they proceed to step 3 else step 2. This approach eliminates the type of error like word saybাহিৰ (out), which is a root word even though case marker suffix is present. If the dictionary is not reviewed in the beginning, than stemmer would remove the suffix of the word which would lead to overstemming. Moreover, the same would be the case for Named Entities like place name: তেজপুৰ (place name). Also, in some cases the term may have been derived from the antonym of the root word. Here, we consider the antonyms as the root word to retrieve sense oriented searched results from an IR. As for example the word in Assamese language- অশুভ (not pleasant) indicates different sense compared to the root form শুভ (pleasant). On knowing the root words at beginning will avoid understemming and overstemming roles of the stemmer and can retrieve sense oriented or meaningful results from the Information retrieval system on firing the query as required by the user.

Step2: Suffix pruning

If the first step fails than step 2 is executed. In this phase, the rule engine generates a list of suffixes in a proper manner that may be attached to the root based on the stemming rules already incorporated in the engine. The generated suffix list must abide by the morphotactic rules for Assamese. A Java program was developed to run this step.

Some rules for stemming are mentioned below in a tabular form: Here, authors have defined 22 rules for stemming Assamese words. Some of the rules are followed by the Assamese grammar book Assamiya Vyakaran by Hem Chandra Baruwa, 2003. As for example, the inflected word is মানুহকেইজনমান. The generated suffix list for the word is মান, কেইজনমান. The list is now transformed to non-increasing order and at first the top one (here কেইজনমান) is being tried to be matched with the already incorporated rules in the engine. Here, the rule *root+ indefinitive + plural* is mapped and the word is stemmed. Here, at the first phase of developing the stemmer, only nouns and verbs are taken into consideration.

Table 2: Morphotactic Rules of Assamese Stemmer

Suffix Type	**Assamese Notation**
Root+casemarker	মানুহ+ৰ
Root+definitive	মানুহ+জন
Root+pleonastic	কৰ+চোন
Root+indefinitive	মানুহ+কেইজন
Root+plural	মানুহ+বোৰ
Root+verb	কৰ+ িছিল
Root+extra	খৰ+কৈ
Root+kinshipnoun	ককা+য়েক
Root+case+extra	মানুহ+ৰ+দৰে
Root+plural+ case+pleo	মানুহ+বোৰ+ ক+হে
Root+Plural+ Case marker	মানুহ+বোৰ+ৰ
Root+Plural+ pleonastic	মানুহ+বোৰ+হে
Root+Definitive+ case	মানুহ+জন+ৰ
Root+Definitive+ pleonastic	মানুহ+জন+হে
Root+Indefinitive+ Plural	মানুহ+কেইজন+ মান
Root+Verb+ pleonastic	পঢ়+ ইলে +গৈ
Root+Casemarker + pleonastic	কৰ+ক+চোন
Root+kinshipnoun+ indefinitive+plural+pleo	নাতিনী+য়েক+ কৈইজনী+মান+হে
Root+pleonastic+ pleonastic	কৰ+গৈ+চোন
Root+plural+ definitive	গৰু+জাক+টো
Root+verb+extra	কৰ+ ি+য়ে
Root+case+ plural+definitive	গৰু+ৰ+জাক+টো

Step3: Exit

4 Performance Analysis

We have implemented both look-up based and rule-based approaches for Assamese stemmer. We evaluated the stemmed output with the morphological root words of Assamese WordNet and Named Entities by computing Hamming distance. It is the number of different position of the bits between two equal length strings. A hamming distance of 0 means the two strings are equal in both position of the character bits and weight. As for example one of the correctly stemmed output is:

Inflected term: মানুহজন

Assamese Stemmer output: w1= মানুহ

Assamese WordNet (ID: 196) w2= মানুহ

Hamming distance=d(w1,w2) =0

Some of the result statistics found while analyzing the performance of the stemmer is shown in a tabular form below:

Table 3: Statistics of stemmer performance

Correctly stemmed	85%
Incorrectly stemmed	15%

5 Stemmer in Assamese IR

Information Retrieval system retrieves relevant and sense oriented information to a user based on the query. Assamese NLP aims to develop a monolingual search engine which will help the web users to retrieve information in ones own native language say Assamese. Only a few (2-3) percent of people of Assamese community knows to speak, read or write English, so retrieving information in own language will be much benefited.

Assamese IR system is technically composed of two parts- Apache Solr & Nutch. Apache Solr is an open source search platform written in JAVA from Apache Lucene project. Some of the major features of Solr are- full text search, real time indexing, dynamic clustering etc. Apache Nutch is also a JAVA coded tool with the crawler feature. The crawler can be biased to fetch important relevant pages at first. We developed Assamese monolingual system considering Solr3.4 and Nutch1.4 as indexer and crawler respectively.

Stemming is an important plug-in of IR. Stemming is performed to an inflected word to avoid mismatches between words that share the same root word. Let us consider a simple example- if we are searching for a document entitled Ways to write a book and the user issues a query writing, than there will be no match with the title. But, if the query is stemmed before than the search system will stem the word writing to write and the retrieval will become easier and successful. Stemming is applied to both Query processing module and IR system module. Both at the indexing time and during processing of the query the stemmer module is added as plug-in to Assamese IR system. Here, we have analyzed the performance of IR system based on two categories-

- IR performance without stemming
- IR performance with stemming

The above two techniques is evaluated with p@k (Precision at k) metric. For modern IR system, recall is meaningless as many numbers of queries retrieves many relevant documents (as of now web-scale) and no user will go through all of them. Here, k=10 and p@10 indicates the number of relevant result of search result page which includes top-ten results of a query. To evaluate our system we tokenized some of the words from Assamese Corpus (size=1.5 million words) developed by (Sarma et al., 2012). The figure 2 indicates higher AP (Average Precision) values of the IR system when performed stemming than without stemming. To evaluate the system we consider 10 Assamese queries Q1 to Q10 those are অমৃতসৰৰ স্বৰ্ণ মন্দিৰ, তিৰুপতি, নালন্দা বিশ্ববিদ্যালয়, কাজিৰঙা ৰাষ্ট্রীয় উদ্যানলৈ, বিহুত, অসমৰ, মাজুলী, তাজমহলত, গড়ঁবোৰ ,

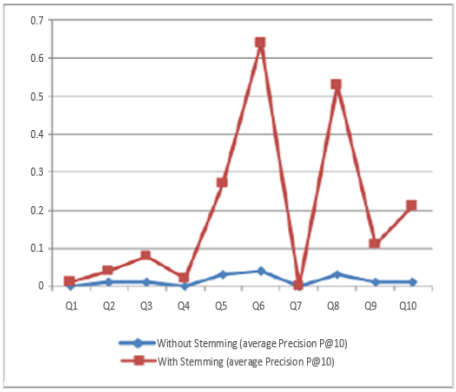


Figure 2: Assamese IR performance: with and without stemming

ব্রহ্মপুত্র নদীত. As the stemmed term indicates larger concept than the original term appears in the document, the stemming increases the number of retrieved relevant documents.

6 Conclusions

The performance of the Assamese stemmer mentioned in this paper shows that it attains a state of art accuracy as a stand along system as well as a component of Information Retrieval system. The proposed technique is Dictionary Look-up and Rule-based approach for this Indo-Aryan language with an acceptable accuracy of 85% and 22 defined morphotactic rules.Increasing the dictionary size will result in more increasing accuracy.

Assamese stemmer is the basic language resource and is used in many applications in the field of Text mining and NLP like IR, MT, Document Classification etc. The accuracy of the stemmer can be improved by defining more stemming rules and increasing the dictionary size with more root words. Moreover, as the IR performance on performing stemming to the inflected terms indicates an overwhelming result, thus stemmer is an important resource for Assamese NLP.

References

Shikhar Kr. Sarma, Moromi Gogoi, Utpal Saikia and Rakesh Medhi 2010. *Foundation and structure of Developing Assamese WordNet*. In Proceedings of 5th International Conference of the Global WordNet Association.

P. Willett 2006. *The Porter stemming algorithm: then and now*. Program: electronic library and information systems, 40 (3).

A. Ramanathan and D. D. Rao 2003. *A Lightweight Stemmer for Hindi*. Workshop on Computational Linguistics for South-Asian Languages, EACL.

N Aswani, R Gaizauskas 2010. *Developing Morphological Analysers for South Asian languages. Experimenting with the Hindi and Gujarati languages*. In Proceedings of the seventh conference on International Language resources and evaluation, Malta.

Md. Redowan Mahmud, Mahbuba Afrin, Md. Abdur Razzaque, Ellis Miller and Joel Iwashige 2014. *A rule based Bengali stemmer*. In Proceedings of the ICACCI.

D Kumar and P Rana 2010. *Design and development of a stemmer for punjabi*. International Journal of Computer Application 11(12) (December 2010).

M. M. Majgaonker and T. J. Siddiqui 2010. *Discovering suffixes- A case study for Marathi language* . International Journal on Computer science and engineering.

U. Prajitha, C. Sreejith and P.C Reghu Raj 2013. *LALITHA: A light Weight Malayalam Stemmer using Suffix Stripping method*. In Proceedings of the ICCC.

M Thangarasu and Dr R Manavalan 2013. *Stemmers for Tamil Language: Performance Analysis*. International Journal Of Computer Science & Engineering Technology, Vol 4.

A . P. Siva Kumar, P. Premchand and A Govardhan 2013. *TelStem: An Unsupervised Telugu Stemmer with Heuristic Improvements and Normalized Signatures*. Journal of Computational Linguistics Research Vol 2 number 1.

Shikhar Kr. Sarma, Himadri Bharali, Ambeswar Gogoi, Ratul Ch. Deka, Anup Kr. Barman 2012. *A Structured Approach for Building Assamese Corpus: Insights Applications and Challenges*. In Proceedings of the 10th Workshop on Asian Language Resources, pp. 21-28, December.

Synthetic, yet natural: Properties of WordNet random walk corpora and the impact of rare words on embedding performance

Filip Klubička[1,3] **Alfredo Maldonado**[2,3] **Abhijit Mahalunkar**[1] **John D. Kelleher**[1,3]
[1]Technological University Dublin, Ireland
[2]Trinity College Dublin, Ireland
[3]ADAPT Centre, Dublin, Ireland
{filip.klubicka,alfredo.maldonado,john.kelleher}@adaptcentre.ie
abhijit.mahalunkar@mydit.ie

Abstract

Creating word embeddings that reflect semantic relationships encoded in lexical knowledge resources is an open challenge. One approach is to use a random walk over a knowledge graph to generate a pseudo-corpus and use this corpus to train embeddings. However, the effect of the shape of the knowledge graph on the generated pseudo-corpora, and on the resulting word embeddings, has not been studied. To explore this, we use English WordNet, constrained to the taxonomic (tree-like) portion of the graph, as a case study. We investigate the properties of the generated pseudo-corpora, and their impact on the resulting embeddings. We find that the distributions in the psuedo-corpora exhibit properties found in natural corpora, such as Zipf's and Heaps' law, and also observe that the proportion of rare words in a pseudo-corpus affects the performance of its embeddings on word similarity.

1 Introduction

A word embedding model maps the words in a vocabulary to dense low-dimensional vectors, by inferring the relative position of each word in a shared multidimensional semantic space from its context of use in a corpus (Mikolov et al., 2013a; Mikolov et al., 2013b). This approach is founded on the distributional hypothesis (Harris, 1954), which states that words which occur in the same contexts tend to have similar meanings. Such word embeddings are created by training a neural network language model on natural language corpora.

While such embeddings have been shown to perform well on semantic relatedness benchmarks (Baroni et al., 2014; Camacho-Collados and Pilehvar, 2018), training on a natural corpus only models one type of semantic relation between words: thematic (i.e. syntagmatic). On the flip side, taxonomic (i.e. paradigmatic) relations are not explicitly contained in natural language corpora, and as such are not included in those embeddings (Kacmajor and Kelleher, 2019). In fact, research suggests that the best measures of taxonomic similarity and thematic relatedness are different in distributional space (Asr et al., 2018). Furthermore, there are many other kinds of relationships between words and concepts that can be found in knowledge engineered resources, such as knowledge bases, ontologies, taxonomies and other semantic networks.

Modelling these relations is an important task in building AI with comprehensive natural language understanding abilities, and there have been many efforts to bring knowledge graphs into an embedding space (see Section 2 for details). One such approach is the WordNet random walk algorithm (Goikoetxea et al., 2015): by randomly walking the WordNet knowledge graph and choosing words from each synset that has been traversed, a pseudo-corpus is generated and used for training word embeddings. The reasoning is that the distributional hypothesis should also apply in this scenario, in the sense that co-occurrence within local contexts in the pseudo-corpus will reflect the connections between words connected in the WordNet graph.

Naturally, the shape of the underlying knowledge graph (in terms of node connectivity: i.e. tree, fully-connected, radial etc.) affects the properties of a pseudo-corpus generated via a random walk over the graph. Developing a better understanding of the relationship between the shape of a knowledge graph, the properties of the resulting pseudo-corpora, and the properties of the resulting embeddings, has the potential to inform how the walk over a given knowledge graph should be tailored to improve embedding performance.

In this paper we provide an analysis of some

of the properties of pseudo-corpora generated using the random walk method, and examine the impact of these properties on embedding performance. We base this analysis on the WordNet taxonomy, because (a) WordNet is one of the most-popular taxonomies in use, and (b) in general, the WordNet taxonomy has a well-understood shape (tree-like) which informs the analysis of our results. We find that the pseudo-corpora synthesized from the WordNet taxonomy are not as artificial as one might expect - they exhibit properties and regularities also found in natural corpora, following natural language laws such as Heaps' law and Zipf's law. Consequently, we hypothesise that word embeddings trained on such corpora might face the same limitations as those trained on natural corpora would. We explore this notion on the case study of rare (i.e. infrequent) words, which are a known problem for word embeddings (Khodak et al., 2018; Pilehvar and Collier, 2017; Pilehvar et al., 2018).

2 Related work

Research on building embeddings from knowledge resources such as WordNet (Fellbaum, 1998), can be broadly categorised into three approaches: i) enrichment, ii) specialisation, and iii) direct learning from knowledge resources.

Both enrichment and specialisation modify precomputed, corpus-based word embeddings with information from a knowledge resource to either augment them (enrichment) or to fit them onto the specific semantic relation described by that knowledge resource (specialisation). Retrofitting (Faruqui et al., 2015) is an example of enrichment: it modifies corpus-based embeddings by reducing the distance between words that are directly linked in resources like WordNet, MeSH (Yu et al., 2016) and ConceptNet (Speer and Havasi, 2012). In our own recent related work, we have explored the impact of corpus size on vector enrichment (Maldonado et al., 2019).

On the other hand, examples of the specialisation approach are PARAGRAM (Wieting et al., 2015), Attract-Repel (Mrkšić et al., 2016), Hypervec (Nguyen et al., 2017) and the work of Nguyen et al. (2016) and Mrkšić et al. (2017) on synonyms and antonyms. Vulić et al. (2018) and Ponti et al. (2018) introduce global specialisation models where vectors for words that are missing in the knowledge resource are also updated.

More related to our work are the approaches to learn directly from knowledge resources. Examples include building non-distributional sparse word vectors from lexical resources (Faruqui and Dyer, 2015), building Poincaré embeddings that represent the structure of the WordNet taxonomy (Nickel and Kiela, 2017) and building embeddings that encode all semantic relationships expressed in a biomedical ontology within a single vector space (Cohen and Widdows, 2017). The latter two methods encode the semantic structure of a knowledge resource in a deterministic manner, while Agirre et al. (2010) follow a stochastic approach based on Personalised PageRank: they compute the probability of reaching a synset from a target word, following a random-walk on a given WordNet relation. Instead of computing random-walk probabilities, Goikoetxea et al. (2015) use an off-the-shelf implementation of the word2vec Skip-Gram algorithm to train embeddings on WordNet random walk pseudo-corpora, changing neither the embedding algorithm nor the objective function[1]. The resulting embeddings encode WordNet taxonomic information rather than natural word co-occurence. An advantage of the embeddings produced by this method is that they can be used as is or can be combined with real-corpus embeddings in order to accomplish enrichment or specialisation (Goikoetxea et al., 2016).

Previous work has analysed semantic properties of word embeddings generated by random walk. Goikoetxea et al. (2016), for example, found WordNet random-walk embeddings to outperform corpus-based word embeddings on the strict semantic similarity (taxonomic similarity) SimLex-999 benchmark (Hill et al., 2015), confirming that they encode taxonomic information better than real-corpus word embeddings. Additionally, other researchers have explored different varieties of the random walk algorithm. Most notably, Simov et al. (2017a) drastically enrich the graph structure by using all available relationships between WordNet synsets, while inferring and adding others from outside resources (Simov et al., 2015; Simov et al., 2017b). However, to the best of our knowledge, there has been no work on analysing the properties of the corpora generated by random-walk processes. In particular, there has been no work on comparing their statistical properties with those of natural corpora.

[1] http://ixa2.si.ehu.eus/ukb/

3 Pseudo-corpora

3.1 Random walk pseudo corpus generation

Our pseudo-corpus generation process is inspired by the work of Goikoetxea et al. (2015). They performed random walks over the full WordNet knowledge base as an undirected graph of interlinked synsets. Their method first chooses a synset at random from the set of all synsets, and then performs a random walk starting from it. They also use a predefined dampening parameter (α) to determine when to stop the walk, so that at each step the walk might move on to a neighbouring synset with probability (α), or might terminate with the probability ($1 - \alpha$). It is usually set to 0.85. Each time the random walk reaches a synset, a lemma belonging to the synset is emitted, using the probabilities in the inverse dictionary. Once the random walk terminates, the sequence of emitted words forms a pseudo-sentence of the pseudo-corpus. The process repeats until a given number of sentences have been generated.

Our pseudo-corpus generation algorithm is similar, however, there are a number of important differences. First, Goikoetxea et al. make use of all available connections in the graph, whereas we only traverse the hypernym/hyponym relationship and ignore non-taxonomic relationship types such as gloss, meronym and antonym relations. This effectively allows us to exclusively traverse WordNet's taxonomic graph, which lets us embed only taxonomic relations. More importantly, this decision is motivated by the fact that we wish to use WordNet's taxonomic graph as a case study of how the underlying structure of a knowledge graph affects the properties of a generated pseudo-corpus. Constraining the random walk to just the taxonomy reduces the graph to a tree shape, which provides an intuitive and transparent understanding of its structure. This restriction to the taxonomic components of the graph has two important implications: (i) it permits us to consider the graph as directed (hypernym/hyponym→up/down), and (ii) it makes the graph quite sparse. The other two significant differences between our algorithm and Goikoetxea et al. are derived from these two implications and are implemented as two new hyperparameters on the algorithm: a directionality and a minimum sentence length parameter.

The directionality parameter constrains the permissible directions that the walk can proceed along as it traverses the tree structure (e.g., only up, only down, both). This hyperparameter permits us to explore the relationship between variations in the random walk algorithm and the number of rare words in the generated corpus (see Subsection 3.2). The minimum sentence length parameter enables us to filter the sentences generated by the random walk algorithm by rejecting any sentence that is shorter than a prespecified length n. The decision to exploit only the taxonomic relations makes the graph quite sparse: a lot of nodes end up disconnected, as some synsets are not part of the WordNet taxonomy, but are connected to it only via non-taxonomic relations. Given that we allow our algorithm to start the random walk anywhere in WordNet, it often begins, and ends, its walk at a disconnected node, which results in a lot of one-word sentences in the synthesized pseudo-corpus. To remedy this, the minimal sentence length hyperparameter disallows generating sentences with only one word, or sentences shorter than the pre-specified value. Section 3.2 contains details on this and other hyperparameters.

In our algorithm[2], the random walk starts at a random synset and chooses a lemma corresponding to that synset based on the probabilities provided by WordNet's inverse mapping from synsets to lemmas. Once the lemma has been emitted, we check if the synset has any hypernym and/or hyponym connections assigned to it (depending on the direction constraint). If it does, we choose one at random with equal probability and continue the walk towards it, choosing a new lemma from the new synset. This process continues until one of two conditions are met: (a) there are no more connections to take, or (b) the process is terminated according to the dampening factor (α). We then restart the process and create a new pseudo-sentence, until we have generated the required number of sentences. Some examples of pseudo-sentences produced by our system:

measure musical notation tonality minor mode

Dutch-processed cocoa powder chocolate milk

[2]Although Goikoetxea et al. provide an implementation of their random walk algorithm, due to the differences outlined above and the special use cases for our research, we have decided to reimplement it in Python and use NLTK's version of WordNet (Bird and Loper, 2004). Our code and generated datasets are being made available online.
`https://github.com/GreenParachute/wordnet-randomwalk-python`

size	direction	min.sent.len.	token count	avg.sent.len.	%same sents	vocabulary	%rare words
500k	up	2w/s	3,515,524	7.03	18.5	64,257	67.35
500k	down	2w/s	1,475,336	2.95	68.56	55,508	53.35
500k	both	2w/s	2,401,498	4.80	20.06	67,049	39.86
500k	up	3w/s	4,011,247	8.02	17.06	63,923	66.48
500k	down	3w/s	2,097,641	4.20	71.01	46,701	52.33
500k	both	3w/s	2,822,171	5.64	12.22	67,353	33.30
1m	up	2w/s	7,041,365	7.04	27.93	66,840	41.84
1m	down	2w/s	2,947,657	2.95	78.57	59,894	40.81
1m	both	2w/s	4,802,354	4.80	28.49	67,647	15.82
1m	up	3w/s	8,032,165	8.03	26.31	66,401	40.52
1m	down	3w/s	4,195,458	4.20	79.46	51,310	43.91
1m	both	3w/s	5,636,469	5.64	18.88	67,683	11.31
2m	up	2w/s	14,079,962	7.04	39.56	67,587	19.32
2m	down	2w/s	5,898,583	2.95	85.91	63,089	30.03
2m	both	2w/s	9,602,490	4.80	37.66	67,756	3.88
2m	up	3w/s	16,061,599	8.03	37.65	67,081	18.20
2m	down	3w/s	8,389,396	4.19	85.92	55,314	35.99
2m	both	3w/s	11,274,757	5.64	26.99	67,757	2.34

Table 1: Statistics of generated random walk corpora

3.2 Pseudo-corpora properties

We controlled the generation of the pseudo-corpora using the following hyperparamters:

1. **Size**. We define corpus size in terms of the number of random restarts, i.e. number of pseudo-sentences generated. We generate pseudo-corpora of sizes 1k, 10k, 100k, 500k, 1m and 2m sentences.

2. **Direction**. As we are only walking the WordNet taxonomy, we define direction as allowing the walk to either only go up the hierarchy, down the hierarchy, or both ways.

3. **Minimum sentence length**. We impose a constraint on minimal sentence length and generate corpora with 2-word and 3-word minimum length sentences.

Combining all the hyperparameters yielded a total of 36 pseudo-corpora of varying sizes, directions and minimal sentence lengths. However, due to space constraints and the fact that the smaller corpora have shown to be too variable to make confident inferences, we only present data and analyses of the three largest corpus groups.

Note that we are not necessarily looking for a combination of hyperparameters that performs best on evaluation tasks, rather we use them as a tool to generate pseudo-corpora with different properties. Following that, for each pseudo-corpus we measure the following statistical properties: total number of tokens, average sentence length (average tokens per sentence), percentage of identical sentences, size of vocabulary, and percentage of rare words in the vocabulary (see Table 1).

From Table 1 it is visible that the number of tokens grows with the size in terms of number of restarts. Interestingly, however, although the average sentence length correlates with absolute number of tokens, it stays constant regardless of the number of restarts, all other things being equal. For example, the average sentence length for the 500k.both.2w/s is 4.8, and the average sentence length for the 2m.both.2w/s corpus is also 4.8 tokens per sentence. This holds for any other analogous combination, further supporting the claim that the underlying graph structure of the corpus is the source of certain word distributions and regularities present in the corpus.

Furthermore, the number of tokens also varies depending on the other two hyperparameters: directionality and minimum sentence length. For example, both average sentence length and absolute number of tokens are sensitive to the direction hyperparameter. Regardless of the number of restarts, corpora generated by only walking up the taxonomy create the longest sentences on average and have the largest number of tokens, while only walking down the taxonomy generates the shortest sentences and the lowest number of tokens.

Such behaviour is a direct consequence of the WordNet taxonomy's structure and the distribution of edges between nodes. The taxonomy is a tree, and as such the vast majority of its nodes are leaf nodes positioned near the bottom. Consequently, each time the random walk restarts, it is far more likely to start somewhere near the bottom

of the taxonomy, rather than at the top. Therefore, if the walk can only go up, on the majority of restarts it will be able to traverse the taxonomy for a large number of nodes before either α kicks in, or it reaches the top and has nowhere to go. Conversely, if the walk is constrained to only move down the taxonomy then on most restarts the walk will only be able to take a few steps before it has nowhere to go and is forced to terminate. Finally, the reason that allowing both directions in the walk generates shorter sentences than going only up is because almost by definition, a synset can have only 1 hypernym, but several hyponyms, so it is more likely to choose a node that is directed downward. In doing so, it behaves more similarly to the algorithm that only goes down and generates shorter sentences than the upward one.

Naturally, the larger the corpus (both in terms of random restarts and tokens), the larger the vocabulary. When comparing the impact of the direction hyperparameter, going down produces corpora with the least WordNet coverage, and going in both directions yields the highest coverage. Again, this is a direct consequence of the structure of the underlying graph. Due to the nature of the random walk going downward the paths are short and there is not much variety, so the vocabulary coverage depends exclusively on the position of the random restarts and is thus significantly lower.

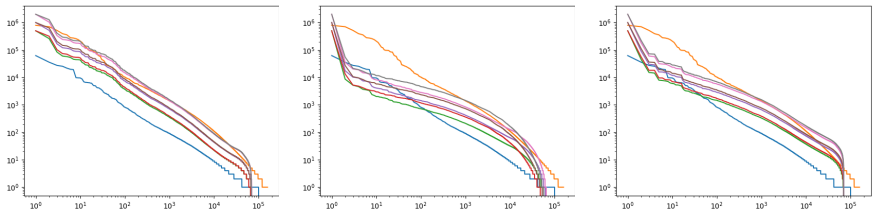
Finally, we look at rare words in the generated corpora. We define a word type as rare if it appears in the corpus less than 10 times. We calculate the percentage of rare words (types/lexemes) versus the full vocabulary. Overall, the percentage of rare words gets smaller as corpus size increases, as more and more words appear over 10 times. However, the hyperparameters seem to have varying effects on this value. For the 500k corpora, the highest percentage of rare words are in corpora generated by only going up, while the lowest percentage are in corpora generated when the walk is allowed to proceed in both directions. All percentages are slightly lower for corpora with a 3-word sentence minimum when compared to corpora with a 2-word sentence minimum. Moving up by one size, corpora with 1m sentences seem to be at a tipping point. Looking at corpora with a 2-word sentence minimum, they follow the percentage of rare words ordering as the 500k corpora of up-down-both, but just barely, and if we look at 3-word sentence minimum corpora the top two rankings switch places. This switch is also apparent in all the 2m-sentence corpora. The percentage of rare words drops off much quicker for corpora generated by only going up compared with corpora generated by only going down. Consequently, even though the up direction generates corpora with the highest percentage of rare words in the smaller sizes, this percentage quickly drops as the corpus size increases. Hence, corpora of 2m sentences generated by only going up have a smaller percentage or rare words compared with the corpora generated by only going down. Likely this is a consequence of the much more drastic increase in absolute number of tokens between the two corpus varieties. The upward corpora consistently have roughly twice as many tokens as the downward corpora, given same number of sentences (i.e. restarts). Overall, the corpus with the smallest percentage of rare words, with only 2.34% rare words in the vocabulary, is the one generated with 2m restarts and allowing the walk to move in both directions. Likely, this is because it is generated from the graph with the most connections, and hence an overall higher coverage; at the size of 2 million sentences, it would have traversed most of the taxonomy several times over, thereby significantly reducing the number of rare words.

3.3 Scaling Linguistic Laws of Natural Languages

The properties described in Subsection 3.2 are a consequence of the corpora being artificially generated from a WordNet's taxonomic graph structure and from the way the random walk algorithm has traversed this graph. However, inspecting word distributions in the corpus showed interesting regularities that seem to indicate similarities with natural corpora. The regularities in the frequency of text constituents have been summarized in the form of *linguistic laws* (Altmann and Gerlach, 2016; Gerlach and Altmann, 2014). Linguistic laws provide insights on the mechanisms of text (language, thought) production. One of the best known linguistic laws is *Zipf's Law* (Zipf, 1949). It states that the frequency, F of the r^{th} most frequent word (i.e. the fraction of times it occurs in a corpus) scales as follows:

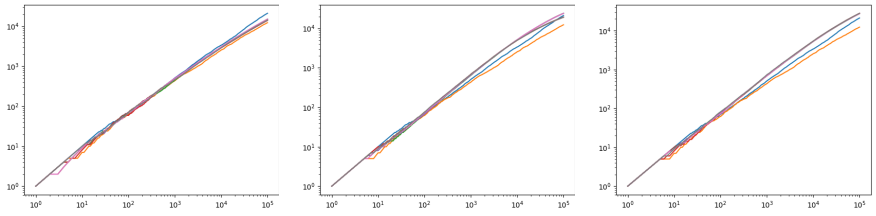
$$F_r \propto r^{-\lambda}, \forall\, r \gg 1 \tag{1}$$

Zipf's Law is approximated by a Zipfian distribution which is related to discrete power law prob-



(a) Direction: up (b) Direction: down (c) Direction: both

Figure 1: Zipf distributions of two natural corpora (shaded blue and orange) and all our pseudo-corpora. We group the three different directions taken by the random walk.



(a) Direction: up (b) Direction: down (c) Direction: both

Figure 2: Heaps' law of two natural corpora (shaded blue and orange) and all our pseudo-corpora. We group the three different directions taken by the random walk.

ability distributions. Here, λ is the scaling exponent and is ≈ 1.0 for natural languages.

Heaps' law is another scaling property and shows how vocabulary grows with text size. Consider n be the length of a text and $v(n)$ be its vocabulary size. Then Heaps' law is formulated as:

$$v(n) \propto n^{\beta}, \forall\ n \gg 1 \tag{2}$$

where the exponent for the Heaps' law is found to be $0 < \beta < 1$ for natural languages.

Here we investigate whether our pseudo-corpora uphold these laws, so as to confirm their naturalness. We employed *Kolmogrov-Smirnov (KS) Distance* to compare the pseudo-corpora against the natural corpora. In our case, we check *KS* distance between the natural and pseudo-corpora for both Zipf's and Heap's law.

Our analysis revealed that the KS distance between our 2 natural corpora is consistent with the distance between the natural and synthetic corpora, indicating consistent variations for both Zipf's and Heaps' law. For both our natural and synthetic corpora, $\lambda \approx 1.1$ and $\beta \approx 0.9$. In this case, it is fair to assume that our pseudo-corpora maintain these properties of natural language. This finding is important because it indicates that embeddings trained on pseudo-corpora will have similar shortcomings to embeddings trained on natural text. For example, past research has highlighted difficulties of learning good embeddings for rare words in natural corpora (Lazaridou et al., 2017; Pilehvar and Collier, 2017).

In addition, in Figures 1 and 2 respectively we also plot Zipf's law and Heaps' law for all our pseudo-corpora, alongside two natural corpora (the Brown corpus (Francis, 1964) and a small chunk of wikitext-2 (Merity et al., 2016)). Though our test of KS distance confirms that all the pseudo-corpora follow Heaps' law and a Zipfian distribution, it is still interesting to note the slight variations in the Zipf curves. Uniformly, the 'up' pseudo-corpora most closely match the natural corpora, the 'down' pseudo-corpora do so to a much lesser degree, and 'both' fall somewhere in the middle. This indicates that the directionality hyperparameter also enables us to simulate slightly different underlying graph structures, in a sense pruning the original graph from the per-

spective of the random walk. These figures reinforce the fact that the nature of the random walk algorithm, the structure of the graph and the paths that are walked have an impact on the resulting pseudo-corpus.

Motivated by these findings, in the next section we will evaluate the performance of a set of embeddings trained on a number of pseudo-corpora and consider the effect of rare words on the performance of these embeddings.

4 Evaluation and analysis

After generating all the corpora, we trained word embeddings and evaluated their performance on the task of word similarity.

4.1 Training

We trained our embeddings using the 2017 version of Pytorch SGNS, a publicly available implementation[3] of the skip-gram with negative sampling (SGNS) algorithm, introduced by Mikolov et al. (2013a). We only made minor data-handling optimisations – the objective function is not modified in any way.

The vectors were computed with SGNS using a window of five words on both sides of a sliding focus word, without crossing sentence boundaries. Twenty words were randomly selected from the vocabulary based on their frequency as part of the negative sampling step of the training. The frequencies in this weighting were smoothed by raising them to the power of $\frac{3}{4}$ before dividing by the total. All vectors produced by the SGNS system had 300 dimensions and trained for 30 epochs. We train separate embeddings on each combination of the three hyperparameters and report scores from the best performing epoch.

4.2 Evaluation

We evaluate the performance of our embeddings on five different benchmarks: the similarity-focused SimLex-999 (Hill et al., 2015); the English test set from the SemEval 2017 Task 2 challenge (Camacho-Collados et al., 2017) (henceforth referred to as SemEval-17); the relatedness dataset WS-353 (Finkelstein et al., 2002); and the Princeton evocation benchmark (Boyd-Graber et al., 2006). However, we suspect none of these benchmarks are ideally suited to the task at hand, as they are all based on human judgements on an often broad idea of word similarity, yet we are specifically modelling taxonomic relations. For this reason, in addition to the above benchmarks, we develop a novel test set, inspired by the work of (Pedersen et al., 2004)[4]: we take the word pairs from SimLex, and replace the human similarity judgements with a WordNet similarity measure (based on the distances in the graph). We refer to this benchmark as WordNet-paths. This serves as a sanity check and an appropriate test set for our taxonomic embedding model.

As is common practice, we evaluate our model by computing a Spearman correlation score between the cosine similarity of the word vectors from our model and the scores in our benchmarks. Table 2 presents the results alongside the percentage of rare words in a given benchmark.

4.3 Discussion

The aim of this experiment is not to beat state of the art scores, but rather to investigate different WordNet taxonomic structures generated by the random walk hyperparameters and their impact on rare words and performance of word embeddings trained on the pseudo-corpora. We hypothesise that the direction constraint of the random walk has an effect on the percentage of rare words in the resulting corpus, which in turn affect the performance of the trained embeddings.

With that in mind, we look at Table 2. Our highest correlation scores come from the WordNet-paths benchmark, which is not surprising as this benchmark reflects most accurately what our models have learned – taxonomic relations in WordNet. The highest overall score comes from the largest corpus, but looking at the different groups of different-sized corpora, the best performing model is always the one allowing both directions in the random walk, which generates the lowest percentage of rare words. Our hypothesis is clearly confirmed on this benchmark, where all the best scores come from corpora with the lowest percentage of rare words, while the lowest scores come from corpora with the highest percentage of rare words in two out of six cases.

In contrast with WordNet-paths, our worst performance is achieved on the evocation benchmark. This is to be expected, as the evocation benchmark models a relationship between words that is very

[3] https://github.com/theeluwin/pytorch-sgns

[4] http://wn-similarity.sourceforge.net

corpus	simlex %rare	simlex score	ws353 %rare	ws353 score	semeval %rare	semeval score	evoc %rare	evoc score	wn-paths %rare	wn-paths score
500k-up-2w/s	2.63	39.03	8.01	39.24	11.81	37.23	5.26	7.93	2.63	52.89
500k-down-2w/s	2.53	19.22	6.86	21.23	10.47	20.46	3.72	4.46	2.53	41.86
500k-both-2w/s	1.14	32.56	2.97	42.76	4.83	38.12	1.31	9.87	1.14	56.31
500k-up-3w/s	2.92	37.07	7.09	34.65	11.60	35.70	4.71	8.61	2.92	50.60
500k-down-3w/s	2.97	31.26	8.70	33.34	10.06	27.51	5.26	4.13	2.97	49.12
500k-both-3w/s	1.04	34.84	2.75	45.53	4.72	40.36	1.10	10.61	1.04	57.00
1m-up-2w/s	1.24	41.73	3.20	43.34	5.85	39.56	2.08	8.61	1.24	53.44
1m-down-2w/s	1.09	30.46	3.43	41.69	6.26	35.09	2.08	6.90	1.09	47.56
1m-both-2w/s	0.50	40.55	0.92	48.25	1.75	40.93	0.44	11.14	0.50	57.60
1m-up-3w/s	1.19	42.28	2.75	39.75	5.85	40.51	2.19	9.75	1.19	54.15
1m-down-3w/s	1.93	36.37	5.03	42.65	8.11	36.19	4.05	5.48	1.93	51.15
1m-both-3w/s	0.35	42.13	0.69	46.59	1.33	39.16	0.33	10.93	0.35	57.73
2m-up-2w/s	0.59	42.58	1.14	44.38	2.77	39.61	0.77	8.63	0.59	53.52
2m-down-2w/s	0.69	34.87	1.14	41.79	4.00	36.75	0.99	5.62	0.69	47.67
2m-both-2w/s	0.15	43.28	0.46	47.03	0.41	40.48	0.22	10.95	0.15	58.00
2m-up-3w/s	0.50	43.40	1.14	43.97	2.46	39.71	0.77	9.65	0.50	54.01
2m-down-3w/s	1.04	36.80	3.43	44.29	5.44	35.17	2.41	4.85	1.04	49.47
2m-both-3w/s	0.05	43.28	0.46	47.51	0.31	40.35	0.22	11.14	0.05	56.55

Table 2: Results for all embeddings trained on various corpora, showing Spearman correlation scores for best epoch per corpus trained on, as well as the percentage of rare words in a given benchmark. Cells shaded green represent the lowest percentage of rare words and the highest Spearman score obtained in the given group of embeddings on a given benchmark. Cells shaded red represent the highest percentage of rare words and the lowest Spearman score on the given group.

different in nature from the purely taxonomic relationship that we model here. This, together with the fact that our best correlation scores come from the WordNet paths benchmark, confirms that our embeddings do indeed reflect a purely taxonomic understanding of words. Yet in spite of the correlation scores being so low, our hypothesis holds here as well – in each group of comparable embeddings, the highest score comes from pseudo-corpora that traversed both directions, and generated the least rare words. The lowest scores stem from corpora with the highest percentage of rare words in five out of six cases.

As expected, we achieve much higher correlations scores on the remaining three benchmarks. Though the highest scores are achieved on WS-353, the overall performances between benchmarks are comparable insofar as they all model word similarity and relatedness. Our hypothesis holds just as consistently when examining the results on SemEval-17 and WS-353, where five out of six times and six out of six times respectively, the best performing model stems from a corpus that yields the lowest percentage of rare words, while the inverse holds four out of six times.

SimLex-999 seems to be somewhat of an outlier among these benchmarks. This is peculiar because, though it is more similarity-focused, the nature of the relations should not be that different from the one in WS-353 and SemEval-17. Our hypothesis still holds in the larger corpora (2m-2w/s, 2m-3w/s and 1m-3w/s), but in the smaller ones the lowest percentage of rare words is produced by the corpora allowing both directions, yet the highest scores actually come from the corpora produced going up. Given that the inconsistencies happen in the smaller corpora, it is possible that this is just an unlucky sample, or that the interplay of confounding factors has a stronger effect in the smaller corpora and negatively affects the performance of the corpora allowing both directions.

Overall, the distribution of best-worst models is fairly consistent across the 5 benchmarks. The best models are those going in both directions, and 2-word sentence minimum models are usually slightly outperformed by 3-word sentence models, though the differences are marginal. Unsurprisingly, models allowing both directions also consistently produce the lowest percentage of rare words. From this, it seems, also follows that more often than not those models have the best scores.

5 Conclusion

In our work we expand our understanding of the random walk algorithm, in terms of the relationship between the structure of the underlying knowledge graph, the properties of the pseudo-corpora generated from the graph, and the performance of the embeddings trained on these pseduo-

corpora. We use the WordNet taxonomy as a case study for our work. We find that all our pseudo-corpora resemble natural corpora at a statistical level. We attribute these properties to the underlying tree structure of the graph from which the pseudo-corpora are built. We also train word embeddings on these corpora to study the impact of these properties on the embedding performance on word similarity evaluation tasks. Our evaluations confirm a successful modelling of taxonomic relations, and on most benchmarks our data supports the hypothesis that the ratio of rare words in a pseudo-corpus affects embedding performance.

Understanding the properties of the pseudo-corpora generated from a knowledge graph structure can inform how the random walk should be designed and run for any graph. E.g. knowing that a tree-like graph structure results in pseudo-corpora exhibiting Zipfian properties is useful as it highlights the presence of rare words in the corpora. As the vocabulary of the lexical resource is finite, the problem of rare words within the generated pseudo-corpora can be addressed by ensuring that the pseudo-corpus is large enough so that even the relatively rare words appear frequently enough to learn adequate embeddings. This perspective helps in answering questions such as: *how large should a pseudo-corpus be?*

Though this might seem obvious, an important takeaway is that the properties of any pseudo-corpus generated from a knowledge graph will be affected by the properties of that graph–its structure and node connectivity will be reflected in the generated corpora, thus impacting the resulting embeddings. We do not claim that any graph structure will exhibit the exact properties we found, but rather that this kind of analysis should be considered when using a random walk algorithm.

As far as future work, there are several exciting avenues that can be explored. Most immediately, it would be important to examine whether the natural properties and rare word percentages in the pseudo-corpora hold when applied to more dense graph structures with connections beyond the WordNet taxonomy, such as WordNet gloss relations, polysemy, antonymy, meronymy, etc. Going further, one could apply the random walk to other knowledge bases to see if the regularities hold there as well. Additionally, combining pseudo-corpora from different knowledge bases, or simply enriching one graph with connections from another, adding additional thematic relations from other knowledge bases. Certainly, this would make the problem more complex, and would render the directionality parameter moot, as a lot of those connections do not have an inherent directionality to them. But this is definitely the next step in improving scores and increasing coverage.

Going even further, it would be beneficial to explore the application of both these taxonomic embeddings, as well as more complex knowledge graph embeddings, on tasks other than word similarity, such as hypernym prediction (which are better suited to exploiting taxonomic knowledge) or perhaps using them to tackle the problem of type and token identification of multi-word expressions.

Acknowledgements

This research was supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Eneko Agirre, Montse Cuadros, German Rigau, and Aitor Soroa. 2010. Exploring knowledge bases for similarity. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'10)*.

Eduardo G. Altmann and Martin Gerlach, 2016. *Statistical Laws in Linguistics*, pages 7–26. Springer International Publishing, Cham.

Fatemeh Torabi Asr, Robert Zinkov, and Michael Jones. 2018. Querying word embeddings for similarity and relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 675–684.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, MD.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*. Citeseer.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver.

Trevor Cohen and Dominic Widdows. 2017. Embedding of semantic predications. *Journal of Biomedical Informatics*, 68:150–166.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional Word Vector Representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 464–469, Beijing.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting Word Vectors to Semantic Lexicons. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1606–1615, Denver, CO.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Winthrop Nelson Francis. 1964. A standard sample of present-day english for use with digital computers.

Martin Gerlach and Eduardo Altmann. 2014. Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, 16:113010, 11.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random Walks and Neural Network Language Models on Knowledge Bases. In *Human Language Technologies: The 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1434–1439, Denver, CO.

Josu Goikoetxea, Eneko Agirre, and Aitor Soroa. 2016. Single or multiple? combining word representations independently learned from text and wordnet. In *AAAI*.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695.

Magdalena Kacmajor and John D. Kelleher. 2019. Capturing and measuring thematic relatedness. *Language Resources and Evaluation*.

Mikhail Khodak, Nikunj Saunshi, Yingyu Liang, Tengyu Ma, Brandon Stewart, and Sanjeev Arora. 2018. A la carte embedding: Cheap but effective induction of semantic feature vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

Angeliki Lazaridou, Marco Marelli, and Marco Baroni. 2017. Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41:677–705.

Alfredo Maldonado, Filip Klubička, and John D. Kelleher. 2019. Size matters: The impact of training size in taxonomically-enriched word embeddings. *Open Computer Science*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, Scottsdale, AZ.

Tomas Mikolov, Ilya Stutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS) In Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe, NV.

Nikola Mrkšić, Diarmuid O Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 454–459, Berlin.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, Copenhagen.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré Embeddings for Learning Hierarchical Representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc., Long Beach, CA.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Nigel Collier. 2017. Inducing embeddings for rare and unseen words by leveraging lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 388–393.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge rare word dataset-a reliable benchmark for infrequent word representation models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1391–1401.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293.

Kiril Simov, Alexander Popov, and Petya Osenova. 2015. Improving word sense disambiguation with linguistic knowledge from a sense annotated treebank. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 596–603.

Kiril Simov, Petya Osenova, and Alexander Popov. 2017a. Comparison of word embeddings from different knowledge graphs. In *International Conference on Language, Data and Knowledge*, pages 213–221. Springer.

Kiril Ivanov Simov, Svetla Boytcheva, and Petya Osenova. 2017b. Towards lexical chains for knowledge-graph-based word embeddings. In *RANLP*, pages 679–685.

Robert Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679—-3686, Istanbul.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-Specialisation: Retrofitting Vectors of Words Unseen in Lexical Resources. In *Proceedings of NAACL-HLT 2018*, pages 516–527, New Orleans, LA.

John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. 2015. From Paraphrase Database to Compositional Paraphrase Model and Back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Zhiguo Yu, Trevor Cohen, Elmer V. Bernstam, Todd R. Johnson, and Byron C. Wallace. 2016. Retrofitting Word Vectors of MeSH Terms to Improve Semantic Similarity Measures. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 43–51, Austin, TX.

George. K. Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*, volume 47. 01.

Augmenting Chinese WordNet semantic relations with contextualized embeddings

Yu-Hsiang Tseng
Graduate Institute of Linguistics
National Taiwan University
seantyh@gmail.com

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

Constructing semantic relations in WordNet has been a labour-intensive task, especially in a dynamic and fast-changing language environment. Combined with recent advancements of contextualized embeddings, this paper proposes the concept of morphology-guided sense vectors, which can be used to semi-automatically augment semantic relations in Chinese Wordnet (CWN). This paper (1) built sense vectors with pre-trained contextualized embedding models; (2) demonstrated the sense vectors computed were consistent with the sense distinctions made in CWN; and (3) predicted the potential semantically-related sense pairs with high accuracy by sense vectors model.

1 Introduction

Chinese Wordnet(CWN) (Huang et al., 2010) has been one of the most important lexical resources in Chinese. Through years of rigorous works from linguists and lexicographers, CWN covers large amount of Chinese words, senses distinctions, and various lexical semantic relations. However, the linguistic knowledge CWN tries to incorporate is far more than a static snapshot of the language usage from a given time. As a lexical resource which aims to facilitate better NLP applications, the current version of CWN has intended to incorporate the complicated and dynamic relations that language implicitly encodes. This is a challenging task for resource maintainer, for they have to manually edit the database, in order to keep up the the neologisms and ever-changing novel word usage.

Recent algorithmic advancements shed lights on how we can augment lexical resources, at least semi-automatically. Thanks to the bloom of internet and social media, voluminous textual data are easily available, where emergent concepts and their relations could be discovered from the real-world and most updated data. This process is further facilitated by recent development of deep learning and machine learning models, such as pre-trained language model (Howard and Ruder, 2018), word embeddings (Joulin et al., 2017), or contextualized embeddings. These computation resources allows us to leverage the ample data, without going through considerable efforts to actually collect, and store the vast amount of data, and setup a model training infrastructure.

In this paper, we took advantage the recent development on contextualized embeddings. Specifically, we used a pre-trained bidirectional encoder representations from transformer (BERT) (Devlin et al., 2018), basing on which we semi-automatically predicted new related senses in CWN. The predictions were only possible with the constraints encoded in Chinese morphology, where the semantic relationship between the whole word and its composing sub-word were suggested (Hsieh and Chang, 2014). We introduced how we applied BERT to construct sense vectors from existing example sentences in each CWN senses, and how to use sense vectors and heuristics rule s regarding Chinese word morphology to semi-automatically generate new relationships (hyponymy/troponymy pairs) among CWN senses. We evaluated these sense vectors with a simulation study and conducted an experiment on the model-predicted sense relation pairs. The procedures described in this paper was shown in Figure 1.

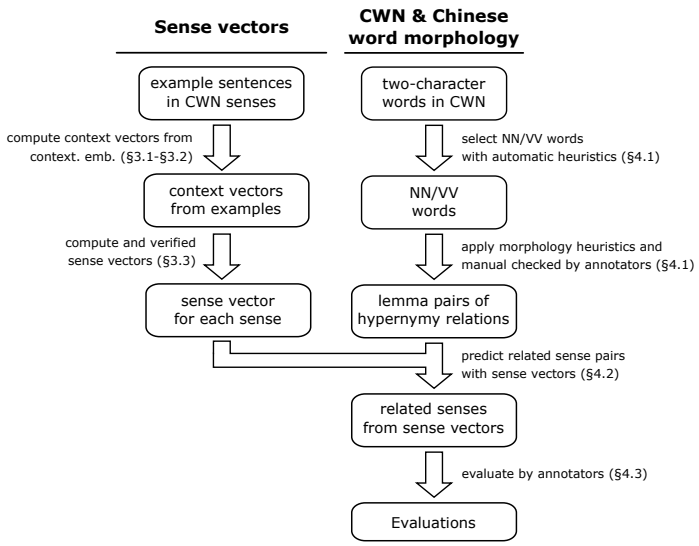


Figure 1: An workflow overview of predicting related senses with sense vectors and Chinese word morphology.

2 Related Works

2.1 Chinese morphology

The concept of word seems to be robust in many language, but remains elusive to languages such as Chinese (Hoosain, 1992). Chinese words were written as a series of Chinese characters, and there is no orthographic cues (such as spaces in English) delineating word boundaries. Therefore, words are instead defined by different theories, focusing on different linguistic aspects, such as their morphological, syntactical or semantic properties. In CWN, words were defined as characters with independent meaning and play a specific syntactic role (CKIP, 1996), and 7 guidelines were introduced to ensure a consistent and meaningful criterion of words.

Most Chinese words are composed of two characters. Characters are the writing units in Chinese, each are written within a square block. Arguably treated as morpheme as its linguistic property by definition, some characters can be used alone, some characters need to combine other characters to form a word, and most of them bring their original meanings into the composition process. For example, 泉水（quán shuǐ, 'spring'）is a word composing of two characters. The second character 水 (shuǐ) can itself be used independently to indicate the meaning of 'water.' For words like 語言 (yǔ yán, 'language'), though the second character 言 (yán) cannot be used independently in contemporary Mandarin Chinese, it still nonetheless contributes an *etymological* meaning of 'speech, speak.'

Unlike inflectional languages, Chinese words do not undergo morphological alternations, such as eat, eats, eaten, eating or eater in English. There are only a few affix-like morphemes in Chinese that account for a small portion of Chinese words. For instance, 師 (-shī) can be attached to a noun as a suffix, indicating a profession, such as 工程 (gōng chéng) means engineering, and attaching the suffix 工程師 (gōng chéng shī) means engineers. However, Chinese do have intricate word morphology, which reflects knowledge about the structure and use of words. For example, 直升機 (zhí shēng jī) means helicopter, and the three characters of which the word are composed literally means vertically(直, zhí) arised(升, shēng) machine(機, jī). Likewise, 汽車 (qì chē) means automobile, the two composing characters could be loosely translated as "gas(汽, qì)-

car(車, chē)." The fact that meanings of word and its composing characters match suggests that Chinese words, through their morphology, reflect systematic knowledge that a native speaker have toward the world. (Packard, 2000)

In order to leverage the copious knowledge encodes within Chinese morphology, previous studies devised heuristic rules to decode the semantic relationships between word and their composing characters (Hsieh and Chang, 2014). The relations decoded provided useful hints for semantic relations, that can be used to expand semantic relations in CWN. Specifically, for some (two-character) words following a *modifier-head* structure, the second component (serving as the head) is the hypernym of the whole word. For example, 書店 (shū diàn) means 'book store', the second component 店 (diàn, 'shop, store') is then inferred to be the hypernym of the whole word (書店). The heuristic rule in application is very effective, for it provide a clear guidance of possible hypernym relations a concept could link to. However, these rules only apply on the lemma level. That is, after the potential hypernyms were identified, the rule cannot provide further guidance on the senses upon which the hypernymy relation should be created.

2.2 Contextualized Embeddings

Vector semantics are models in which researcher use a formal mathematical structure (i.e. vectors) to represent how lexical meanings of words reside in a vector space. The vectors representing each words also encode, to some extent, their mutual semantic relations in that space. This general approach, while being a heated topic in recent years (Landauer and Dumais, 1997; Griffiths et al., 2007; Mikolov et al., 2013; Peters et al., 2018), could be traced back to mid-20th century (Firth, 1957). The idea was to explore the co-occurrence of the words in context (sentences, or a groups of preceding and following words), and use the context to determine the *location* of a word vector in semantic space, where thus location could best reflect the relationships with other words.

While models of vector semantics enjoyed great successes in various NLP tasks, even were indispensable constructs in virtually all deep learning models, challenges emerged when they came to WordNet. WordNet, as a lexical resource of word senses and linguistic knowledge, make intricate distinctions on word senses and the synsets among them. However, vector semantics models had a major limitation of meaning conflation deficiency (Camacho-Collados and Pilehvar, 2018), namely they conflate multiple meanings of a word (lemma) into one representation. For example, in word2vec model (Mikolov et al., 2013), vectors of target word were constructed through the task of predicting the target word with surrounding word vectors (continuous bag of words, CBOW), or, conversely, predicting surrounding words with the target word vectors (skip gram). Different word contexts were independent samples in training, they are not explicitly used by the model. The resulting word vectors were therefore undifferentiated representations of word senses.

Other models have the potential to accommodate, or even represent, word senses information, but not without caveats. For example, latent Dirichlet allocation (LDA) (Griffiths et al., 2007), representing meanings of each word as a probability distribution over different topics, could describe each word sense as a mixture of different topic components. But the problems remains on how to relate latent topics with the word senses. Other endeavors relies on a sense-disambiguated corpus (Iacobacci et al., 2015), and inferred the sense vectors through the disambiguated context. But this approach required a mature word sense disambiguation (WSD) algorithm or sense-tagged corpus with given sets of word sense distinctions. Chinese WSD is an active and productive research topic, but the word sense disambiguation on CWN word senses remains a challenging task.

Instead of relying on sense-disambiguated corpus, recent models tried to incorporate word context into deep learning models and construct contextualized vectors (Peters et al., 2018; McCann et al., 2017). Inspired by the deep learning models in computer vision, these models represent word contexts as an abstract information built upon the basic word embeddings in a language modeling task. Specifically, a model was trained to predict the

next word in a sentence based on the words previously seen. The models used word vectors as input, but the embeddings layers (i.e. word vectors) stacked upon were deep layers tried to encode the contextual information. The outputs of these deep layers were used to complete the prediction task in training; and additionally, they represented the context vectors the words occurred in. Recent deep learning researches provided multiple choices of such layers, like bidirectional LSTM used in ELMo (Peters et al., 2018) and decoder transformer used in OpenAI transformer (Radford et al., 2018). These models, instead of treating each word as a static vector, could generate a contextualized vector for each word in any given contexts. However, as these models were trained on language modeling tasks, only either preceding or succeeding word contexts were exploited to build context vectors.

Bidirectional-encoder representation (BERT) (Devlin et al., 2018) employed different task to train models making use of surrounding word contexts to generate context vectors. As other contextualized vector models, BERT also uses word vector as its input, but the deep layers stacked upon them were layers of encoder transformers (Vaswani et al., 2017). In order to allow encoder to consider the surrounding word contexts without peeking into the predicting targets in the same time, BERT used a cloze task in its training stage. In the cloze task, each word in the whole sentence was available to model, with only the clozed word (the target) masked out. The model then learned to construct a context vector with the surrounding words, and predicts the clozed word with the context vectors. The contextualized vectors trained on this model had wide range of applicability. It had been shown that without substantial modification, the model achieved superior performance on NLP tasks, such as question answering and language inferences.

This paper aims to investigate whether the model of contextualized embeddings can help researchers to identify the semantic relations between word senses defined in CWN or not. The goals of present paper are as follows: (1) Examine how the sense vectors computed by contextualized vector models (i.e. BERT) differentiated the word sense distinctions made in CWN. (2) Predict possible hypernymy-hyponymy relations among sense pairs from sense vectors, guided by Chinese morphology. (3) Evaluate the predictions made by the model with human annotations.

3 Building sense vectors

Word sense is closely related to the context the word resides in, and the contextualize embeddings is meant to encode the context. If we can characterized the context through contextualized embeddings, the context vector was then a formal representation of a word sense. We therefore computed sense vectors from the contextualized embeddings of the target word located in an disambiguated context.

In this section, we first identified the lemmas (and their senses) to be included in current analysis and the experiment in following section. Secondly, we built sense vectors from example sentences of each sense. Thirdly, in order to explore the nature of the sense vectors, we conducted a simulation study over the computed sense vectors.

3.1 Extracting example sentences

We first selected 1,815 lemmas from CWN. [1] These lemmas satisfied following criterion: (1) they are two-character lemmas; (2) each of the composing character is itself a lemma in CWN; (3) all senses of each lemma (both two-character lemma and one-character lemma) must have at least two example sentences. The complete lemmas hence included 2,897 lemmas, which were comprised of two-character lemmas, and their 1,082 unique composing characters as one-character lemma.

These lemmas were related to 11,521 senses (40.0% of all CWN senses) in CWN, and 37,976 example sentences were extracted from these sentences.

3.2 Computing sense vectors

We used BERT (pre-trained on Chinese Wikipedia data dump) as the model of contextualized embeddings. The model had 12

[1] Note that *homonyms* are treated as separate words in CWN, e.g., 打 ('punch' and other derived senses) and 打 ('dozen') are the same lemma used as two words. In this experiment, homonyms are considered as different word senses.

layers, each having 768 hidden states. In this analysis, we concatenated the hidden states of the last 4 layers as the contextualized embeddings. The resulting contextualized embedding dimensions ($\mathrm{CE}_{\mathrm{dimension}}$) was 3,072. The context vector of target lemma in the sentence was then selected from the contextualized embeddings obtained from BERT model. The context vector of example j of sense i, denoted by s_{ij}, can be written as:

$$s_{ij} = \underbrace{\mathbf{1}_{\mathrm{target}}}_{1\times T} \underbrace{\mathbf{CEs}([w_{ij}^{(1)}, \ldots, w_{ij}^{(t)}, \ldots, w_{ij}^{(T)}])}_{T\times \mathrm{CE}_{\mathrm{dimension}}} \tag{1}$$

where T denoted the number of tokens in the example sentences, $w_{ij}^{(t)}$ was the t^{th} token in the example sentences, and $\mathbf{1}_{\mathrm{target}}$ is a vector with each of its element an indicator function:

$$\mathbf{1}_{\mathrm{target}} = \begin{cases} 1, & w_{ij}^{(t)} \text{ is the target lemma} \\ 0, & \text{otherwise} \end{cases}$$

The sense vector, μ_j, of sense j for a given lemma μ, was computed as the centroid of context vectors in all $n_e^{(j)}$ example sentences:

$$\mu_j = \sum_i^{n_e^{(j)}} s_{ij}/n_e^{(j)} \tag{2}$$

The sense vectors were computed for respective senses in selected CWN lemmas. However, these sense vectors were only a linear combinations of the context vectors, which were generated by an intricate deep learning model. The possibility exists that the context BERT trying to represent might be an abstract concept independent from the word context referred in language usage. In order to further investigate the nature of these sense vectors, we carried out following simulation study.

3.3 Sense vector simulation

The purpose of the simulation study was to verify the sense vectors came from groups respecting sense distinctions made in CWN. We compared the grouping patterns of sense vectors and two others from simulated conditions, to demonstrate the sense vectors reflected different contexts of word senses, instead of coming from random patterns.

We first devised a statistic to quantify how clear-cut the groups context vectors formed into, where the sense vectors were computed from. For a given lemma μ, to describe how well the context vectors, s_{ij} were "grouped together" within different senses, we calculated two scores, $\mathrm{MS}_k^{(\mathrm{senses})}$ and $\mathrm{MS}_k^{(\mathrm{error})}$, based on the euclidean distance between s_{ij}, their sense vector μ_j, and the centroid of all sense vector, $\overline{s_{..}}$:

$$\mathrm{MS}_k^{(\mathrm{senses})} = \frac{\sum_j n_j \|\mu_j - \overline{s_{..}}\|^2}{m-1} \tag{3}$$

$$\mathrm{MS}_k^{(\mathrm{error})} = \frac{\sum_i \sum_j \|s_{ij} - \mu_j\|^2}{N_k - m} \tag{4}$$

The ratio of these two scores measured the extent to which the sense vectors distanced from each other, by comparing with the sense vectors distanced from their respective context vectors. This ratio, ζ_k, was computed as:

$$\zeta_k = \frac{\mathrm{MS}_k^{(\mathrm{senses})}}{\mathrm{MS}_k^{(\mathrm{error})}}$$

Intuitively, a small ζ_k indicated the sense vectors themselves were not clearly grouped, since the distance between the sense vectors was similar with the distance between the context vectors used to calculate the sense vectors. This ratio was closely related to F statistic, which was often in comparing two sample variances. However, two caveats existed kept us from directly proceeding to hypothesis testing with F statistic. (1). The explicit distribution of sense vectors as a random variable was not readily available, it is unclear if ζ_k still followed F-distribution under null hypothesis. (2). The simulation was to compare all lemmas in CWN. That is, each lemma was itself a sample in the simulation. However,each lemma has different number of sense vectors and number of examples, a normalized index was then needed to describe ζ_k from different lemmas.

To normalize ζ_k from different lemmas with different senses and examples, we defined π_k, which was the area under the right-tail of ζ_k

in the probability density function of F distribution.

$$\pi_k = 1 - \int_0^{\zeta_k} F_{pdf}(x; df_1, df_2)dx \quad (5)$$

$$df_1 = m - 1 \quad (6)$$

$$df_2 = N_k - m \quad (7)$$

$$(8)$$

where F_{pdf} denoted the probability density function of a given F distribution, N_k denoted total number of examples in lemma$_k$.

Since ζ_k may not follow F distribution, the value of π_k was just a score indicating the "well-groupness" of the senses in lemma k. Smaller π_k signified more clear-cut grouping. The resulting π_k from actual sense vectors had mean of 0.14, standard deviation of 0.10 (Figure 2).

In order to better interpreted the π_k from actual sense vectors, we compare the π_k with two other simulated conditions: (1) random Gaussian vectors and (2) permuted vectors. The first simulated condition was to replace all context vectors with random standard Gaussian vectors of the same length. This condition provided a random baseline of how π_k distributed if context vectors were random noises. The second simulated conditions permuted the actual context vectors. The context vectors were randomly shuffled, and randomly assigned to each word senses, while the sense number and the number of examples of each sense remained the same. The underlying rationale was if the context vectors from the same sense were closer together, then a permuted version of which would destroy the patterns.

Figure 2 showed the results of simulations and the sense vectors. Patterns of π_k in random condition ($M = 0.41$, $SD = 0.05$) was similar to those in permuted condition ($M = 0.42$, $SD = 0.09$). Importantly, the distribution of π_k of actual sense vectors were smaller than any of the simulated conditions. These patterns showed the computed sense vectors had clear grouping structures and the groupings were consistent with sense distinctions in CWN.

4 Experiment

With sense vectors as a computable representation of word senses, we aimed to semi-automatically discover potential hypernymy-hyponymy sense relations in CWN, guided by Chinese morphology. Previous study argued that Chinese two-character words with inner structure of two nouns and two verbs, were likely a hyponymy of the second character (when used as a one-character word). That is, at lemma level, we could discover semantic relations leveraging Chinese word morphology. However, semantic relations in WordNet are relationship among word senses. Given there are multiple word senses in each lemma, manually found them would be a daunting task. With help of sense vectors, we could try to find senses among which relations existed.

4.1 Selecting candidate lemma

To find out candidate hypernymy-hyponymy lemma pairs, we first used heuristic rules to automatically select words composing of two nouns (NN) or two verbs (VV). The heuristic rules were to determine the part-of-speech of composing character, basing on the dictionary data compiled by the Ministry of Education of Taiwan. Three criterion were applied consecutively: (1) excluding senses from classical Chinese, compare the number of senses a POS have, the POS with more sense count was the POS of the character; (2) if sense counts of different POS were equal, compare the frequency sum of the example words (as listed in sense entries) of that sense in a corpus; (3) if the frequency sum were equal, compare the sense counts of POS in CWN. These three criterion labeled 99% words in 1,815 two-character words (the same set of words in analyzing sense vectors). POS of the remaining words were assigned manually. There were respectively 824 and 362 words of NN and VV structures selected.

Three graduate students in Graduate Institute of Linguistics, National Taiwan University examined these N_1N_2 and V_1V_2 words, labeling words (W) with hyponymy relations (W is a kind of N_2) or troponymy relations (V is a way of doing V_2). Since determining the relations were relatively straightforward given the words and composing character , each item

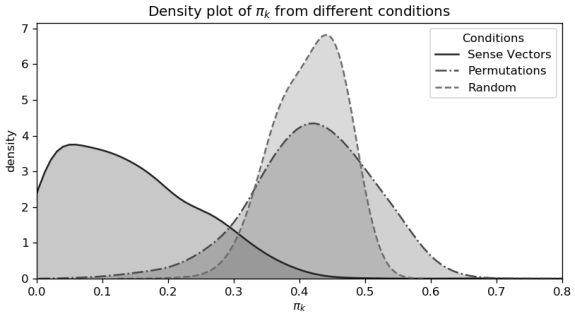


Figure 2: Distribution of sense vectors statistics, π_k.

was only annotated by one annotator. The resulting word list comprised 337 NN words and 150 VV words.

4.2 Predicting related senses

We used sense vectors computed in previous section to predict which sense were related in the lemma pair (i.e. the whole word lemma and the N_2/V_2 lemma). Given a pair of lemmas, μ_j was the sense vector computed of lemma μ and ν_j were of lemma ν. We predicted the related senses as the nearest sense vectors between two set of lemma senses. The distance measure, $d_{i,j}$, was the euclidean distance between the sense vectors:

$$d_{i,j} = \|\mu_i - \nu_j\|^2$$

All distances between the sense pairs in lemma μ and lemma ν formed a distance matrix $\mathbb{D}$:

$$\mathbb{D}_{(\mu,\nu)} = \begin{array}{c} \\ \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_m \end{array} \begin{array}{c} \begin{array}{cccc} \nu_1 & \nu_2 & \cdots & \nu_n \end{array} \\ \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,n} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,n} \\ d_{3,1} & d_{3,2} & \cdots & d_{3,n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m,1} & d_{m,2} & \cdots & d_{m,n} \end{pmatrix} \end{array}$$

The predicted sense pairs were the senses pairs of smallest d_{ij}:

$$\text{Related sense pair } (\mu_i, \nu_j) = \underset{i,j}{\operatorname{argmin}} \left\{ d_{i,j} \,\middle|\, d_{i,j} \in \mathbb{D}_{(\mu,\nu)} \right\}$$

The calculations were performed on all 487 lemma pairs. Two of the lemmas had format errors in the example sentences, and had no sense vectors. Therefore 485 sense pairs predictions were made.

Word Structure	NN n=337	VV n=148	Overall n=485
Baseline			
Random	0.12	0.21	0.15
First Sense	0.40	0.46	0.42
Model Prediction			
Top 1	0.81	0.83	0.82
First 5	0.96	0.94	0.96
First 10	0.99	0.97	0.98

Table 1: Accuracy of related sense pairs predicted by model and baseline performance.

4.3 Evaluation

Model-predicted related sense pairs were equally divided into three parts and each part was evaluated by an annotators. Annotators marked whether the predicted sense pairs were actually hyponymy/troponymy pairs. If they found erroneous predictions, correct sense pairs would be added, and these data were further used in evaluation. The results were shown in Table 1.

The overall accuracy of model predictions was 0.82, with similar performance on either NN or VV constructions. To better illustrate the nature of the predicting task, two baseline performance were provided: (1) a random baseline was the performance the model was random guessing; (2) 'first sense strategy' was the model always picked the first sense listed in CWN. Compared with the accuracy of

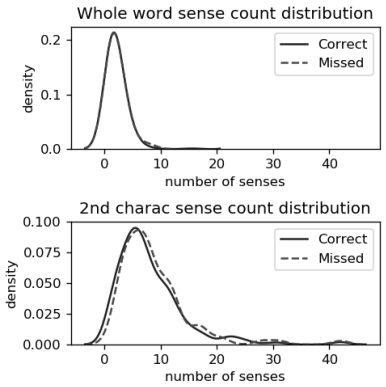


Figure 3: Sense counts on correct and missed-linked senses

these baselines, present model provides valuable suggestions on potential sense pairs.

Table 1 also shows the prediction ranks of the correct sense pairs. That is, if the correct sense pairs were not the nearest one in the distance matrix, would the correct pairs rank in first 5 or 10 pairs in the distance matrix. The results indicated there were 96% of correct pairs were ranked within the first 5 pairs.

To further investigate the errors made by the model, Figure 3 shows the sense counts distribution of the whole word and the second composing character (N_2 / V_2), on correct and missed predictions. From Figure 3, the distribution of the second character when missed predicted, was marginally more than the correct ones; while the distribution was virtually the same in whole word. The latter pattern was expected since the Chinese two-character words generally had fewer word senses.

This experiment demonstrated how to leverage Chinese word morphology and sense vectors to discover potential hypernymy or troponymy relations in CWN. The evaluation also showed this semi-automatically procedure suggest valuable sense pairs.

5 Conclusion

This paper combines recent advancements of contextualized embeddings and existing CWN resources to build sense vectors. We have demonstrated these sense vectors followed the sense distinctions made in CWN, and showed sense vectors, guided by Chinese morphology, provided valuable suggestion discovering hypernymy/troponymy. The semi-automatic procedures greatly facilitate the on-going development of CWN in the fast-paced language environment.

Acknowledgements

This work was supported by Ministry of Science and Technology (MOST), Taiawan. Grant Number MOST 108-2634-F-001-006. We thank Yong-Fu Chao, Ying-Yu Chen, Chiung-Yu Chiang, and Yi-Ju Lin (National Taiwan University) for their assistance on data preprocessing and annotations.

References

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, December.

CKIP. 1996. *Study on Chinese word boundaries and computational standard in segmentation.* CKIP Technical Reports. Institute of Information Science, Acaemia Sinica.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

J. R. Firth. 1957. Applications of general linguistics. *Transactions of the Philological Society*, 56(1):1–14, November.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Rumjahn Hoosain. 1992. Psychological reality of the word in chinese. In *Language Processing in Chinese*, pages 111–130. Elsevier.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July. Association for Computational Linguistics.

Shu-Kai Hsieh and Yu-Yun Chang. 2014. Leveraging morpho-semantics for the discovery of relations in chinese wordnet. In *Proceedings of the Seventh Global Wordnet Conference*, pages 283–289, Tartu, Estonia.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and

Shen-Wei Huang. 2010. Constructing chinese wordnet: Design principles and implementation. (in chinese). *Zhong-Guo-Yu-Wen*, 24:2:169–186.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China, July. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *CoRR*, abs/1708.00107.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

J.L. Packard. 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *NAACL-HLT*, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Visualising WordNet Embeddings: some preliminary results

Csaba Veres
Department of Information Science and Media Studies
The University of Bergen, Norway
csaba.veres@uib.no

Abstract

AutoExtend is a method for learning unambiguous vector embeddings for word senses. We visualise these word embeddings with t-SNE, which further compresses the vectors to the *x,y* plane. We show that the t-SNE co-ordinates can be used to reveal interesting semantic relations between word senses, and propose a new method that uses the simple *x,y* co-ordinates to compute semantic similarity. This can be used to propose new links and alterations to existing ones in WordNet. We plan to add this approach to the existing toolbox of methods in an attempt to understand learned semantic relations in word embeddings.

1 Introduction

There is currently a great deal of interest in the representations of words as continuous, real valued vectors, or *embeddings*. Various popular methods produce a single vector for each word form in the training set, for example GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013a), and SVD (Levy et al., 2015).

These methods could be regarded as modern day experiments inspired by Zellig Harris' hypotheses about the distributional structure of language. Harris proposed that word meanings give rise to observable distributional patterns in language, such that two semantically unrelated words A and C would be less likely to be found in common linguistic contexts as two semantically related words A and B (Harris, 1954). Modern machine learning techniques have made it computationally possible to *embed* very high dimensional distributional patterns in a much lower dimensional vector space, in which the distances between any given vectors is related to the similarities of context in which the corresponding words are found in the training set. Semantic relatedness is therefore correlated with the calculated distance (e.g. cosine distance) between vectors, although the precise nature of the relatedness is not well understood. One of the long term motivations behind the work reported in this paper is to develop a methodology for investigating the nature of the semantic relationships discovered by various methods of context embedding.

A general problem with current methods of single layer *embeddings* is that they treat each word-form as a single word in a *bag of words* model. Thus the embedding for each word-form conflates contexts over every sense of ambiguous words. There have been proposals to discover unique vectors for the different senses of ambiguous words, typically by using clusters of words related to the different senses, either before (Reisinger and Mooney, 2010) or after training (Schütze, 1998).

In this paper we investigate semantic relationships between WordNet synsets using word embeddings. The most convenient resource for this are the vectors trained with AutoExtend (Rothe and Schütze, 2015). This method uses structural information from WordNet to learn new embeddings for synsets and lexemes from non-disambiguated word vectors. Their insight is to use the constraints detailed in WordNet[1], and to formalise those constraints with respect to the embeddings. For example, the learned embedding for the word-form *W/suit* is formally related to two lexemes, one *L/suit (S/suit-of-clothes)*, and the other *L/suit (S/lawsuit)*, where the *S* prefix denotes that the lexeme is a part of the synset *S/*. Further, the embedding for the lexeme *L/suit (S/lawsuit)* is connected to the embeddings for the lexemes *L/case (S/lawsuit)* and *L/lawsuit (S/lawsuit)* because they are elements in the synset *S/lawsuit*.

[1]The technique is not restricted to WordNet, but could be used with any other resource that defines structural constraints between senses.

Finally, these lexemes are themselves aligned with the words *W/case* and *W/lawsuit*, for which embeddings have been learned (see (Rothe and Schütze, 2015), figure 1). The goal is to learn embeddings for the lexemes and synsets from the embeddings of the words and the formal constraints taken from the resource, in this case WordNet.

The main goal in this paper is to explore semantic relationships in the vector space of lexemes created by the disambiguation algorithm. We compare these to the baseline embeddings created with the word2vec skip-gram model (Mikolov et al., 2013b). To the best of our knowledge the semantics of vector similarities in embedding space have not been subject to rigorous linguistic investigation. We think that investigating semantic relations using the lexemes learned through the AutoExtend framework will provide important data for understanding the relations captured by word embedding techniques in general. We begin with some visualisations before moving on to some more quantitative accounts. The experiments reported in this paper are at an early stage, mostly aimed at gathering observations rather than finding explanations for them.

2 Lexeme Visualizations

In these experiments we used AutoExtend to learn vectors for 73747 lexemes from embeddings generated with the GoogleNewsCorpus, and WordNet3.0. The first experiment was to visualize the whole set with the T-distributed Stochastic Neighbor Embedding (t-SNE) method (van der-Maaten and Hinton, 2008), which is a nonlinear dimensionality reduction technique that attempts to keep the relative distances in the high dimensional space intact during the low dimensional transformation. Perhaps not surprisingly the visualisation of the entire set was not terribly useful because of its very high density of points, and is not reproduced here.

The second experiment was to visualize a meaningful sub set of the embeddings that illustrate a sub domain of interest[2]. We took the meaningful subset from an experimental semantic bookmarking platform, LexiTags (Veres, 2013; Veres, 2011), in which users assign WordNet lexemes as *tags* to their bookmarks. The tags are meaningful because they are used to describe web resources of interest to users of the platform. We collected 248 tags and constructed a t-SNE plot of the corresponding WordNet embeddings (figure 2).

Some interesting relations are immediately apparent. For example the tag *boring* is used in an uncommon sense denoted by the lexeme {boring.n.02: (the act of drilling a hole in the earth in the hope of producing petroleum}, which in the visualization is closely related to {extraction.n.03: the action of taking out something (especially using effort or force)}. However in the baseline word2vec embeddings only the more common adjectival sense is available, with the related words being {uninteresting, depressing, and dull}.

There also appears to be a cluster that captures an interesting progression from {crime.n.01: an act punishable by law} to {corruptness.n.01: the state of being corrupt}, {government.n.01: the organization that is the governing authority of a political unit} and finally to the result, a {revolution.n.02: the overthrow of a government by those who are governed}. Perhaps a sense of causality between the lexemes has been captured.

Additionally there are some interesting relationships between lexemes from different word classes, for example the actions {synchronize.v.01: make synchronous and adjust in time or manner}, and {install.v.01: set up for use} when used in the domain of computer science often involves in the creation of a {backup.n.04: a copy of a file or directory on a separate storage device}. Again this might be an act of causation.

3 Sense Clusters

The visualisations suggest some interesting patterns in the relationships. However a more systematic study will require better ways to quantify observations. To this end we propose a unique method for using the t-SNE results which, to our knowledge, has never been reported.

Recall that the t-SNE algorithm compresses the 300 dimensional vectors into two points $(x_1, y_1), (x_2, y_2)$ for visualisation, where the distance $d = |\mathbf{x} - \mathbf{y}|$ is optimised to preserve the neighbourhood relations in the original high dimensional vector space. Thus the distance d is construed as the semantic distance between the two points. We propose to use these distances directly in calculating the semantic similarities between lexemes, to take the place of cosine similar-

[2] a common approach in practice according to an article by Sergei Smetanin in Medium: Towards Data Science `https://is.gd/UyUrhP`

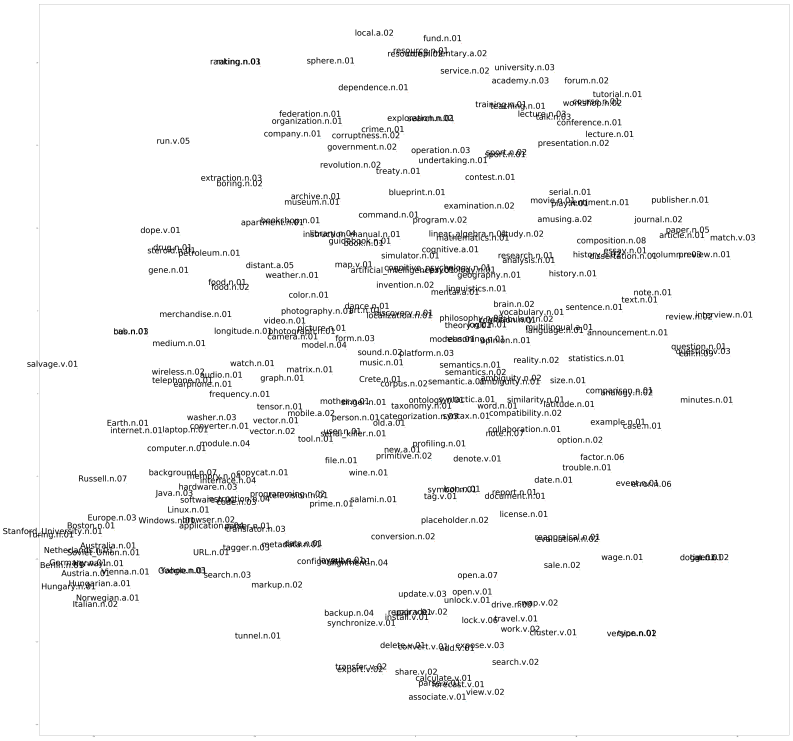


Figure 1: Visualisation of the selected *tag* lexemeses

ity in the original vector space. Thus, we have two measures of similarity, which might reveal different clusters.

In order to discover clusters in the *x,y* co-ordinate space we used the divide and conquer approach to the closest pair of points problem, where the closest pair is recursively identified by finding the closest pair in one half of the gradually diminishing problem space[3]. We used a python implementation of the algorithm[4] to find the closest pair of points, then found the five closest points to the first in the pair. Then we deleted one of the closest match points from the initial pair and repeated the divide and conquer algorithm to find the next closest pair of points from the remaining set. In the end this gave us a large set of clusters formed by the closest points in the entire co-ordinate system, and the five closest points to those.

Table 1 shows some hand selected examples of the closest points, together with their neighbours in the two dimensional t-SNE space, the original 300 dimensional AutoExtend space, as well as the word2vec embedding.

It is clear that both sets of results based on the AutoExtend vectors are better able to capture the precise meaning of the search terms, and return more relevant neighbours than word2vec. Common embedding techniques such as word2vec can return words in the result set that are either irrelevant, relevant along some obscure semantic dimension, or simply morphological derivatives of the search term. There are examples of each of these in our result set.

Looking at the two result sets from the lexeme embeddings it appears that the t-SNE results are superior, at least for these examples, to cosine similarity measures. More of the results seem to capture the precise meaning of the particular lexeme. For the *opposite* example, the t-SNE results better capture the sense that *opposites* are *different*. AutoExtend also captures this but to a lesser extent, where the closest neighbour is *identical*, which is the opposite of *opposite*. Right semantics, wrong polarity.

Another interesting observation is that the t-SNE results might be useful in identifying synsets with very similar meanings in WordNet, which is necessary for creating new versions with less fine-grained meaning distinctions (e.g. (Snow et al., 2007)). Again in the *opposite* example the second and fifth meaning of *different* appear as if they could be merged. The rule would be to merge the synsets for lexemes of the same word form in a cluster.

The next steps in this research is to quantify the relationship between the lexemes in the t-SNE clusters and existing WordNet links. It seems clear from the examples that the embedding relations are not identical to the relations already in WordNet, but can potentially reveal interesting, additional thematic links. This can be used to propose new links in WordNet.

4 Conclusion

In conclusion this very brief look at the results shows that the t-SNE compression provides a very interesting set of results to complement the study of semantic relations. As far as we know these are novel ideas which have not been investigated.

We plan to use these results to modify WordNet by merging similar synsets, and by including new thematic links.

Clearly the work is at an early stage, but we are excited at the possibilities presented by these preliminary results.

References

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Neural and Information Processing System (NIPS)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

J Reisinger and RJ Mooney. 2010. Multi-prototype vector-space models of word meaning.

[3] https://en.wikipedia.org/wiki/Closest_pair_of_points_problem

[4] http://www.rosettacode.org/wiki/Closest-pair_problem

S Rothe and H Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics*, 24:97–123.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *EMNLP-CoNLL*.

Laurens van derMaaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *JournalofMachine-LearningResearch*, 9:2579–2605.

Csaba Veres. 2011. Lexitags: An interlingua for the social semantic web. In *Proceedings of the 4th International Workshop on Social Data on the Web, SDoW@ISWC 2011, Bonn, Germany, October 23, 2011*.

Csaba Veres. 2013. Crowdsourced semantics with semantic tagging: "don't just tag it, lexitag it!". In *Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web, Sydney, Australia, October 19, 2013*, pages 1–15.

Table 1: Selected lexemes and their closest neighbours in the t-SNE compression. Also shown are the nearest nedighbours in the original AutoExtend embeddings, and the closest neighbours in word2vec. The first row is the target word, neighbours ordered by descending similarity.

t-SNE most similar words	*most similar words in AutoExtend vector space*	*word2vec most similar words*
opposite.s.04 the other one of a complementary pair; ”the opposite sex”		
opposite.s.03 moving or facing away from each other; ”looking in opposite directions”	**opposite.s.01** being directly across from each other; facing	perpendicular
different.s.02 distinctly separate from the first; ”that’s another (or different) issue altogether”	**identical.s.02** being the exact same one; not any other	side
different.s.05 distinct or separate; ”each interviewed different members of the community”	**vocationally.r.01** affecting the pursuit of a vocation or occupation	inwards
face-to-face.r.02 directly facing each other	**variant.s.01** differing from a norm or standard	diagonally
other.a.01 the act of tying or binding things together	**different.s.02** distinctly separate from the first	right
listening.n.01 the act of hearing attentively		
sensing.n.02 becoming aware of something via the senses	**percussion.n.04** tapping a part of the body for diagnostic purposes	listened
taste.n.07 a kind of sensing distinguishing substances by means of the taste buds	**auscultation.n.01** listening to sounds within the body (usually with a stethoscope	listens
lipreading.n.01 perceiving what a person is saying by observing the movements of the lips	**moralism.n.02** judgments about another person’s morality	listener
fingering.n.02 touching something with the fingers	**lipreading.n.01** perceiving what a person is saying by observing the movements of the lips	hear
swell.n.03 a crescendo followed by a decrescendo	**rehearing.n.01** the act of hearing again	vocalizing

The Making of Coptic Wordnet

Laura Slaughter,[♠] Luis Morgado da Costa,[♦] So Miyagawa,[♣]
Marco Büchler,[♣] Amir Zeldes,[♥] Hugo Lundhaug,[♠] Heike Behlmer [♣]

[♠] University of Oslo, Norway
[♦] Nanyang Technological University, Singapore
[♣] Georg-August-Universität Göttingen, Germany
[♥] Georgetown University, USA

Abstract

With the increasing availability of wordnets for ancient languages, such as Ancient Greek and Latin, gaps remain in the coverage of less studied languages of antiquity. This paper reports on the construction and evaluation of a new wordnet for Coptic, the language of Late Roman, Byzantine and Early Islamic Egypt in the first millenium CE. We present our approach to constructing the wordnet which uses multilingual Coptic dictionaries and wordnets for five different languages. We further discuss the results of this effort and outline our on-going/future work.

1 Introduction

This paper reports on the process of constructing a wordnet(WN) for the Coptic language. Coptic belongs to the Egyptian branch of the Afroasiatic language family, spoken in Egypt mainly in the first millennium CE and written in an extended form of the Greek alphabet (see Section 1.2). Together with its precursor Ancient Egyptian written in Hierolgyphic, Hieratic and Demotic scripts, Coptic forms part of the longest continuously documented language on Earth, spanning over four millennia. Despite its importance for historical and comparative linguistics, as well as ancient history, Coptic remains comparatively low in digital resources when compared to contemporary languages of the Ancient and Early Medieval Mediterranean such as Latin and Ancient Greek. With the recent launch of an open source Coptic Dictionary Online (Feder et al., 2018) with an interface for human reading, this project aims to follow with the next logical step in machine readable resources for Coptic: providing a wordnet for the language, which will also be the first wordnet for the Egyptian branch of the Afroasiatic languages.

Wordnet projects aim to provide a machine-tractable lexical resource for automated processing of texts. The purpose of a wordnet for the Coptic language is in the first instance to support digital scholarship on the language. The Coptic language has fewer lexical resources than Greek and Latin and the manuscripts written in Coptic (mainly between the 4th and 12th centuries CE) have received less attention, meaning there is much room for studying their transmission history, an effort that can benefit from a wordnet, for example in recognizing non-verbatim textual reuse.

In this paper, we will present our work on constructing the Coptic Wordnet and outline the goals for this on-going project, as well as an evaluation of its current coverage.

1.1 Background

A number of wordnets already exist for ancient languages: Ancient Greek (Bizzoni et al., 2014, AGWN), Latin (Minozzi, 2009), Sanskrit (Kulkarni et al., 2010), Middle Ancient Chinese (Zhang et al., 2014, MidacWN), and Pre-Qin Ancient Chinese (Zhang et al., 2017, PQACWN). Constructing a wordnet can be extremely time-consuming when done manually, so most wordnets are bootstrapped using another existing wordnet which is referred to as the "pivot language"; usually this is done using the English language Princeton WordNet (Fellbaum, 1998, PWN). The bootstrapping approach to construction is called the "expansion" approach and manual construction is referred to as the "merge" approach (Vossen, 1998). The ancient language wordnets listed above were all boot-

strapped using PWN with the exception of Sanskrit which used the Hindi Wordnet (Bhattacharyya, 2017). Latin Wordnet used two wordnets as pivot languages, Italian WN (Bhattacharyya, 2017) and PWN.

There are both advantages and disadvantages to using pivot languages to bootstrap new wordnets (Bond et al., 2016). One primary advantage is that the 'expand' approach provides immediate multilingual links. The disadvantage of the approach is that the concepts which are not in the pivot language(s) cannot be expressed and are omitted until they are added manually. This problem could be exacerbated for ancient languages since concepts that were expressed in ancient times can lack modern-day equivalents. Conversely, linking to modern terminology can result in a connection to a modern idea that is misleading or has no relevance. Some synsets in modern wordnets do not fit ancient living environments such as those having to do with modern science and technology. This particular challenge is covered in the paper describing Ancient Greek Wordnet issues concerning modern concepts that evolved from ancient concepts (Bizzoni et al., 2014).

Due to the limited number of contexts attested in ancient languages, we expect not to cover a hierarchy of terms as rich as the one that can be seen in modern language resources. To illustrate, PWN has over 10 levels of hypernyms, including terms available to discuss the taxonomy of "sheep" using modern rank-based scientific classification. Many of these categories are informed by the modern understanding of biology, as we have the benefit of scientific contributions impacting how we talk about the world, starting with Linnaeus' work on taxonomies in the 1750s. In the ancient world, we do not have evidence that words were available to cover all of these levels, differentiating, e.g. between placental mammals, monotremes, and marsupials. This issue surrounding the hierarchies is addressed in the paper describing the construction of the Sanskrit wordnet (Kulkarni et al., 2010), which points to the challenge of traditional Sanskrit texts on philosophy and medicine containing many discussions on ontological categories and hierarchies that differ from those in the modern Hindi wordnet.

Even though we see that the issues presented above could provide motivation for choosing the "merge" approach, the immediate multilingual links do provide the needed resources to applications and research within the Digital Humanities, particularly with an aim to study the relationship between Coptic texts and parallel or contemporary texts in other ancient languages. In addition, using a pivot language (such as English, through PWN) is an intermediate step to link directly to the Collaborative Interlingual Index (Bond et al., 2016, CILI), which allows concepts to link across languages without necessarily subscribing to any one wordnet's hierarchy.

1.2 The Coptic Language

The Coptic language is the last stage of the Egyptian language which has been recorded in writing for more than 5,000 years. Pre-Coptic Egyptian language was the vehicle of the culture, politics and religions of the Ancient Egyptian civilization and written in three scripts: Hieroglyphic, Hieratic and Demotic (the latter from 700 BCE).

After the conquest of Egypt in 332 BCE, the Egyptian language borrowed a considerable number of words from Ancient Greek. As early as the 1st and 2nd centuries CE, there had been attempts to write the Egyptian language with the Greek alphabet.

From the 2nd-3rd century, writing the Egyptian language with the Greek alphabet and several Demotic phonograms became common and standardized. This writing system is now known as the Coptic alphabet, and a variant of the Egyptian language which is written in this alphabet is called the Coptic language. The major Coptic dialects include: Sahidic, Boharic, Fayyumic, Mesokemic, Akhmimic, and Lycopolitan. The current version of the Coptic WN contains only the Sahidic dialect, which was the main vehicle of Coptic literature in the first millennium CE and is often considered the 'classical' form of the language. However, there are plans are to extend it to include other dialects in the future. This dialect was chosen primarily based on immediate research needs for processing text reuse cases.

Typologically, Coptic departs from earlier synthetic (highly inflectional) Middle Egyp-

tian, and more analytic (or periphrastic) Late Egyptian, developing instead an agglutinative morphology, in which pronouns and auxiliaries are fused to associated verbs, substantially complicating morphological analysis and the ability to recognize variant forms of Coptic words in running text. The language also allows object incorporation into verbs (similar to English forms such as 'to name-call', but much more frequent), as well as fusion of Greek-origin and native Egyptian lexical items (Grossman, 2014).

There is generally no word division in Coptic writing (*scripto continua*) in Late Antiquity, though modern conventions spell Coptic with spaces between word groups known as bound groups. A bound group contains a content lexeme that is usually a noun or a verb, along with clitic articles, auxiliaries, prepositions and object or possessor pronouns. Coptic is a head-initial, Subject-Verb-Object (SVO), in which nouns carry grammatical gender (M/F), and adjectival senses are generally supplied by nouns ('person of wisdom' means 'wise person') or verbs (e.g. for color terms, a verb meaning 'become white' or 'be white'), with a very small closed class of lexical adjectives remaining from older Egyptian.

As of April 2019, there are 22,777 known Coptic sources (e.g. fragments, codices, epigraphical items, etc.) indexed by the Trismegistos database.[1] The effort to digitize these sources is still on-going and the volume of available digitized text is steadily growing. While most Coptic manuscripts are still waiting to be digitized, a number of projects/sites are contributing to this effort, including: Coptic Scriptorium (Schroeder and Zeldes, 2016), the Corpus dei Manoscritti Coptic Letterari [2], the St. Shenouda the Archimandrite Coptic Society, the Editio Critica Maior of the Greek New Testament,[3] the Digital Edition of the Coptic Old Testament[4], the Marcion project[5], and the Marc Multilingue project[6].

[1]http://www.trismegistos.org/
[2]http://www.cmcl.it/
[3]https://www.uni-muenster.de/intf/ecm.html
[4]http://coptot.manuscriptroom.com/
[5]http://marcion.sourceforge.net/
[6]http://www.safran.be/marcmultilingue/

1.3 Motivation

Like most other wordnets, the motivation behind this project is to perform automatic analysis of texts, including: classic uses in NLP, word similarity tasks, classification of texts, and enhancing the performance of information retrieval. One of the major motivations behind the construction of the Coptic wordnet in particular was to use the hierarchies for text reuse in TRACER (Büchler et al., 2014), but applications for searching and hyperlemmatization using senses (discussed further in Kučera (2007)) are conceivable as well. The currently available NLP pipeline for Coptic (Zeldes and Schroeder, 2016) already offers lemmatization to base dictionary entries, but automatically linking word forms to wordnet entries could make comparisons of automatically analyzed texts to existing texts in Coptic, as well as other languages with aligned wordnets, much easier.

2 Methods

Our automated method for building a new wordnet requires two main types of resources: (1) bilingual dictionaries or any other source providing candidate lemmas aligned with translations, and (2) matching wordnets, sharing a common structure – PWN, in our case. Ideally there should be at least one high coverage wordnet for each of the languages that candidate lemmas are aligned to. Unfortunately, we know that this is rarely the case, and different languages have wordnets of different sizes, which can be a bottleneck for our automated method.

2.1 Dictionaries

The lemma alignments for Coptic were extracted from three sources: the Coptic Dictionary Online (Feder et al., 2018, CDO)[7], Marcion's dictionary[8], and a subset of data from the Database and Dictionary of Greek Loan Words in Coptic (DDGLC)[9] to which we were granted access, and which contains Greek loan words used in Coptic and their respective translations/definitions in English. Both the CDO and Marcion are based on Crum's Coptic

[7]https://coptic-dictionary.org/
[8]http://marcion.sourceforge.net
[9]https://www.geschkult.fu-berlin.de/en/e/ddglc

dictionary (Crum, 1939). The CDO provides trilingual translations in English, French, and German. Less is known about the construction of Marcion, however, which provides translations in English, Czech and Greek.

A summary of the number of Coptic lemmas and the number of translations available in each language is provided in Table 1. These numbers include several preprocessing steps of cleaning and splitting data (e.g. translations often contained multiple lemmas separated by commas or semicolons that were split; parenthetical notes were removed; etc.).

2.2 Wordnets

Concerning the second type of resources, wordnets, we were fortunate to be able to find resources for all languages available in our translations. The automated process (see Section 2.3, below) was done in two stages. For the first stage we collated wordnet data for English, Greek, Czech, German and French from multiple sources, namely: the Princeton Wordnet (Fellbaum, 2017), GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010), the Open German Wordnet[10], WOLF: Wordnet Libre du Français (Sagot and Fišer, 2008), and the Greek Wordnet (Stamou et al., 2004). In addition, data for these five languages was also collected from the Extended Open Multilingual Wornet (Bond and Foster, 2013, OMW), which offers automatically collected linked-data from Wiktionary and the Unicode Common Locale Data Repository (CLDR), and from the English subset of the NTUMC Wordnet (Tan and Bond, 2014; Seah and Bond, 2014; Morgado da Costa and Bond, 2016), which includes a few thousand new senses for English, including pronouns, exclamatives and number of other basic senses missing from the Princeton Wordnet.

All this data was linked through a locally built copy of the OMW, linking all wordnets through the structure of the Princeton Wordnet. Table 2 shows the number of senses available for each language in the small multilingual wordnet built for this project, at Stage I and Stage II of the building process.

The second stage of the construction of the Coptic WN consisted of applying the same method over an improved collection of data. This included both better preprocessing of the dictionary data and the addition of two new wordnets to the local multilingual wordnet used for the automated construction: the Ancient Greek Wordnet (Bizzoni et al., 2014) and an unreleased open and improved version of the Czech Wordnet (Pala and Smrž, 2004). Although technically different languages (with different language codes), the Ancient Greek Wordnet and the Greek Wordnet were merged into a single 'Greek' lexicon to facilitate the linking process. Table 2 shows that the addition of these two wordnets significantly boosted the number of available senses for both Greek and Czech which, in turn, helped to produce an improved version of the Coptic WN (see Section 3).

2.3 Automated Construction Method

Our method follows the basic assumptions of the expansion approach, leveraging on the structure of the Princeton Wordnet as reference, but gathering new senses through a naive algorithm inspired by the idea of multilingual sense intersection (Bonansinga and Bond, 2016; Bond and Bonansinga, 2015) to determine potential senses of a new wordnet.

The idea of multilingual sense intersection has a simple logical foundation. Through this approach, the semantic space of a polysemous word in any language can be constrained by aligned translations of the same word in other languages. This technique has been used for Word Sense Disambiguation (WSD) of parallel text, and words alignments across an increasing number languages have been shown to incrementally constrain the semantic space of a word. Figure 1 shows a conceptualization of this logic, for three languages.

In our case, instead of parallel text (which often requires statistical methods to produce word alignments), we use the word-aligned dictonary data produced between Coptic and the five other languages mentioned above: English, Greek, French, German, and Czech (see section 2.1).

The data produced by this technique can be sorted in multiple ways. One of the most meaningful ways to sort this data is by the number of languages that suggest any given

[10] https://github.com/hdaSprachtechnologie/odenet

Resource	Coptic	English	Greek	Czech	German	French
Marcion	7,069	15,748	9,674	13,726	-	-
CDO	4,362	10,021	-	-	10,021	10,435
DDGCL	4,850	9,227	4,854	-	-	-

Table 1: Lemma Alignments by Resource

Language	Senses (Stage I)	Senses (Stage II)
Czech	16,079	**63,198**
English	209,787	209,787
French	130,420	130,420
German	145,420	145,420
Greek	37,765	**114,383**

Table 2: Wordnet Senses

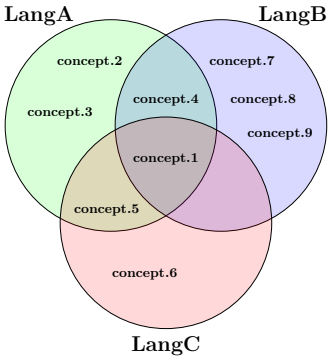


Figure 1: Sense Intersection

concept (i.e. in Figure 1 *concept.1* would be suggested by three languages, while *concept.4* and *concept.5* would be suggested by alignments in two languages). Concepts suggested by more languages have, empirically, a higher likelihood of being correct.

Within concepts suggested by the same number of languages, the algorithm we used employs other metrics to rank candidates: number of individual lemmas matched in each language; part-of-speech congruency, ambiguity of each lemma, and lemma-concept saturation level (i.e. for each concept being suggested, what percentance of lemmas was seen to inform the same concept, per language). This algorithm also performs some language specific string normalization (removal of the infinitival 'to'; removal of determiners preceeding nouns such as 'a' or 'the', case normalization – i.e. for English but not for German).

The development of this system is still ongoing and a full description of its workings is outside the scope of this paper.

2.4 Output and Data Sampling

The output of our system is exemplified in Table 3. In addition to the columns shown in Table 3, the system also outputs a sum-score of multiple other checks mentioned above. Each result row shows, in order, a reserved space for the human validation, the number of languages used to inform this result, the lemma that will be added to the candidate concept, all the translations that were matched to the candidate concept, the PWN offset of the candidate concept and English lemmas, definitions and examples, provided by the PWN.

Two Coptic scholars examined 300 rows (i.e. senses) from our results, with the goal of clarifying the true relationship between the scoring assigned and the mapping of senses to the wordnet. The evaluation task consisted of a three-way decision to be recorded in the first column of each row. This three-way decision comprised: attesting the existence of the candidate sense (i.e. the lemma was known to include the meaning proposed by the Candidate Synset) – marked with *1*; revealing uncertainty about whether the Candidate Lemma could have the proposed sense – marked with *?*; and rejecting the possibility that the Candidate Lemma could be used in the candidate sense – marked with *0*.

The initial sample of 300 senses was done under the assumption that the sum-score mentioned in Section 2.3 would outperform the simple metric of 'number of languages that suggested the concept'. Under this assumption, we selected two groups of 150 sequential sense candidates – one group with high ranked sum-scores and another with medium ranked sum-scores. Upon a closer inspection of the results (which will be discussed in detail in Section 3), we realized that the simple metric of calculating the number of overlapping languages suggesting any given concept was actu-

0/1/?	No. Langs	Candidate Lemma	Matched Translations	Candidate Synset	English Lemmas, Definitions and Examples
1	2	ϭωπε	'fra|saisir|n', 'fra|saisir|v', 'eng|seize|v', 'eng|seize|n'	02273293-v	confiscate; attach; impound; seize; sequester [take temporary possession of as a security, by legal authority] The police confiscated the stolen artwork

Table 3: Manual Checking (example)

ally a better predictor of correct senses.

3 Results

3.1 Data Sampling

The human evaluation task (detailed in Section 2.4) focused on a blind review of 300 senses. The agreement of both reviewers over this task was 68% (i.e. 204/300 senses). This number refers to agreement in either accepting or rejecting a candidate sense.

To better understand these numbers, one important note to take into consideration, for this evaluation, is the fact that there are no native speakers of Coptic. Because of this, the Coptic knowledge of even the most expert scholar must be considered fragmentary. The amount of exposure to the language most certainly leads to some assumptions about how the language works, including the possible senses a word can have. In addition, the Zipfian nature of language distribution further corroborates our empirical understanding that being exposed to different Coptic texts most certainly has an impact on sense knowledge. In other words, some obscure senses for a given Coptic word might appear so rarely that only scholars who have read certain documents can know about it. This is also why many wordnet projects resort to sense-tagging corpora in order to further evaluate and improve their wordnets. Unfortunately, in such an early stage of our project, we have not yet been able to include this method in our evaluation.

Following the discussion in the paragraph above, we calculated two different measures to evaluate our automated construction method: the percentage of senses accepted by either of the reviewers (i.e. union), and a stricter measure reporting only the percentage of senses accepted by both reviewers (i.e. intersection). These results are presented in Table 4.

No. Langs	Correct(%) Union	Correct(%) Intersect.
1	(n=119) 25%	7%
2	(n=134) 89%	49%
3	(n=40) 98%	63%
4	(n=7) 100%	100%
Total	62% (n=300)	34% (n=300)

Table 4: Human evaluation of the results (union and intersection), by language overlap

Union was calculated by identifying when either of the reviewers assigned a *1* (correct), regardless if the second reviewer assigned *0* (incorrect) or *?* (uncertain). This measure always rewards the user who claims to know the existence of a sense, since the other reviewer might assume or not know of its existence. Intersection was calculated by only counting answers when both reviewers provided answers compatible with the inclusion of that sense. In both measures, when one reviewer assigned a *?* (uncertain), the second reviewer's response was considered the default – in other words, the answer *?* (uncertain) is compatible with both accepting or rejecting an answer, taking the other reviewer's response as final. For example, if one reviewer attested the existence of a sense, but the second reviewer was uncertain, we counted this as "correct" (for both union and intersection measures). In this sample, there was no instance where both reviewers were uncertain.

In addition to the total scores, Table 4 also presents scores grouped by the number of intersected languages that informed each candidate sense. We consider these numbers to be very positive, as they show that the overlap of two or more languages gives a union baseline score of 89%. The intersection of 3

or more languages gives a baseline score of 98% for union (and 63% for intersection). Finally, senses informed by four languages, predict candidate senses 100% of the time.

Despite an unbalanced sample, the numbers still show that our method is principled. The higher the number of intersected languages, the better the prediction accuracy of our method. Furthermore, the overlap of just two languages appears to already be quite informative – reaching a high boundary union score of 89% and a low boundary intersection score of 49%. Even assuming that the union score might include some false positives, a value within this range would suggest a prediction well above chance.

3.2 Wordnet Statistics and Coverage

A summary of the final results produced by our method can be found in Table 5. In total, the second stage of our wordnet includes 218,677 automatically inferred Coptic senses, which is a decent increase from what was generated during the first stage (with less data). In addition, and following the discussion on confidence scores in the section above, Table 5 also shows the number of available senses sorted by the number of languages that intersected that sense.

No. Langs	No. Senses (Stage I)	No. Senses (Stage II)
1	182,883	184,657
2	19,967	30,207
3	3,329	3,575
4	183	238
Total	206,362	218,677

Table 5: Senses per Language Overlap

While the majority of senses was informed by only one language, 34,020 senses (Stage II) are the result of the intersection of two or more languages. If the numbers from Table 4 are confirmed in our ongoing evaluation experiment, then these senses would be expected to have a confidence score of 89% and above.

Table 6 presents how these 218,677 senses are distributed among synsets and parts of speech. In total, the senses are distributed among 25,871 synsets, and fairly well distributed across different parts of speech. On average, there are 7 senses per nominal synset,

POS	No. synsets	No. senses
nouns	13,904	97,527
verbs	7,491	92,019
adjective	3,488	20,723
satellite adj	229	587
adverb	737	7,373
non-referential	22	448
Total	25,871	218,677

Table 6: WN Coverage: Coptic (Sahidic)

and about 12.2 senses per verbal synset. Although many of these senses might not be correct, the high number of senses might also be explained by the many forms a single Coptic lemma can take – which were listed in the dictionaries we used. Many of these forms are, in fact, motivated by morphology, while others are motivated mostly by spelling variation. In the future we would like to dedicate some time to better classify and tag these forms.

The 25,871 synsets cover about 77.4% of the list of 5000 "core" word senses in Princeton WordNet (Boyd-Graber et al., 2006) – a usual measure for coverage of wordnet resources. Further evaluations of coverage at such an early stage of our project might be somewhat difficult. Nevertheless, we decided to test how our wordnet fared in a task of sense matching over open text. A small corpus of $52,789$ word tokens was used, and $20,235$ $(38,3\%)$ out of all tokens were able to find a compatible entry in the Coptic WN. While this coverage may seem low, it fits with other similar experiments done for Ancient Greek (34%) and Latin (33%) (Moritz et al., 2016).

3.3 Release

This Coptic Wordnet is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0)[11]. We have produced OMW tsv files, which can also be used in the Python Natural Language Toolkit (Bird et al., 2009). In addition, and keeping up with the recent requirements to belong to the OMW, we will also release this data using the WN-LMF format[12]. The use of WN-LMF will be essential to access the new Collaborative Interlingual Index (CILI) (Bond et al., 2016) – a language agnostic, flat-structured in-

[11] https://creativecommons.org/licenses/by/4.0/
[12] https://github.com/globalwordnet/schemas

dex to link wordnets across languages. The Coptic WN data can be found on GitHub at https://github.com/coptic-wordnet.

4 Discussion and Future Work

The results from the method of constructing the Coptic WN are promising. We have introduced the method of sense intersection to construct a wordnet which jumpstarts the process of producing a wordnet that is useful for digital humanities tasks. One of the current limitations relates to the expansion approach that uses only dictionary sources. We plan to create and annotate a sense-tagged corpus, alongside the wordnet, so that we can also gain word frequency information, test for coverage and review concepts in context.

We have also argued for the use of union between reviewers as a valid metric since reviewers will not have the same experience with the language. Several different reviewers can positively identify attested concepts and this is in no way a reflection that they do not agree. It can indicate, however, that there is debate within the scholarly community. Because of this, we would like to invite more Coptic scholars into this project, so that the full lexical semantic knowledge can be captured within this resource.

We are currently discussing ways to link to the Coptic Dictionary Online (CDO). This would require following practices of Linked Open Data, where the Coptic WN can be connected to CDO's entries (e.g. via URIs) and, conversely, CDO could be extended to link to related entries from Coptic WN.

Additional on-going work relates to the Collaborative Interlingual Index (CILI). Within the domain of Religious Studies, the PWN has shown numerous shortcomings, including badly formed definitions and an inconsistent hierarchical structure (Slaughter et al., 2018). Following this, we believe that the development of the Coptic WN can be used to contribute to on-going Digital Humanities work within the domain of Religious Studies. This is especially true since much of the content of Coptic sources is primarily religious or theological in nature.

We also believe that the Coptic WN can be a useful resource to further inform multiple Coptic (pre-)processing tools, and help in tasks such as part-of-speech tagging and lemmatization. One such example would be the tools available through the Coptic Scriptorium (Schroeder and Zeldes, 2016; Zeldes and Schroeder, 2016) which includes multiple Coptic processing tools.

The Coptic WN is relevant to the study of purely linguistic research topics, including but not limited to research in lexical semantics. We would like to extend the work of the Etymological Wordnet (de Melo, 2014) to provide a tool for the study of Coptic-related language evolution – including the problems of concept drift (Fokkens et al., 2016) and diachronous meaning shift, concerning how concepts travel through space and time (crossing dialects and even languages), taking slightly different meanings as they move.

Finally, as it was mentioned above, one of the major motivations behind the construction of the Coptic WN was to use its hierarchy for text reuse. In essence, this task is designed to capture short snippets of text similarity (e.g. quoting, summarizing, paraphrasing, translation). TRACER is a system capable of using multiple algorithms to find text reuse across large corpora – which is accomplished by word replacement. Our wordnet can be used to generate possible word replacements including synonyms, hypernyms, hyponyms, or cohyponyms. We are currently exploring hierarchy traversal and replacement strategies that best produce accurate examples of text reuse.

Acknowledgments

We would like to thank the Database and Dictionary of Greek Loanwords in Coptic (DDGLC) Project Coordinator, Tonio Sebastian Richter and Scientific-Technical Staff, Katrin John.

Many thanks to Adam Rambousek for providing the work *in progress* of the future version of the Czech Wordnet.

The TRACER part of this work has been made available by the early career research group eTRAP (No. 01UG1409, 01UG1509) funded by the German Ministry of Education and Research.

We would also like to acknowledge the project "A Linked Digital Environment for Coptic Studies" (NEH Grant No. HAA-261271-18).

References

Pushpak Bhattacharyya. 2017. IndoWordNet. In *The WordNet in Indian Languages*, pages 1–18. Springer.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit NLTK*. O'Reilly Media, Inc.

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1140–1147, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Giulia Bonansinga and Francis Bond. 2016. Multilingual sense intersection in a parallel corpus with diverse language families. In *Proc. of the 8th Global WordNet Conference*, pages 44–49.

Francis Bond and Giulia Bonansinga. 2015. Exploring cross-lingual sense mapping in a multilingual parallel corpus. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 56–61, Trento.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In Christiane Fellbaum Verginica Barbu Mi-titelu, Corina Forăscu and Piek Vossen, editors, *Proceedings of the Global WordNet Conference*, pages 50–57, Bucharest, Romania.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In *Proceedings of the Third International WordNet Conference*, pages 29–36.

Marco Büchler, Philip R Burns, Martin Müller, Emily Franzini, and Greta Franzini. 2014. Towards a historical text re-use detection. In *Text Mining*, pages 221–238. Springer.

Walter E. Crum. 1939. *A Coptic Dictionary*. Oxford University Press, Oxford.

Gerard de Melo. 2014. Etymological Wordnet: Tracing the history of words. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1148–1154, Paris, France, May. European Language Resources Association (ELRA).

Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T Schroeder, and Amir Zeldes. 2018. A linked Coptic dictionary online. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 12–21, Santa Fe, NM.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Christiane Fellbaum. 2017. Wordnet: An electronic lexical resource. *The Oxford Handbook of Cognitive Science*, pages 301–314.

A.S. Fokkens, S. ter Braake, E. Maks, and D. Ceolin. 2016. On the semantics of concept drift: Towards formal definitions of semantic change. In S. Darányi, L. Hollink, A. Meroño Peñuela, and E. Kontopoulos, editors, *Proceedings of Drift-a-LOD*, pages 247–265.

Eitan Grossman. 2014. Transitivity and valency in contact: The case of Coptic. In *47th Annual Meeting of the Societas Linguistica Europaea*, Poznań, Poland, 9. Talk given at a workshop on Transitivity and Valency in Contact: A Cross-Linguistic Perspective (convened by Susanne Michaelis and Eitan Grossman).

Birgit Hamp and Helmut Feldweg. 1997. GermaNet-a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, Spain.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT-the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, Valletta, Malta.

Karel Kučera. 2007. Hyperlemma: A concept emerging from lemmatizing diachronic corpora. *In: Levicka, J., Garabik, R. (eds): Computer Treatment of Slavic and East European Languages*, pages 121–125.

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharyya. 2010. Introducing Sanskrit Wordnet. In *Proceedings on the 5th Global Wordnet Conference (GWC 2010), Narosa, Mumbai*, pages 287–294.

Stefano Minozzi. 2009. The Latin WordNet Project. In *In Anreiter, P. and Kienpointner, M., editors, Latin Linguistics Today. Akten des 15. Internationalem Kolloquiums zur Lateinischen Linguistik, Innsbrucker Beitrage zur Sprachwissenschaft*, volume 137, pages 707––716.

Luís Morgado da Costa and Francis Bond. 2016. Wow! What a useful extension! Introducing non-referential concepts to WordNet. In *Proceedings of the 10th edition of the International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4323–4328, Portorož, Slovenia.

Maria Moritz, Andreas Wiederhold, Barbara Pavlek, Yuri Bizzoni, and Marco Büchler. 2016. Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1859, Austin, Texas, November. Association for Computational Linguistics.

Karel Pala and Pavel Smrž. 2004. Building Czech WordNet. *Romanian Journal of Information Science and Technology*, 7(1-2):79–88.

Benoît Sagot and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Caroline Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *DHQ: Digital Humanities Quarterly*, 10(2). `http://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html(visited:2019-06-28)`.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 82–88.

Laura Slaughter, Wenjie Wang, Luis Morgado da Costa, and Francis Bond. 2018. Enhancing the collaborative interlingual index for digital humanities: Cross-linguistic analysis in the domain of theology. In P. Vossen (Eds.) F. Bond, C. Fellbaum, editor, *The 9th Global WordNet Conference (GWC 2018)*, pages 8–12.

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring Balkanet shared ontology for multilingual conceptual indexing. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 781–784, Lisbon.

Liling Tan and Francis Bond. 2014. NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89.

Piek Vossen. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer, Dordrecht: Kluwer Academic Publishers.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for Coptic. In *Proceedings of the 10th ACL SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH2016)*, pages 146–155, Berlin.

Yingjie Zhang, Bin Li, Xiaoyu Wang, Xueyang Liu, and Jiajun Chen. 2014. Mapping word senses of Middle Ancient Chinese to WordNet. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 446–450. IEEE.

Yingjie Zhang, Bin Li, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2017. PQAC-WN: Constructing a wordnet for Pre-Qin Ancient Chinese. *Language Resources and Evaluation*, 51(2):525–545.

Evaluating the Wordnet and CoRoLa-based Word Embedding Vectors for Romanian as Resources in the Task of Microworlds Lexicon Expansion

Elena Irimia
RACAI
Bucharest, Romania
elena@racai.ro

Maria Mitrofan
RACAI
Bucharest, Romania
maria@racai.ro

Verginica Barbu Mititelu
RACAI
Bucharest, Romania
vergi@racai.ro

Abstract

Within a larger frame of facilitating human-robot interaction, we present here the creation of a core vocabulary to be learned by a robot. It is extracted from two tokenised and lemmatized scenarios pertaining to two imagined microworlds in which the robot is supposed to play an assistive role. We also evaluate two resources for their utility for expanding this vocabulary so as to better cope with the robot's communication needs. The language under study is Romanian and the resources used are the Romanian wordnet and word embedding vectors extracted from the large representative corpus of contemporary Romanian, CoRoLa. The evaluation is made for two situations: one in which the words are not semantically disambiguated before expanding the lexicon, and another one in which they are disambiguated with senses from the Romanian wordnet. The appropriateness of each resource is discussed.

1 Introduction

The work presented in this paper was carried out in the broader frame of the ROBIN[1] project, whose aim is to develop systems and services for using robots in various contexts occasioned by the emerging digital society we live in. Focused on different types of robots, from those specialized in assisting elderly people to software robots dedicated to autonomous or semi-autonomous car-driving, ROBIN has a sub-component that deals with the essential function of human-robot language communication in Romanian (ROBIN-Dialog[2]). The prototype system for verbal interaction with the robot was restricted to several microworlds (see section 3) and the Romanian language resources and tools that are under development at the moment are specifically targeted at describing and serving these microworlds. As a result, the robot should be able to communicate successfully with the human users on topics concerning the specified microworlds and to perform some tasks designated to it, all these activities involving spoken Romanian.

The robot used in this project is Pepper, created by Softbank Robotics[3] and designed to be mass-produced and to become an important actor, improving human everyday life by assisting in different activities. Therefore, Pepper was intended to receive widespread acceptance in society and its shape, size, look and behavior were customized to emulate sociability (Pandey and Gelin, 2018).

A system able to ensure the dialog between a robot and a human user combines different modules dedicated to automatic speech recognition (ASR) (translating the human's vocal message into text), natural language processing (NLP) with its tasks of analysis and synthesis, a dialog management (DM) system and automatic speech generation from text (text-to-speech, TTS) (Tufiş et al., 2019). Except for the DM module, all the other components are language dependent, thus they need training on Romanian data and use (at run-time) Romanian acoustic and language models and a Romanian lexicon enhanced with information about stress, syllabification and phonetic transcription. In the context of the ROBIN-Dialog project, the acoustic and language models could benefit from all the available bimodal training data (see (Barbu Mititelu et al., 2018) for the description of the speech component of the Reference Corpus of the Contemporary Romanian Language - CoRoLa), but tailoring the system to the specific

[1] http://aimas.cs.pub.ro/robin/
[2] http://www.racai.ro/p/robin/
[3] https://www.softbankrobotics.com/us/pepper

microworlds is necessary for preventing semantic ambiguities and misleading. This can be done by designing a wide enough lexicon to cover various ways of expressing the semantic content possible in the targeted microworlds but limited to the semantic fields of interest (e.g., avoiding out-of-context senses for polysemous words). The process of constructing a lexicon - balancing all the needs of the dialog system modules in this specific context - is the focus of this paper.

2 Related work

One of the important steps in human-robot language communication is addressing the problem of creating exhaustive lexicons on different topics, so as to enable the robot to process different ways of expressing the same topic. WordNet is one of the main resources used for the enrichment of different domain specific vocabularies. Hiep Phuc Luong et al. (2009) presented a semi-automatic approach used to disambiguate the senses present in WordNet in order to enrich the vocabulary for ontology concepts in the domain of amphibians.

Other important resources used in expanding the lexicons are the word embeddings vectors extracted from different corpora. The main hypothesis on which the current models of semantic word representations are based is that words occurring in similar contexts have similar meanings (Clark, 2015). Moreover, such representations, most of the times, get closer to human intuitions (Agirre et al., 2009). Therefore, pretrained word embeddings vectors are used to a wide variety of NLP tasks, including vocabulary expansion. For example, Leeuwenberg et al. (2016) and Pennington et al. (2014) demonstrated that word embeddings are able to capture synonyms and analogies. Ono et al. (2015) used synonym and antonym information extracted from thesauri together with distributional information obtained from large scale unlabelled data in order to train word embeddings to capture antonyms.

We are not aware of any work in which the results obtained by using these two resources to be evaluated and this is one of our aims in this paper.

In what follows we define a microworld (section 3), describe the extraction of the lexicon from the screenplays based on two microworlds (section 4), we explain how we have expanded this lexicon using the Romanian wordnet and CoRoLa-based word embeddings (section 5), then we analyze the results obtained and discuss their relevance (section 6) before concluding the paper.

3 Designing microworlds

We define a microworld as an extremely reduced universe that is confined to a well-delimited space, is anchored in time, contains a finite set of objects, is populated by some people and the robot, among which verbal exchanges occur. These exchanges are on topics connected to the microworld. These people know how to collaborate with the robot, while the robot is meant to learn how to collaborate with the people. The learning phase of the robot needs to cover the following topics: the space topology, recognizing the people in the microworld, understanding natural language and reacting to it, which presupposes the ability to formulate an oral response to a human's command or to execute the command within the microworld.

For the present paper, we focus on two microworlds imagined for the interaction with the robot, which is attributed an assistive role: a private home and a research laboratory. In the former, the robot will help people to take care of themselves: undergo some measurements of relevance for their condition (e.g., measure the blood glucose), communicate the value of a certain measurement at a specific time, keep track of them during a longer period of time, display their evolution for a specified period of time, remind people what medication to take, when to do it and even where the medication is, etc. In the latter microworld, the robot will be the host for visitors of the laboratory, greeting, welcoming known people, introducing itself to the new visitors. It will also transmit verbal messages from one person to another, provided that they are present in the laboratory.

After designing the microworlds, a first preparatory step in the process of teaching the robot to interact with people is the creation of a screenplay for each microworld, with verbal interactions and actions. Our focus here is the former, namely the possible dialogues in natural language (Romanian) between people and the robot. A set of possible actions the robot could do in each microworld was identified and possible topics for verbal exchanges corresponding to them were created. This is to be understood at a conceptual level, while all the possible ways of expressing these topics are registered as their lexicalized forms. The robot must be able to understand them all, that is why it has

to be taught a large vocabulary and numerous syntactic structures. For example, the following ways of asking the robot if it knows the person called George were identified in Romanian: "Îl ştii pe George?" / "Ştii cine e/este George?" / "Îl cunoşti pe George?".

The human-robot communication is confined to the entities and possible activities in the respective microworlds. The robot will never initiate the dialogue. It is there to help the human by answering a question or carrying out a task. The robot is not to be understood as a repository of world knowledge. It can only answer questions about the entities in the respective microworld (such as "Este George în sala 306?" (Is George in room 306?), iff George is known to the robot, i.e. the latter was trained to recognize John's face, and the robot knows where room 306 is in the space whose topology it was taught). The human will give the robot as much information as necessary for performing the task, will formulate it concisely, clearly, avoiding obscurity and ambiguity.

Consequently, the vocabulary used in such dialogues cannot be conceived as specific to a domain. Terms specific to a domain, such as "blood pressure", "blood glucose", etc. may occur in the microworld with Pepper playing an assistive role in a private home. However, they are terms that have penetrated the general language, are familiar to every speaker, thus their domain specificity being drastically reduced.

4 Extracting the lexicon

Based on the screenplays mentioned in section 3, an initial list of lemmas was created. The screenplays serve as a corpus that was processed with the TEPROLIN platform (Ion, 2018), using the TTL module to normalize, sentence split, tokenize, POS-tag and lemmatize the data. Then, a list of all the unique lemmas in this annotated corpus was extracted, to serve as a starting point for the enhancement process described in the next sections. We treat differently the content words and the function words: we teach the robot the limited list of all the function words in Romanian, but we want to control the (virtually unlimited) set of content words the robot has to deal with, to stay in the discourse microworlds. Therefore, we set apart a list of 190 content lemmas to work with in our experiments.

For the final form of the extended ROBIN lexicon (containing a comprehensive list of lemmas that need to be represented in our resource), we added:

- all the morphological variants (i.e., inflected forms) of the words that were in the initial lexicon and of the words that were extracted using the two resources, by looking-up in an in-house extensive lexicon of Romanian (TBL, comprising 1.2 million hand-validated entries);
- all the Romanian function words (pronouns, determiners, articles, prepositions, conjunctions and some numerals, recovered also from TBL, 2382 entries);
- the information about stress, syllabification and phonetic transcription, generated with the TTS (see (Stan et al., 2011)) module from TEPROLIN.

5 Expanding the lexicon

The lexicon extracted from screenplays, as presented in section 4, was expanded with the purpose of enhancing it with words capable of capturing the lexical and syntactic varieties of the language. In order to extend the lexicon, two resources were used: the Romanian WordNet (RoWN) (Tufiş and Barbu Mititelu, 2015) and precalculated word embeddings vectors based on the CoRoLa corpus (Păiş and Tufiş, 2018). From the initial lexicon we chose only those lemmas that occur both in RoWN and in CoRoLa. We call this subset L and it contains 178 content lemmas. The difference between the whole set of content lemmas extracted from the screenplays and L is represented by foreign words (e.g. *cool*), proper nouns (e.g. *George*), and several content words not implemented in RoWN (e.g. the adverb *româneşte* "in a Romanian way").

We ran another experiment in which we semantically disambiguated the words in L. For six of them, no sense implemented in RoWN is the one with which the respective words are used in the screenplays. Consequently, we obtained a smaller set of 172 disambiguated words, which we call L' ($L' \subset L$).

5.1 Using RoWN

RoWN has been created since the BalkaNet project (Tufiş et al., 2004). During this project, the aim was to cover the initial Base Concepts set

from EuroWordNet (Vossen, 2002). All their hyperonym synsets from Princeton WordNet (Miller, 1995; Fellbaum, 1998) (PWN) were implemented into RoWN. The literals are translated and their list is enriched with the help of synonymy and other dictionaries; the synsets glosses are mainly taken from the corresponding Romanian explanatory dictionary entries or, when such definitions could not be found to match exactly the PWN sense, the Romanian glosses were the translation of the English ones. More than 400 concepts considered specific to the Balkan area were included in the BalkaNet wordnets as synsets for which a hypernym was found among the synsets already implemented in the wordnets (Tufiş et al., 2004). The further quantitative enrichment of RoWN targeted the lexical coverage of various corpora collected over time (Tufiş and Barbu Mititelu, 2015). At the moment RoWN contains 59,348 synsets in which 85,277 literals (representing 50,480 unique ones) occur, out of which 20,031 (i.e., 17,816 unique ones) are multiword literals, accounting for 23.5% of the total number of literals (i.e., 35.3% unique ones). The qualitative enrichment focused on in-line importing of the SUMO/MILO concept labels (Niles and Pease, 2001), connotation vectors for synsets (Tufiş and Ştefănescu, 2012), derivational relations (Barbu Mititelu, 2013) and annotation of verbal synsets with labels specific to various types of multiword expressions, adopting the same framework (the PARSEME annotation guidelines) (Barbu Mititelu and Mitrofan, 2019). RoWN can be queried at `http://relate.racai.ro/` and at `http://dcl.bas.bg/bulnet/`, the latter offering also the possibility of visualizing aligned wordnets (Rizov et al., 2015).

Since wordnets are rich knowledge bases in which words and synsets are linked by lexical and semantic relations, we used the Romanian wordnet to attain broader lexical and semantic coverage of the scenarios created for the two microworlds, by extracting from it words semantically related to the ones in the screenplays. We call *semantically related words* those words occurring in synsets that establish one of the following relations with the synset(s) to which the words in L or L' belong: hypernym, cause, entailment, similar_to, verb_group, also_see, near_participle, near_derived_from, near_eng_derivat, near_pertainym, near_antonym. All relations whose name is prefixed with *near_* are considered language specific. They exist in PWN without this "prefix", i.e. they are participle, derived_from, eng_derivat, pertainym, antonym, respectively. When transferred into the RoWN this prefix served as a way of signaling that for Romanian the relation may not hold, although some semantic relatedness exists.

We disregarded for our task the following relations: hyponym, instance_hypernym, instance_hyponym, member_holonym, part_holonym, substance_holonym, member_meronym, part_meronym, substance_meronym, attribute, domain_TOPIC, domain_REGION, domain_member_USAGE, domain_member_REGION, domain_USAGE, domain_member_TOPIC. The reason for disregarding hyponymy is that a hyponym cannot replace its hypernym (Cruse (1986) showed that implication is unilateral in the case of hyponymy). Kleiber and Tamba (1990) showed that in the case of holonymy-meronymy, the relation of implication holds only when the predicate expresses location or time: in the following examples, (1) implies (2) and both of them express location. However, (3) does not necessarily imply (4), where the same words are used without reference to a place.

(1) The fly is on the child's *elbow*.

(2) The fly is on the child's *arm*.

(3) The child's *elbow* is on the table.

(4) The child's *arm* is on the table.

That is why we disregarded all types of holonymy and meronymy in wordnet. Instances are not relevant for our microworlds, just like all domain-related relations: the scientific domain to which a word belongs (the domain_TOPIC and domain_member_TOPIC relations), the geographical or cultural domain of a concept (the domain_REGION and the domain_member_REGION relations) or the usage of a word (the domain_USAGE and the domain_member_USAGE relations)[4]. As can be noticed in the definition of our understanding of semantically related words, we do not explore the wordnet graph on more than one level to look for related words, so that to avoid expanding the lexicon with too general words or with words seman-

[4]Some of these relations are language-specific, so there is no need to consider them; they were automatically transferred from PWN, without checking their applicability to Romanian data.

tically too distant from the ones in *L*.

5.2 Using Word Embeddings Vectors

It is known that neural word representations have the ability to capture useful semantic properties and linguistic relationships between words (Bakarov, 2018). On the basis of the Romanian reference corpus CoRoLa, which contains almost 1 billion words distributed in different text types and domains, and using distributed neural language model word2vec (Mikolov et al., 2013), high quality word embeddings vectors were generated (Păiş and Tufiş, 2018). We extracted and used the first 10 nearest neighbours to a given lemma in the word embedding space (semantically similar lemmas). The neighbours were obtained by computing a similarity score between the given lemma and the rest of the words in the vocabulary. The similarity score was obtained by the calculation of the cosine of the angles between two vectors; the closer the score is to 1, the more similar the two lemmas are.

6 Analysis of the Words Extracted from the Two Resources

The aim of this analysis is to discuss the relevance of the words extracted using the resources described in section 5 above for the task of extending the lexicon coverage for the two screenplays. A word is considered *relevant* if one can imagine a sentence that could fit the screenplays either for rephrasing an existing sentence or for completing the screenplay with further exchanges.

Both resources face the challenge of overgeneration: words tend to have more senses in corpora, while in wordnets they occur with many if not all their senses. However, in the screenplays they are mostly used with one of their senses, as the microworld could be thought of as a closed, limited domain. Having the expansion of *L* as a purpose, we discuss the results obtained without semantically disambiguating the words in the initial lexicon and then the results obtained after semantically disambiguating them (L') .

For the sake of clarity, let:

- *n*=178 be the number of lemmas in *L*

- n'=172 be the number of lemmas in L'

- *A* be the set of *n* lemmas from *L* together with the set of lemmas of their related words in RoWN (the number of related words for each initial lemma varies and depends on the number of synsets identified as relevant and on their length):

L_lemma_1: $rownlemma_{1,1}$, ..., $rownlemma_{1,i}$, ...

L_lemma_2: $rownlemma_{2,1}$, ..., $rownlemma_{2,j}$, ...

...

L_lemma_n: $rownlemma_{n,1}$, ..., $rownlemma_{n,k}$, ...

- similarly, A' be the set of n' lemmas from L' together with the set of lemmas of their related words in RoWN;

- *B* be the set of *n* lemmas from *L* that were identified in CoRoLa together with the set of lemmas extracted from the word-embedding vectors (the number of related words for each initial lemma is set to 10, see section 5.2):

L_lemma_1: $welemma_{1,11}$, ..., $welemma_{1,10}$

L_lemma_2: $welemma_{2,1}$, ..., $welemma_{2,10}$,

...

L_lemma_n: $welemma_{n,1}$, ..., $welemma_{n,10}$

- similarly, B' be the set of n' lemmas from L' that were identified in CoRoLa together with the set of lemmas extracted from the word-embedding vectors.

We applied the following set operations to the two resources in order to find:

1. lemmas that could be obtained from both resources ($A \cap B$, and $A' \cap B'$ respectively): $\forall$ L_lemma_i, $\forall$ $rownlemma_{i,j}$, $rownlemma_{i,j}$ is in $A \cap B$ if $\exists$ *k* so that $rownlemma_{i,j} = welemma_{i,k}$;

2. lemmas that were obtained from RoWN but not from word embeddings vectors ($A \setminus B$, $A' \setminus B'$ respectively) and lemmas obtained using word embeddings vectors but not RoWN ($B \setminus A$, $B' \setminus A'$ respectively): e.g. $\forall$ L_lemma_i, for each $rownlemma_{i,j}$, $rownlemma_{i,j}$ is in $A \setminus B$ if there is no *k* so that $rownlemma_{i,j} = welemma_{i,k}$;

In what follows we discuss the results of these set operations.

6.1 Relevance of different word types for the screenplays

In the process of expanding the initial lexicon with new words, different types of words can prove their usefulness. The relevance of synonyms is self-evident. Hypernyms are known to replace a word in a context (Cruse, 1986), so their relevance is also clear. As far as antonyms are concerned,

they may allow for rephrasing the sentence with a negative form of the verb in Romanian: here is an example with the antonyms *continua* (go on) and *înceta (stop)*:

(5) a. *Continuă* să mergi! (*Go on* walking!)

b. *Nu înceta* să mergi! (*Don't stop* walking!)

Here is a set of examples showing the relevance of words derived from the word in the initial lexicon: the pair is *căuta* (verb, *to search*) - *căutare* (noun, *search, searching*), where the latter is derived from the former:

(6) a. Am *căutat* în camera 3316. (I *searched* in room 3316.)

b. Am făcut *căutarea* în camera 3316. (I made the *search* in room 3316.)

6.2 Words found in both resources

In this section we look, on the one hand, at the intersection of the sets *A* and *B*, showing the results without previous semantic disambiguation of the words in *L* (see column A ∩ B), and, on the other hand, of the sets A' and B' (see column $A' \cap B'$). We started our analysis with this step because we assumed words identified by both resources are probably the most interesting ones in terms of similarity with the initial lemmas, as they are enforced by both resources. Initial lemmas in Table 1 are words from L, respectively from L', for which the intersection of the set of words extracted from RoWN and of the set of words extracted from word embeddings is not null. Comparing the number of initial lemmas in the set intersections with the number of elements in *L* (178) and L' (172), we notice that for only 64% (Table 1 line 1 column 2) of the words in *L* and for 49% (Table 1 line 1 column 3) of the words in L' we found words common to the both resources. This brings us to the conclusion that the two resources complement each other, rather than confirming each other's decisions in our task.

Validating the words found in the two resources, we notice that the rate of acceptance is quite high (95% and 100%, respectively - see no. of validated words from the no. of found words in Table 1), which confirms our intuitions that words identified by both resources are highly probable candidates. Adding the disambiguation criterion brings the probability of finding a good word in the intersection almost to 100%, eliminating all the bad results.

The validated words for the experiment involving semantic disambiguation are a subset of the validated words in the experiment without disambiguation. One might have expected these sets to be identical, i.e. only the synsets to which the disambiguated words belong offer relevant related words. However, the explanation for accepting (in the non-disambiguated setting) related words to other senses of the initial lemmas is that we understand synonymy in a broader way: any word that may imply *any* syntactic reorganization of the sentences in the screenplay, as long as the compositional meaning of the sentences is *almost* the same[5].

Types of words	**A ∩ B**	**A′ ∩ B′**
no. of initial lemmas	114	85
no. of found words	211	140
no. of validated words	201	140
% of validated words	95	100
no. of validated empty lists	0	0
no. of synonyms	103	73
no. of antonyms	25	14
no. of derivations	54	42
% of synonyms	51	52
% of antonyms	12	10
% of derivations	27	30

Table 1: Nondisambiguated vs. disambiguated sets intersection.

Regarding the types of words that are found, most of them are synonyms of the words in the scenarios. More synonyms are found in the first experiment, which means that senses that were not chosen in the word sense disambiguation phase of our work could also contribute relevant words, even synonyms. For example, for the verb *considera* (consider) the following related words were found and validated in the first experiment: *aprecia*, *susține*, *crede*, whereas after disambiguating the initial lemmas, the only related word found was *crede*. However, although *susține* could be accepted only for some contexts, we consider that *aprecia* is definitely worth being included in the lexicon. One explanation for this situation is the fine granularity of wordnets, which makes some senses to be too closely related and expressed by the same words. As a consequence of this granularity, several senses of a word should have

[5]Compare this with the definition of synonyms in (Miller, 1995): "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value".

been accepted in the disambiguation task, while at the RoWN level, the synsets should have been richer, sharing more literals. Besides synonyms, antonyms[6] were also found, although with a low rate. The high number of derived words reported for both experiments shows the importance of derived words in rephrasing the same semantic content, recognized by the two resources.

6.3 Words found only in RoWN

The next group of results we present are those that were found using RoWN, but not in the word embeddings. The data is summarized in Table 2.

Types of words	**A \ B**	**A′ \B′**
no. of initial lemmas	178	172
no. of extracted words	5130	1651
no. of validated words	843	840
no. of validated empty lists	26	33
no. of synonyms	563	469
no. of antonyms	27	31
no. of derivations	45	48
% of synonyms	66	55
% of antonyms	3	3
% of derivations	5	5

Table 2: Related words found only in RoWN.

A first remark is the large number of related words extracted from RoWN: for each word around 27 words, on average (see line 2 in Table 2), were extracted, due to the high number of relations used. However, many useless words (84%, see the no. of validated words as a percent of the no. of found words in Table 2) were extracted in the first experiment, whereas, as expected, the situation improved in the second experiment, in which only half of the extracted words were useless. We analyzed the invalidated words extracted: some of them are extracted by means of lexical not of semantic relations (see the discussion about relations prefixed with *near_* in the RoWN in subsection 5.1). Others are hypernyms that would seem unnatural in the screenplays, contrary to the linguistic expectations. The same inadequate usage characterizes some verbs from the same group as some initial verbal lemma. Although with these relations we also extract words that are useless for our task, we cannot eliminate them from the list of relations we need for expanding the lexicon, because they also return good words. We could not come up with any heuristic for deciding when to accept such relations and when to neglect them.

It is noteworthy that synonyms represent more than half of the total number of useful related words found in RoWN. Given the reduced average synset length in RoWN (that is 1.46, see (Tufiş et al., 2013)), we infer that the words in L and L' belong to longer synsets. This is something one could have expected, given the rather general character of most words occurring in the screenplays (see section 3 above for a discussion about the vocabulary of microworlds). Such words, belonging to the core vocabulary used by all people, are known to develop synonyms, derived words, to enter more expressions, to be semantically rich.

From the number of validated empty lists in Table 2 we understand that for those words no extracted word could be accepted as semantically related.

6.4 Words found only with word embeddings vectors

B \ A statistics	**B \ A**	**B′ \A′**
no. of initial lemmas	178	172
no. of extracted words	1600	1554
no. of validated words	737	656
no. of validated empty lists	21	19
no. of synonyms	46	40
no. of antonyms	47	32
no. of derivations	101	81
% of synonyms	6	6
% of antonyms	6	5
% of derivations	14	12

Table 3: Nondisambiguated vs. disambiguated B-A statistics.

For 178 initial lemmas, $B \setminus A$ extracted 1660 (and $B' \setminus A'$ extracted 1554) supposedly similar words from CoRoLa using word embeddings, from which 737 (and 656 respectively) were validated. We notice that although the number of extracted words is reduced considerably compared to the ones extracted from wordnet in the non-disambiguated setting, the number of validated words is lower, but close (737 vs. 843, 656 vs. 840). This implies that the two resources quantitative contribution to expanding the lexicon is similar, and, if done in the disambiguated setting, in-

[6]See (Ono et al., 2015) for extracting antonyms using word embeddings.

volves much less validation effort. While the differences in numbers and percents for the contribution of antonyms is negligible, what is evident in the data is that most of the synonyms come from the wordnet (see the 66% percent from Table 2 versus the 6% percent from Table 3) and most of the derivations come from the corpus (see 12-14% in Table 3 versus 5% in Table 2). Examples of initial lemmas whose list of extracted words abounds in derivated words are "robot" (robot) and "cântări" (weigh)[7]:

- *robot*: *robotiza*, computer, *robotic*, *roboţel*, *robotizat*, *robotică*, *robotizare*;

- *cântări*: *recântări*, gram, *recântărire*, greutate, *cântărit*, *cântărire*.

7 Conclusions

The experiments presented here prove the adequacy of RoWN and CoRoLa-based word embeddings for expanding a lexicon so as to ensure a wider lexical and syntactic coverage, meant to ensure the ability of a robot to understand humans in specific microworlds.

We worked with a list of 178 non-disambiguated initial lemmas (L) and with a list of 172 disambiguated initial lemmas (L') and we obtained a number of 1,694 unique lemmas ($A \cap B$) and, respectively, a number of 1,287 unique lemmas ($A' \cap B'$), extracted from RoWN and CoRoLa. The amount of validation work is substantially decreased in the disambiguated setting (even with the supplementary disambiguation costs) and, while such a solution is preferable in similar tasks, the loss in interesting, valid extracted words corresponding to different senses of the lemmas has to be taken into account. A solution would be to accept more senses for a specific lemma in the disambiguation phase, when the human validator considers it necessary. Words identified as related by the two resources are most probably good candidates, while in the disambiguated setting the probability of their usefulness is close to 100%.

As far as the contribution of different relations in wordnet is concerned, the way in which the task was formulated seems to have determined the acceptance of mainly synonyms (even if in a larger sense than that accepted by the wordnet projects), antonyms and words derived from the initial ones. Although a hyponym can be replaced by its hypernym, the need for precision can prevent this, whereas larger contexts would encourage this replacement as a means of avoiding repetition, which was not our concern in this experiment, as we did not focus on context, but on single sentences. The majority of synonyms was extracted from the wordnet, while the derivatives are mostly obtained from the corpus.

[7] Only the italicized words are derivationally related to the given ones.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 19–27, 2009.

Amir Bakarov. 2018. A survey of word embeddings evaluation methods. In *arXiv preprint arXiv:1801.09536*, 2018.

Verginica Barbu Mititelu. 2013. Increasing the effectiveness of the Romanian Wordnet in NLP applications. *CSJM*, vol. 21, no. 3, 320–331.

Verginica Barbu Mititelu, Dan Tufiş, and Elena Irimia. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of LREC 2018*, Japan, p.1178-1185.

Verginica Barbu Mititelu and Maria Mitrofan. 2019. Leaving No Stone Unturned When Identifying and Classifying Verbal Multiword Expressions in the Romanian Wordnet. In *Proceedings of the 10th Global WordNet Conference*, Wroclaw, Poland, (this volume).

Alan D. Cruse. 1986. *Lexical Semantics*. Cambridge, CUP.

Christiane Fellbaum. 1998, ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Paul Grice. 1975. Logic and conversation. In Cole, P.; Morgan, J. *Syntax and semantics. 3: Speech acts*. New York: Academic Press. pp. 4158.

Radu Ion. 2018. TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In *Proceedings of the 13th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Iaşi, 22-23 November 2018

Georges Kleiber, and Irène Tamba. 1990. L'hyponymie revisité: inclusion et hiérarchie. *Langages*, no. 98: L'hyponymie et l'hyperonymie, Larousse.

Artuur Leeuwenberg, Mihaela Vela, Jon Dehdari, and Josef van Genabith. 2016. A minimally supervised approach for synonym extraction with word embeddings. In *The Prague Bulletin of Mathematical Linguistics*, 111-142, 2016.

Hiep Phuc Luong, Susan Gauch, and Mirco Speretta. 2009. Enriching concept descriptions in an amphibian ontology with vocabulary extracted from wordnet In *22nd IEEE International Symposium on Computer-Based Medical Systems*, 1-6, 2009.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

George A. Miller. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM Vol. 38, No. 11: 39-41.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, 2–9.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 984–989.

Amit Kumar Pandey and Rodolphe Gelin. 2018. A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of its Kind. *IEEE Robotics Automation Magazine*: 40–48.

Vasile Păiş and Dan Tufiş. 2018. Computing Distributed Representations of Words using the CoRoLa Corpus. In *Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science*, vol. 19: 185–191.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543, 2014.

Borislav Rizov, Tsvetana Dimitrova, and Verginica Barbu Mititelu. 2015. Hydra for Web: A Multilingual Wordnet Viewer. In *Proceedings of the 11th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, Iaşi, Romania, 19–30.

Rion Snow, Sushant Prakash, Daniel Jurafsky, Andrew Y. Ng. 2007. Learning to Merge Word Senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*: 1005–1014.

Mirco Speretta, and Susan Gauch. 2008. Using text mining to enrich the vocabulary of domain ontologies In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, 549–552, 2008.

Adriana Stan, Junichi Yamagishi, Simon King, and Matthew Aylett. 2011. The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate In *Speech Communication* vol.53 442-450.

Dan Tufiş, Dan Cristea and Sofia Stamou. 2004. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.* Journal on Information Science and Technology, Special Issue on BalkaNet, Romanian Academy, 7 (1-2), 7–41.

Dan Tufiş and Dan Ştefănescu. 2012. Experiments with a differential semantics annotation for WordNet 3.0. In *Decision Support Systems* vol.53, no. 4, 695–703.

Dan Tufiş, Verginica Barbu Mititelu, Dan Stefanescu, and Radu Ion. 2013. The Romanian wordnet in a nutshell. *Language Resources and Evaluation*, 47: 1305–1314.

Dan Tufiş, and Verginica Barbu Mititelu. 2015. The Lexical Ontology for Romaniany. In Nuria Gala, Reinhard Rapp and Gemma Bel-Enguix (eds.), *Language Production, Cognition, and the Lexicon*: 491–504.

Dan Tufiş, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Radu Ion, and George Cioroiu. 2019. Making Pepper Understand and Respond in Romanian. In *Proceedings of CSCS22* (in press).

Piek Vossen. 2002. *EuroWordNet general document version 3.* Report, University of Amsterdam.

Towards linking synonymous expressions of compound verbs to Japanese WordNet

Kyoko Kanzaki
Toyohashi University of Technology
Aichi Japan
kanzaki@imc.tut.ac.jp

Hitoshi Isahara
Toyohashi University of Technology
Aichi, Japan
Isahara@tut.jp

Abstract

This paper describes our project on Japanese compound verbs. Japanese "Verb (adnominal form) + Verb" compounds, which are treated as single verbs, frequently appear in daily communication. They are not sufficiently registered in Japanese dictionaries or thesauri. We are now compiling a list of the synonymous expressions of compound verbs in "compound verb lexicon" built by the National Institute of Japanese Language and Linguistics. We extracted synonymous words and phrases of compound verbs from five hundred million Japanese web corpora. As a result, synonymous expressions of 1800 compound verbs were obtained automatically among 2700 in the "compound verb lexicon". From our data, we observed that some compound verbs represent not only motion but also additional nuances such as an emotional one. In order to reflect the abundant meanings that compound verbs own, we will try to think of a link of synonymous expressions to Japanese wordnet. Concretely, in the case of synonymous phrases, we try to link adverbial expressions which are a part of phrases to the adverbial synset in Japanese wordnet.

1 Introduction

Japanese "Verb (adnominal form) + Verb" compounds, which are treated as single verbs, frequently appear in daily communication, however, they are not sufficiently registered in Japanese dictionaries or thesauri.

The Japanese "compound verb lexicon" was constructed by the National Institute for Japanese Language and Linguistics (NINJAL) (https://db4.ninjal.ac.jp/vvlexicon/). It has the meanings, example sentences, syntactic patterns and actual sentences from the corpus that they possess. However, it has no relation with another words, such as synonymous words and phrases.

We detect them automatically as much as possible in order to help humans find out synonymous expressions that they may fail to bring to mind and then manually compile a lexicon of synonymous expressions of Japanese compound verbs.

In this paper, firstly we explain how to build the list of compound verbs and their synonymous words and phrases, and then consider what should be considered for linking to the Japanese wordnet based on our obtained result.

2 Related researches

So far, in NLP domain researches on complexed verbal meaning have treated multi word expressions in order to distinguish a literal meaning with the metaphoric meaning, but their purposes are word sense disambiguation or generation of compounding (Sag et.al.2002; Hashimoto and Kawahara 2008 and so on). In Japanese, Uchiyama and Ishizaki (2003), and Uchiyama and Baldwin (2004) investigated ambiguities of compound verbs and tried to find the generation rules. As a resource on phrases, Tanabe et.al (2014) built Japanese Dictionary of Multi word Expressions.

However, works on the organization of words and phrases are few. Our goal is to compile a list of words and verbal phrases with linking similar relations by using both automatic and manual ways.

3 Japanese compound verbs

The morphological form of a compound verb is a combination of a first verb in an adnominal form and a second verb coming after it, as in *hikari* (adnominal form)-*kagayaku* (give.off.light

& shine) ‘shine like the sun’, *nage* (adnominal form)-*ireru* (throw & put.in) ‘throw in’.

Japanese compound verbs are divided into two types in terms of syntactic and morphological analysis; syntactic compound verbs and lexical compound verbs (Kageyama 1993).

Kageyama (1993) says that syntactic compound verbs are easily recognizable and interpretable due to some characteristics, that is, a limitation of a variety of the second verbs, no restriction on the first verbs and so on. For example, “*utai_hajimeru* (sing & start, ‘start singing’” “*hanashi_hajimeru* (speak & start, ‘start speaking’)” “*hashiri_hajimeru* (run & start, ‘start running’)” and so on. We can generate varieties of “Verb_*hajimeru* (Verb & start, ‘start V_ing’)”. The second verbs of syntactic compound verbs are mainly aspectual verbs and also are limited to 30 verbs which are classified into 9 categories; inception, continuation, completion, incompletion, excessive action, habitual, reciprocal action and potential.

We exclude the syntactic compound verbs and treat only lexical compounds which tightly combine two verbs as one word and not productive compared to syntactic compound verbs.

4 Extracting synonymous expressions from corpus

4.1 Data

We use “five hundred million Japanese texts gathered from web” produced by Kawahara et.al. (2006) as corpus for extracting synonymous words and phrases. The data has been processed into morphologically analyzed data.

As for compound verbs for an extraction of synonymous expressions, we dealt with compound verbs registered in “compound verb lexicon” built by NINJAL. The total number of compound verbs in this lexicon is 2700, and each one has meanings, syntactic patterns and example sentences.

4.2 Procedure

For the first step, we extracted synonymous words and phrases of compound verbs from corpora.

Step1: Preprocessing

Some compound verbs can be paraphrased into phrases. Therefore we concatenated modification relations between verbs and adverbial words and made them into units which we treated as “verbs” (e.g. correctly / understand >>> “correctly understand”). Also compound verbs which are not registered in a dictionary of a morphological analyzer need to combine two verbs (verb in adnominal form + verb (*nage* ‘throw’/ *ireru ‘put in’* >>> *nageireru ‘throw in’*)).

For the first experiments, we had put all words segmented by Japanese morphological analyzer and calculated the similarity between compound verbs and another verbs by cosine similarity measure, but the result was not good. We obtained many unrelated words for each compound verb. Therefore, we decided to exclude the passive and causative form and so on which make an alternation of case markers.

After that, we generate the list of the sets of a noun, a verb and a case marker, which is an input data for vectorization.

Step2: Vectorization and cosine similarity

We performed vectorization of all verbs and nouns in the web corpus by using word2vec (Mikolov 2013), one of the deep learning methods. The learning model of word2vec that we used is CBOW (contiguous bag of words). Then we explored the semantic distance between verbs (including verbal phrases) by cosine similarity. For each compound verb, the verb and verbal phrases were arranged in descending order from the highest score.

Step3: Creating a list of candidates of synonymous expressions

For each compound verb, 2000 similar expressions were chosen in order from the highest score of cosine similarity. Here, the lists of synonymous expressions for each compound verb were created. However, in this list, the polysemy of compound verbs was not taken into account. That is, the synonymous expressions of compound verbs were stored together without distinction of their polysemous meaning in this list.

Step4: Shrinking synonymous expressions and getting clusters for each compound verb

A rough diagram of the process to get categories for each compound verb is shown below.

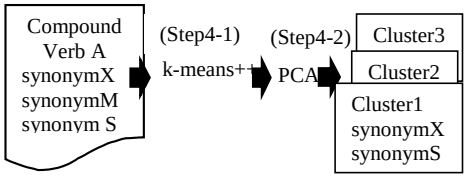


Figure1. The process of getting clusters

Step4-1) Decreasing candidates of synonymous expressions

At the beginning, each compound verb has candidates of 2000 synonymous expressions extracted from the web corpus. To easily determine plausible synonymous expressions, we decrease 2000 synonyms in a step-by-step approach by iterating the k-means++ (Arthur and Vassilvitskii 2007). As for input data for k-means++, we used vectors of synonymous expressions obtained by word2vec. Centroids were calculated by Euclidian distance. Our procedure is described below.

P1. Firstly, we set 64 clusters for the k-means++. In this stage, 2000 expressions are classified into 64 clusters.

P2. For each cluster, we extract 10 expressions with the highest cosine similarity values. In this stage, we narrow down to 640 expressions (64 clusters * 10 expressions). We considered that cosine similarity values between a compound verb and other expressions would be plausible to choose synonym candidates for a compound verb.

P3. We iterate the same process as the step P2 for 640 expressions. In this stage, we set 10 clusters. For each cluster, we extract 10 expressions with the highest similarity values in the list. As a result, we obtained 100 synonym expressions classified into 10 clusters (10 clusters * 10 expressions). This data is used as input data for Principal Component Analysis (PCA).

Step4-2) Finding the number of senses for each compound verb by PCA

The 100 synonymous expressions are now classified into 10 clusters for each compound verb. However, the number of senses of a compound verb differs from each other. We tried to detect the appropriate numbers of senses of each compound verb by using principal component analysis (PCA). We implemented PCA with 100 expressions for each compound verb.

5 Result

As a result, synonymous expressions for 1800 compound verbs were obtained automatically.

The following is a result of "持ち込む (*mochi_komu*)" derived from PCA. The synonymous expressions that we decided are surrounded with circles.

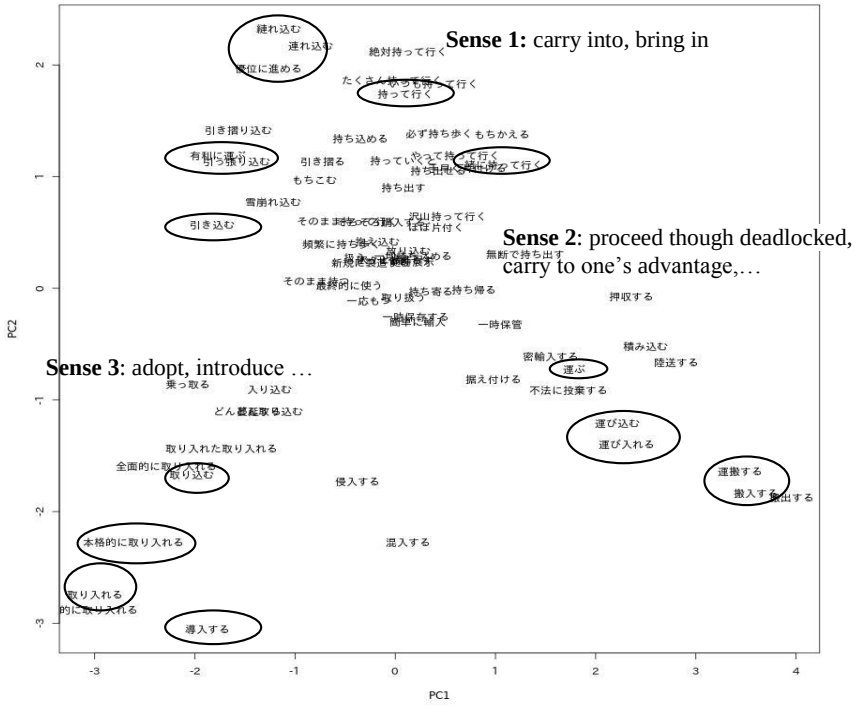


Figure 2. A distribution of synonymous expressions of "持ち込む (*mochikomu*) " derived from PCA

The list of synonymous expressions of “持ち込む (*mochikomu*)” is as follows.

Sense1 :
運び入れる (*hakobiireru*, ‘carry in’)
搬入する (*han’nyu_suru*, ‘carry to’)
運び込む (*hakobikomu*, ‘carry into’)
運ぶ (*hakobu*, ‘carry’)
持って行く (*motte_iku*, ‘take’)
一緒に持っていく (*isshoni motteiku*, ‘bring in’, ‘take … with [person]’.)
Sense2:
もつれこむ (*motsurekomu*, ‘to proceed though deadlocked’)
優位に進める (*yuui_ni* (adj in adverbial form) *susumeru*, ‘advance [the match]’)
有利に運ぶ (*yuuri_ni*(adj in adverbial form) *hakobu*, ‘carry to one’s advantage’)
引っ張り込む (*hipparikomu*, ‘pull’)
引き込む (*hikikomu*, ‘pull’)
Sense3:
取り込む (*torikomu*, ‘incorporate’)
取り入れる (*toriireru*, ‘incorporate’),
導入する (*dounyuusuru*, ‘introduce’)

4.1. Evaluation for 40 compound verbs

In order to predict how many suitable synonyms and clusters we’ve semi-automatically obtained by our method, we evaluate our results manually. For 40 compound verbs which are the most frequent compound verbs in our corpus, 4 examinees evaluated the suitability for synonymous expressions classified in each cluster. We evaluated the expressions for each cluster by comparing them to sense descriptions of the compound verb in CVL. As a result, 59% of extracted words are evaluated as synonyms. And we evaluated the suitability of clusters created by our method. We compared the clusters to sense descriptions of the compound verb in CVL. As a result, 65% of extracted clusters are evaluated as representing the proper meaning of the compound verb. For example, “*Furikaeru* (*Furu-+kaeru*)” has a single meaning like “look behind with twisting body” in CVL. Our method could extract another meaning, i.e. “think back on the previous episode.”

In terms of a recall, the total number of meanings of 40 compound verbs registered in CVL is 64. Among them 14 meanings could not be obtained by our methods (22%). These 14 meanings are included in 13 compound verbs.

5.1 Add more synonymous expression to the list

We selected these synonymous expressions from 100 synonyms candidates whose similarity score is 10 from the top in 10 clusters obtained from k-means++ (referred to step4-1). They are distributed on the PCA (step4-2). By performing this process, we could easily find senses for each compound verb from the distribution generated by PCA. On the other hand, when we observed candidates with a similarity score lower than Top 10, we found some examples which seem to be appropriate. Because of replenishing more synonymous expressions, we decided to check all the candidates in the 10 clusters for each compound verb.

For example, we added some examples to the list of “持ち込む (*mochikomu*)”. They appear lower than Top 10.

Sence1: 持参する (*jisan*_suru, ‘bring … with [person]’)

Sense2: 何とか制す (*nantoka* (adverb) *seisu*, ‘manage to get through’), 粘り勝つ (*nebari_katsu*, ‘compete tenaciously with each other, and finally win’)

5.2 Consideration

From our result, Japanese lexical compound verbs are found to be deeply related with adverbs and adverbial expressions. One of the reasons for this is that a compound verb represents a verbal meaning with the speaker’s emotional expressions. For example, “持ち込む (*mochikomu*)” in Sense2 implies that it’s not easy to realize a good result. Even “持ち込む (*mochikomu*)” in Sense1 has sometimes a meaning of an emphasis.

Also compound verbs are sometimes paraphrased into not only words but also phrases. Japanese compound verbs stand on a border of words and phrases.

6 Japanese compound verbs and Japanese wordnet

We try to incorporate a synonym list that we compiled into Japanese wordnet.

Currently, in Japanese wordnet, 2584 compound verbs are registered. In our experiment, we obtained 1800 compound verbs. If those compound verbs and their synonymous expressions are registered, it would be useful for not

only natural language processing like information retrieval, but also linguistic researches and language learning.
Our plan is:

(1) As for compound verbs registered in Japanese wordnet, we add or modify synonymous expressions of compound verbs and reconsider senses based on our result and Japanese dictionaries.
(2) As for compound verbs which are not in Japanese wordnet, we register synonymous expressions and then link them to the corresponding synsets in Japanese wordnet.
(3) We consider that the emotional and sensory meanings which compound verbs have are interesting and important information. The adverbial expressions included in synonymous expressions would be deeply related to compound verbs. We would like to try to put the adverbial expressions included in synonymous expressions into the Japanese wordnet. This means linking phrasal meanings to wordnet.

We show “持ち込む (*mochikomu*)” as an example in Figure3. Sense1 and Sense3 are registered in Japanese wordnet. Sense2 is not registered. As for Sense1, for the sake of a precise meaning of “持ち込む(*mochikomu*)”, not only verb but also preposition “in”, a kind of modification of a verb, is added and linked to the Japanese wordnet.

On the other hand, Sense2 is a new meaning that we obtained from the data. The following expressions with underlines and those in boldfaces mean adverbial expressions which represent emotional nuance.

もつれこむ (***motsure****komu*, ‘to proceed **though deadlocked**’)
優位に進める (***yuui_ni***(adj in adverbial form) *susumeru*, ‘advance [the match]’)
有利に運ぶ (***yuuri_ni***(adj_in adverbial form) *hakobu*, ‘carry to **one’s advantage**’)
引っ張り込む (*hipparikomu*, ‘pull in’)
引き込む (*hikikomu*, ‘pull in’)
何とか制す (***nantoka*** (adverb) *seisu*, ‘**manage to get through**’)
粘り勝つ (***nebari****katsu*, ‘compete **tenaciously** with each other, and finally win’)

We will link not only verbs like “proceed” “carry” “get through” and “win” but also a kind of modification for verbs like “through deadlocked” “advantage” “manage to” “tenaciously” to the Japanese or English wordnet because they’re important to understand a nuance of the compound verb “持ち込む (*mochikomu*)”. As an example, one of synonymous expressions “粘り勝つ (*nebarikatsu*), ‘compete tenaciously with each other and finally win’ ” are shown in Figure 4.

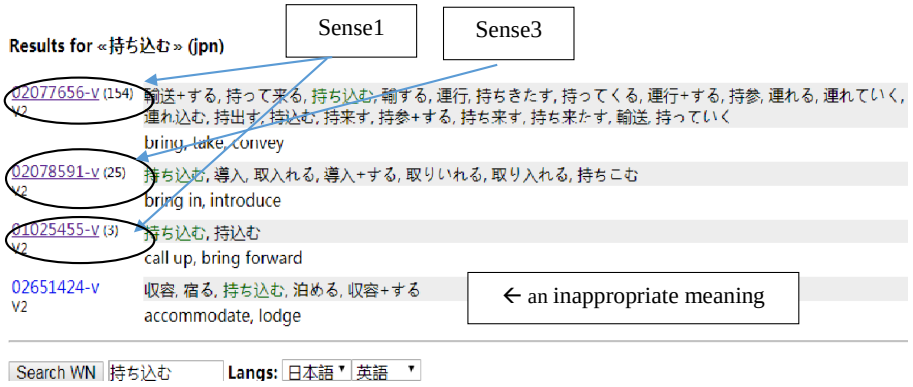


Figure 3. Comparison between senses that we obtained and synset of “持ち込む (*mochikomu*)” registered in Japanese wordnet

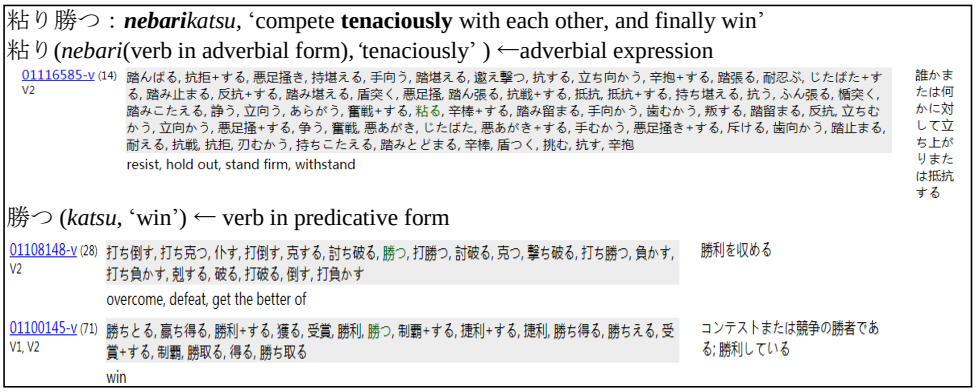


Figure 4. Link of one of synonymous expressions of “持ち込む (*mochikomu*)”: “粘り勝つ (***nebari**katsu*), ‘compete **tenaciously** with each other and finally win’ ”

7 Conclusion

In our work, first, we compiled a list of synonymous expressions of compound verbs by extracting from corpora semi-automatically and then try to link them to Japanese wordent. Japanese compound verbs have characteristics between words and phrases. We would like to consider how to combine phrasal expressions to wordnet. In addition, some compound verbs are deeply related with sense modalities. Therefore, it would be important to treat the adverbial meaning which it implies. If we registered a link of not only words but also phrasal expressions, Japanese wordnet would be useful for cross lingual works like linguistic researches, education and also, information retrieval.

References

David Arthur, Sergei Vassilvitskii. 2007. k-means++: The Advantages of Careful Seeding. In Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms, 1027-1035.

Chikara Hashimoto and Daisuke Kawahara. 2008. Construction of an idiom corpus and its application to idiom identification based on WSD in corporating idiom-specific features. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing(EMNLP 2008). 992-1001.

Taro Kageyama(1993), *Bunpō to Gokeisei* [Grammar and Word Formation], Tokyo: Hituzi Syobo

Daisuke Kawahara and Sadao Kurohashi. 2006. A Fully-lexicalized Probabilistic Model for Japanese syntactic and Case Structure Analysys. In Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2006), NY, USA, 176-183.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In proceedings of 27th Annual Conference on Neural Information Processing Systems,3111–3119.

Ivan A.Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickenger. 2002. Multiword Expressions: A pain in the Neck for NLP. In CICLing '02 Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, 1-15.

Toshifumi Tanabe, Masahito Takahashi and Kimiaki Shudo. 2014. A lexicon of multiword expressions for syntactically precise, wide coverage natural language processing, Computer Speech and Language, vol.28. No.6, 1317-1339, Elsevier.

Kiyoko Uchiyama and Shun Ishizaki. 2003.The Method on the Semantic Analysis for disambiguation of compound verbs. In proceedings of the 9th annual conference of Natural Language Processing,163-166.

Kiyoko Uchiyama, Timothy Baldwin,.2004. Automatic Disambiguation of Compound Verbs by Machine Learning. In proceedings of the 10th annual conference of Natural Language Processing, 741-744.

Language Resource References

National Institute for Japanese Language and Linguistics. Compound Verb Lexicon.(2013-2015) http://vvlexicon.ninjal.ac.jp/en/

Acknowledgements

This work was supported by JSPS KAKENHI （Grant-in-Aid for Scientific Research (C)）Grant Number JP 16K02727.

Thinking globally, acting locally – progress in the African Wordnet Project

Marissa Griesel
University of South Africa (UNISA)
Pretoria, South Africa
griesm@unisa.ac.za

Sonja Bosch
University of South Africa (UNISA)
Pretoria, South Africa
boschse@unisa.ac.za

Mampaka L. Mojapelo
University of South Africa (UNISA)
Pretoria, South Africa
mojapml@unisa.ac.za

Abstract

The African Wordnet Project (AWN) includes all nine indigenous South African languages, namely isiZulu, isiXhosa, Setswana, Sesotho sa Leboa, Tshivenda, Siswati, Sesotho, isiNdebele and Xitsonga. The AWN currently includes 61 000 synsets as well as definitions and usage examples for a large part of the synsets. The project recently received extended funding from the South African Centre for Digital Language Resources (SADiLaR) and aims to update all aspects of the current resource, including the seed list used for new development, software tools used and mapping the AWN to the latest version of PWN 3.1. As with any resource development project, it is essential to also include phases of focused quality assurance and updating of the basis on which the resource is built. The African languages remain under-resourced. This paper describes progress made in the development of the AWN as well as recent technical improvements.

1 Introduction

The African Wordnet Project (AWN) has seen various phases of development with different funding cycles and collaborators (see Bosch & Griesel, 2017 for a comprehensive breakdown of previous phases). The most recent cycle is funded by the South African Centre for Digital Language Resources (SADiLaR)[1] and will run from 2018 to the end of February 2020, with an extension to 2022 currently under consideration. The most notable change to the project in the past two years is the addition of four further languages to include the full range of nine indigenous South African languages, namely isiZulu (ZUL), isiXhosa (XHO), Setswana (TSN), Sesotho sa Leboa (NSO), Tshivenḓa (VEN), Siswati (SSW), Sesotho (SOT), isiNdebele (NDE) and Xitsonga (TSO). The number of synsets, usage examples and definitions for all languages included in the AWN have also been substantially increased. As with any resource development project, it is essential to include phases of focused quality assurance and updating of the basis on which the resource is built. For the AWN, this meant reassessing several core aspects, including the seed terms used for further development, software to assist linguists to develop and structure the wordnets, as well as the process by which development is managed.

The African languages remain under-resourced despite progress being made with a resource catalogue hosted by the Resource Management Agency of SADiLaR. Currently there are still no freely available dictionaries for any of the languages and as Oliver (2014:7) notes: "The most commonly used strategy within the expand model is the use of bilingual dictionaries". In this paper, key aspects pertaining to the development of a multilingual wordnet for such under-resourced languages will be highlighted and our solutions to challenges that emerged as a result of the growth in the scope of the project, will be discussed. The last section of the paper will mention smaller challenges and project specific matters that might be of interest to other projects with similar restrictions.

2 Recent progress

The AWN team first began the development of wordnets for South African languages in 2010 and has grown slowly but consistently. Currently, the AWN includes 61 000 synsets across the nine identified languages. The number of synsets per language varies from nearly 17 000 for Setswana, to only 600 each for isiNdebele and Xitsonga. This variation is due to the amount of time linguists have available to work on the project as well as the incremental addition of languages to the project (see below). In addition to the basic synsets, the AWN also includes 26 500 definitions

[1] https://www.sadilar.org/

and 37 000 usage examples across the nine languages.

One of the most significant expansions to the AWN over the past two years has been the addition of four new languages. This means that all nine indigenous South African languages are now represented[2], although not in equal numbers yet. The current funding phase will see isiNdebele and Xitsonga also grow to 1 000 synsets each, with definitions and usage examples.

In addition to this, the AWN has also added definitions to the synsets already captured in previous phases. Where the developers initially focussed only on synsets with their usage examples, feedback from the South African Digital Humanities and Human Language Technology communities indicated that definitions would make the AWN even more useful in language learning applications – an ever-growing research and development area given the multilingual nature of the country. An initial experiment into this application is described in Bosch and Griesel (2018). In the second application, data from the AWN has also been used experimentally in the Kamusi GOLD project[3] to populate an online dictionary for which definitions and usage examples are important.

Section 3 describes another significant decision regarding the content of the AWN – moving away from relying on the Eurocentric core base concepts[4] (CBC) to a more localised wordlist to be used as seed terms for new synsets.

3 The SIL list as seed terms

3.1 Contextualisation

The SIL Comparative African Wordlist (SILCAWL) was compiled in 2006 by Keith Snider (SIL International and Canada Institute of Linguistics) and James Roberts (SIL Chad and Université de N'Djaména). It is a list of lexical data consisting of 1 700 words with both English and French glosses which resulted from linguistic research in Africa. The items are organised semantically under 12 main headings which generally move on a continuum from items relating to human domains on the one extreme, via animate domains, to items relating to non-human domains on the other extreme, and then from concrete items to more abstract items. The following are the 12 main headings:

1. Man's physical being	7. Plants
2. Man's non-physical being	8. Environment
3. Persons	9. Events and actions
4. Personal interaction	10. Quality
5. Human civilisation	11. Quantity
6. Animals	12. Grammatical items

Table 1. Headings in the SIL CAWL list

Each of the above headings is then subdivided into second and third level headings. For instance, in the case of Persons, the following first level headings are distinguished: STAGES OF LIFE, BLOOD RELATIONS, MARRIAGE RELATIONS, RELATIONS, EXTENDED AND SOCIAL, and PROFESSIONS. A third level, for example, in the case of PROFESSIONS includes divisions such as: farmer, fisherman, hunter, blacksmith, potter, weaver, medicine man etc. The parts of speech covered in the SIL list are nouns, verbs, adjectives, adverbs, pronouns, interrogatives and conjunctions. Although Snider and Roberts (2006:4) concede that they still notice "imperfections and room for improvement (e.g. words that could be deleted, words that could be added, words that could be moved to different semantic domains etc"), the SIL list has proven to be a welcome improvement on the CBC list used in the past in the development of the AWN that follows the expand model (Vossen, 1998) and is based on the English Princeton WordNet (PWN) (Fellbaum, 1998). The most significant improvement is observed against the background of localisation where the content (of the entries) would be lexicalised within an African environment.

3.2 Comparison of the SIL list to the core base concepts

The CBC list is a combination of seed lists extracted from European language corpora for the EuroWordNet and BalkaNet projects (see a description of the core base concepts list at http://globalwordnet.org/gwa-base-concepts/). The CBC aims at covering terms that display many relations with other terms (synsets) and are also placed high in the semantic structure of a wordnet. It includes very basic terminology such as "light" (noun), "Earth" (noun), "catch" (verb) and "shake" (verb), but also less frequently used terms such as "actinic radiation" (noun) and "protozoan" (noun). As discussed in Bosch and

[2] An Afrikaans wordnet already exists independently from this project but is not currently under active development. See https://hdl.handle.net/20.500.12185/158.

[3] https://kamusi.org

[4] As found on http://globalwordnet.org/?page_id=68

Griesel (2017), these unfamiliar terms caused some problems for the African language team, resulting in wasted time and lost momentum in the early phases of development and the team decided to investigate an alternative seed list such as the SIL list described in Section 3.1, drawn up from local African sources.

All terms in the CBC can be found in the PWN and therefore have a direct mapping to the larger wordnet structure with a unique identifying number. The SIL list, however, includes 41 terms that have no equivalents in the PWN. These terms are not necessarily foreign to an English native speaker but might be more frequently used in the African context. They include terms such as "cooking stone" (noun) and "thorn tree" (noun).

Another noteworthy category of terms that is present in the SIL list but not in the CBC includes various terms where African languages make a distinction based on usage that other languages might not make but are well known (lexicalised) to native speakers. The South African languages, for instance, distinguish between harvesting by digging up versus harvesting by cutting or plucking, etc. The subtle differences between these terms in isiZulu and Sesotho sa Leboa are illustrated in Table 2.

SILCAWL	ZUL	NSO
0757 harvest (maize) (v)	*ukuvuna* *ukukwica* *ukucasa* (while still green, harvest green corn before it has hardened) *Comment:* synonyms, or near synonyms in the case of *ukuvuna* and *ukucasa*	*buna* *Comment:* general concept related to the time of harvest
0758 harvest, dig up (yams)	*ukuvuna* *Comment:* same as harvesting crops that grow above the ground	*bupula* *Comment:* harvest groundnuts
0760 harvest, collect (honey from hive)	*ukuthapha* *Comment:* extract, take out honey from a hive.	*rafa* *Comment:* extract honey from a hive.

Table 2. Harvesting in isiZulu and Sesotho sa Leboa

Kinship terms are another instance of a very intricate system in the African languages as illustrated in a few examples in Table 3.

SILCAWL	ZUL	NSO
BLOOD RELATIONS		
0348 father's brother (uncle)	*ubabamkhulu* (big father) 'father's elder brother' *ubabomncane* (small father) - 'father's younger brother'	*ramogolo* 'father's elder brother' *rangwane* 'father's younger brother'
0351 father's sister (aunt)	*ubabekazi* (female father) 'father's sister'	*rakgadi* 'father's sister'
0349 mother's brother (uncle)	*umalume* (male mother) 'mother's brother'	*malome* 'mother's brother'
0350 mother's sister (aunt)	*umamekazi* (female mother) or *umame* 'mother's sister'	*mmamogolo* 'mother's elder sister' *mmane* 'mother's younger sister'
MARRIAGE RELATIONS		
0365 father-in-law	*ubabezala* 'father-in-law' used by Zulu-speaking woman *umukhwe* 'father-in-law' used by Zulu-speaking man	*ratswale* 'father-in-law'
0366 mother-in-law	*umkhwekazi* 'mother-in-law' used by Zulu-speaking man *umamezala* 'mother-in-law' used by Zulu-speaking woman	*mmatswale* / *mogwegadi* 'mother-in-law' (man speaking – dialectal) *mmatswale* 'mother-in-law' (woman speaking)
0367 brother-in-law	*umfowethu* 'husband's brother' *umkhwenyawethu* 'sister's husband' (man speaking) *umlamu* 'wife's brother' *umkhwenyana* 'sister's husband' (woman speaking)	*molamo, sebara* 'sister's husband' (man and woman speaking) *molamo, sebara* 'wife's brother' (man speaking)
0368 sister-in-law	*udadewethu* 'husband's sister'	*mogadibo*

	umakoti, umlobokazi, umkami 'brother's wife' (man speaking) *umlamu* 'wife's sister' *umakoti womfowethu, umakoti womnewethu* 'brother's wife' (woman speaking)	'husband's sister'/ 'brother's wife'

Table 3. Kinship terms in isiZulu and Sesotho sa Leboa

3.3 Translation of the expanded SIL list

One of the advantages of a common seed list such as the SIL list across all the languages in the AWN, is that it enables the creation of a parallel corpus within the larger wordnet structure. Parallel synsets are not only useful for language learners, but also in applications such as multilingual information retrieval, semantic analysis and machine translation. The AWN team therefore decided to incorporate the terms in the SIL list using the following steps: first, the English term in the SIL list was compared to the PWN and an ID to the corresponding synset was added to each term in the SIL list. If available, the definition and usage example from the PWN was also extracted to a simple spreadsheet. This document was next presented to an expert South African English lexicographer to a) fill in any gaps there might still be so that each term has a part of speech tag, definition and usage example; and b) edit the existing PWN data to fit the South African context better.

The African language translators were briefed on the nature of the project and specifically on the unique characteristics of a wordnet with a strict protocol to follow. Translation of the first 1 000 synsets from the expanded SIL list dataset took roughly five months, including internal quality assurance. The output of this process was a multilingual parallel corpus of common terms, each with a clear definition, usage example and part of speech tag. This is already a valuable resource, but for inclusion into the AWN, we will now need to incorporate this data into the hierarchical structure of a wordnet, identify the relations within this structure and perform formal quality assurance. This process is currently ongoing.

4 Visualisation of the AWN in WordnetLoom

Developing data to populate a wordnet offline in spreadsheets has its advantages, most notably fast tracking of development because it is a familiar process for inexperienced linguists, ease of applying spell checking or other quality assurance, no delays due to interruptions in internet connectivity or access to a central server, etc. However, it is very difficult to see the true nature of a wordnet with connecting relations and multilingual similarities. The AWN previously used the DEBVisDic editor (see Rambousek & Horak, 2016) to facilitate development and align work across the different languages. At the onset of the current phase, however, it became clear that a focus on quality assurance of, especially the semantic relations, was needed and it was decided to port the AWN to WordnetLoom (WNL) (cf. Naskret *et al.*, 2018) – an editor with advanced visualisation of wordnets. While preparing the data for use in this tool, the AWN was also mapped to the PWN 3.0 to ensure the latest format and most up to date English equivalents. To move from DEBVisDic to WNL involved extracting the AWN database in LMF format, whereafter a programmer could map the AWN to PWN 3.0 using an in-house script. Where there was no PWN 3.0 equivalent or ID, the PWN 2.0 ID was retained.

Advantages of this tool are that it speaks to the organic growth development style that the AWN teams have always favoured (see Bosch & Griesel, 2017) and also adds the ability to perform more productive searches when working in a specific domain or looking for a specific semantic relation. The addition of a multilingual relation also means that specific senses in different languages can be connected to each other without having to connect the entire synset. The subtle differences between isiZulu and Sesotho sa Leboa verbs and kinship terms mentioned in Section 3.2 can, for instance, be represented more accurately. Discussions with the development team behind this state-of-the-art software tool led to an intense two-day workshop in South Africa, facilitated by members of the WNL and Polish Wordnet development team where linguists were introduced to WNL and its many advanced features. The workshop, which was hosted by SADiLaR, was attended by at least two linguists from each of the nine languages included in the AWN and took on a very hands-on approach. Figure 1 shows the "harvest" example from Table 2 as it was added to the Sesotho sa Leboa wordnet using WNL.

5 Conclusion and future work

Many challenges, including the low resource nature of the languages in the AWN, restraints on funding, a part-time development team etc. were reported on extensively in Bosch and Griesel (2014). The AWN team have managed to mitigate these risks to a large extent and porting development of the AWN to WNL played a large part in this. The porting process described in Section 4 did not come without some initial challenges and adaptations needed. Most notable is that the visualisation of the AWN now draws our attention to the lack of proper definition and application of the semantic relations between terms. Relations were previously automatically carried over from the PWN as is common when following the expand model. The AWN team was always aware that this method was not fool proof and that relations would need revision. WNL enables linguists to see immediately all synsets connected with any of the predefined relations as well as the lacking relations within the South African context. Some terms also need to be moved from an independent synset to a more accurate embedded synonym position and vice versa. Lexical gaps between the (American) English PWN and the African languages can now also be addressed more effectively by eliminating the need to link a synset in an African language to a synset in the PWN as synsets can either stand independently in WNL or be linked to another African language. Again, the visualisation of the synsets within the larger structure is key in this process of identifying the lexical gaps, as can be seen in Figure 2, a representation of the isiZulu marriage relations discussed in Table 3 above. These aspects will receive priority attention during the quality assurance phase that is underway.

With continued research, collaboration with other developers and an invested interest in growing the African languages as digital language resources, we believe that this project will soon be of significant academic and industrial interest to members of the global wordnet community.

Acknowledgements

The African Wordnet project (AWN) was made possible with support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR (www.sadilar.org) is a research infrastructure established by the Department of Science and Technology of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

References

Bosch, Sonja and Griesel, Marissa. 2018. African Wordnet: facilitating language learning in African Languages. *Proceedings of Ninth Global WordNet Conference 2018 (GWC2018*), 12 January 2018, Nanyang Technological University (NTU), Singapore. Available at http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_22.pdf

Bosch, Sonja and Griesel, Marissa. 2017. Strategies for building wordnets for under-resourced languages: the case of African languages. *Literator* 38(1), a1351. https://doi.org/10.4102/lit.v38i1.1351

Fellbaum, Christiane, (ed), 1998. *Wordnet: An electronic lexical database*. The MIT Press, Cambridge, Mass.

Griesel, Marissa and Bosch, Sonja. 2014. Taking stock of the African Wordnet project: 5 years of development. *Proceedings of the Seventh Global WordNet Conference 2014* (GWC2014), pp. 148-153. Tartu, Estonia. Available at http://gwc2014.ut.ee/proceedings_of_GWC_2014.pdf

Naskręt, Tomasz, Dziob, Agnieszka, Maciej Piasecki, Maciej, Saedi, Chakaveh and Branco, António. 2018. WordnetLoom - a Multilingual Wordnet Editing System Focused on Graph-based Presentation. *Proceedings of Ninth Global WordNet Conference 2018 (GWC2018*), 12 January 2018, Nanyang Technological University (NTU), Singapore. Available at http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/gwc-2018-proceedings.pdf

Oliver, Antoni. 2014. WN-Toolkit: Automatic generation of WordNets following the expand model. *Proceedings of the Seventh Global WordNet Conference 2014* (GWC2014), Tartu, Estonia. Available at http://gwc2014.ut.ee/proceedings_of_GWC_2014.pdf

Princeton University. 2017. WordNet – A lexical database for English. https://wordnet.princeton.edu/ Accessed on 12 March 2019.

Rambousek, Adam and Horák, Aleš. 2016. DEBVisDic: Instant Wordnet building. In V. Mititelu, C. Forăscu, C. Fellbaum and P. Vossen (eds.), *Proceedings of the Eighth Global WordNet Conference 2016* (GWC2016), Bucharest, Romania, January 25–29, pp. 317–321.

Snider, Keith and Roberts, James. 2006. *SIL Comparative African Wordlist (SILCAWL).* Available at https://www.eva.mpg.de/lingua/tools-at-lingboard/pdf/Snider_silewp2006-005.pdf)

Vossen, Piek. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic, Dordrecht.

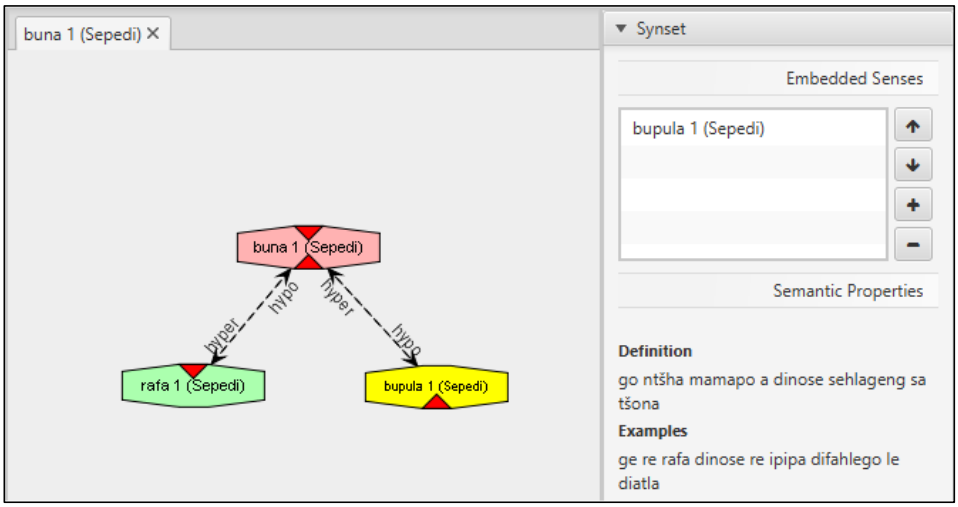


Fig. 1. "Harvest" (buna) as included in the Sesotho sa Leboa wordnet

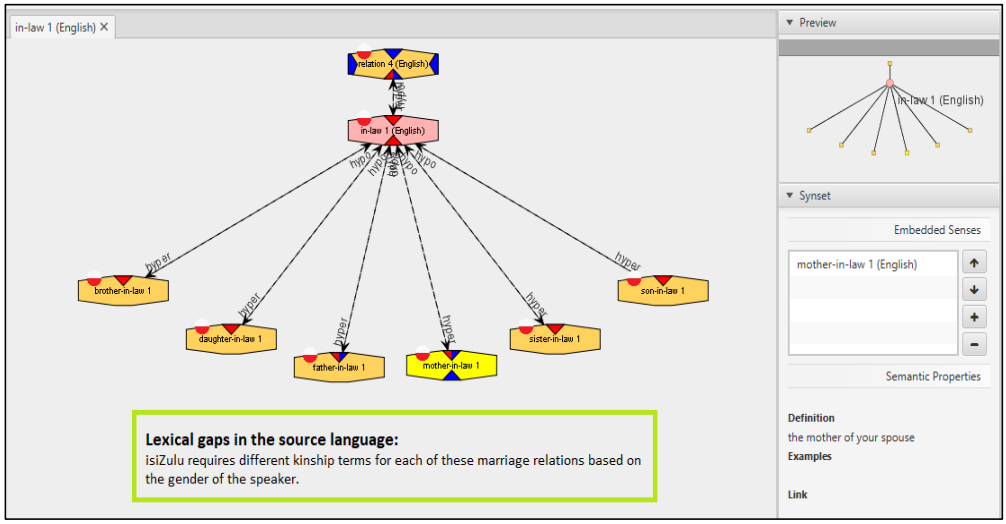


Fig. 2. Lexical gaps between English and isiZulu for kinship terms

Commonsense Reasoning Using WordNet and SUMO: a Detailed Analysis

Javier Álvez LoRea Group
Itziar Gonzalez-Dios Ixa Group
German Rigau Ixa Group
University of the Basque Country (UPV/EHU)
{javier.alvez,itziar.gonzalezd,german.rigau}@ehu.eus

Abstract

We describe a detailed analysis of a sample of large benchmark of commonsense reasoning problems that has been automatically obtained from WordNet, SUMO and their mapping. The objective is to provide a better assessment of the quality of both the benchmark and the involved knowledge resources for advanced commonsense reasoning tasks. By means of this analysis, we are able to detect some knowledge misalignments, mapping errors and lack of knowledge and resources. Our final objective is the extraction of some guidelines towards a better exploitation of this commonsense knowledge framework by the improvement of the included resources.

1 Introduction

Any ontology tries to provide an explicit formal semantic specification of the concepts and relations in a domain (Noy and McGuinness, 2001; Guarino and Welty, 2002; Guarino and Welty, 2004; Gruber, 2009; Staab and Studer, 2009; Álvez et al., 2012). As with other software artefacts, ontologies typically have to fulfil some previously specified requirements. Usually both the creation of ontologies and the verification of its requirements are manual tasks that require a significant amount of human effort. In the literature, some methodologies collect the experience in ontology development (Gómez-Pérez et al., 2004; Guarino and Welty, 2004) and in ontology verification (Gangemi et al., 2006).

In order to evaluate the *competency* of SUMO-based ontologies in the sense proposed by Grüninger and Fox (1995), Álvez et al. (2019) propose a method for the semi-automatic creation of *competency questions* (CQs). Concretely, they adapt the methodology to evaluate the ontologies so that it can be automatically applied using automated theorem provers (ATPs). The construction of CQs is based on several predefined *question patterns* (QPs) that yield a large set of problems (dual conjectures) by using information from WordNet and its mapping into SUMO. For example, the synsets $machine^1_v$ and $machine^1_n$ are related by the morphosemantic relation *instrument* in the *Morphosemantic Links* (Fellbaum et al., 2009) of WordNet, as depicted in Figure 1. In the same figure, the mappings of the synsets are also provided: $machine^1_n$ and $machine^1_v$ are connected to $Machine_c=$ and $Making_c+$, where the symbol '$=$' means that $machine^1_n$ is semantically equivalent to the $Machine_c$, while '$+$' means that the semantics of $Making_c$ is more general than the semantics of $machine^1_v$. Hence, it is possible to state the the relationship of $machine^1_n$ and $machine^1_v$ in terms of SUMO as follows:

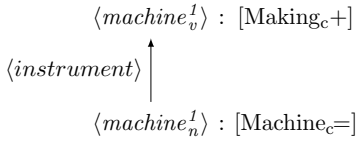


Figure 1: An example of WordNet and its mapping to SUMO

```
(forall (?Y)                                        (1)
   (=>
      (instance ?Y Machine)
      (exists (?X)
         (and
            (instance ?X Making)
            (instrument ?X ?Y)))))
```

The problem that results from Figure 1 consists of the above conjecture, which is considered to

be true according to our commonsense knowledge, and its negation, which is therefore assumed to be false.

State-of-the-art ATPs for first-order logic (FOL) such as Vampire (Kovács and Voronkov, 2013) or E (Schulz, 2002) have been proved to provide advanced reasoning support to large FOL conversions of expressive ontologies (Ramachandran et al., 2005; Horrocks and Voronkov, 2006; Pease and Sutcliffe, 2007; Álvez et al., 2012). However, the semi-decidability of FOL and the poor scalability of the known decision procedures have been usually identified as the main drawbacks for the practical use of FOL ontologies. In particular, given an unsolved problem (i.e. a problem such that ATPs do not find any proof for its pair of conjectures) it is not easy to know if (a) the conjectures are not entailed by the ontology or (b) although some of the conjecture is entailed, ATPs have not been able to find the proof within the provided execution-time and memory limits. On the contrary, given a solved problem, it is hard to know whether the solution is obtained for a good reason, because an expected result does not always indicate a correct ontological modelling.

In this paper, we provide a detailed analysis of the large commonsense reasoning benchmarks created semi-automatically by (Álvez et al., 2017; Álvez et al., 2019). The aim of this analysis is to shed light on the commonsense reasoning capabilities of both the benchmark and the involved knowledge resources. To that end, we have randomly selected a sample of 169 problems (1% of the total) following a uniform distribution and manually inspected their source knowledge and results. By means of this detailed analysis, we are able to evaluate the quality of automatically created benchmarks of problems and to detect hidden problems and misalignments between the knowledge of WordNet, SUMO and their mapping.

Outline of the paper. In order to make the paper self-contained, we first introduce the ontology, the mapping to WordNet and the evaluation framework in Section 2. Next, we provide a full summary and the main conclusions obtained from our manual analysis in Section 3. Then, we individually examine some of the problems in Section 4. Finally, we provide some conclusions and discuss future work in Section 5.

2 Commonsense Reasoning Framework

In this section we describe briefly the whole commonsense reasoning reasoning framework. First, we present the knowledge resources needed and then the reasoning framework.

2.1 Resources

The resources we present in this section are FOL-SUMO, WordNet and the semantic mapping between them.

SUMO[1] (Niles and Pease, 2001) is an upper level ontology proposed as a starter document by the IEEE Standard Upper Ontology Working Group. SUMO is expressed in SUO-KIF (Standard Upper Ontology Knowledge Interchange Format (Pease, 2009)), which is a dialect of KIF (Knowledge Interchange Format (Genesereth et al., 1992)). The syntax of both KIF and SUO-KIF goes beyond FOL and, therefore, SUMO axioms cannot be directly used by FOL ATPs without a suitable transformation (Álvez et al., 2012).

To the best of our knowledge, there are two main proposals for the translation of the two upper levels of SUMO into a FOL formulae that are described in Pease and Sutcliffe, (2007), Pease et al. (2010) and Álvez et al. (2012) respectively. Both proposals have been developed under the *Open World Assumption* (OWA) (Reiter, 1978) and are currently included in the *Thousands of Problems for Theorem Provers* (TPTP) problem library[2] (Sutcliffe, 2009). In this paper, we use Adimen-SUMO v2.6, which is freely available at `https://adimen.si.ehu.es/web/AdimenSUMO`. From now on, we refer to Adimen-SUMO v2.6 as FOL-SUMO.

The knowledge in SUMO, and therefore in FOL-SUMO, is organised around the notions of *instance* and *class*. These concepts are respectively defined in SUMO by means of the predicates *instance* and *subclass*.[3] Additionally, SUMO also differentiates between *relations* and *attributes*, which are organized using the predicates *subrelation* and *subAttribute* respectively. For simplicity, from now on we denote the nature of SUMO concepts by adding as subscript the symbols o (SUMO instances that are neither relations nor attributes), c (SUMO classes that are nei-

[1] `http://www.ontologyportal.org`

[2] `http://www.tptp.org`

[3] It is worth noting that term *instance* is overloaded since it denotes both the SUMO predicate and the SUMO concepts that are defined by using that predicate.

ther classes of relations nor classes of attributes), r (SUMO relations) and a (SUMO attributes). For example: *Waist*$_o$, *Artifact*$_c$, *customer*$_r$ and *Female*$_a$.

WordNet (Fellbaum, 1998) is linked with SUMO by means of the mapping described in Niles and Pease (2003). This mapping connects WordNet synsets to terms in SUMO using three relations: *equivalence*, *subsumption* and *instantiation*. We denote the mapping relations by concatenating the symbols '=' (*equivalence*), '+' (*subsumption*) and '@' (*instantiation*) to the corresponding SUMO concept. For example, the synsets *horse*^1_n, *education*^4_n and *zero*^1_a are connected to *Horse*$_c$=, *EducationalProcess*$_c$+ and *Integer*$_c$@ respectively. *equivalence* denotes that the related WordNet synset and SUMO concept are equivalent in meaning, whereas *subsumption* and *instantiation* indicate that the semantics of the WordNet synset is less general than the semantics of the SUMO concept. In particular, *instantiation* is used when the semantics of the WordNet synsets refers to a particular member of the class to which the semantics of the SUMO concept is referred.

The mapping between WordNet and SUMO can be translated into the language of SUMO by means of the proposal introduced in Álvez et al. (2017). This translation characterises the mapping information of a synset in terms of SUMO instances by using equality (for SUMO instances) and the SUMO predicates *instance*$_r$ and *attribute*$_r$ (for SUMO classes and attributes respectively). For example, the noun synsets *smoking*^1_n and *breathing*^1_n are respectively connected to *Smoking*$_c$= and *Breathing*$_c$=. Thus, the SUMO statements that result by following the proposal described in Álvez et al. (2017) is:

$$(\text{instance ?X Smoking}) \quad (2)$$

$$(\text{instance ?X Breathing}) \quad (3)$$

2.2 Evaluation Framework

The competency of SUMO-based ontologies can be automatically evaluated by using the framework described in Álvez et al. (2019) and the resources mentioned above. This framework is based on the use of *competency questions* (CQs) or *problems* (Grüninger and Fox, 1995) derived from the knowledge in WordNet and its mapping to SUMO by means of several predefined *question patterns*. In this paper, we have considered the following QPs:

WordNet Relation	QP	Problems
Hyponymy	Noun #1	7,539
	Noun #2	1,944
	Verb #1	1,765
	Verb #2	304
Antonymy	#1	91
	#2	574
	#3	2,780
Morphosemantic Links	Agent	829
	Instrument	348
	Result	788
Total	**–**	**16,972**

Table 1: Creation of problems on the basis of QPs

- The four QPs based on *hyponymy* —2 QPs for nouns and 2 QPs for verbs— and the three QPs based on *antonymy* introduced in Álvez et al. (2019).
- The three QPs based on the *Morphosemantic Links agent*, *instrument* and *result* introduced in Álvez et al. (2017).

In Table 1, we report on the number of CQs/problems that results from each QP.

$$\langle breathing^1_n \rangle : [\text{Breathing}_c=]$$
$$\langle hyp \rangle \uparrow$$
$$\langle smoking^1_n \rangle : [\text{Smoking}_c=]$$

Figure 2: Example for Noun #2: *smoking*^1_n and *breathing*^1_n

For example, the second QP based on *hyponymy* focuses on pairs of hyponym synsets *(hypo,hyper)* such that the hyponym $hypo$ is connected to SUMO using *equivalence*. In those cases, the semantics of $hypo$ is equivalent to the semantics of the SUMO statement that results from its mapping information. Further, the semantics of $hyper$ is more general than the semantics of $hypo$. Consequently, we can state that the set of SUMO instances related to $hyper$ is a superset of the set of SUMO instances connected to $hypo$. In particular, the noun synset *smoking*^1_n ("*the act of smoking tobacco or other substances*") is hyponym of *breathing*^1_n ("*the bodily process of inhalation and exhalation; the process of taking in oxygen from inhaled air and releasing carbon dioxide by exhalation*") (see Figure 2). By the instantiation of the

second QP based on hyponymy using statements (2-3) which result from their mapping information, the following CQ that states that every instance of *Smoking*$_c$ is also instance of *Breathing*$_c$ can be obtained:

```
(forall (?X)                                    (4)
    (=>
        (instance ?X Smoking)
        (instance ?X Breathing)))
```

Given a set of CQs and an ontology, the evaluation framework proposes to perform two dual tests using FOL ATPs for each CQ: the first test is to check whether, as expected, the conjecture stated by the CQ is entailed by the ontology (*truth-test*); the second one is to check its complementary (*falsity-test*). If ATPs find a proof for either the truth- or the falsity-test, then the CQ is classified as *solved* (or *resolved*). In particular, the CQ is *passing*/*non-passing* if ATPs find a proof for the truth-test/falsity-test. Otherwise (that is, if no proof is found), the CQ is classified as *unresolved* or *unknown*. In this last case, we do not know whether (a) the conjectures are not entailed by the ontology or (b) although (some of) the conjectures are entailed, ATPs have not been able to find the proof within the provided execution-time and memory limits.

3 Detailed Analysis of the Experimental Results

In this section, we report on a detailed and manual analysis of the experimental results obtained from a small number of the CQs described in Section 2.

From this experimentation, we have randomly selected a sample of 169 problems (1% of the total) following a uniform distribution and analysed the results obtained for those problems by focusing on two questions: 1) we analyse the quality of mapping of the involved synsets and 2) we analyse the knowledge required for solving the problems.

Regarding the quality of the mapping (first question), we classify the mapping of synsets as either *correct* or *incorrect* according to the following criteria: a mapping is classified as *correct* if the semantics associated with the SUMO concept and with the synset are compatible, and it is classified as *incorrect* otherwise. For example, both the verb synset *machine*^1_v and the adjective synset *homemade*^1_a are connected to *Making*$_c$+, where the semantics of the SUMO class *Making*$_c$ is *"The subclass of Creation*$_c$ *in which an individual Artifact*$_c$ *or a type of Artifact*$_c$ *is made"*. Since the semantics of the verb synset *machine*^1_v is *"Turn, shape, mold, or otherwise finish by machinery"*, we classify the mapping of *machine*^1_v as *correct*. On the contrary, the semantics of the adjective synset *homemade*^1_a is *"made or produced in the home or by yourself"*. Thus, we classify the mapping of *homemade*^1_a as *incorrect*.

In addition, synsets with a correct mapping are classified as either *correct and precise* or *only correct*: a correct mapping is also considered as *correct and precise* if the semantics of the synset and the SUMO concept are equivalent, and it is classified as *only correct* (that is, correct but not precise) if the semantics of the SUMO concept is more general than the semantics of the synset. For example, the mapping of *machine*^1_v to *Making*$_c$ is classified as *only correct* since the semantics of *Making*$_c$ is more general than the semantics of *machine*^1_v. By contrast, the mapping of the noun synset *machine*^1_n to *Machine*$_c$= is classified as *correct and precise* since the semantics of *machine*^1_n is *"Any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks"* and the semantics of *Machine*$_c$ is *"Machine*$_c$*'s are Device*$_c$*'s that that have a well-defined resource and result and that automatically convert the resource into the result"*.

Regarding the required knowledge (second questions), we distinguish three cases:

- If the problem is solved, then we classify the knowledge in the proof provided by ATPs as either *correct* or *incorrect* depending on whether it matches our world knowledge or not.
- If the problem is unsolved and the mapping of the two involved synsets is correct, then we manually check whether the problem can be entailed by the knowledge in the ontology.
- If the problem is unsolved and the mapping of some of the involved synsets is incorrect, then the knowledge in the problem does not match our world knowledge and, consequently, it is not subject of classification.

It is worth noting that, in the case of unsolved problems such that the required knowledge is classified as existing, ATPs cannot find a proof for its

Question Pattern	#	Entailed					Incompatible					Unsolved			Total				
		S	CM	IM	CK	IK	S	CM	IM	CK	IK	U	CM	IM	S CM	IM	CK	IK	U
Noun #1 (7,539)	80	39	33 (5)	6	39	0	15	7 (0)	8	15	0	26	19 (0)	7	59 (5)	21	54	0	26
Noun #2 (1,944)	15	9	9 (5)	0	9	0	2	2 (2)	0	2	0	4	3 (2)	1	14 (9)	1	11	0	4
Verb #1 (1,765)	13	5	3 (1)	2	5	0	0	0 (0)	0	0	0	8	6 (0)	2	9 (1)	4	5	0	8
Verb #2 (304)	2	0	0 (0)	0	0	0	0	0 (0)	0	0	0	2	2 (1)	0	2 (1)	0	0	0	2
Antonym #1 (91)	1	0	0 (0)	0	0	0	0	0 (0)	0	0	0	1	0 (0)	1	0 (0)	1	0	0	1
Antonym #2 (584)	6	1	0 (0)	1	1	0	0	0 (0)	0	0	0	5	3 (1)	2	3 (1)	3	1	0	5
Antonym #3 (2,780)	33	9	4 (0)	5	9	0	0	0 (0)	0	0	0	24	7 (0)	17	11 (0)	22	9	0	24
Agent (829)	5	1	1 (0)	0	1	0	0	0 (0)	0	0	0	4	3 (1)	1	4 (1)	1	1	0	4
Instrument (348)	2	0	0 (0)	0	0	0	0	0 (0)	0	0	0	2	2 (2)	0	2 (2)	0	0	0	2
Result (788)	12	1	1 (0)	0	1	0	0	0 (0)	0	0	0	11	6 (4)	5	7 (4)	5	1	0	11
Total problems (16,972)	**169**	**65**	**51 (11)**	**14**	**65**	**0**	**17**	**9 (2)**	**8**	**17**	**0**	**87**	**51 (11)**	**36**	**111 (24)**	**58**	**82**	**0**	**87**

Table 2: Detailed analysis of problems

truth- or falsity-test because of the lack of time or memory resources.

In Table 2 we summarise some figures of our detailed analysis, where problems are organised according to their QP. The name of the QP and the number of resulting CQs is given in the first column (Question Pattern column) and the remaining columns are grouped into five main parts. In the first part (#, one column), we provide the number of problems of each category that have been randomly chosen. In the second and third parts (Entailed and Incompatible, five columns each), we provide the result of our quality analysis for the solved problems that have been classified as entailed (its truth-test has been proved) and incompatible (its falsity-test has been proved) respectively. Concretely we show:

- The number of solved problems (S column).
- The number of solved problems with a correct (CM column) and incorrect mapping (IM column). Additionally, in the CM column we provide the number of solved problems with a correct and precise mapping between brackets.
- The number of solved problems that have been proved on the basis of correct (CK column) and incorrect knowledge (IK column).

In the fourth part (Unsolved, three columns), we provide the result of our analysis for the unsolved problems:

- The number of unsolved problems (U column).
- The number of solved problems with a correct (CM column) and incorrect mapping (IM column). As before, in the CM column we provide the number of solved problems with a correct and precise mapping between brackets.

Finally, in the last part (Total, five columns) we summarise the result of our analysis:

- The number of problems with a correct (correct and precise between brackets) and incorrect mapping (CM and IM columns).
- The number of solved problems (S columns) that have been proved on the basis of correct (CK column) and incorrect knowledge (IK column).
- The number of unsolved problems (U column).

In total, the synsets in 111 problems (66 %) are decided to be correctly connected to SUMO and, among them, the synsets in 24 problems (14 %) are decided to be precisely connected. Thus, some of the synsets are not correctly connected to SUMO in 58 problems (34 %). Further, 82 problems (49 %) are solved and the knowledge of the ontology that is used in the proofs reported by ATPs is decided to be correct (100 %) according to our world knowledge. Among solved problems, 65 problems (79 %) are classified as entailed and 17 problems (21 %) are classified as incompatible. By manually analysing incompatible problems, we have discovered that the knowledge of WordNet and SUMO related to all the problems with a correct mapping is not well-aligned. Thus, we can conclude that this reasoning framework also enables the correction of the alignment between WordNet and SUMO. For example, *cloud*^1_n ("*any collection of particles (e.g., smoke or dust) or gases that is visible*") is hyponym of *physical_phenomenon*^1_n ("*a natural phe-*

nomenon involving the physical properties of matter and energy") in WordNet. However, $cloud_n^1$ and $physical\_phenomenon_n^1$ are respectively connected to $Cloud_c{=}$ and $NaturalProcess_c{=}$, which are inferred to be disjoint classes in FOL-SUMO.

Further, the mapping of the involved synsets is classified as correct in 51 of 65 entailed problems (78 %), while only 14 problems (22 %) are classified as entailed with an incorrect mapping. By contrast, the percentage of problems with an incorrect mapping is much higher among incompatible and unsolved problems: 42 % (8 of 17 entailed problems) and 41 % (36 of 87 unsolved problems) respectively. This is especially the case of the problems from the antonym categories: 26 of 40 antonym problems (65 %) have an incorrect mapping. This fact reveals the poor quality of the mapping of SUMO to WordNet adjectives. Finally, we have manually checked that 45 of the 51 unsolved problems with a correct mapping (88 %) cannot be entailed by the knowledge in SUMO, which sets an upper bound on the number of problems that can be classified as solved although augmenting the knowledge of the ontology and correcting the mapping and the alignment between WordNet and SUMO.

Next, we summarise the main conclusions drawn from our detailed analysis:

- The solutions of all the solved problems (with either correct or incorrect mapping) are based on correct knowledge of the ontology (CK columns). This means that we have not discovered incorrect knowledge in the ontology by inspecting the proofs provided by ATPs.

- The mapping of a half third of the problems is classified as incorrect (58 of 169 problems) and, among them, almost a half of the problems belong to the antonym categories (26 of 58 problems). This is mainly due to the poor quality of the mapping of SUMO to WordNet adjectives because many of them are connected to SUMO processes instead of SUMO attributes. Further, the number of problems with a precise mapping among the problems with a correct mapping is very low (24 of 111 problems). However, this is not surprising due to the large difference between the number of concepts defined in the core of SUMO (around 3,500 concepts) and WordNet (117,659 synsets).

- Among incompatible problems, the ones with a correct mapping (9 of 17 problems) enable the detection of misalignments between the knowledge of WordNet and SUMO.

- The number of solved problems among the *Morphosemantic Links* problems with a correct mapping is very low (only 2 of 13 problems), which reveals that FOL-SUMO lacks the required information about processes in SUMO.

- Most of the unsolved problems with a correct mapping —45 of 51 problems (88 %)— are due to the lack of information in the core of SUMO. However, we have also discovered 6 problems for which either its truth- or falsity-test is entailed by knowledge in the core of SUMO although it cannot be proved by ATPs within the given resources of time and memory. Thus, ATPs are able to solve 82 of 88 the problems (93 %) that are entailed by the current knowledge of the ontology.

4 Exhaustive Analysis of some Problems

In this section, we present a detailed analysis of some of the examples that have been reported in Table 2.

4.1 Examples of Entailed Problems

Next, we present two examples among the 65 problems that are classified as entailed. The mapping information is correct in the first example, while it is incorrect in the second one.

4.1.1 Case 1: Correct mapping

The first example we present involves the synsets $army_n^1$ ("*a permanent organization of the military land forces of a nation or state*") and $armed\_service_n^1$ ("*a force that is a branch of the armed forces*"). These synsets are respectively mapped to the SUMO classes $Army_c$ and $MilitaryService_c$ by equivalence.

In WordNet $army_n^1$ is hyponym of $armed\_service_n^1$ and in SUMO $Army_c$ is subclass of $MilitaryService_c$. In this case, the knowledge in both resources and in the mapping is correctly aligned, so we get an entailed problem. In Table 2, we report 51 entailed problems with a correct mapping.

4.1.2 Case 2: Incorrect mapping

The second example of entailed problem involves the synsets *atmospheric_electricity*^1_n ("*electrical discharges in the atmosphere*") and *electrical_discharge*^1_n ("*a discharge of electricity*"). These synsets are respectively mapped to the SUMO classes *Lightning*$_c$ and *Radiating*$_c$ by subsumption.

These synsets are related by hyponymy-hyperonymy in WordNet and by subclass in SUMO, as in the previous case. But, the mapping seems misleading for *electrical_discharge*^t_o *Radiating*$_c$: ("*Processes in which some form of electromagnetic radiation, e.g. radio waves, light waves, electrical energy, etc., is given off or absorbed by something else.*"). However, this case is resolved because by chance the knowledge in WordNet and in the incorrect mapping to SUMO is aligned.

We have discovered 14 entailed problems with an incorrect mapping.

4.2 Examples of Incompatible Problems

Next, we present three examples of problems that are classified as incompatible due to several reasons.

4.2.1 Case 1: Knowledge misalignment

The first example we present involves the SUMO classes *Smoking*$_c$ and *Breathing*$_c$ and the synsets *smoking*^1_n ("*the act of smoking tobacco or other substances*") and *breathing*^1_n ("*the bodily process of inhalation and exhalation; the process of taking in oxygen from inhaled air and releasing carbon dioxide by exhalation*") .

The synset *smoking*^1_n is hyponym of *breathing*^1_n in WordNet, which are respectively connected to *Smoking*$_c$= and *Breathing*$_c$=. These classes are disjoint in SUMO. That is, instances of *Smoking*$_c$ cannot be instances of *Breathing*$_c$. So, according to the knowledge in SUMO, it is not possible to breath and smoke at the same time, but, according to WordNet smoking is a subtype of breathing. In this case we have, therefore, a knowledge misalignment problem: the knowledge in one of the resources contradicts the knowledge in the other one.

Another example of this type of cases involves the SUMO classes *Cloud*$_c$ and *NaturalProcess*$_c$ and the synsets *cloud*^1_n ("*any collection of particles (e.g., smoke or dust) or gases that is visible*") and *physical_phenomenon*^1_n ("*a natural phenomenon involving the physical properties of matter and energy* "), which are mapped to SUMO respectively by equivalence and subsumption.

In WordNet *cloud*^1_n is hypomym of *physical_phenomenon*^1_n, but in SUMO they belong to different hierarchies: *Cloud*$_c$ is subclass of *Substance*$_c$ and *NaturalProcess*$_c$ is subclass of *Process*$_c$, and these classes are disjoint as in the previous example.

From the incompatible problems reported in Table 2, the knowledge is misaligned in 5 problems.

4.2.2 Case 2: Imprecise mappings

The next example involves the SUMO classes *Transfer*$_c$ ("*Any instance of Translocation where the agent and the patient are not the same thing.*") and *Removing*$_c$ ("*The Class of Processes where something is taken away from a location. Note that the thing removed and the location are specified with the CaseRoles patient and origin, respectively.*"). The involved synsets are *fetch*^1_v ("*go or come after and bring or take back*") and *carry_away*^1_v ("*remove from a certain place, environment, or mental or emotional state; transport into a new location or state*"). *fetch*^1_n is mapped to *Transfer*$_c$ via equivalence while *carry_away*^1_n is mapped to *Removing*$_c$ via subsumption.

fetch^1_v and *carry_away*^1_v are antonyms in WordNet, but their corresponding SUMO classes are related via subclass in SUMO: *Removing*$_c$ is subclass of *Transfer*$_c$. In our opinion this is a case of imprecise mapping, although correct, the class *Transfer*$_c$ is too general for the synset *fetch*^1_v.

In Table 2, we report two incompatible problems with a correct but imprecise mapping.

4.3 Examples of Unresolved Problems

Next, we present two examples of problems that are unresolved due to different causes.

4.3.1 Case 1: Lack of knowledge

The first example corresponds to the problems that are not solved due to the lack of knowledge in the ontology and involves the synsets *machine*^1_v (*Making*$_c$+) and *machine*^1_n (*Machine*$_c$=). These synsets are related via morphosemantic relation *instrument*. However, there is no similar knowledge encoded in SUMO, so this example remains unresolved.

In Table 2, we report 45 problems with a correct mapping that are unresolved due to the lack of knowledge in the ontology.

4.3.2 Case 2: Lack of resources

The second example corresponds to resolvable problems that remain unresolved due to the lack of resources (mainly time) of ATPs. This example involves the synset $male_a^3$ linked to $Male_c+$ and the sysnset $female_a^1$ linked to $Female_c=$ as antomyms. In this case, although all the knowledge is correct the ATPs cannot find the prove for it.

Among the problems reported in Table 2, we have found 6 problems with a correct mapping that can be solved but that remain unresolved due to the lack of resources of ATPs.

5 Conclusion and Future Works

In this paper we have presented a detailed analysis of a sample of large benchmark of commonsense reasoning problems that has been automatically obtained from WordNet, SUMO and their mapping.

Based on this analysis, we can detect that although the framework enables the resolution of around 49 % of the total problems, only 36 % of the total are resolved for the good reasons: 60 problems resolved with a correct mapping. We have also detected that the mapping requires a general revision and correction: in particular, in the case of adjectives. On the contrary, the knowledge in SUMO involved in the revised proofs seems to be correct according to our commonsense knowledge. Further, the problems classified as incompatible enable the detection of misalignments between WordNet and SUMO, while the problems classified as unknown can be taken as a source of knowledge for the augmentation of SUMO. Actually, we are planning to develop an automatic procedure for the augmentation of SUMO on the basis of the knowledge in WordNet. Finally, we have detected some problems that can be solved on the basis of the knowledge of SUMO but that are not solved due to the limitation of resources of ATPs.

Acknowledgments

This work has been partially funded by the the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government, the projects CROSSTEXT (TIN2015-72646-EXP) and GRAMM (TIN2017-86727-C2-2-R) supported by the Ministry of Economy, Industry and Competitiveness of the Spanish Government, the Basque Project LoRea (GIU18/182) and BigKnowledge – *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018*.

References

J. Álvez, P. Lucio, and G. Rigau. 2012. Adimen-SUMO: Reengineering an ontology for first-order reasoning. *Int. J. Semantic Web Inf. Syst.*, 8(4):80–116.

J. Álvez, P. Lucio, and G. Rigau. 2017. Black-box testing of first-order logic ontologies using WordNet. *CoRR*, abs/1705.10217.

J. Álvez, P. Lucio, and G. Rigau. 2019. A framework for the evaluation of SUMO-based ontologies using WordNet. *IEEE Access*, 7:1–19.

C. Fellbaum, A. Osherson, and P.E. Clark. 2009. Putting semantics into WordNet's "Morphosemantic" Links. In Z. Vetulani and H. Uszkoreit, editors, *Human Language Technology. Challenges of the Information Society*, LNCS 5603, pages 350–358. Springer.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann. 2006. Modelling ontology evaluation and validation. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, LNCS 4011, pages 140–154. Springer.

M. R. Genesereth, R. E. Fikes, D. Brobow, R. Brachman, T. Gruber, P. Hayes, R. Letsinger, V. Lifschitz, R. Macgregor, J. Mccarthy, P. Norvig, R. Patil, and L. Schubert. 1992. Knowledge Interchange Format version 3.0 reference manual. Technical Report Logic-92-1, Stanford University, Computer Science Department, Logic Group.

A. Gómez-Pérez, M. Fernández-López, and O. Corcho-García. 2004. Ontological engineering. *Computing Reviews*, 45(8):478–479.

T. Gruber. 2009. Ontology. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 1963–1965. Springer.

M. Grüninger and M. S. Fox. 1995. Methodology for the design and evaluation of ontologies. In *Proc. of the Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI 1995)*.

N. Guarino and C. A. Welty. 2002. Evaluating ontological decisions with OntoClean. *Commun. ACM*, 45(2):61–65, feb.

N. Guarino and C. A. Welty. 2004. An overview of OntoClean. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, pages 151–171. Springer.

I. Horrocks and A. Voronkov. 2006. Reasoning support for expressive ontology languages using a theorem prover. In J. Dix et al., editor, *Foundations of Information and Knowledge Systems*, LNCS 3861, pages 201–218. Springer.

L. Kovács and A. Voronkov. 2013. First-order theorem proving and Vampire. In N. Sharygina and H. Veith, editors, *Computer Aided Verification*, LNCS 8044, pages 1–35. Springer.

I. Niles and A. Pease. 2001. Towards a standard upper ontology. In Guarino N. et al., editor, *Proc. of the 2nd Int. Conf. on Formal Ontology in Information Systems (FOIS 2001)*, pages 2–9. ACM.

I. Niles and A. Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In H. R. Arabnia, editor, *Proc. of the IEEE Int. Conf. on Inf. and Knowledge Engin. (IKE 2003)*, volume 2, pages 412–416. CSREA Press.

N. F. Noy and D. L. McGuinness. 2001. Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05 and SMI-2001-0880, Stanford Knowledge Systems Laboratory and Stanford Medical Informatics.

A. Pease and G. Sutcliffe. 2007. First-order reasoning on a large ontology. In Sutcliffe G. et al., editor, *Proc. of the Workshop on Empirically Successful Automated Reasoning in Large Theories (CADE-21)*, CEUR Workshop Proceedings 257. CEUR-WS.org.

A. Pease, G. Sutcliffe, N. Siegel, and S. Trac. 2010. Large theory reasoning with SUMO at CASC. *AI Communications*, 23(2-3):137–144.

A. Pease. 2009. Standard Upper Ontology Knowledge Interchange Format. Retrieved June 18, 2009, from `http://sigmakee.cvs.sourceforge.net/sigmakee/sigma/suo-kif.pdf`.

D. Ramachandran, R. P. Reagan, and K. Goolsbey. 2005. First-orderized ResearchCyc: Expressivity and efficiency in a common-sense ontology. In P. Shvaiko et al., editor, *Papers from the Workshop on Contexts and Ontologies: Theory, Practice and Applications (AAAI 2005)*, pages 33–40. AAAI Press.

R. Reiter, 1978. *On Closed World Data Bases*, pages 55–76. Springer, Boston, MA.

S. Schulz. 2002. E - A brainiac theorem prover. *AI Communications*, 15(2-3):111–126.

S. Staab and R. Studer. 2009. *Handbook on Ontologies*. Springer, 2nd edition.

G. Sutcliffe. 2009. The TPTP problem library and associated infrastructure. *J. Automated Reasoning*, 43(4):337–362.

Building the Cantonese Wordnet

Joanna Ut-Seong Sio
Palacky University
Olomouc, the Czech Republic
joannautseong.sio@upol.cz

Luis Morgado da Costa
Nanyang Technological University
Singapore
luis.passos.morgado@gmail.com

Abstract

This paper reports on the development of the Cantonese Wordnet, a new wordnet project based on Hong Kong Cantonese. It is built using the expansion approach, leveraging on the existing Chinese Open Wordnet, and the Princeton Wordnet's semantic hierarchy. The main goal of our project was to produce a high quality, human-curated resource – and this paper reports on the initial efforts and steady progress of our building method. It is our belief that the lexical data made available by this wordnet, including Jyutping romanization, will be useful for a variety of future uses, including many language processing tasks and linguistic research on Cantonese and its interactions with other Chinese dialects.

1 Introduction

1.1 Chinese and its Dialects

Chinese is generally treated as one language with many dialects for both cultural and political reasons. The dialects are spoken by people who mostly identify as a single nationality with a shared cultural history. Linguistically speaking, this unifying view is problematic as the dialects are not always mutually intelligible. Chinese is, more accurately, a family of genetically-related languages most probably descended from a form of late Old Chinese dating from the Han Dynasty or slightly earlier (with the possible exception of Min (Handel, 2015)). Various dialects, including Cantonese, had also been importing grammatical elements from neighboring languages (Yue-Hashimoto, 1991), creating dialectal variations that are more than the sum of language-internal changes. An arguably less confusing term 'Sinitic' is often used to refer to the Chinese languages (Handel, 2015). The term 'topolects' is coined by Mair (1991) to refer to Chinese dialects or, more generally, to speech varieties where the label of either 'language' or 'dialect' would be controversial. Nevertheless, for the purpose of this paper, we will continue to use the term 'Chinese' to refer to this family of languages and the term 'dialects' to refer to its variants while being fully aware of the complexity involved.

There are seven most recognised dialectal groups: Mandarin (or Northern Chinese), Xiang, Gan, Wu, Yue, Hakka and Min (Handel, 2015). Norman (1988) classifies the traditional seven dialectal groups into three larger groups: Northern (Mandarin), Central (Wu, Gan, and Xiang) and Southern (Hakka, Yue, and Min). Cantonese belongs to Yue, the Southern group, and it is often used as an alternative name for this whole group. The variety this Cantonese Wordnet is based on is Hong Kong Cantonese. Hong Kong Cantonese is often considered a prestige variety due to its association with the prosperous southern provinces as well as with the Cantonese culture of films and popular music.

1.2 Project Motivation

A few wordnets exist for Chinese languages. These efforts include some work on Pre-Qin Ancient Chinese (Zhang et al., 2017), Middle Ancient Chinese (Zhang et al., 2014), as well as multiple wordnets for Mandarin Chinese, namely: the Sinica Bilingual Ontological Wordnet (Huang, 2003; Huang et al., 2004, BOW), the Southeast University Chinese WordNet (Xu et al., 2008, SEW), the Chinese WordNet (Huang et al., 2010, CWN) and the Chinese Open Wordnet (Wang and Bond, 2013, COW). The Chinese Open Wordnet is the best and most recent effort to produce a

high quality wordnet for Mandarin Chinese, learning from previous analogous experiences and developed alongside a sense-tagged corpus (Tan and Bond, 2014; Wang and Bond, 2014; Seah and Bond, 2014).

Unfortunately, scholarly efforts often seem to forgo Cantonese, as there is a chronic absence of digital resources to study and process this important Chinese dialect. To further stress this problem, it is important to note that the significant differences between Mandarin, for which there are plenty of resources, and Cantonese make the idea of using Mandarin resources to process Cantonese fairly useless.

It was this chronic absence of Cantonese digital resources that ultimately fed our motivations to build the Cantonese Wordnet. Our motivation spawns from our belief that Cantonese should have plenty of open, computational tractable and linguistically rich resources, such as wordnets and corpora, that support scholarly work, as well as this language's maintenance and preservation – similar to what happens with Mandarin Chinese.

We would like our Cantonese Wordnet to support many Natural Language Processing tasks, such as speech recognition, word sense disambiguation, machine translation or information retrieval. And, at the same time, to also support the study of purely linguistic research topics, such as lexical semantics, tonal patterns, verb subcategorization, etc.

2 Cantonese: an Overview

Cantonese is the second most widely known Chinese dialect after Mandarin (Matthews and Yip, 1994). It is spoken in Guangdong Province, Guangxi Province, the Special Administrative Regions of Hong Kong and Macau, as well as diaspora communities in North America, Australia, Malaysia, Singapore, etc. According to Ethnologue,[1] there are 73 million Cantonese speakers worldwide. But despite the large number of speakers, credible online resources on Cantonese, free or otherwise, are limited, especially in comparison with Mandarin.

There is a considerable lexical overlap between Cantonese and Mandarin. Snow (2004, 49) mentions that the difference between Cantonese and Mandarin vocabulary ranges from 30-50%. Ouyang (1993, 23) estimates that about 1/3 of the lexical items used in regular Cantonese speech is not found in Mandarin. To give an example for a very common item,'umbrella' is yǔsǎn 雨傘 in Mandarin but ze1 遮 in Cantonese. In cases where they share the same lexical item, the item is always pronounced differently in the two dialects. For example, 'teacher' 老師 is pronounced as lǎoshī in Mandarin and lou5si1 in Cantonese. The vowels of the first syllables in each case are different, and the onsets of the second syllables are also different, not to mention tonal differences. Note that the romanization system is different here as well. Mandarin uses pīnyīn, which is based exclusively upon the pronunciation of the Beijing dialect. In Cantonese, Jyutping is used, a point we will come back to later.

In the existing Mandarin Chinese Wordnet, simplified characters are used. The simplified script, adopted in 1949, aims to alleviate some of the difficulty associated with use of the traditional script, as a measure to eradicate illiteracy. In Hong Kong, Macau and Taiwan, the traditional script is used, though in the former two, changes are happening rapidly since their return to China in 1997 and 1999, respectively.

Cantonese is primarily a spoken variant. A lot of lexical items, excluding those shared with Mandarin, do not have fixed agreed upon characters, these are often called 'characterless' words. It is not always easy to determine which character to use as there is no standardization. In some cases, multiple options are available, while in some other cases, no options are available. We will pick up on this issue in Section 4.3.

2.1 Jyutping Romanization System

Pīnyīn is the official romanization system of Mandarin Chinese or Pǔtōnghuà (lit. 'common speech'). And since Mandarin Chinese/Pǔtōnghuà and Cantonese have different phonological systems, a different romanization system is needed for Cantonese. Many romanization systems exists for Cantonese (e.g., Jyutping, S.L. Wong, Sidney Lau, Yale, the Government System, etc.) (Cheng and Tang, 2016). We adopt the Jyutping system, 粵拼, for the Cantonese Wordnet. Jyutping was

[1] https://www.ethnologue.com/language/yue

developed by the Linguistic Society of Hong Kong (LSHK) in 1993. Its formal name is The Linguistic Society of Hong Kong Cantonese Romanization Scheme.[2] Since its inception, it is used widely in academic papers as well as social media.

Cantonese syllables contain onset and rime. The rime can be further divided into the nucleus and coda. The lists of possible onset, nucleus and coda in Jyutping are shown in Table 1. /m/ and /ng/ are syllabic nasals, meaning they can appear on their own to form a syllable. Kataoka and Lee (2008) provide the correspondence between Jyutping, the International Phonetic symbol (IPA) and other Cantonese romanization systems.

Jyutping	phonemes
Onset	b, p, m, f, d, t, n, l, g, k, ng, h, gw, kw, w, z, c, s, j
Nucleus	aa, i, u, e, o, yu,oe, a, eo
Coda	p, t, k, m, n, ng, i, u

Table 1: Jyutping Syllable Struture

In Jyutping, tones are expressed numerically, using numbers 1 to 6. Table 2 shows how these numbers relate to their respective tonal contour using Chao's number (1 is the lowest and 5 is the highest) together with their description.

Jyutping	Chao's	description
1	53/55	high falling/high level
2	35	mid rising
3	33	mid level
4	21	low falling
5	13	low rising
6	22	low level

Table 2: Cantonese Tones

Traditional Chinese philology treats syllables with final stops (p, t, k) as distinct tone classes (checked tones), yielding a nine-tone system. Until recently, there was also a contrast between high level (55) and high falling (53). However, this distinction has collapsed for most speakers today.

Cantonese has a lot of homophones, characters that have the same pronunciation but have different meanings. To uniquely identify a lemma, both its Jyutping representation and its graph (character) are needed. For example, sing1 can mean 'to rise' 升 or 'star' 星. Without the character, it is ambiguous.

[2]https://www.lshk.org/jyutping

3 Methodology

There are two main methods to build wordnets (Vossen, 1998). The first method is known as the 'expansion' approach, where the structure of another wordnet is used as 'pivot', and the main work is essentially a translation effort – conserving the structure of the pivot wordnet and translating nodes of the hierarchy. The Princeton Wordnet (Fellbaum, 1998, PWN) is, by far, the most frequently used 'pivot' for projects that employ this approach. The second method is known as the 'merge' approach. This is usually a slower method, since no pivot structure is assumed, but it ensures a higher degree of freedom to more carefully model the structure of the wordnet based on the language in question, without depending on pre-assumed semantic relations. One of the immediate benefits of this approach is the ability to add new concepts that are not part of the 'pivot' language, a problem many wordnet projects that followed the 'expansion' approach have struggled with.

And while the 'merge' approach is perhaps more principled in theory, the major drawback from this approach is that it does not benefit from the parallel translations available from all other projects that used the same pivot. The best example of this benefit is the Open Multilingual Wordnet (Bond and Foster, 2013, OMW), a project that links dozens of open wordnets using PWN as the common structure. This language alignment is very useful for many NLP tasks, such as Machine Translation and Word Sense Disambiguation.

A recent addition to this discussion is the conception of the Collaborative Interlingual Index (Bond et al., 2016, CILI) – an open, language agnostic, flat-structured index that links wordnets across languages without imposing the hierarchy of any single wordnet. And even though PWN was the main contributor to the initial set of concepts present in CILI, this set is no longer constrained by it – multiple projects are now able to contribute to CILI's set of concepts, and gain the ben-

efits of multilingual alignments without the penalty of being frozen within some imposed structure. To the best of our knowledge, the quickest and easiest way to link to CILI and to access these language alignments without an imposed structure is, interestingly enough, to use the expansion approach with PWN hierarchy has pivot because all PWN concepts have direct links to CILI. And this was what we decided to do.

As we had no urgent need for a high coverage wordnet, for which multiple bootstrapping techniques are available to quickly create high coverage lower quality resources, we decided to build a high quality resource fully checked by native speakers. And although we knew from the start that building a wordnet from scratch would be very time-consuming, without going against our commitment to high quality, we decided to ease our task by leveraging on existing resources as much has possible.

We used the Chinese Open Wordnet (Wang and Bond, 2013, COW) as pivot. COW is a high quality hand-checked resource for Mandarin Chinese that was also created through the expansion approach (using PWN as pivot). This means that by linking our wordnet to COW, we would have easy access to PWN's concept IDs and, as a result, also to CILI.

The basic assumption of our method was that while Mandarin Chinese and Cantonese are fairly different languages, and it was clear from the start that resources from one language would not perform well in tasks for the other language, there is still a fair amount of overlap in the lexical usage. This is not without caveats, since Cantonese uses traditional characters and Mandarin Chinese uses simplified characters – and conversion from simplified to traditional characters is inherently lossy. That being said, we decided to automatically convert the lemmas in simplified Mandarin Chinese to traditional Chinese, and use this to jumpstart the manual construction of our Cantonese Wordnet.

Since the other Mandarin Chinese wordnets such as the Sinica Bilingual Ontological Wordnet (Huang, 2003; Huang et al., 2004, BOW), the Southeast University Chinese Wordnet (Xu et al., 2008, SEW) and the Chinese Wordnet (Huang et al., 2010, CWN) included lemmas that were not present in COW, we started by building a small 'Chinese Wordnet' from the union of all four Chinese wordnets: COW, BOW, SEW and CWN. This was fairly easy since all these wordnets were linked to PWN's hierarchy. Next, we used Hanziconv[3] to convert all lemmas from simplified Mandarin Chinese to traditional characters. And finally, we generated a list of all candidate senses (with lemmas converted into traditional characters) that satisfied any of these three criteria:

- Senses that belong to the 4,960 'core' concepts in Princeton WordNet (Boyd-Graber et al., 2006) – a usual measure for coverage of wordnet resources;
- Senses from all concepts in two sense tagged Sherlock Holmes stories, as reported by NTUMC (Tan and Bond, 2014); and
- Senses from any concept with sense sum-frequency score of one or higher, as reported by the PWN (i.e. most concepts yield sum-score of 0);

3.1 Human Validation and Jyutping

The data generated by the process explained above generated a list of 47,499 candidate senses, spanning over 9,340 synsets. Based on this information, we created a spreadsheet for our human validation task. As of this moment, a single Cantonese native speaker, who is also a trained linguist with extensive work on Cantonese language is manually checking, correcting and adding to this data.

An example of this spreadsheet is shown in Table 3. This spreadsheet contains the candidate Cantonese lemmas (converted to traditional characters from one of the existing Mandarin lemmas), English lemmas (provided by the PWN), Mandarin lemmas (provided by the collection of Chinese wordnets), English definitions and examples (provided by the PWN), and the synset ID of the PWN3.0.

The Jyutping romanization is not produced automatically. It is, in fact, being added by hand by the lexicographer. To our knowledge, there are no open Jyutping dictionaries available under an open license. For this reason, we decided to include this valuable resource in our wordnet. Having Jyutping romanization

[3]https://pypi.org/project/hanziconv/

Cantonese Lemma	English Lemmas	Mandarin Lemmas	English Definitions	English Examples	Synset
今夜 [deleted]	[deleted]	[deleted]	[deleted]	[deleted]	[deleted]
今晚, gam1 maan5	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r
今晚, gam1 maan1	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r
今 晚 黑, gam1 maan5 hak1	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r
今 晚 黑, gam1 maan1 hak1	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r

Table 3: Human Validation and Jyutping (example)

for each sense will not only facilitate searching, but can also be very useful for a variety of other tasks, such as speech recognition or even for educational purposes. We will make use of the new structure provided by the WN-LMF format to cluster Jyutping romanizations as variants inside the canonical lemma (i.e. the traditional Chinese characters).

The process explained in the section above generated one candidate Cantonese sense for each available Mandarin sense inside each concept. In the example shown in Table 3, the concept 00079499-r contained two Mandarin senses: jīn yè 今夜, and jīn wǎn 今晚. Both these lemmas have the same form in simplified and traditional Chinese. This resulted in two lines produced (the top two lines in Table 3). The human validation task comprised:

1. asserting if the candidate sense in each line provided was a correct Cantonese sense – incorrect senses would be deleted (see Table 3, line 1);
2. adding Jyutping romanization for each correct sense – senses with more than one pronunciation required the line to be copied and the corresponding romanization added to the new line (see Table 3, lines 2-3);
3. adding any missing senses that were not suggested by the conversion of the Mandarin lemmas. This was a non-exhaustive search, and it depended on the lexicographer's ability to recall missing senses (see Table 3, lines 4-5);

At this moment, our lexicographer has hand-checked 18,168 (38.25%) of the total set of candidate senses (i.e. 47,499 senses). Out of the total number of candidate senses checked, 8,295 (45.7%) were kept (i.e. the conversion of Mandarin lemmas was correct), which is in line with Snow's (2004, 49) predictions. In addition to these converted senses, a total of 3,797 new senses were added by the lexicographer (i.e. that were not suggested by the conversion from simplified Mandarin Chinese) – this comprises about 31.4% of the total number of senses we currently have in our wordnet, and which is in line with Ouyang's (1993, 23) predictions concerning the ratio of exclusive Cantonese senses. In total, our wordnet currently has 12,092 senses (a summary of this release's statistics is provided in Section 5).

4 Issues

4.1 Separated and Intervening Lexemes

What is represented by one lemma in English sometimes requires two lexemes separated from each other with an intervening lexeme in Cantonese. For example, 'to punch' in the sense of 'to deliver a quick blow' is expressed as [daai2...jat1kyun4], literally 'hit...one punch' (打... 一拳), where ... is the slot for the recipient of the punch, the object of the verb. Another example is 'to fire' in the sense of 'terminate the employment of', which can be expressed as [gaak3...zik1] (one of the many options in Cantonese), 'remove...duty' (革... 職), where ... is the slot of the person being fired, the object. This is essen-

tially different from the English 'pick up' and 'pick...up' cases (where ... is the object) as [daai2jat1kyun4 + 'object'] and [gaak3zik1 + 'object'] are both ungrammatical – the separation is obligatory. In view of this, we have used separated lexemes (with ...) whenever it is necessary to be faithful to the English concept, a practice also adopted by COW.

4.2 Compositionality of Telic Verbs

In many cases, the translated term in Cantonese is compositional. For example 'to remember' in the sense of 'recall knowledge from memory', is nam2hei2 諗起 in Cantonese, where a post-verbal particle hei2 meaning 'up' is needed. Sybesma (1997) points out that Mandarin does not have monomorphemic counterparts for English verbs like 'see', 'hear', and 'find', which qualify as achievements (indeed he claims that Chinese has no inherently telic verbs at all). The Mandarin counterparts of these verbs are compound verbs, where the second constituent expresses the attainment of the result ('phase complement' in Chao (1968); see also Li and Thompson (1981)), e.g., kàndào 看到 'look-arrive > see'; kànjiàn 看見 'look-see > see'; tīngjiàn 聽見 'listen-see > hear'; zhǎodào 找到 'look for-arrive > find'. The situation is the same in Cantonese. To ensure we have high quality translation equivalents, these particles are included in the lemmas (the same procedure is adopted by COW). The consequence is that such entries can be analyzed as compositional.

4.3 The Lack of Standardization in Written Cantonese

Cantonese is primarily a spoken dialect. Cantonese has never been subjected to rigorous and formal standardization, despite efforts of lexicographers which resulted in a few Cantonese-standard Chinese dictionaries and Cantonese word lists (Li, 2000). Cantonese school children are not taught how to read or write Cantonese. The knowledge of written Cantonese among its speakers arises informally through exposure to its pervasive use (Bauer, 2018).

Written Cantonese is mainly used for informal or less serious kind of communication (Snow, 2004, 18), but is not uncommon. It is used regularly in advertising (e.g. signs, posters, novels) as well as newspapers (e.g. Apple Daily, a popular newspaper in Hong Kong). Written Cantonese conveys a greater degree (compare with standard Chinese) of 'informality, directness, intimacy, friendliness, casualness, freedom, modernity and authenticity' (Bauer, 2018, 4). At least partly due to the special situation in Hong Kong for a long time, where children speak Cantonese but write in standard Chinese (the situation has changed since the handover in 1997), written Cantonese ranges over a continuum. On the one end, there are texts that are essentially standard Chinese but with a few Cantonese items, on the other end are texts that are written entirely in Cantonese (Snow, 2004, 60-61).

There is substantial overlap between Mandarin and Cantonese vocabulary. For shared vocabulary items, e.g., 飯 'rice', fàn in Mandarin and faan6 in Cantonese, the traditional version of the same character is used, and with a different pronunciation.

It is estimated that about one-third of the lexical items in Cantonese are not shared with Mandarin (Ouyang, 1993, 23). This also includes some very basic vocabulary, such as the negator, which is 不 bù in Mandarin and 唔 m4 in Cantonese, or very basic content words like 'see', which is 看 kàn in Mandarin, but 睇 tai2 in Cantonese. For Cantonese-specific lexical items, the choice of the characters is not always obvious due to the lack of standardization.

The standardization of written Cantonese lexical items exhibits a gradience, ranging from items like the negator 唔 m4 and 'see' 睇 tai2, which are not controversial, to items which are regularly represented phonetically with English letters in its written forms in online forums, e.g. *hea* he3 'to laze around'. In-between the two extremes, there are many cases where two or more characters are used to represent the same lexical item. For example the word bei2 'to give' can be written with 4 different characters, 比, 俾, 畀, 被 (Bauer, 2018, 135). For this first version of the Cantonese Wordnet, items which are only represented by English letters are not listed. For cases where multiple characters are used, all options will be given whenever possible. For discussion on strategies on how Cantonese

characters are formed, see Li (2000) and Bauer (2018).

4.4 Alternation in Pronunciation

In the Cantonese Wordnet, there are many cases where a particular character is given multiple pronunciations. The two common causes for alternation is *pinjam* 變音 'changed tone' and laan5jam1 懶音 'lazy pronunciation'.

Many morphological constructions in Cantonese are expressed solely or partly by tone change (Yu, 2009). Traditional descriptive linguistic literature of Cantonese refers to this *pinjam* 變音 process. Table 4 shows some examples of tone change cases in deverbal nominalization (Yu, 2009).

character	verb	noun
掃	'to sweep' sou3	'broom' sou2
磅	'to weight' bong6	'scale' bong2
油	'to grease' jau4	'oil' jau4

Table 4: Cantonese Tone Change (I)

The term laan5jam1 ('lazy pronunciation', lǎnyīn in Mandarin, literally meaning 'lazy pronunciation') has been used in recent years to refer to ongoing sound changes in Hong Kong Cantonese. This term designates the use of a variety of consonant variants in the speech of younger native speakers of Hong Kong Cantonese (Ding, 2010). One example is syllable-initial /n/ and /l/ merger (/n/ > /l/), a phenomenon that started around the 70s. This is shown in the Table 5. There are many other examples of 'lazy pronunciation' (e.g., /ng/ > /m/) in Cantonese.

character	meaning	jyutping
男	'male'	naam4 **or** laam4
女	'female'	neoi5 **or** leoi5
呢度	'here'	nei1 dou6 **or** lei1 dou6

Table 5: Cantonese Tone Change (II)

In addition to *pinjam* 變音 'changed tone' and laan5jam1 懶音 'lazy pronunciation', there are also cases of tone change, which are not clear what the motivation is. Nevertheless, whenever possible, all options were captured by our wordnet.

4.5 The Continuum between Spoken and Written Cantonese

Cantonese has different registers (e.g., everyday conversation vs. news report). A lot of words which are too formal to use in regular conversation might appear in TV broadcast, or formal speeches and thus some more formal versions of such terms (as long as they are deem possible in Cantonese) are also included in our wordnet with the aim of covering the range of registers. The consequence is that the boundary is not always clear. When in doubt, the decision was always to include such items.

The question as to what to include can be determined in a more objective way in the future. We would like to experiment with Cantonese texts of various registers, using both the Cantonese and Mandarin wordnets in parallel to help better understand and identify words that were not included as part of the Cantonese Wordnet. In time, we hope to establish the extent of shared vocabulary items between Mandarin and Cantonese, as well as to identify uniquely Cantonese items.

5 Statistics

Table 6 provides a summary of the current state of the Cantonese wordnet.

POS	No. synsets	%	No. senses	%
nouns	1,830	(0.52)	5,114	(0.42)
verbs	975	(0.28)	3,227	(0.27)
adjective	565	(0.16)	3,044	(0.25)
adverb	163	(0.05)	707	(0.06)
Total	3,533	-	12,092	-

Table 6: WN Statistics

In total, the first version of our wordnet covers a bit over 3,500 concepts using over 12,000 senses. The part-of-speech distribution is generally in sync with other projects, such as the PWN – with perhaps a weaker dominance of nominal senses and concepts to a slight heavier presence of their verbal counterparts. Our current version covers 35.81% (n = 1,776) of the 'core' PWN concepts.

Since our wordnet is currently pivoting on the hierarchy provided by PWN, through COW, we have no information about semantic relations to report. In further stages of

our project, however, we might revise this position and consider taking advantage of CILI to adapt our wordnet's semantic hierarchy to better fit the assumptions of Cantonese native speakers.

As mentioned above, in Section 3.1, the process of human validation is still ongoing, and we expect to provide an update to these statistics in the camera-ready version of this paper.

6 Release

This Cantonese Wordnet will be released under a Creative Commons Attribution 4.0 International License (CC BY 4.0)[4].

Keeping up with the recent changes and requirements of the OMW, the Cantonese Wordnet will be primary released and supported for the recent WN-LMF format,[5] developed and maintained by the Global WordNet Association. The use of WN-LMF is not only required by the most recent version of the OMW, but is also an essential vehicle to access the new Collaborative Interlingual Index (Bond et al., 2016, CILI). Once linked to CILI, our wordnet will be able to contribute with new concepts, present only in Cantonese such as sap1jit6 濕熱, an adjective with the literal meaning of 'hot wet' (it describes a general negative health condition resulting from an unhealthy lifestyle, e.g. smoking, sleep deprivation, etc.), or gung1zyu2beng6 公主病, a noun that literally means 'princess disease'. It describes girls who are over-confident, over-reliant and demand princess-like treatment.

In addition, this release will also include the tab-separated-value format used by the original OMW specifications. These files are still very useful for their size, simplicity, and legacy compatibilities with existing systems. One such example is the use of this data through NLTK: Python Natural Language Toolkit (Bird et al., 2009) – which currently still uses this legacy format. However, the simplicity of this format doesn't come without a cost. Due to the flatter nature of this format, the Jyutping romanization of Cantonese lemmas will be added as separate lemmas (i.e. effectively doubling the number of words and senses within this format).

The data for this wordnet is available on Github[6].

7 Conclusions and Future Work

This paper presented the ongoing efforts to build a Cantonese Wordnet. We have motivated this project with the lack of digital resources available for Cantonese – a major Chinese dialect. We have introduced our methodology, which is to use existing Mandarin wordnets to project Cantonese candidate senses. So far our wordnet includes over 3,500 concepts and over 12,000 senses. We have discussed some specific challenges encountered while building our wordnet and how we addressed them. We hope that this new open resource will promote a variety of future uses, including language processing tasks and linguistic research.

We would like to continue our efforts to improve the coverage and quality of our Cantonese Wordnet. This would include:

- finish validating and revising the list of candidate senses generated through the methods explained in Section 3 (so far we have completed 38.25% of this validation);
- add example sentences for each sense, which would be the start of an open, sense-tagged Cantonese corpus;
- given that Cantonese is predominantly used in speech, we would also like to add audio recording for each pronunciation of each lemma;

Once the Cantonese wordnet reaches a sufficient coverage, we would like to use it to research a variety of topics, including:

- study the amount of Mandarin words that have entered common Cantonese speech and writing and, conversely, when and why some Mandarin words are never used Cantonese;
- study the morphologically conditioned tone changes in Cantonese such as *pinjam* and other less understood phenomena; and
- shed some light on the potential relation between register (formal register is often tied to written Chinese, which is based

[4]https://creativecommons.org/licenses/by/4.0/
[5]https://github.com/globalwordnet/schemas
[6]https://github.com/lmorgadodacosta/CantoneseWN

on Mandarin) and tone change (a speech phenomenon);

Acknowledgments

We thank the reviewers for their comments. Joanna Ut-Seong Sio gratefully acknowledges the research funding received from the Faculty of Arts, Palacky University in Olomouc through its Research Grant Scheme (funding cycle 2019-2021).

References

Robert Bauer. 2018. Cantonese as written language in hong kong. *Global Chinese*, 4(1):103–142.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*, pages 29–36.

Yuen Ren Chao. 1968. *Language and symbolic systems*, volume 457. CUP Archive.

Siu-Pong Cheng and Sze-Wing Tang. 2016. *Cantonese Romanizaton*. Routledge.

Picus Sizhi Ding. 2010. Phonological change in hong kong cantonese through language contact with chinese topolects and english over the past century. *Marginal dialects: Scotland, Ireland and beyond*, pages 198–218.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Zev Handel. 2015. The classification of chinese. *The Oxford handbook of Chinese linguistics*, page 34.

Chu-Ren Huang, Ru-Yng Chang, and Hshiang-Pin Lee. 2004. Sinica bow (bilingual ontological wordnet): Integration of bilingual wordnet and sumo. In *LREC*.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Chu-Ren Huang. 2003. Sinica bow: integrating bilingual wordnet and sumo ontology. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 825–826. IEEE.

Shin Kataoka and Cream Yin-Ping Lee. 2008. A system without a system: Cantonese romanization used in hong kong place and personal names. *Hong Kong Journal of Applied Linguistics*, 11(1):79–98.

Charles Li and Sandra Thompson. 1981. A functional reference grammar of mandarin chinese. *Berkeley, CA: University of California Press. Find this author on*.

David CS Li. 2000. Phonetic borrowing: Key to the vitality of written cantonese in hong kong. *Written Language & Literacy*, 3(2):199–233.

Victor H Mair. 1991. *What is a Chinese" dialect/topolect"?: Reflections on some key Sino-English Linguistic Terms*.

Stephen Matthews and Virginia Yip. 1994. *Cantonese*. Routledge.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

Jueya Ouyang. 1993. Putonghua guangzhouhua de bijiao yu xuexi [comparison and study of putonghua and cantonese].

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 82.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

Rint Sybesma. 1997. Why chinese verb-le is a resultative predicate. *Journal of East Asian Linguistics*, 6(3):215–261.

Liling Tan and Francis Bond. 2014. NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Shan Wang and Francis Bond. 2014. Building the sense-tagged multilingual parallel corpus. In *LREC*, pages 2403–2409.

Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual chinese-english wordnet. In John Domingue and Chutiporn Anutariya, editors, *The Semantic Web*, pages 302–314, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alan Yu. 2009. Tonal mapping in cantonese vocative reduplication. In *Annual Meeting of the Berkeley Linguistics Society*, volume 35, pages 341–352.

Anne Yue-Hashimoto. 1991. The yue dialect. *Journal of Chinese Linguistics Monograph Series*, (3):292–322.

Yingjie Zhang, Bin Li, Xiaoyu Wang, Xueyang Liu, and Jiajun Chen. 2014. Mapping word senses of middle ancient Chinese to WordNet. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 446–450. IEEE.

Yingjie Zhang, Bin Li, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2017. Pqac-wn: constructing a wordnet for pre-qin ancient chinese. *Language Resources and Evaluation*, 51(2):525–545.

Deep Learning in Event Detection in Polish

Łukasz Kobyliński, Michał Wasiluk
Institute of Computer Science,
Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
lkobylinski@ipipan.waw.pl, m.wasiluk89@gmail.com

Abstract

Event detection is an important NLP task that has been only recently tackled in the context of Polish, mostly due to lack of language resources. The available annotated corpora are still relatively small and supervised learning approaches are limited by the size of training datasets. Event detection tools are very much needed, as they can be used to annotate more language resources automatically and to improve the accuracy of other NLP tasks, which rely on the detection of events, such as question answering or machine translation. In this paper we present a deep learning based approach to this task, which proved to capture the knowledge contained in the training data most effectively and outperform previously proposed methods. We show a direct comparison to previously published results, using the same data and experimental setup.

1 Introduction

The task of identifying events in natural language has a direct impact on the effectiveness of many other tasks in the area of natural language processing. An obvious example is the task of question answering, where the knowledge base has the form of a collection of texts in natural language (Saurí et al., 2005). The answer to the question *When was the current president elected?* requires recognition of the current system time, determining who the current president (of Poland, by implicit assumption) is and identifying the event of election. Other NLP tasks directly influenced by the results of event detection include summarization (Filatova and Hatzivassiloglou, 2004), (Vanderwende et al., 2004), (Li et al., 2006) and machine translation (Horie et al., 2012). In the first case, the events identified in the text allow organizing the content of the summarized document by topics and ordering them chronologically. In the case of machine translation, event detection may be used to create the intermediate knowledge representation layer that is independent of any natural language, which is then used to form the final translation.

In the case of the Polish language, there are only a few published papers on the identification of temporal expressions in natural language text. This is largely due to the current lack of resources, enabling this type of study. For example the authors of (Jarzębowski and Przepiórkowski, 2012) use parallel corpora and annotation projection to Polish to gather the necessary evaluation material. They use the National Corpus of Polish (Przepiórkowski et al., 2012), which contains the basic annotation of simple temporal expressions. Specifically, the manually annotated subcorpus of the NCP includes such tags as: *date* (calendar dates, such as *24 October, 1945*) and *time* (hours, minutes and seconds, e.g. *five after twelve*).

The recently published subcorpus of the KPWr corpus (Kocoń and Marcińczuk, 2015) has been specifically annotated with temporal expressions and events, using an adaptation of the TimeML specification (Saurí et al., 2006). This collection of annotated texts along with additional dictionaries has been used in (Kocoń and Marcińczuk, 2016) to train a CRF-based classifier for the task of identifying events.

2 Event Detection Task

We define the task of detecting events in text as a problem of identifying tokens or token sequences, which should be annotated as an event mention according to the TimeML specification, adapted to Polish by (Marcińczuk et al., 2015). As in the original TimeML specification, we understand events as situations that happen or occur, an

"event is anything that takes place in time (date, time and/or duration) and space (has a location), may involve agents (executor or participants), may contain or be part of other events and may produce some outcome (object)." (Marcińczuk et al., 2015). We aim to classify identified events into one of the following categories, defined by the specification:

- **action** (a dynamic situation which occurs in time and space),
  e.g. *run, fly, hit*,
- **state** (a static situation, which does not change over a period of time),
  e.g. *stand, sit, remain*,
- **reporting** (a dynamic situation where an agent informs about an event or narrates an event),
  e.g. *explain, tell, inform*,
- **perception** (a physical perception of an event by an agent),
  e.g. *see, hear, observe*,
- **aspectual** (indicates a change of a phase of another event),
  e.g. *begin, start, interrupt*,
- **i_action** (intensional action, a situation, where an agent declares his or her will to perform an action or give a command),
  e.g. *try, promise, delay*,
- **i_state** (intensional state, a possible action or state; an agent refers to some possible event, which may or may not occur in the future),
  e.g. *believe, fear, wish*.

The goal of the task is thus to create an annotation layer, which associates event category labels with corresponding tokens. Below is an example annotation, taken from the training corpus (other annotation layers not shown here for readability):

(1.) Po tym **zwycięstwie**$_{\text{action}}$ MKS został liderem grupy 2.

(1.) *After this* ***victory***$_{action}$ *MKS became the leader of group 2.*

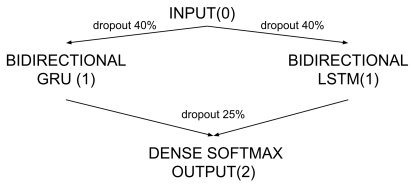


Figure 1: Branched bi gru-lstm architecture.

3 Deep Learning Approach to Event Detection

Preprocessing In the first stage of the proposed method we preprocess the available data and generate feature vectors for the neural network. We scan through the text using a fixed-length processing window: for each token in a sentence a sequence composed of this token (in the center of the window) and its W nearest neighbors within the sentence is generated. Thus, the sequence has a length of $2W + 1$, where W is called the *window size*. The neural network takes as an input a sequence of feature vectors of individual tokens and classifies the central token into one of previously described categories, with an additional *not event* class for not relevant tokens.

Features We use two kinds of embeddings for the real-valued feature vector generation:

1. Simple indexed embeddings, which turn positive integers (indexes) into dense vectors of fixed size by means of simple matrix multiplication:
   — **struct** — structure of a token (vector size: 5) - a token string with all digits replaced by 'd', lowercased characters replaced by 'x', uppercased characters replaced by 'X' and any other character replaced by '-' ("Warszawa-2017" → "Xxxxxxx-dddd"). A packed structure is a structure with all neighbouring duplicate code characters removed ("Xxxxxxx-dddd" → "Xx-d"),
   — **position** — position of a token in a sequence (3).
2. Pretrained Word2vec (Mikolov et al., 2013) embedding models:
   — **orth** — trained on orthographical word forms from National Corpus of Polish and the Polish Wikipedia (vector size: 300),

Annotation	action	aspectual	i_action	i_state	perception	reporting	state
Number	12861	316	717	1205	149	341	1318

Table 1: Annotations in KPWr-540 by category.

architecture	accuracy	F1						
		action	aspectual	i_action	i_state	perception	reporting	state
br bi gru-lstm	96.291	86.06	74.46	55.73	80.39	90.82	77.60	74.92
br bi lstm-lstm	96.282	86.18	74.22	57.89	77.62	88.28	77.21	74.58
br bi gru-gru	96.229	85.83	72.03	56.77	78.92	89.97	78.55	73.57
br gru-lstm	96.181	85.75	72.68	54.76	77.11	88.20	77.30	73.65
br gru-gru	96.174	85.75	73.80	55.55	76.89	87.33	76.82	71.91
br lstm-lstm	96.162	85.66	73.24	54.87	76.94	85.93	75.33	73.16
bi gru	96.117	85.44	72.97	53.32	79.39	88.42	75.28	72.15
bi lstm	96.098	85.47	71.34	52.55	76.44	87.78	76.19	73.72
lstm	95.937	84.87	71.88	47.57	75.83	74.07	71.91	72.14
gru	95.834	84.47	71.20	47.82	75.20	73.17	71.28	69.38

Table 2: A comparison of network architectures, ordered by overall accuracy (80—20 data split, average from 5 tests, KPWr-540, W = 1, dropout = 0.4, {'hypernym-1', 'lemma', 'orth', 'class'} embeddings).

— **base** — trained on lemmatized word forms (300),
— **class** — trained on POS classes of words from National Corpus of Polish and the Corpus of Polish language of the 1960s (PL1960)[1] (30),
— **ctag** — trained on POS tags of words (300),
— **hypernym-1** — trained on hypernyms of words taken from plWordNet[2] (100),
— **synonym** — trained on synonyms of words taken from plWordNet (100).

In the case of word sense ambiguity during generation of plWordNet-based features (several matching synonyms or hypernyms), the first base form common to all synonyms from all matching synsets is chosen (in alphabetically sorted order). If there is no common base form, or there is no match, the base form of the original token is selected.

Word2vec embedding models were trained in two main steps:

1. Replacement of all tokens in the corpus with corresponding values of the given feature.

2. Training of the w2v model on this newly created corpus using the gensim library[3].

Word2vec feature vectors are assigned to individual tokens by computing given feature value (lemma, hypernym etc.), which then is directly mapped to corresponding feature vector. Eg. *ludzie -> człowiek -> [feature vector]*.

The input vector of an individual token is a concatenation of all component feature vectors. The size of the input vector of the individual token in a sequence with all described features included was 1138 elements.

Network architecture Based on preliminary experiments (described in the Experimental Results section), we have chosen a network consisting of two distinct subnetworks as the most promising for further experiments. The network is split into two branches, Bi-LSTM and Bi-GRU subnetworks. Each of these subnetworks takes the same input, but with a different random dropout applied to it. Bi-LSTM and Bi-GRU can simultaneously model word representation with its preceding and following information. They are composed of two LSTM/GRU neural networks with a hyperbolic (tanh) and hard sigmoid activation functions. The forward LSTM/GRU allows to model the preceding contexts, and a backward LSTM/GRU to model the following contexts respectively.

[1] http://clip.ipipan.waw.pl/PL196x
[2] http://plwordnet.pwr.wroc.pl/wordnet/, (Maziarz et al., 2016)
[3] https://radimrehurek.com/gensim/

dropout	accuracy	F1						
		action	aspectual	i_action	i_state	perception	reporting	state
0.4	96.29	86.06	74.46	55.73	80.39	90.82	77.60	74.92
0.5	96.28	86.14	74.00	56.84	79.42	89.45	76.89	73.84
0.3	96.27	86.00	73.40	56.64	80.11	89.43	78.07	73.91
0.6	96.24	86.06	73.19	56.62	78.72	89.51	77.94	73.96
0.2	96.17	85.68	71.91	54.53	78.57	87.14	77.21	72.97
0.1	96.12	85.36	70.97	53.19	78.98	86.32	76.19	73.61
0.7	96.08	85.62	72.37	54.15	77.45	88.63	76.75	73.12
0.0	96.03	85.11	71.35	50.45	78.59	81.79	74.25	71.57
0.8	95.72	84.26	71.30	49.85	77.22	84.80	73.27	70.56
0.9	95.14	82.38	71.85	35.05	75.69	28.42	67.38	61.40

Table 3: The influence of the input dropout parameter on network accuracy (80-20 data split, average from 9 tests, branched bi gru-lstm architecture, KPWr-540, W = 1, {'hypernym-1', 'lemma', 'orth', 'class'} embeddings).

We flatten and concatenate the bidirectional sequence features learned from the subnetworks and apply random dropout to the result. Then, we use a dense softmax approach to perform final classification. The architecture of the network has been presented on Figure 1.

We train our model with categorical cross-entropy loss function and Adam optimizer (Kingma and Ba, 2014) with small learning rate decay. For GRU and LSTM we use glorot (Glorot and Bengio, 2010) and orthogonal (Saxe et al., 2013) initializers.

4 Experimental Results

Data The KPWr-540 corpus (Kocoń and Marcińczuk, 2015) has previously been used to train machine learning methods for the task of event detection. Here we use the same dataset to allow direct comparisons with previously published approaches. The dataset contains 540 documents, 6 915 sentences (948 sentences without any event utterance) and 121 747 tokens. In total, there are 17 078 human-made annotations in the corpus. The breakdown of the annotation types has been presented in Table 1. The annotations consist predominantly of a single token, only 4 annotations have a token span length of 2.

Preliminary experiments To determine the appropriate network architecture for the stated problem we have conducted a series of preliminary experiments on the available dataset, using a 80—20 split between train and test data. The most representative differences between network architectures, as measured during these experiments, have been presented in Table 2. The accuracy column represents overall classification accuracy of the network (*no event* class included).

In further experiments we have measured the influence of the dropout parameter on classification accuracy (the results are presented in Table 3) and we have found the optimal set of features (the results are presented on Figure 2).

Evaluation The final evaluation of the proposed method accuracy has been performed using a 10-fold cross-validation on the available data. In these experiments each fold's training set was additionally split into 2 parts: train (80%) — used for neural network training and validation (20%) — used for early stopping and best model selection. We have also evaluated two approaches to the train set splitting: performing a simple split and a balanced split with preserved ratio of event category samples. The weights were balanced for each class.

The results of the final comparison of the proposed method to the CRF-based approach presented in (Kocoń and Marcińczuk, 2016) has been shown in Table 4. The presented deep-learning approach proved to perform better for each event category, as measured by the F1 score.

5 Conclusions and Future Work

In this paper we have applied a deep-learning approach to the problem of detecting events in text. As in many other NLP tasks, modern neural networks proved to perform very well in this domain and outperformed the previously proposed method, which was based on Conditional Random Fields.

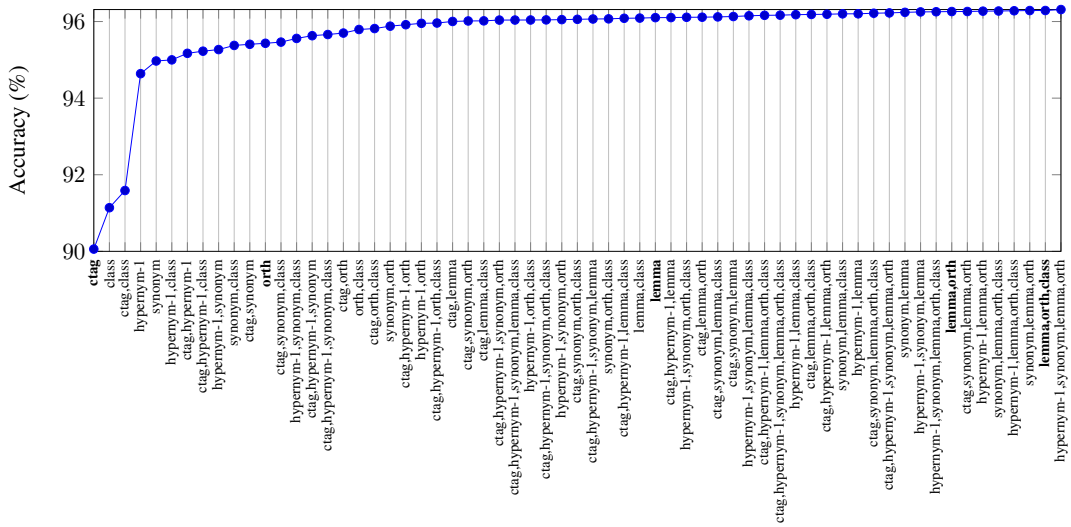


Figure 2: The impact of word embeddings configuration on overall classification accuracy (80-20 data split, average from 5 tests, KPWr-540, W = 1, dropout = 0.4).

Category	branched bi gru-lstm W=1			Liner2 w/o dictionaries			Liner2 with dictionaries (*)		
	P	R	F1	P	R	F1	P	R	F1
action	**84.90**	**88.33**	**86.57**	82.51	84.90	83.69	82.49	83.87	83.18
aspectual	85.87	**72.96**	**78.67**	87.56	60.13	71.29	**87.58**	59.24	70.68
i_action	66.89	**58.58**	**62.12**	**67.48**	42.54	52.18	63.56	40.92	49.79
i_state	84.35	**82.60**	**83.38**	84.35	78.26	81.19	**85.19**	77.56	81.20
perception	85.17	**75.61**	**79.33**	**97.53**	53.02	68.70	85.90	55.37	67.34
reporting	69.29	**66.65**	**67.11**	**75.00**	57.18	64.89	71.13	51.30	59.61
state	**73.03**	**69.09**	**70.86**	71.84	61.15	66.07	68.10	62.17	65.00

Table 4: Comparison of the best performing network architecture against the previously proposed CRF-based approach. Ten-fold cross-validation on the KPWr-540 corpus. (*) Results taken from (Kocoń and Marcińczuk, 2016).

It is interesting to note that features based on words (**orth**, **lemma** in Figure 2) influenced the resulting accuracy the most, proving that such embeddings carry essential information for this task. On the other hand, features based on part-of-speech tags (**ctag**, **class**) were among the least informative. A characteristic feature of processing inflected languages is the importance of lemmatization and including lemmas in the feature set. The large number of inflected word forms in languages such as Polish (and other Slavic languages), makes it more difficult for word-form embeddings to capture information that is properly generalized. Generating embeddings from lemmas helps to solve the problem, as long as the lemmatization is accurate and does not introduce additional disambiguation difficulties.

In future work we would like to tackle a more general task of event recognition in text, including the identification of textual arguments of events. This may include such entities as place names (where the event takes place), time and date specifications (when it takes place), or person names (agents or beneficiaries of an event). We would also like to analyze the relationships occurring between several events recognized in a text fragment (e.g. event identity).

Acknowledgements

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In *Proceedings of the ACL Workshop on Summarization*, pages 104–111.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May. PMLR.

André Horie, Kumiko Tanaka-Ishii, and Mitsuru Ishizuka. 2012. Verb temporality analysis using Reichenbach's tense system. In *Proceedings of COLING 2012: Posters*, pages 471–482, Mumbai, India, December. The COLING 2012 Organizing Committee.

Przemysław Jarzębowski and Adam Przepiórkowski. 2012. Temporal information extraction with cross-language projected data. In Hitoshi Isahara and Kyoko Kanzaki, editors, *Advances in Natural Language Processing: Proceedings of the 8th International Conference on NLP, JapTAL 2012, Kanazawa, Japan, October 22-24, 2012*, volume 7614 of *Lecture Notes in Artificial Intelligence*, pages 198–209. Springer-Verlag, Heidelberg.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jan Kocoń and Michał Marcińczuk. 2015. KPWr events. CLARIN-PL digital repository.

Jan Kocoń and Michał Marcińczuk, 2016. *Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents*, pages 12–19. Springer International Publishing, Cham.

Wenjie Li, Mingli Wu, Qin Lu, Wei Xu, and Chunfa Yuan. 2006. Extractive summarization using inter- and intra- event relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 369–376, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michał Marcińczuk, Marcin Oleksy, Tomasz Bernaś, Jan Kocoń, and Michał Wolski. 2015. Towards an event annotated corpus of Polish. *Cognitive Studies*, 15:253–267, 12.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. PlWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. 2005. Evita: A robust event recognizer for qa systems. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 700–707, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Saurí, Jessica Littman, Bob Knippen, Andrea Gaizauskas, Robert abd Setzer, and James Pustejovsky. 2006. Timeml annotation guidelines.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6120.

Lucy Vanderwende, Michele Banko, and Arul Menezes. 2004. Event-centric summary generation. In *Working Notes of the 2004 Document Understanding Conference (DUC'04)*.

Textual genre based approach to use wordnets in language-for-specific-purpose classroom as dictionary

Itziar Gonzalez-Dios, German Rigau
Ixa Group
University of the Basque Country (UPV/EHU)
{itziar.gonzalezd,german.rigau}@ehu.eus

Abstract

When teaching language for specific purposes (LSP) linguistic resources are needed to help students understand and write specialised texts. As building a lexical resource is costly, we explore the use of wordnets to represent the terms that can be found in particular textual domains. In order to gather the terms to be included in wordnets, we propose a textual genre approach, that leads us to introduce a new relation *term_used_in* to link all the possible terms/synsets that can appear in a text to the synset of the textual genre. This way, students can use wordnet as dictionary or thesaurus when writing specialised texts. We explain our approach by means of the logbooks and terms in Basque. A side effect of this works is also enriching the wordnets with new variants and synsets.

1 Introduction

Language for specific purposes (LSP) is a subfield of applied linguistics that studies language in different contexts e.g. language for business, language for engineering, etc. The work in this area has been mainly done in the field of terminology, but nowadays theory-building data analysis and classroom/workplace practice have an important role in the development of the field (Gollin-Kies et al., 2015).

In this paper, we propose to use wordnets as a lexical/terminological references to consult in LSP teaching. Exactly, we present a method that combines textual genre analysis together with classroom practice in order to compile terms to be included in wordnets. The final aim is to provide students with a multilingual and semantically rich consult resource that will gather of the terms that can be used in a specific textual genre.

We have decided to use wordnets as a basis resource because they offer rich semantic information liked to different languages and we think it is appropriate to centralise all the resources. Moreover, its relations are helpful for students when looking for similar words, related concepts, etc. That is why, we propose a new relation: the *term_used_in*. This relation will link all the terms/synsets that can be used in a textual genre, without altering the hierarchy of wordnet.

The context of this research is Basque as LSP for sea studies. Currently, many subjects are taught in Basque at university level, but it is still a language under normalisation (the standard variant was officially created in 1968) and this fact influences the corpus and the resources we can use: there is no specialised corpus on some fields of knowledge and lexicographic/terminological data is sparse. Moreover, as in the case of fishing or farming, the specialised variant has been oral. That is why we propose to base on textual genres, standard models of text types. Following Cabré (1999), we also think that specialised texts meet certain norms that vary depending on the domain. Indeed, textual genres are a key component on specialised discourse (Gotti, 2008). Moreover, During this work, as side-effect we are also enriching wordnets, in our case, Basque WordNet (BWN) (Pociello et al., 2011).

To illustrate our approach, we report on case study about logbooks, a nautical textual genre that compiles terms from different domains such as metrology, meteorology, geography among others. We will work on terms on Basque language, a language under normalisation that is developing its specialised languages.

This paper is structured as follows: in Section 2 we sum up the context of our work; in Section 3 we present our approach and we show an example of its practical application in Section 4. After that, in Section 5 we discuss some issues relating

the process and we conclude and outline the future work in Section 6.

2 Domains, specialised knowledge and textual genres

Domains are usually defined as unitary areas of knowledge (specialised or not) and are related to semantic fields, subject matters, broadtopics, subject codes, subject domains, categories... In WordNet (Fellbaum, 1998) we can find the *Domain of synset/Member of this domain*, where synsets are linked with a category, region or usage pointer (domains) and the domains are linked with synsets. In WordNet Domains (Magnini and Cavaglia, 2000; Bentivogli et al., 2004) synsets have been semi-automatically annotated with one or more domain labels from a set of 165 hierarchically organised labels, contrasted to the Dewey Decimal Classification (DDC) system. In eXtended WND (González-Agirre et al., 2012) a graph-based approach was carried out to improve WordNet Domains by means of a simple inheritance process through the nominal and verbal hierarchies and applying UKB to propagate the domain information. BabelDomains (Camacho-Collados and Navigli, 2017) are automatically created by combining distributional and graph-based approaches and are based on Wikipedia categories for the featured articles. These hierarchical approaches are related to classical terminology work.

Specialised knowledge is the principles and techniques that are acquired in a particular discipline. According to Cabré (2003), specialised knowledge is transferred by terminological units (terms) and this transfer occurs during the specialised communication, in the discourse produced in each situation by the experts (communicative approach to terminology). A way of studying the specialised communication is through the corpus analysis. Indeed, many works dealt with terminology extraction from corpus e.g. Alegria et al. (2004).

A key component of specialised discourse is the textual genre, a prototypical type of discourse. Cabré (2005) points out that documents corresponding to textual genres are used in every professional domain, and that students should know their standard features and characteristics in order to be able to write them. These standards include format, phraseology and vocabulary. In other words, each textual genre will be marked by its own terms.

3 Approach for gathering terms

The approach we propose is conceived for environments where no corpus or few texts exist and the lexicographic/terminological resources are sparse and scattered. Next, we explain the proposed approach.

- **Critical overlook of the existing and referential resources**: before we start working on any target field it is important to know which are the lexical/terminological resources we can consult and reuse. Moreover, it is also convenient to analyse how the terms in the target domain are represented in general-purpose dictionaries/ terminological databases.

- **Analysis of the communicative needs and textual genres**: in order to choose a textual genre, we need to make an analysis of the the communicative needs, that is, we need to know which textual genres are the most used. Classroom practice is important in this step, getting to know which texts are most used and most difficult to write can be decisive to choose the textual genre. Another option is the one presented by da Cunha and Amor Montane (2019) where they make questionnaires to domain experts to know which are the most used and most difficult texts to write. This step could also be automatised if specialised corpora were available.

- **Term compilation and representation in wordnets**: in order to compile the terms, we need to consult in the existing and previously analysed resources the terms that can be used in the target textual genre. Then, we will include the terms in Basque WordNet because of its reusalibity as variants in their respecting synset. We will add the relation *term_used_in* to the hypermyn of the synsets to link it to the text genre.

We propose to create the *term_used_in* relation in order to offer students LSP students help when writing and consulting specialised vocabulary and terminology.

4 Practical application of the approach: Basque nautical terms and logbooks

In this section we describe a practical application of the above presented methodology. In this case study we will report on the logbooks and Basque terms.

4.1 Critical overlook of resources

In this section we present the resources where we can find nautical terminology in Basque.

Relating the general resources, Euskalterm[1] is the main terminological database for Basque. When looking for nautical terminology in Euskalterm, they appear under other subjects such as 1) Fishing, 2) Sports, Games and Leisure, 3) Industry, 4) Law 5) Geology and Meteorology or 6) Education and pedagogy. For instance, the term *ziaboga* (turning, a basic manoeuvre) can be found in the sports and leisure domain, because it has been compiled in the rowing dictionary.

Another general resource is *Zientzia eta Teknologia Hiztegi Entziklopedikoa*[2], a dictionary of Science and Technologies. In this dictionary, there are three categories where nautical terms can be found: sea, oceanographic and meteorology. Moreover, terms related to sea engineering can be found in categories such as general technology, electric technology and mechanic technology.

The last general resource we want to mention is WordNet. Using the synset *seafaring* and the *domain term category* relation we can find nautical terms and in WordNet Domains we also do find the nautical category. But due to the size of the BWN not all the English words are covered by the Basque version. For example, if we look for the hyponyms of the word *itsasontzi* (stands for vessel, watercraft), there are four synsets in Basque (*belaontzi* sailing ship, *galera* galley, *arrantzaontzi* fishing boat and *yate* yacht) whereas there are fourteen for English.

Relating the maritime specific resources, the most important is "*Itsasontziaren Eskuliburua*" (The Manual of the Vessel) (Sotés et al., 2015), a manual that has been written by professors of sea studies and it is conceived as a photo-dictionary. It is divided in three topics: the vessel, the port and the containerisation and it includes four term lists (Basque-Spanish, Spanish-Basque, Basque-English and English-Basque). In the book, the terms related to the previously mentioned topics are explained and illustrated with figures. This resource is so far the best for the nautical terminology and it is being integrated in *Terminologia Zerbitzurako Online Sistema (TZOS)* (Arregi et al., 2013), the terminological database of academic Basque.

Moreover, there are some resources in Basque related to navigation and the sea e.g. dictionaries such as the "Fishing dictionary", "Transport and Logistics dictionary", "Maritime Law dictionary or "Astronomy dictionary" included in EuskalTerm or independent dictionaries such as the "Dictionary of the Port of Pasaia", "the Activity Book of the Port of Bilbao", "Regatta dictionary", "Biscayan fishermen dictionary", or fishmongers dictionaries. There are also PhD theses on the fishermen speech and vocabulary of certain towns.

Finally, MARITERM (Marinelli et al., 2004) is a maritime lexical database structured as WordNet that contains the specialised lexicon of navigation and maritime transport. It can be considered a domain adaption of WordNet, with its peculiarities to the nautical domain. The lexicon includes also terms of other domains such as meteorology, geography, cartography, astronomy, law related to the sea and maritime contracts, sailing races or publications.

In conclusion, the shortfalls of the general resources are that a) nautical terms are spread in different categories (terms are scattered) and b) the coverage is low. The main problem of the maritime resources is that c) their texts and wordlists are difficult to process computationally due to their format (some of them are not even digital) and, that some of them are not available or have the reusable licenses.

4.2 Analysis of the communicative needs and textual genres

In order to analyse the communicative needs we have examined the documentation that needs to be carried on the ships. In the case of vessels with Spanish ensign, the documentation is specified by the law 14/2014, in the articles 78-87 of the chapter second chapter. According to this law, the documents that must be carried on the ship are the certificate of enrolment, navigation certificates, ensign, crew list, logbook and the bell book (logbook concerning the machines). In our opinion, the linguistically and terminologically most inter-

[1] http://www.euskadi.eus/euskalterm/
[2] https://zthiztegia.elhuyar.eus/

esting documents are the logbooks. Moreover, this textual genre is one of the most used by students and professionals.

Logbooks are the documents that the captain writes and must compile every eight hours with all the important events relating the nautical and meteorological incidents in the navigation. So, in this textual genre we will find terms about measures, size, coordinates, directions, meteorologic phenomena and places, which, in our opinion, makes it to be a very rich textual genre on nautical terminology.

As a curiosity, we want to mention that in Paleoclimatology based on the logbooks from Catalan seafarers dating from the 17th century, Prohom (2002) have rebuilt the Atlantic ocean climate. Therefore, this textual genre is not only interesting for linguistic studies but also for historical and climatological ones.

4.3 Term compilation and representation in BWN

Even though logbooks are symbolically written in vessels, the purpose of the term compilation and representation is to provide students how can they use the terms in Basque. To that end, we have looked for the terms that can be used in the logbooks in the Basque referential resources. Following, we list the the hypernyns of the terms we have gathered. The list of all Basque terms is shown in Gonzalez-Dios (2019).

- Magnitudes (4 terms)
- Cardinal and intercardinal directions (16 terms)
- Meteorological phenomena:
  - Wind: Beaufort scale, wind oscillation and wind speed (24 terms)
  - Sea: Douglas scale, form of the waves and *galerna* types (16 terms)
  - Clouds: types, forms, distribution and moisture (28 terms)
  - Precipitations: types, intensity, amount of liquid (different for rain and snow), types of storms (21 terms)
  - Temperature (10 terms)

We have included all the terms that were not already covered and have an equivalent synset in English e.g. *abiadura_angeluar* linked to *angular_velocity* in BWN. We have decided not to include geographic terms, because so far entities have not been added to BWN. Dates and hours have also not been added.

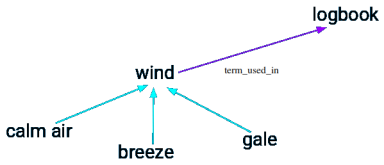


Figure 1: Example of the *term_used_in* relation

Finally, the hypernyns are linked to the synset *logbook* via the *term_used_in*. An example of this is shown in Figure 1, where a synset (*wind*) is connected to the synset of the textual genre where it is used (*logbook*). This way, students can consult which terms can appear in this textual genre.

5 Discussion

Following the presented approach, we have gathered terms and included in a semantically rich resource such as BWN, and tried to avoid the dispersion of terms, an important problem with Basque nautical terminology as shown in Section 4.1. In addition, we have provided LSP students an improved and centralised resource to help write the specialised texts.

However, when trying to represent these terms in BWN we have found some issues we will like to discuss. The first is about the conceptualisation: in logbooks and referential resources some terms are organised in a different way from WordNet and sometimes that classification was more detailed than the WordNet hierarchy e.g the types of the clouds (Figure 2) were organised in our resource taking the levels low, mid, high... (in green) into account whereas in WordNet all of them are together (in black) .

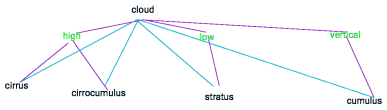


Figure 2: Example of categorisation of clouds

Secondly, many terms are not in English.

For instance, local meteorologic phenomena like *galerna*, or international conventions such as the Douglas scale, shapes of waves, etc. So the need of new language dependant concepts are necessary. That is, the need of CILI (Bond et al., 2016) is remarkable in this work. In fact, many of these terms are international and other wordnets would profit from these new synsets.

Thirdly, as we have seen, several domains are linked by gathering terms/synsets approach. In the case of the logbooks, moreover, it is remarkable that, although it is a text from the nautical domain, most of its words are not included in the *nautical* domain of WordNet Domains hierarchy. This makes us think of bigger domains, domains where knowledge from different areas meets. Indeed, this is related to communicative bottom-up approaches (Zabala et al., 2018).

Finally, we would like to encourage the use of the proposed relation *term_used_in* so that all these variants can be related. Indeed, we think it can be a step towards the characterisation of professional textual genres in wordnets. Moreover, as textual genres are *international* models, this approach can help to improve the recall of wordnets, since it allows to detect missing synsets, that is, words that are in certain texts, but not yet in WordNet.

6 Conclusion and Future Work

In this paper we have presented a method to get specialised knowledge by gathering terms and to include it in wordnets. Moreover, we want to encourage the use of wordnets in LSP classrooms as a dictionary, that can be useful for less-resourced specialised languages. To that end, we rely on textual genres as basis for term/synset gathering to be included in BWN. Indeed, textual genres have been proven to be useful to compile terms that would not appear in traditional hierarchies since they belong to different domains. We have explained our approach by means of the case of logbooks in Basque, a professional textual genre with terms from different domains and a language which is developing its specialised languages. Moreover, we have proposed a new relation called *term_used_in* for wordnets through which students can consults terms that can be used in a certain textual genre. As future work, we plan to analyse other textual genres from the engineering domain and keep on adding terms to Basque WordNet and, thus, enriching it.

Acknowledgments

This work has been partially funded by the the projects DeepReading (RTI2018-096846-B-C21), CROSSTEXT (TIN2015-72646-EXP) and BigKnowledge – *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018*.

References

Iñaki Alegría, Antton Gurrutxaga, Pili Lizaso, Xabier Saralegi, Sahats Ugartetxea, and Ruben Urizar. 2004. Linguistic and Statistical Approaches to Basque Term Extraction. *GLAT-2004: The Production Of Specialized Texts*.

Xabier Arregi, Ana Arruarte, Xabier Artola, Mikel Lersundi, and Igone Zabala. 2013. TZOS: An On-Line System for Terminology Service. *Centro de Lingüística Aplicada*, pages 400–404.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the Wordnet Domains Hierarchy: Semantics, Coverage and Balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 101–108. Association for Computational Linguistics.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*, volume 2016.

María Teresa Cabré. 1999. *La terminología: representación y comunicación*. Institut Universitari de Lingstica Aplicada. Universitat Pompeu Fabra.

María Teresa Cabré. 2003. Theories of terminology: their Description, Prescription and Explanation. *Terminology*, 9(2):163–199.

María Teres Cabré. 2005. Recursos lingüísticos en la enseñanza de lenguas de especialidad. In *V Jornada-Coloquio de la Asociación Española de Terminología (AETER): Comunicar y enseñar a comunicar el conocimiento especializado*.

Jose Camacho-Collados and Roberto Navigli. 2017. Babeldomains: Large-scale domain labeling of lexical resources. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 223–228.

Iria da Cunha and M Amor Montane. 2019. Textual Genres and Writing Difficulties in Specialized Domains. *Signos*, 52(99):4–30.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Sandra Gollin-Kies, David R Hall, and Stephen H Moore. 2015. *Language for Specific Purposes*. Palgrave Macmillan.

Aitor González-Agirre, German Rigau, and Mauro Castillo. 2012. A Graph-based Method to Improve WordNet Domains. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 17–28. Springer.

Itziar Gonzalez-Dios. 2019. Nautikako terminologia biltzen testu-generoak abiapuntu: nabigazio-egunerokoen eredua. In Itziar Aduriz and Ruben Urizar, editors, *Hizkuntzalari euskaldunen III. topaketa. Zer berri?*, pages 79–91. Udako Euskal Unibertsitatea.

Maurizio Gotti. 2008. *Investigating specialized discourse*. Peter Lang.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC*, pages 1413–1418.

Rita Marinelli, Adriana Roventini, and Alessandro Enea. 2004. Building a maritime domain lexicon: a few considerations on the database structure and the semantic coding. In *LREC 211 Fourth International Conference on Language Resources and Evaluation, held in Memory of Antonio Zampolli*. Citeseer.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and Construction of the Basque WordNet. *Language resources and evaluation*, 45(2):121–142.

Marc J. Prohom. 2002. El uso de los diarios de navegación como instrumento de reconstrucción climática. *Investigaciones Geográficas*, 28:89–104.

Iranzu Sotés, Iñaki Alcedo, Imanol Basterretxea, Aingeru Basterretxea, and Xabier Sotés. 2015. *Itsasontziaren Eskuliburua*. Euskal Herriko Unibertsitateko Argitalpen Zerbitzua.

Igone Zabala, Izaskun Aldezabal, María Jesús Aranzabe, Jose Maria Arriola, Itziar Gonzalez-Dios, and Mikel Lersundi. 2018. Corpus-driven Terminology Work for Describing Basque Academic Terminology: the Weaving Terminology Networks programme (TSE programme). In *EAFT Summit*.

Fitting Semantic Relations to Word Embeddings

Eric Kafe
MegaDoc
Charlottenlund, Denmark
kafe@megadoc.net

Abstract

We fit WordNet relations to word embeddings, using *3CosAvg* and *LRCos*, two set-based methods for analogy resolution, and introduce *3CosWeight*, a new, *weighted* variant of *3CosAvg*. We test the performance of the resulting semantic vectors in *lexicographic semantics tests*, and show that none of the tested classifiers can learn symmetric relations like *synonymy* and *antonymy*, since the source and target words of these relations are the same set. By contrast, with the asymmetric relations (*hyperonymy / hyponymy* and *meronymy*), both *3CosAvg* and *LRCos* clearly outperform the baseline in all cases, while *3CosWeight* attained the best scores with *hyponymy* and *meronymy*, suggesting that this new method could provide a useful alternative to previous approaches.

1 Introduction

Analogy is the prototypical formulation of any relation: *a is to a' as b is to b'* means that the relation between a and a' is the same as the relation between b and b'. Thus, the analogy establishes a paradigmatic relation between a class of source items (a and b) and a class of target items (a' and b'), and all relations are special cases of analogy.

Both morphological analogies like $(car, cars) \approx (apple, apples)$, and semantic analogies like $(man, woman) \approx (king, queen)$ have been shown to hold in vector-space representations of words, derived from cooccurrence matrices in large corpora (Mikolov et al., 2013c). This approach has proven useful in many applications, in particular machine translation, where it reveals analogies across languages (Mikolov et al., 2013a), although more complex morphology or deeper semantic relations cause a drop in accuracy (Köper et al., 2015).

The original method (Mikolov et al., 2013c), which is now called *3CosAdd*, resolved *analogy completion* tasks like $(man, king) \approx (woman, ?)$ by searching for the most similar vector to $woman + king - man$, using *cosine similarity*, with $queen$, as result.

3CosAdd:

$$b' = argmax_{b' \in V}(cos(b', b + a' - a)) \qquad (1)$$

Alternative methods like *PairDistance* and *3CosMul* have been shown to occasionally perform slightly better (Levy et al., 2015).

Very often, the most similar target word b' is likely to be one of the already given words a, a' and especially b, so these are always discarded from the searched vocabulary V, which should, more precisely, be understood as $C^V_{a,a',b}$ (the complement set of the three premisses in the vocabulary). Otherwise, test accuracy often drops to *zero* (Linzen, 2016), raising questions about the proper interpretation of these vector-space operations (Rogers et al., 2017; Schluter, 2018).

However, the limits of *pair-based* approaches became clear with the *Bigger Analogy Test Sets (BATS)* (Gladkova et al., 2016), where, in particular, a series of *Lexicographic semantics* tests proved very difficult. These tests consist in ten series of questions, covering seven semantic relations (hypernyms, hyponyms, three kinds of meronyms, synonyms and antonyms). The first example from each series is shown in Tab. 1, where we can see that the expected answer often differs from the corresponding WordNet target. In particular, four out of these ten examples do not have a solution in WordNet 3.1, which adds to the difficulty of solving these tests.

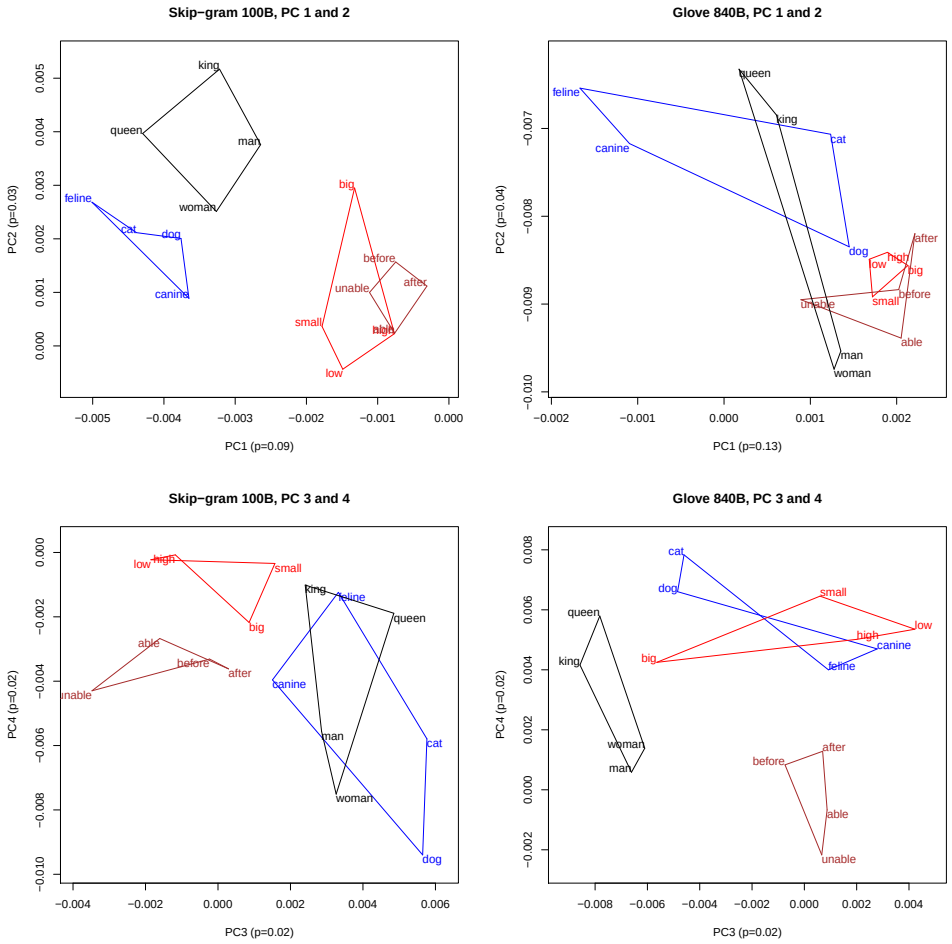


Figure 1: Word Analogies in Skip-gram and Glove models (Principal Components)

A new standard was introduced with the *set-based* methods *3CosAvg* and *LRCos* (Drozd et al., 2016). Instead of relying only on two pairs of words, these methods solve analogies by learning from several pairs, which was shown to clearly outperform all previous methods, although the performance on the *Lexicographic semantics* tests remained modest.

On the other hand, using the semantic knowledge from WordNet relations as a training objective of word embeddings has been shown to improve their performance on semantic tasks (Yu and Dredze, 2014), and the hypernymy and meronymy relations of the Polish wordnet have been successfully used to train linear classifiers (Czachor et al., 2018). A complementary approach consists in retrofitting the embeddings to the semantic relations, which improved on previous baselines (Faruqui et al., 2014), although it seems unlikely that retrofitting can benefit other words than those that were retrofitted.

In this study, we apply the set-based approach to the WordNet relations (Fellbaum, 1998), by using *3CosAvg* and *LRCos* to fit WordNet relations to word embeddings, and test the performance of the resulting vectors on the *Lexicographic semantics tests* from BATS.

Table 1: Lexicographic test examples from BATS

TEST	QUESTION	ACCEPTED ANSWERS	WORDNET 3.1	
L01 [HYPERNYMS - ANIMALS]	*allosaurus*	*dinosaur, reptile, bird, archosaur, archosaurian, archosaurian reptile,*	HYPERNYM	bird-footed dinosaur, theropod, theropod dinosaur
L02 [HYPERNYMS - MISC]	*armchair*	*chair, seat, piece of furniture, article of furniture, furnishing, artifact, artefact, unit, object, physical object, physical entity, entity*	HYPERNYM	chair
L03 [HYPONYMS - MISC]	*backpack*	*daypack, kitbag, kit bag*	HYPONYM	kit bag, kitbag
L04 [MERONYMS - SUBSTANCE]	*atmosphere*	*gas, oxygen, hydrogen, nitrogen, ozone*	HAS SUBSTANCE	∅
L05 [MERONYMS - MEMBER]	*acrobat*	*troupe*	IS MEMBER	∅
L06 [MERONYMS - PART]	*academia*	*college, university, institute*	HAS PART	college, university
L07 [SYNONYMS - INTENSITY]	*afraid*	*terrified, horrified, scared, stiff, petrified, fearful, panicky*	SYNONYM	∅
L08 [SYNONYMS - EXACT]	*airplane*	*aeroplane, plane*	SYNONYM	aeroplane, plane
L09 [ANTONYMS - GRADABLE]	*able*	*unable, incapable, incompetent, unequal*	ANTONYM	unable
L10 [ANTONYMS - BINARY]	*after*	*before, earlier, previously*	ANTONYM	∅

2 Methods

2.1 Set-based analogy resolution

We test the set-based methods *3CosAvg* and *LRCos* (Drozd et al., 2016), and compare their performance with the *Only-B* baseline (Linzen, 2016), and with a new, weighted formulation of *3CosAvg*, which we call *3CosWeight*.

Only-B (Linzen, 2016) is a very appropriate baseline, because it simply disregards the training set, so it allows to precisely gauge the advantage obtained from set-based approaches:

$$b' = argmax_{b' \in V}(cos(b', b)) \quad (2)$$

As always, words that are already known (here only b) need to be discarded from the searched vocabulary V.

Add-Opposite (Linzen, 2016) tests the opposite direction of *3CosAdd* (Eq. 1):

$$b' = argmax_{b' \in V}(cos(b', b + a - a')) \quad (3)$$

3CosAvg (Drozd et al., 2016) is an extension of *3CosAdd*, which, instead of a single word pair, uses the difference between the overall average of the source and target classes:

$$b' = argmax_{b' \in V}(cos(b', b + avg\_offset)) \quad (4)$$

$$avg\_offset^1 = \frac{\sum_{i=0}^{m} a'_i}{m} - \frac{\sum_{i=0}^{n} a_i}{n} \quad (5)$$

A slightly different variation of *3CosAvg* calculates *avg_offset* as the average of the vector differences in each (source,target) pair instead of the difference between the overall class averages (Bouraoui et al., 2018). Thus, the practical implementation of *3CosAvg* is open to various interpretations and extensions, as we will see next.

3CosWeight is a new, weighted formulation of *3CosAvg*, where we multiply the previously defined *avg_offset* with a weight *w*:

$$b' = argmax_{b' \in V}(cos(b', b + (w * avg\_offset))) \quad (6)$$

[1]Thanks to Aleksander Drozd, who gave us permission to correct the order of the subtraction in the *avg_offset* formula (Eq. 5). The formula printed in the original article (Drozd et al., 2016) unfortunately presents this subtraction in the opposite order (a *minus* a').

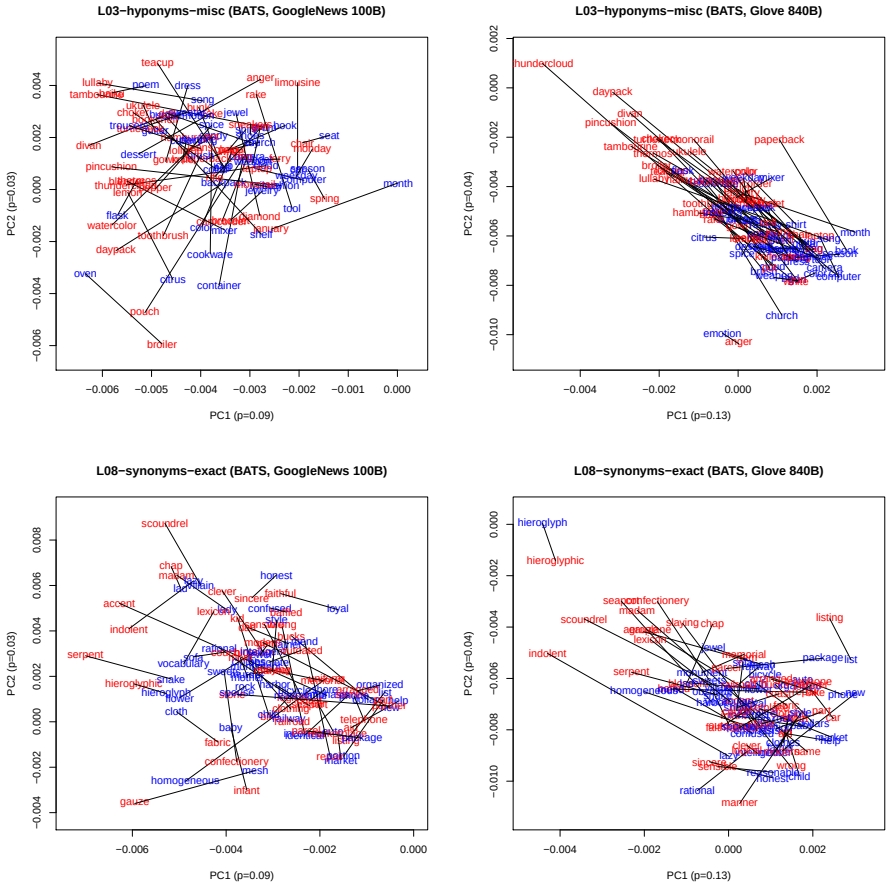


Figure 2: BATS relation pairs in GoogleNews and Glove (Principal Components)

It follows from this definition that *3CosWeight* is identical to *3CosAvg* when multiplying the averaged vector by $w = 1$, and that the result is identical to the *Only-B* method when $w = 0$, while multiplying by $w = -1$ is identical to adding the opposite vector, like in the *Add-Opposite* method. In this study, we try whole integer values of *w* in the range [-2,+5], in order to test whether the weight *w* can boost the performance of the averaged vectors.

Last, we compare these results with *LRCos* (Drozd et al., 2016).

LRCos uses logistic regression to calculate the probability that b' belongs to the target class:

$$b' = argmax_{b' \in V}(P_{(b' \in target\_class)} * cos(b', b)) \quad (7)$$

2.2 Implementation

We downloaded two widely-known sets of embeddings, which have emerged as the best performers in various benchmarks, and are freely available online. Both rely on very large corpora and consist in word vectors with 300 dimensions, meaning that each vector is an array of 300 floating-point numbers in the interval [-1, +1].

The *GoogleNews-vectors-negative300* embeddings [2] are *Skip-gram vectors* (Mikolov et al., 2013b), representing a corpus of 100 billion words, while the *glove.840B.300d* [3] embeddings

[2] https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz

[3] http://nlp.stanford.edu/data/glove.840B.300d.zip

consist in *Global Vectors* (Pennington et al., 2014), derived from a corpus of 840 billion words.

For each of the *Lexicographic semantics* relations in BATS, we produced a two-column text database with the word pairs from the corresponding WordNet 3.1 relation converted to lowercase.

We used the open-source *Vecto v. 0.2* [4] software package to load and process the embeddings, and perform the BATS tests. First, we applied Vecto's *filter_by_vocab* function in order to restrict the embeddings to the set union of all WordNet relations (147478 words) and the words in the BATS *Lexicographic semantics* tests (4126 words), converted to lowercase, yielding a vocabulary of 147620 words, of which 54697 were present in the GoogleNews embeddings, and 65066 in Glove. Thus, although both of the original embeddings include over two million "words" (many of which are noise), they actually cover less than half of the WordNet vocabulary.

We wrote a small Python dictionary called *bats2wn*, which links the adequate WordNet relation (hypernyms, hyponyms, meronyms, synonyms or antonyms) to each of the *Lexicographic semantics* tests in BATS (cf. Tab. 1), so that this data can be processed by the analogy resolution methods, where we simply replace the BATS training set by the corresponding WordNet relation pairs. This only required very small additions to the original Python code in *Vecto*.

Contrary to the BATS pairs, where each target is a list, in our WordNet relation pairs, each target is only a single word. So although the current version *v. 0.2* of Vecto uses a heuristic to speed up learning by only considering the first valid word in each target list, this short-cut has no effect here, because each relation pair only contains one local target, so all targets of each source word are actually used. This allows to preserve the symmetry of the symmetrical relations (synonymy and antonymy), which would otherwise be compromised by the arbitrary loss of some targets.

We became aware of this potential problem by first using WordNet relation pairs converted to the BATS target list format, and realizing that the results did not have the expected properties: hypernymy and hyponymy could not be recognized as inverse relations, and synonymy and antonymy were not symmetric. So this problem was solved by presenting the relation data as word pairs instead of target lists, without modifying Vecto, which would require removing a *break* statement in the *3CosAvg* implementation, and merging the target lists for *LRCos*.

It is important to note that the current (*v. 0.2*) Vecto implementation of *avg_offset* differs from the article formula (Drozd et al., 2016) by also averaging over the *m* local targets of each source word, before calculating the global difference of averages (Eq. 8). More precisely formulated, the global target class average is thus the average of the local averages.

$$avg\_offset = \frac{\sum_{i=0}^{n} \frac{\sum_{j=0}^{m} a'_j}{m} - a_i}{n} \quad (8)$$

Normally, this detail would result in small variations, compared to implementations that only subtract the global averages. However, the current Vecto implementation only picks one word in the target list, so the local averaging has no effect, since it only averages over a single word. In our setup, each relation pair is also presented with only one target, but all target words are used, so the result is actually equivalent to the original formula (Eq. 5), and the mathematical properties of the studied relations are preserved.

With some tests, the *LRCos* precision could vary by a few percent between subsequent runs, because Vecto's standard implementation relies on random words for the negative examples used for training the classifier. Specifically, Vecto (version 0.2) uses the target word of each relation pair as *positive* examples, while the *negative* examples consist in four copies of the source words of the relation, plus a set of random words of the same size as the set of source words. Since the arbitrary random choices can be fortunate for one embedding and unlucky for another, the standard implementation of LRCos does not allow fair comparisons. So we also tested a deterministic variant of LRCos, where we simply removed the random part of the *negative* examples.

We used the default settings in Vecto to perform series of *Leave-one-out* cross-validations, where each question is answered after training on all the (source, target) pairs in the tested semantic relation, where the question word is not a source word.

[4] https://github.com/vecto-ai

Table 2: WordNet relations fitted with *3CosAvg* to 300-dim. Skip-gram and Glove vectors

	SKIP-GRAM					GLOVE				
dim.	1	2	...	299	300	1	2	...	299	300
HYPERNYM	-0.001645	0.000994	...	-0.000552	-0.009935	-0.011362	0.011182	...	0.003938	0.006483
HYPONYM	0.001644	-0.000996	...	0.000557	0.009932	0.011361	-0.011181	...	-0.003940	-0.006482
HASSUBSTANCE	-0.007939	0.002562	...	0.003922	0.004445	-0.012542	0.015891	...	0.001864	0.018110
ISMEMBER	-0.005050	-0.002306	...	0.018941	-0.006816	-0.003275	-0.000562	...	-0.000785	-0.003439
HASPART	-0.005742	-0.001865	...	-0.004137	0.003926	0.015685	-0.003739	...	0.005990	-0.015377
SYNONYM	-0.000001	-0.000001	...	-0.000001	-0.000004	0.000002	0.000009	...	0.000002	-0.000001
ANTONYM	-0.000008	0.000009	...	-0.000021	0.000016	-0.000008	-0.000015	...	-0.000029	-0.000008

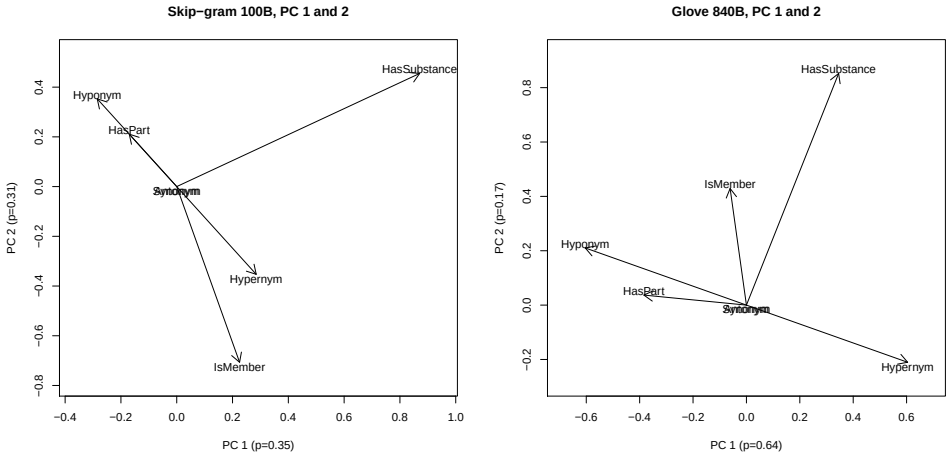


Figure 3: Fitted WordNet relation vectors (Principal Components of Tab. 2)

3 Results

3.1 WordNet vectors

Applying *3CosAvg* (Eq. 4) on WordNet relation pairs in the Skip-gram and Glove embeddings produced the set of semantic *WordNet vectors* shown in Tab. 2.

Like the word vectors, each WordNet relation vector is a list of 300 real numbers in the interval [-1,+1], representing the average projection from the relation source words to their related targets. In both cases we see that each value in the respective vectors of the inverse relations (hypernymy and hyponymy) are the negative of each other up to the fifth decimal, while the sixth decimal shows a spurious divergence, due to the inherent inaccuracy of floating-point arithmetics. As mentioned earlier, this important property of the inverse relations may be lost when using heuristics to prune the training set, which we avoided here by presenting the relations as word pairs instead of target lists.

In theory, the vectors for the symmetric relations (synonymy and antonymy) should contain only zeroes, since the set of source words is identical to the set of target words, so the difference of their respective set averages is expected to be exactly *zero*. In practice, the *synonymy vectors* contain only zeroes up to the fifth decimal (cf. Tab. 2), while the sixth decimals reveal errors introduced by floating-point operations. By contrast, with arbitrary target list pruning, the non-zero values would already appear at the third decimal. Exceptionally, the *antonymy vectors* contain a few non-zero values at the fifth decimal, thus revealing a small error in WordNet 3.1, where a few antonym pairs (for ex. *have* vs. *lack* and *lack* vs. *miss*) do not have a symmetric variant.

3.1.1 Visualizing the vectors

We performed a Principal Components Analysis of the subset of the Glove and GoogleNews em-

Table 3: Precision with WN 3.1 vectors (percent)

	SKIP-GRAM										GLOVE									
	3CosWeight								*LRCos*		*3CosWeight*								*LRCos*	
weight	-2	-1	0	1	2	3	4	5	det	rnd	-2	-1	0	1	2	3	4	5	det	rnd
L01: *Hypernym*	6	8	10	16	20	26	30	36	50	46	4	6	8	10	20	34	36	30	**66**	56
L02: *Hypernym*	0	0	2	4	4	6	6	6	20	14	2	6	10	12	16	16	22	24	**36**	**36**
L03: *Hyponym*	18	22	28	30	30	26	30	32	20	22	22	22	24	32	38	**42**	38	30	32	30
L04: *Substance*	0	0	0	2	2	2	0	0	4	6	6	8	8	10	10	8	8	**12**	2	**12**
L05: *Member*	4	4	4	6	6	10	10	10	6	12	2	6	8	8	12	12	**14**	12	8	10
L06: *Parts*	6	6	6	6	6	6	6	6	12	14	2	2	2	6	8	10	14	**18**	16	16
L07: *Synonym*	26	26	26	26	26	26	26	26	26	**28**	22	22	22	22	22	22	22	22	22	24
L08: *Synonym*	28	28	28	28	28	28	28	28	28	36	44	44	44	44	44	44	44	44	44	**46**
L09: *Antonym*	18	18	18	18	18	18	18	18	18	**22**	14	14	14	14	14	14	14	14	14	14
L10: *Antonym*	38	38	38	38	38	38	38	38	32	30	**48**	**48**	**48**	**48**	**48**	**48**	**48**	**48**	**48**	36
mean	14.4	15	16	17.4	17.8	18.6	19.2	20	21.6	23	16.6	17.8	18.8	20.6	23.2	25	26	25.4	**28.8**	28

beddings used in this study, i. e. the union set of the WordNet and BATS vocabularies. Fig. 1 shows some well-known **word analogies** plotted onto their principal components. The proportion of variance explained by each component is indicated in parentheses, and we see that it is low. For example, with the two first components (PC1 and PC2) of the Skip-gram model, the cumulated proportion of the explained variance amounts to 12% (0.09 + 0.03), so this plot provides only a correspondingly limited representation of the data. The same concern applies to the representation of the relation pairs in Fig. 2, which are rarely parallel nor have the same length. Nevertheless, some analogies present a clearly square-like shape, as noted in several articles (Mikolov et al., 2013c). We also plotted the same analogies on the third and fourth components (PC3 and PC4), revealing other shapes, where some are also square. This indicates that many more principal components than just the first two would be necessary in order to obtain a faithful representation of the word analogies as well as the semantic relations.

By contrast, we also performed a Principal Component analysis of Tab. 2, i.e. the **WordNet vectors** fitted by *3CosAvg*, and plotted the two first components in Euclidean space. (Fig. 3). The proportion of variance explained by each Principal Component (PC) is reported in the parentheses, and we see that these two-dimensional plots provide a very reasonable representation of the 300-dimensional vectors, since they explain a large part of the overall variance (35%+31% for Skip-gram, and 64%+17% for GloVe). In fact, with Glove, the majority of the variance is already explained by the first PC, which is very close to the axis formed by the hypernymy and hyponymy vectors. The overall structure of both models is essentially similar: in both cases the hypernym vector is the exact opposite of the hyponym vector. Also, in both cases, the antonym and synonym vectors are very close to the center, which is not surprising since the theory predicts that *3CosAvg* should yield only zero for all the parameters of symmetric relations.

3.2 Performance

The percentages shown in Tab. 3 are even numbers, because each test consists in fifty questions, so we measure *precision* by simply doubling the number of correct answers, which is a whole number between zero and fifty. A correct answer means that the best ranking prediction is a member of the set of accepted answers.

Overall, the Glove model outperformed Skip-gram with almost all relations and methods. We observe that both the random (*rnd.*) and the deterministic (*det.*) variants of *LRCos* outperform *3CosAvg* (*weight=1*) by a wide margin, while the latter only slightly improves on the *Only-B* (*weight=0*) baseline. But increasing the weight in *3CosWeight* improved the results for all *asymmetric* relations in both models: higher weights (like 3, 4 and 5) thus clearly improved over *3CosAvg*, while reducing the distance to LRCos. Moreover, *3CosWeight* provided the best results for *hyponymy* completion with both Skip-gram and Glove, and the best results for all the three kinds of *meronymy* overall. However, the optimal weight differs for each relation, suggesting a need for more research, in order to explain these variations.

Previous overall precision for the same *Lexi-*

cography tests and *3CosAvg* was 13% with GloVe and 9.6% with Skip-gram, while *LRCos*, also then, showed clearly superior performance, with 16.8% and 15.4% respectively (Gladkova et al., 2016). These results cannot be directly compared with ours, since they were obtained with other embeddings, but they show the same main trends, especially concerning the superiority of Glove over Skip-gram and of LRCos over 3CosAvg.

A striking observation is that the performance curve is completely flat across all the deterministic methods, applied to the symmetric relations (antonymy and synonymy). In this case, neither *3CosAvg* nor the deterministic *LRCos* can improve on the *Only-B* baseline, although the random variant of *LRCos* shows small occasional improvements or degradations obtained by chance, and thus unlikely to be consistently reproducible or predictive of performance on downstream tasks.

4 Discussion

4.1 Symmetry and asymmetry

Our results confirmed that *symmetry* and *asymmetry* are important mathematical properties of some WordNet relations, which determine the performance of the classification methods used in this study. *Synonymy* and *antonomy* are perfectly symmetric relations in WordNet, since every (a,a') pair is reversible, so the a class is identical to the a' class. Hence, their class-wise averages are also identical, and the difference of both averages is zero in theory, though in practice floating-point arithmetics represent the result as a very small number (cf. Tab. 2). For this reason, the *3CosAvg* method actually reduces to *Only-B*, when applied to symmetric WordNet relations. In the BATS, the same relations are not symmetric, which explains why results obtained by training on BATS alone are unlikely to transfer well to downstream tasks. Likewise, when the symmetry is lost due to implementation heuristics, the result cannot be expected to adequately handle real-world data.

With *asymmetric* relations, the set of source words may overlap to some extent with the set of target words. In particular, many words have both *hypernyms* and *hyponyms*, and contribute to the average of both classes. So, for these relations, the class-wise difference of averages only stems from the top and leaf words in the relation graph.

4.2 Polysemy

WordNet 3.1 distinguishes between thirteen senses of *man*, two of which are *antonyms* of two senses out of the four senses of *woman*, while one of the ten senses of *king* ("a male sovereign") is an antonym of *queen* ("A female sovereign ruler"), though in another sense ("a competitor who holds a preeminent position"), *king* and *queen* are synonyms.

Standard word embeddings express all the different senses of the same word with only one vector, but use different vectors for each morphological form of the same lemma. On the contrary, WordNet collapses the different word forms into one lemma, but distinguishes between the various senses of each word. Thus, WordNet fits with the word embeddings through the particular word forms, which correspond to only one morphological variant of their lemma, but aggregate all of its senses indiscriminately.

This structural discrepancy between both word models may be a major reason for the relatively low performance of standard word embeddings on *lexicographic semantics* tasks. Then it should be possible to obtain better results with lemma-based embeddings, and even better performance could be expected from word-sense vectors (Arora et al., 2018).

4.3 Future Work

The retrofitting of embeddings to semantic relations (Faruqui et al., 2014) is compatible with our method, because it is possible to fit relations to embeddings that were retrofitted to the same relations. However, we do not know if the respective benefits of both approaches could accumulate. Retrofitting brings related vectors closer together, and thus further apart from unseen words, although these could potentially be related as well, in which case we may suspect that the downstream performance actually could degrade.

A more promising approach consists in pursuing three distinct optimization goals simultaneously (Bouraoui et al., 2018): the *(source, target)* pair should belong to the given relation, while the *source* word should be a member of the *source* class (in analogies this is already known), and the *target* word be a member of the *target* class. *3CosAvg* tests the first goal, the second is always true in analogy completion tasks, and *LRCos* tests the third. Combining these objectives has been

shown beneficial with the BATS relations as training set (Bouraoui et al., 2018).

However, the BATS relations do not provide enough examples to train a classifier that can generalize adequately to downstream tasks. In particular, the lack of symmetry in the BATS synonyms and antonyms does not allow to recognize important mathematical properties of these relations. More semantic tests are needed, and the BATS is still too small. Larger tests derived from WordNet itself seem promising (Piasecki et al., 2018), though these would be limited to the word pairs known in WordNet, resulting in a limited ability to predict the performance on related pairs outside WordNet.

More successful detection of hypernyms and meronyms has been achieved using k-means clustering with the Polish wordnet (Czachor et al., 2018), so for these relations it might be possible to improve our results with similar techniques. In particular, the present study does not include *indirect relations*, although augmenting the hypernym training set with the *transitive hypernyms* would very probably be an advantage, since the BATS answer sets includes them.

5 Conclusion

We fitted WordNet relations to word embeddings, using *3CosWeight*, a new, *weighted* variant of *3CosAvg*, which allows to emulate well-known methods like *3CosAvg*, *Only-B* and *Add-Opposite*.

We showed that none of the tested classifiers can learn to distinguish between source and target classes of symmetric relations like *synonymy* and *antonymy*, since these classes are identical.

This study confirmed the superiority of *LRCos* over *3CosAvg* for learning *hyperonymy*, while *3CosWeight* was more successful with *hyponymy* and *meronymy*, suggesting that *3CosWeight* can provide a useful alternative to the other methods.

Still, the performance of these methods remains modest, and might eventually benefit from being applied to semantically disambiguated word-sense embeddings, or combined with complementary approaches.

Acknowledgments

Thanks to the anonymous reviewers and the participants at GWC 2019 in Wroclaw, and in particular to Christiane Fellbaum, Hugo Gonçalo Oliveira and Maciej Piasecki for their insightful comments and suggestions, which helped to improve this article.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Zied Bouraoui, Shoaib Jameel, and Steven Schockaert. 2018. Relation induction in word embeddings revisited. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1627–1637, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Gabriela Czachor, Maciej Piasecki, and Arkadiusz Janz. 2018. Recognition of hyponymy and meronymy relations in word embeddings for polish. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, page 254.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3519–3530.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv:1411.4166*.

Christiane Fellbaum. 1998. *WordNet, An Electronic Lexical Database*. MIT Press, Cambridge.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. 2015. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 40–45. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. *arXiv:1606.07736*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Maciej Piasecki, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski, and Paweł Kedzia. 2018. Wordnet-based evaluation of large distributional models for polish. In *Proceedings of the 9th Global Wordnet Conference (GWC 2018)*.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148.

Natalie Schluter. 2018. The word analogy testing caveat. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 242–246, June.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 545–550.

Building The Mongolian WordNet

Khuyagbaatar Batsuren[§†*], **Amarsanaa Ganbold**[†], **Altangerel Chagnaa**[†], and **Fausto Giunchiglia**[§]
[§]KnowDive Group, DISI, University of Trento, Italy
[†]Machine Intelligence Laboratory, DICS, National University of Mongolia, Mongolia
{k.batsuren, fausto.giunchiglia}@unitn.it
{amarsanaag, altangerel}@num.edu.mn

Abstract

This paper presents the Mongolian Wordnet (MOW), and a general methodology of how to construct it from various sources e.g. lexical resources and expert translations. As of today, the MOW contains 23,665 synsets, 26,875 words, 2,979 glosses, and 213 examples. The manual evaluation of the resource[1] estimated its quality at 96.4%.

1 Introduction

Language resources are crucial in the research of computational linguistics e.g., information retrieval, document classification, query answering. In recent years, world languages are divided in two groups: highly-resourced languages (e.g., English or Chinese) and under-resourced languages (e.g., Kazakh or Uyghur). Due to the lack of language resources, the second group of languages displays more mediocre performance than the first group. Mongolian was one of the under-resourced languages.

This paper describes a general methodology by which we built the Mongolian WordNet (MOW), a high-precision wordnet-like lexical resource. Our main technical contributions are (1) a general method to extract high-precision wordnet translations from a bilingual dictionary, (2) a medium-scale lexical resource for the Mongolian language.

The paper is organized as follows. Section 2 presents state-of-the-art methods. Section 3 provides the main methodology how the MOW is built, and Section 5 describes the automatic algorithm to extract the wordnet translations from a bilingual dictionary. We evaluated the results of this method in section 6. Finally, section 7 concludes the paper.

2 State of the Art

Princeton WordNet (PWN) has been a primary lexical resource for most researches involved in lexical semantics, from Computational Linguistics to Semantic Web. Examples of particular applications are word sense disambiguation (Navigli, 2009) and ontology research (Oltramari et al., 2002). This successful case for English inspired many researchers to build wordnets for other languages. Given the awareness of the structural and semantic diversity across languages (Giunchiglia et al., 2017), mono-lingual wordnets have been developed in two ways: the expansion method from PWN and the merge method with PWN.

- The *expansion method* – researchers first accept that the semantic structure of PWN should be more or less similar to their language's semantic network, and translate English synsets to that of a target language.
- The *merge method* – researchers first create a semantic network for their language, and develop its synsets by adding words and definitions. In a final round, they merge their semantic network with PWN by linking[2] synsets with PWN.

To our knowledge, a vast majority of the wordnets have been developed by using the expansion method (Bond and Paik, 2012), while very few wordnets including Open Dutch

[*] This work has been done during internship at National University of Mongolia
[1] https://milab.num.edu.mn/research/monwordnet/
[2] Hereby, a linking is a manual finding of an equivalent meaning between synsets of two resources.

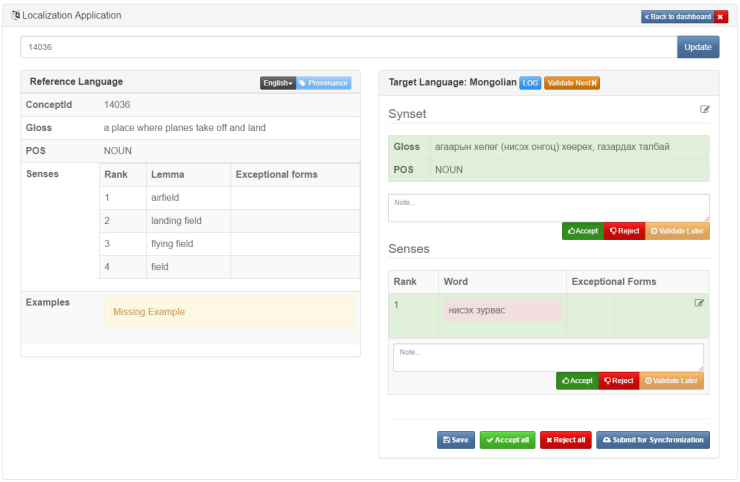


Figure 1: A screenshot of validator user interface

WordNet (Postma et al., 2016), Hindi WordNet(Bhattacharyya, 2017), Polish WordNet have used the merge method. The obvious obstacle is the cost of human labor and the deep expertise of several different domains and cultures, needed in the development of a semantic network.

Researchers in comparative linguistics state that the semantic space of languages are vast and very differential from one another (Von Fintel and Matthewson, 2008) (Giunchiglia et al., 2018). This is because of the differences between speakers of languages, e.g., culture, geographic environment. This is the primary condition underlying the actual choice of the *merge* method because of the importance of individual culture is a fundamental to their wordnet-like lexical resource.

Early linguists (Youn et al., 2016) revealed that an universal structure of lexical semantics exists across all languages at least between basic concepts, and it is why the majority of wordnet developers selected intuitively the *expand* method. Later on, the Global WordNet Association recommended that the monolingual semantic network should be extended by adding cultural synsets under the coordinated usage of the global wordnet grid between wordnets(Vossen et al., 2016).

3 Methodology

In terms of *Wordnet development*, we adopted the expansion method. In the future, we are planning to change and expand the core semantic structure by adding more cultural concepts under the coordination of the global wordnet grid (Vossen et al., 2016). Our wordnet project has two main stages of development: (1) expert translation and (2) automatic translation.

In the *expert translation*, the project has been running since 2016 by employing only expert linguists to translate PWN to Mongolian (Section 4). In the *automatic translation*, we have used a freely, available bilingual Mongolian dictionary to translate PWN to Mongolian (Section 5).

4 Expert Translation

The expert translation method generally follows ontology localization (Espinoza et al., 2009) (Das and Giunchiglia, 2016) which adapts an existing ontology in a language to another by using translation of terms. In this method (Ganbold et al., 2014) (Giunchiglia et al., 2015) (Huertas-Migueláñez et al., 2018), recruited linguistic experts and asked them to provide synsets, in the target language that properly represent a concept denoted by a synset in the source language. The main idea is to find out the most suitable words for

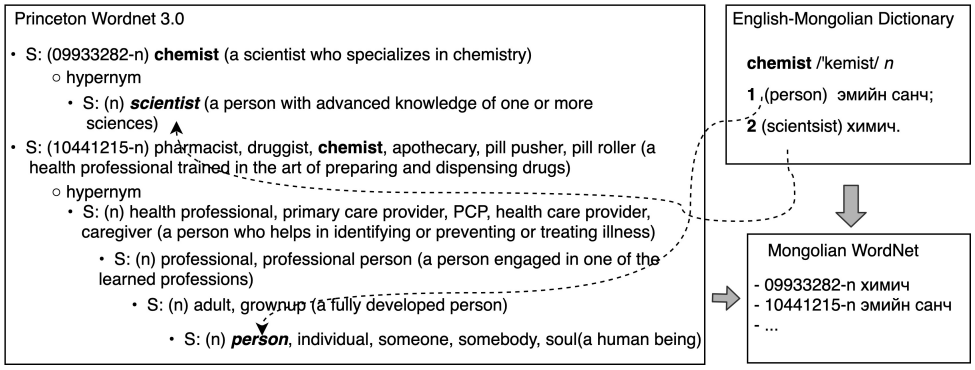


Figure 2: The hypernym-based translation between Princeton WordNet and Bilingual dictionary on a given word "chemist"

the concept in terms of linguistic context use rather than word-for-word translation between synsets.

This method consists of two main tasks: a) translation and b) validation. In the translation task, a language translator provides synset words, its gloss, and example sentences in the target language after she fully understands the meaning of a given synset to localize. If the translator assumes the concept does not exist in the target language, she should mark it as a lexical gap, which means a free combination of words represents the concepts. In this way, we avoid literal translations which may produce a wrong or unwanted result. In the validation task, a language validator evaluates all the elements of the given synset, provided by the translator. The validator either confirms each element or rejects elements one by one with feedback. In the case of a lexical gap, she can accept as it is or suggest word(s) for the synset where she denies it as a gap. When the translator receives feedback, he/she accommodates comments if she agrees with the validator. Alternatively, she can reject the evaluation with comments. Upon reaching an agreement between the translator and the validator, we believe this process produce target language synset with high-quality at the end.

Tasks for translators and validators are assigned by a language manager who manages overall translation activity. Tasks are grouped into a subset of wordnet hierarchy, called subtree, which allows the linguistic experts to understand what they translate/validate. It helps to differentiate concepts by exploring their hyponym/hypernym or sibling relations. The walk-through of tasks is breadth-first.

The linguistic experts use an expert sourcing tool whose screenshot of a validation process is shown in Figure 1. Several volunteered (Ganbold and Chagnaa, 2015) (Ganbold et al., 2018) and paid experts with this tool produced 12,141 synsets, 24,277 senses, and 12,830 words so far.

5 Automatic Translation

Given the two resources PWN and bilingual dictionary below, the main task is to find automatically a set of pairs of $<c, s>$ where c is a synset id from PWN and s is a sense instance of the dictionary. Our method in Algorithm 1 is based on the multiple intuitive criteria:

- if a collocate noun of the sense s maps into one of hypernyms of the synset c then s can express the meaning of the synset c. The example of hypernym-based translations is shown in Figure 2.

- if a given word w has one sense for both dictionary and PWN, the dictionary sense is equivalent to the PWN synset. For example, for the noun word 'mimic,' both PWN and dictionary has only one sense. This intuition of *monosemy translation* has been used to build a French WordNet (Sagot and Fišer, 2008) and Thai WordNet (Sathapornrungkij and Pluempitiwiriyawej, 2005).

The algorithm is structured with three main steps as follows.

Algorithm 1: WordNet Retrieval Algorithm

```
   Input  : w, an english word
   Input  : R, a lexical resource PWN
   Input  : D, a bilingual dictionary
   Output : M, a set of pairs of <id_R, w_D>
 1 C ← Synsets(R, w);
 2 S ← Senses(D, w);
 3 M ← ∅;
 4 if |C| == 1 and |S| == 1 then
 5     for one synset c ∈ C and one sense s ∈ S do
 6         if pos(c) ≠ pos(s) then
 7             continue;
 8         M ← M ∪ <c, words(s)>;
 9 else
10     for each synset c ∈ C do
11         for each sense s ∈ S do
12             if pos(c) ≠ pos(s) then
13                 continue;
14             if μ(collocate(s), c) then
15                 M ← M ∪ <c, words(s)>;
16 return M;
```

Step 1: Initialization (Lines 1–3). C is initialized with a list of synsets which are expressed by the input word w in the lexical resource $\mathcal{R}$ as PWN (line 1). S is initialized with a list of the Mongolian senses which are contained by the input word w in the bilingual dictionary $\mathcal{D}$ (line 2).

Step 2: Monosemy translation (Lines 4–8). In this step, it first checks if the lexical resource R and the bilingual dictionary D have one-to-one mapping between them for the input word w (line 4). if so, in the line 5, it assigns the corresponding one synset from $\mathcal{R}$ into c and the corresponding one sense from $\mathcal{D}$ into a sense instance s (line 5). Then it checks if the synset and the sense share same part of speech (line 6). Then if it succeeds it adds $<c,\ words(s)>$ into the answer set M where $words(s)$ returns only words of the sense s in the bilingual dictionary $\mathcal{D}$.

Step 3: Hypernym-based translation (lines 10–15). In this step, the algorithm iterates each possible pair of a synset c from C and a sense s from S. Then for each pair, if the synset c and the sense s share same part of speech (line 12). If so, the function μ checks if the collocate noun of the dictionary sense s is a hypernym of the synset c in the lexical resource $\mathcal{R}$. If it succeeds it adds $<c,\ words(s)>$ into the answer set M where $words(s)$ returns only words of the sense s in the bilingual dictionary $\mathcal{D}$.

Finally, in Line 16, the algorithm returns the answer set M.

5.1 English-Mongolian Bilingual Dictionary

This bilingual dictionary between English and Mongolian contains over 43,442 English headwords (including compound words) that are translated into 79,299 Mongolian words (or senses). For each english word, the dictionary provides its related senses with their mongolian words. For example, given a word "chemist", the dictionary stores an information as follows:

chemist /'kemist/ *n* **1**. (person) эмийн санч; **2**. (scientist) химич.

where the numbers represent each meaning and it is followed by the collocates (e.g. person or scientist) that are used to distinguish the meanings. Let the 3-tuple $a = <w,\ p,\ S>$ be the headword instance where w represents a head word, p represents a part of speech of the word w, S is a set of senses expressed by the word w. Let the sense instance, s, is the three tuple of $<id,$ col, $w_m>$ where id represents a sense number of s, col is a collocate noun to distinguish s from other meanings, and w_m is a mongolian translation word.

For the above example, the headword instance h is $<$'*chemist*', '*noun*', $S>$ where $S = \{<1,$ '*person*', 'эмийн санч'$>$; $<2,$ '*scientist*', 'химич'$>\}$.

6 Results and Evaluation

PWN has 133974 English words and then given in input to the algorithm 1, which, in turn, generated two sets of 3652 synsets and 7872 synsets from the two automatic methods of *hypernym* translation and *monosemy* translation respectively. For each of the three translations, 200 cases were randomly selected, which were equally selected across four parts of speech. Three linguists were selected to evaluate the samples. They were also provided with the corresponding English glosses and words for the synsets involved, and they were asked the following question: "Do you think meanings of the English synset s_e and the Mongolian synset s_m are equivalent?", and they had to provide a yes/no answer.

Table 1: The results of the three translations: *expert*, *monosemy*, and *hypernym-based* translations.

#	Method	Synsets	Senses	Words	Core Coverage	Accuracy
1	Expert translation	12141	24277	12830	41.1	99.0
2	+ monosemy translation	7872	11038	10235	8.1	98.2
3	+ hypernym-based translation	3652	5629	3792	12.4	92.1
Total	Mongolian Open WordNet	23665	40944	26857	61.6	Avg. 96.4

Table 2: The best twenty wordnets ranked by a number of synsets (Note: we only consider the wordnets that are publicly available and linked to PWN)

#	Language	Synsets	Senses	Words	Examples	Glosses	References
1	English	109942	191523	133974	48459	109942	(Miller, 1995)
2	Finnish	107989	172755	115259	0	0	(Lindén and Carlson, 2010)
3	Chinese	98324	123397	91898	17	541	(Wang and Bond, 2013)
4	Thailand	65664	83818	71760	0	0	(Thoongsup et al., 2009)
5	French	53588	90520	44485	0	0	(Sagot and Fišer, 2008)
6	Romanian	52716	80001	45656	0	0	(Tufiş et al., 2008)
7	Japanese	51366	151262	86574	28978	51363	(Bond et al., 2009)
8	Catalan	42256	66357	42444	2477	6576	(Gonzalez-Agirre et al., 2012)
9	Slovene	40233	67866	37522	0	0	(Fišer et al., 2012)
10	Portuguese	38609	60530	40619	0	0	(de Paiva et al., 2012)
11	Spanish	35232	53140	32129	651	17256	(Gonzalez-Agirre et al., 2012)
12	Polish	35083	87065	59882	0	0	(Piasecki et al., 2009)
13	Italian	33560	42381	29964	1934	2403	(Emanuele et al., 2002)
14	Indonesian	31541	92390	24081	0	3380	(Noor et al., 2011)
15	Malay	31093	93293	23645	0	0	(Noor et al., 2011)
16	Basque	28848	48264	25676	0	0	(Pociello et al., 2011)
17	Dutch	28253	57706	40726	0	0	(Postma et al., 2016)
18	**Mongolian**	23665	40944	26857	213	2976	our resource
18	Croatian	21302	45929	27161	0	0	(Oliver et al., 2016)
19	Persian	17705	30365	17544	0	0	(Montazery and Faili, 2010)
20	Greek	17302	23117	17278	0	0	(Stamou et al., 2004)

Table 1 provides accuracy values for the three translations. The average accuracy for all the translations is 96.4, and the inter-annotator agreement between three annotators was 98.1.

The Mongolian WordNet now contains 23665 synsets, 40944 senses, and 26857 words as a result of the combination of all the above methods. As can be seen from Table 1, the resource is covering the 61.6 percents of 4960 "core" synsets derived from (Boyd-Graber et al., 2006).

7 Conclusion

We described how Mongolian WordNet is created by using three types of translation: *expert*, *monosemy*, and *hypernym-based* translations under the expansion method of PWN. Our main goal was to create a high-quality lexical resource, so that in automatic translations, we only selected the intuitive patterns (*monosemy* and *hypernym*) which are ensuring high quality in principles.

Mongolian WordNet contains 23665 synsets, 40944 senses, and 26857 words. There are 15976 nouns, 3791 verbs, 601 adverbs, and 3037 adjectives. In addition, it has 213 examples and 2976 glosses. The average polysemy is 1.52. The resource is delivered in the tab-separated format (Bond and Foster, 2013) under the CC BY-NC-SA 4.0 license[3].

8 Acknowledgements

The research has received funding from the Mongolian Science and Technology Fund under grant agreement SSA_024/2016. The result described in this paper is part of the Mongolian Local Knowledge Core project, partially supported by the National University of Mongolia under grant agreement P2017-2383. The first author is supported by the Cyprus Center for Algorithmic Transparency, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 810105.

[3]https://creativecommons.org/licenses/by-nc-sa/4.0/

References

Pushpak Bhattacharyya. 2017. Indowordnet. In *The WordNet in Indian Languages*, pages 1–18. Springer.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1352–1362.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small*, 8(4):5.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the japanese wordnet. In *Proceedings of the 7th workshop on Asian language resources*, pages 1–8. Association for Computational Linguistics.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*, pages 29–36. Citeseer.

Subhashis Das and Fausto Giunchiglia. 2016. Geoetypes: Harmonizing diversity in geospatial data (short paper). In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 643–653. Springer.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

Pianta Emanuele, Bentivogli Luisa, and Girardi Christian. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Mauricio Espinoza, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez. 2009. Ontology localization. In *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*, pages 33–40.

Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. slownet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117.

Amarsanaa Ganbold and Altangerel Chagnaa. 2015. Crowdsourcing Localization of Ontology and Geographical Names. In *The Eighth International Conference on Frontiers of Information Technology*, pages 120–124, Jilin, China.

Amarsanaa Ganbold, Feroz Farazi, Moaz Reyad, Oyundari Nyamdavaa, and Fausto Giunchiglia. 2014. Managing language diversity across cultures: The english-mongolian case study. *International Journal on Advances in Life Sciences*, 6(3-4).

Amarsanaa Ganbold, Altangerel Chagnaa, and Gábor Bella. 2018. Using Crowd Agreement for Wordnet Localization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Fausto Giunchiglia, Mladjan Jovanovic, Mercedes Huertas-Migueláñez, and Khuyagbaatar Batsuren. 2015. Crowdsourcing a large scale multilingual lexico-semantic resource. In *The Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP-15), San Diego, CA*.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4009–4017.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed Alhakim Freihat. 2018. One world—seven thousand languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018*.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Mercedes Huertas-Migueláñez, Natascia Leonardi, and Fausto Giunchiglia. 2018. Building a lexico-semantic resource collaboratively. In *The XVIII EURALEX International Congress*, page 148.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mortaza Montazery and Heshaam Faili. 2010. Automatic persian wordnet construction. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 846–850. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10.

Nurril Hirfana Bte Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open wordnet bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*.

Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2016. Automatic expansion of croatian wordnet. In *Međunarodni znanstveni skup Hrvatskoga društva za primijenjenu lingvistiku*.

Alessandro Oltramari, Aldo Gangemi, Nicola Guarino, and Claudio Masolo. 2002. Restructuring wordnet's top-level: The ontoclean approach. *LREC2002, Las Palmas, Spain*, 49.

Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2):121–142.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open dutch wordnet. In *Proceedings of the Eighth Global WordNet Conference*, page 300.

Benoît Sagot and Darja Fišer. 2008. Building a free french wordnet from multilingual resources. In *Proceedings of the Ontolex 2008 Workshop*.

Patanakul Sathapornrungkij and Charnyote Pluempitiwiriyawej. 2005. Construction of thai wordnet lexical database from machine readable dictionaries. *Proc. 10th Machine Translation Summit, Phuket, Thailand*.

Sofia Stamou, Goran Nenadic, and Dimitris Christodoulakis. 2004. Exploring balkanet shared ontology for multilingual conceptual indexing. In *LREC*.

Sareewan Thoongsup, Kergrit Robkop, Chumpol Mokarat, Tan Sinthurahat, Thatsanee Charoenporn, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai wordnet construction. In *Proceedings of the 7th workshop on Asian language resources*, pages 139–144. Association for Computational Linguistics.

Dan Tufiş, Radu Ion, Luigi Bozianu, Alexandru Ceauşu, and Dan Ştefănescu. 2008. Romanian wordnet: Current state, new applications and prospects. In *Proceedings of 4th Global WordNet Conference, GWC*, pages 441–452.

Kai Von Fintel and Lisa Matthewson. 2008. Universals in semantics. *The linguistic review*, 25(1-2):139–201.

Piek Vossen, Francis Bond, and J McCrae. 2016. Toward a truly multilingual global wordnet grid. In *Proceedings of the Eighth Global WordNet Conference*, pages 25–29.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

English WordNet 2019 – An Open-Source WordNet for English

John P. McCrae
Data Science Institute
Insight Centre for Data Analytics
National University of Ireland Galway
john@mccr.ae

Alexandre Rademaker
IBM Research and FGV/EMAp
Brazil
alexrad@br.ibm.com

Francis Bond
Nanyang Technological University
bond@ieee.org

Ewa Rudnicka
Wroclaw University of Technology
ewa.rudnicka@pwr.edu.pl

Christiane Fellbaum
Princeton University
fellbaum@princeton.edu

Abstract

We describe the release of a new wordnet for English based on the Princeton WordNet, but now developed under an open-source model. In particular, this version of WordNet, which we call English WordNet 2019, which has been developed by multiple people around the world through GitHub, fixes many errors in previous wordnets for English. We give some details of the changes that have been made in this version and give some perspectives about likely future changes that will be made as this project continues to evolve.

1 Introduction

WordNet (Miller, 1995; Fellbaum, 1998) is one of the most widely-used language resources in natural language processing and continues to find usage in a wide variety of applications including sentiment analysis (Wang et al., 2018), natural language generation (Juraska et al., 2018) and textual entailment (Silva et al., 2018). However, in the recent few years there has been only one update since version 3.0 was released in 2006, in spite of its wide use and the interest in the data. In the meantime, a number of other wordnet teams working with the WordNet data have proposed modifications or extensions to its latest release. These two facts have provided the chief motivation for our present initiative, namely developing an open-source WordNet for English on the basis of Princeton WordNet (to be released under the name English WordNet 2019).

In order to allow for meaningful comparisons of performance on tasks using WordNet as a component, it is important to maintain a single (or very few) wordnets as a standard and reference.

One of the core issues preventing further development of the original WordNet model has been the question of how to ensure the resource maintains its quality. The Princeton WordNet team has followed a model that requires an expert lexicographer to review and implement all changes. In this paper, we discuss the development of Open English WordNet, which instead follows a methodology of quality assurance that is based on those typically used for open-source projects, especially those connected to the Linux operating system. In particular, we can consider this to be an application of Linus's Law ("given enough eyeballs, all bugs are shallow") to the development of WordNet, similar to other open source orientated projects such as OpenWordNetPT (Paiva et al., 2012) and the recently announced Global FrameNet project[1]. Still, we will do our best to make new data or proposed changes verified by expert lexicographers or developers whenever possible.

We have implemented this in terms of a new 'fork' of Princeton WordNet, and have released a new version of WordNet that fixes many of (mostly trivial) errors, such as spelling mistakes, and thus improves the quality of the resource. We take inspiration from other forks such as the MariaDB fork of MySQL and aim to make this a 'drop-in' replacement for Princeton WordNet. This is achieved by ensuring that that data is available in a wide range of formats, including those used by Princeton to publish the resource and standards promoted by the Global WordNet Association so that existing projects can use these changes without updates to their workflows. In particular, we continue to follow the basic conception of Princeton WordNet and do not introduce changes that would fundamentally affect the nature of the wordnet. Instead, our focus for this release is on fixing more minor errors and for future releases we plan to extend this to principally adding new synsets and relations, using the existing structure

[1] https://www.globalframenet.org/

as a guide. As an open-source project we expect that the community will create synsets that reflect their views, and that this may in the long run lead to more significant divergences from the Princeton WordNet model,

Moreover, we also present a new website and project that allows the resources to be queried at http://en-word.net, which presents the most recent changes in a dynamic manner as they are updated on the GitHub website. To indicate that this is a clearly new version of WordNet we have termed this version the 2019 edition of English WordNet and provide a clear and auditable list of changes that have been made such that it would be possible for the Princeton WordNet to use these changes in any future versions they make.

This paper is structured as follows: first, we will present some other efforts to extend the Princeton WordNet for English and then we will describe the kinds of changes that we have made for this release. We will then provide a brief discussion of the open issues that will be handled in the next version and how they may be handled. We will then briefly describe the release and the implementation of the user interface.

2 Background

Princeton WordNet (Miller, 1995; Fellbaum, 2010) is the first wordnet for English, however it is not the only one that has been developed for this language. Moreover, it has been the case that during the development of several wordnets for other languages signficant changes and/or additions were made to the underlying structure and content of the English section of the wordnet. In at least one case, namely the development of the Polish wordnet, plWordNet, the additions to the underlying English wordnet have been so numerous that they were released as a new wordnet, enWordnet (Rudnicka et al., 2015; Maziarz et al., 2016). These involved the addition of new lemmas (over 11k), lexical units (over 11k) and synsets (7.5k). The latter were linked to WordNet 3.1 synsets via hyponymy relation. Still, no alterations to the original WordNet synsets or relations were made within this project. Currently, enWordnet is only available as part of the plWordNet project and does not constitute a 'drop-in' replacement for Princeton WordNet.

Some projects have attempted to expand Princeton WordNet with new terminology in other directions, for example the Colloquial WordNet project (McCrae et al., 2017), has been working on adding new terms that are used in social media, and this is available using the same GWC formats (McCrae et al., 2019) as this work; a similar project called SlangNet (Dhuliawala et al., 2016) seems to be unavailable now. There have also been a number of attempts to extend WordNet in terms of the kinds of annotation that it contains, such as the addition of sentiment and emotion information (Strapparava et al., 2004) or combining it with a upper-level ontology (Niles and Pease, 2003).

Another significant direction has been the automatic extension of WordNet and several projects have been published based on extending WordNet with information from other resources, especially Wiktionary and Wikipedia. One of the most prominent of such resources is BabelNet (Navigli and Ponzetto, 2012), which combines multiple methods using machine learning based methods, which have been shown to have a precision of up to 89.7%. A similar effort was carried out by the UKP group and led to the Uby resource (Gurevych et al., 2012), who report similar levels of accuracy in the mapping. While such automatically constructed resources may be valuable for a large number of applications, they cannot replace WordNet for applications that require a gold standard lexicon or very high precision. Further, many of these resources have taken WordNet as is, and have often repeated the same design and frequently copied many of the minor errors into their own resources.

3 The Open English WordNet Project

The Open English WordNet Project[2] takes the form of a single Git repository, published on GitHub, and consisting for the most part of a collection of XML files describing the synsets and lexical entries in the resource. These XML files represent each of the lexicographer file sections of the original resource and a simple script is provided to stitch them together into a single XML file. The XML files are compliant with the GWC LMF model (McCrae et al., 2019)[3], which is itself based partially on the LMF model (Francopoulo et al., 2006) and in particular the WordNet (a.k.a

[2] https://github.com/globalwordnet/english-wordnet

[3] https://globalwordnet.github.io/schemas/

Kyoto) LMF variant (Soria et al., 2009). Due to its basis on LMF, a particular challenge was that the entire wordnet should be represented as a single XML document. However, due to the relative verbosity of the LMF format, the final data ended up as 97 MB, exceeding the upload limits of GitHub, so instead the single XML file was divided by lexicographer sections. Even still, this creates several very large files (over 10 MB) and this has resulted in some challenges for those working on the project[4], which may be solved by the adoption of a less verbose format.

The model for contributing to this work is similar to that of other large open-source projects, where a small number of trusted developers are able to make changes to the code directly to the source of the wordnet, while submissions may be proposed by any user registered with GitHub in two principal channel:

Issues Any user may log an issue with the system, describing the changes that they would like to make to the wordnet, along with technical information including the identifier of the synset and the type of proposed change (e.g., 'merge synset'). Issues are then assigned to a trusted developer and implemented by them.

Pull request Technically-inclined users may make the changes directly to the XML and propose them for review by one of the trusted developers. This method generally leads to faster acceptance of changes.

In both cases, changes are covered by contribution guidelines[5], which also maintain the integrity of the project in terms of fostering an inclusive, kind, harassment-free, and cooperative community. Currently, this combination of technical hurdles and clear guidelines has prevented any cases of politically motivated or otherwise inappropriate changes being proposed to the wordnet.

In addition to the raw data itself, a number of scripts have been introduced that can be used with the model. These include a 'post-receive' hook that takes the most recent changes to the WordNet and immediately converts it into other formats including RDF based on OntoLex-Lemon (Cimiano et al., 2014) as well as in the WNDB formats used for previous versions of WordNet, allowing English WordNet to be a 'drop-in' replacement for Princeton WordNet. Furthermore, this update is used to populate the searchable frontend, which is available at http://en-word.net/.

4 Scope of Changes

One of the first major class of errors that we attempted to fix were simple spelling errors that occur particularly in the definitions and the examples of the synsets. In most cases these were entirely obvious errors for example the following definition:

> habitually do something or be in a certan[6] state or place (use only in the past tense)

This change in a few cases also affected the lemmas in the resource, for example the lemma 'poetic jstice' was corrected. In a few cases, there was uncertainty as the spelling variant was non-standard, for example in 3 cases the word 'Moslem' was used as opposed to the 115 cases of the far more common variant 'Muslim', so these were corrected to a single spelling form.

A second major source of errors was that many examples did not use any lemmas from the synset and as such could not be considered examples of the synset. We used a simple edit distance based approach to identify 434 synsets for which this appeared to be an issue. Of those we found that 341 represented a clear error that was easy to be fixed. For these various strategies were followed:

- The example was deleted as there were other examples in the synset that exemplified the meaning better
- A new example was found by conducting a GDEX (Kilgarriff et al., 2008) search of a EnTenTen15 web corpus provided by the Sketch Engine tool[7].
- The example was modified by replacing a word not in the synset with a synset member or by providing a suitable modification, for example the example of 'double negative' was 'I don't never go' and was updated to 'double negative such as 'I don't never go" to include the lemma.

[4] Issue #31: https://github.com/globalwordnet/english-wordnet/issues/31

[5] https://github.com/globalwordnet/english-wordnet/blob/master/CONTRIBUTING.md

[6] corrected to 'certain'

[7] https://www.sketchengine.eu/

- An issue was logged, as it was identified that this example shows a more signficant change. This was often the case when the example used a lemma or a hypernym and it was not clear if the distinction between synsets was meaningful.

A third major change was to introduce new synset members based on a previously calculated WordNet-Wikipedia mapping (McCrae, 2018). In particular, if this mapping, which has already been manually verified, linked to a page title that did not match the lemma, the page title was added as a new lemma to the synset. This was, as with all changes, manually verified in its entirety before the change was made.

Finally, the repository has been open to new suggestions of changes and there have been many suggestions already contributed about sporadic and various changes to the wordnet. A sample of these include:

- The sense of 'threepenny' as a size was incorrect in the actual length in inches of a threepenny.
- Grammatical errors were fixed, such as in the definition '(of) or pertaining to the Corinthian style of architecture' of 'Corinthian' the first word was missing.
- The death dates and birth dates of various famous figures. Notably the change to the synset for 'William A. Cragie' was accepted into the Princeton WordNet and is the only change from this project that has been taken up to date.

5 Ambition

Our ambition for this project is to have annual releases and as such we detail some of the changes that we plan to make that would not fundamentally change the nature of the resource, and these changes will likely be the basis of the releases for the next couple of years. We then look into more significant extensions that would be planned for releases in the long-term.

5.1 Non-trivial fixes

Currently, there are 113 open issues listed on the project and this is due to a clear plan that the project would only deal with issues for the 2019 release that are unlikely to have any effect on any projects that are dependent on Princeton WordNet. This precludes making certain changes involving deleting or adding new synsets, however this restriction is intended to be relaxed for the 2020 release. A summary of the kinds of errors is given in Table 1, and these are categorized by the likely changes that would need to be made.

Change Type	Issues Reported
Synset Duplicate	45
Synset Split	7
New Synset	22
Synset Members	10
Delete Synset	8
Add Relation	3
Change Relation	14
Definition	18
Example	1

Table 1: The current list of issues that have been reported but not implemented in this version of the resource

Synset duplicate It appears that two synsets refer to the same concept. For example, currently the wordnet has entries for both 'Aram Kachaturian' and 'Aram Khachaturian'[8], in both cases referring to an Armenian composer with the same date of birth. In this case one of the synsets will be deleted and all synset links merged.

Synset split In some cases it has been suggested that a synset represents two distinct concepts. For example, the synset for 'Dharma'[9] is defined as 'basic principles of the cosmos; also: an ancient sage in Hindu mythology worshipped as a god by some lower castes', and it is clear that these two definitions are not compatible. These cases are harder to solve, as it is unclear whether a single new concept should be introduced or whether the original should be deleted and two new concepts introduced.

New synset Here obvious gaps have been discovered in WordNet. For example, the synset for 'jackal' also identifies the synset by its

[8] https://github.com/globalwordnet/english-wordnet/issues/66

[9] https://github.com/globalwordnet/english-wordnet/issues/113

	Princeton WordNet 3.1	English WordNet 2019 (Change)
Synsets	117,791	117,791
Lemma	159,015	159,789 (+797, -23)
Senses	207,272	208,353 (+1,081)
Synset Relations	285,668	285,666 (-2, 662 changed)
Sense Relations	92,535	92,535
Definitions	117,791	117,791 (925 changed)
Examples	47,539	48,419 (-237, +1117)

Table 2: Comparative size of Princeton WordNet 3.1 and English WordNet 2019

Latin name 'Canis aureus'[10]. However, in fact 'jackal' is a term for four closely related Canis species, suggesting that all four should have synsets with a single upper concept for all jackals.

Synset members In this case, one of the synset members is incorrect and could be updated. This is often reported alongside a second issue above (synset split).

Delete synset In general, we would prefer not to remove synsets from the WordNet, however there are several synsets in Princeton WordNet that do not seem to meet the requirements for inclusion. An example of this is 'de-ionate', which while clear in its meaning, does not, according to searches of Sketch Engine's large EnTenTen15 Web Corpus, appear to be in use in any domain. There is still an open question as to whether we should delete such rare or incorrect words, however we do notice that on a Google search for this term, the few usages we can find appear to be cases where 'deionized' was likely intended, and so omitting incorrect words may help users to identify errors in their usage of the language.

Add relation This indicates a relation between two synsets is missing.

Change relation The type or target of a relation is incorrect. A number of clearer errors of this type were fixed in the 2019 release (e.g., the use of `hypernym` in place of `instance_hypernym`) and others are scheduled for 2020, for example the inclusion of 'impressionist' as a direct hyponym of 'painter' suggesting that impressionist art was only carried out through the medium of painting.

Definition/example These represent the largest class of changes in the 2019 release as they only affected issues with the textual definition of synsets and most of these could be implemented without any semantic change to the synset. More of these changes are planned for the 2020 version of English WordNet.

5.2 Extending WordNet

As described in the introduction, there are a number of resources that have made extensions to WordNet and there seems to be no strong reason that the results of these projects could not be included within the English WordNet. Firstly, the Colloquial WordNet project (McCrae et al., 2017) uses the same form of data as English WordNet and many of its entries could be easily included in the context of English WordNet. However, as the resource was mostly created by a single annotator the quality control issues are not clear. Furthermore, by the nature of the resource, it follows that some of the entries may be too vulgar or ephemeral to be worthy of inclusion in English WordNet, however these are marked in the original resource.

Another large resource with many extra English synsets is enWordNet (Maziarz et al., 2016) and this consists of many extensions to WordNet, which could be introduced into English WordNet. Although the format used for enWordNet is different to that of English WordNet (and in fact conceptually differs in some ways from that of Princeton WordNet), many of the definitions introduced appear to be drawn from Wikipedia and this may require the project to adopt the more restrictive CC-BY-SA license of Wikipedia. Moreover, it is not

[10] https://github.com/globalwordnet/english-wordnet/issues/125

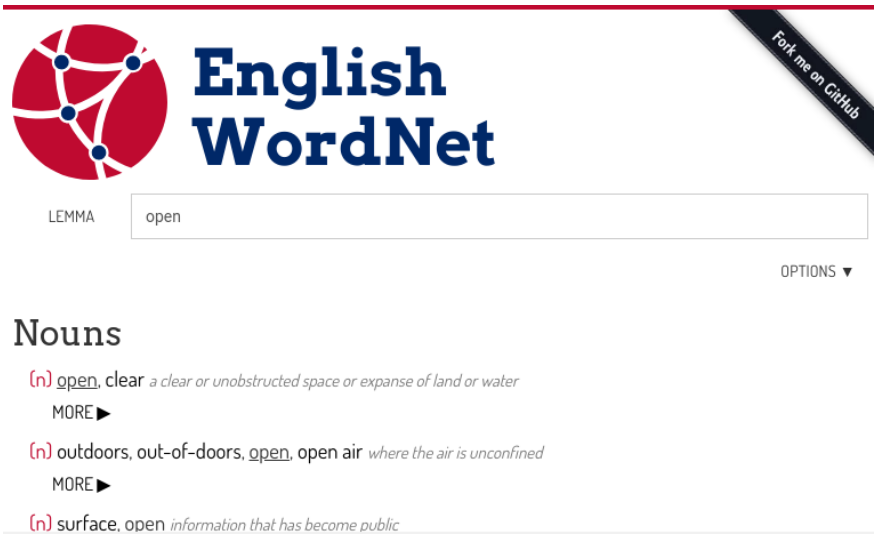


Figure 1: Screenshot of the new English WordNet interface

clear how many of the entries have been reviewed by native speakers of English.

Finally, a long term goal would be to introduce a principled method for introducing new synsets, which are of high quality and this would have to involve reviewing of all the links between synsets that have been introduced. It is expected that this could be achieved by a semi-automatic procedure where potential links are learnt from text (Espinosa-Anke et al., 2016) combined with a crowd-sourced reviews. Another important aspect of each synset is also its definition and as many of the definitions in WordNet are of poor quality (McCrae and Prangnawarat, 2016), it is necessary to adopt some general guidelines for writing definitions that can ensure high quality, such as those defined for ontological definitions (Seppälä et al., 2017). Further, we will implement and further extend the validations that are available and automate the checking such that it is clear if any changes are breaking issues. In particular, we currently implement simple DTD validation of the merged XML, which also catches many other issues, such as senses without synsets, but we are working to extend this validation to include issues, such as hypernyms without hyponyms, etc.

In order to achieve this, it is important that strong tools are available for the creation and maintenance of the resource and it is likely that tools coming out of the ELEXIS project (Krek et al., 2018; Pedersen et al., 2018) will be adapted to this task.

6 Results for this release

This release represents a mostly maintenance release where obvious errors have been fixed. In Table 2 we see that most of the updates are to the definitions and examples used to describe the synsets in English WordNet. There have also been a number of removals relative to the previous version of Princeton WordNet: mispelled lemmas were removed and replaced with a correctly spelled variant and these were counted as both a removal and addition of a lemma. Secondly, due to an issue[11] two links were removed as they were deemed clearly incorrect. These changes in total 2,002 synsets which means changes in 1.70% of synsets over the most recent version of Princeton WordNet.

7 Interface to English WordNet

In addition to the development of a new resource, we have also developed a new interface to the resource, which is available at `https://en-word.net`. This interface is developed using the latest Web technologies including the

[11] https://github.com/globalwordnet/english-wordnet/issues/11

use of AngularJS[12] and the use of Rocket[13], a Rust-based framework for Web applications. This interface is also open-source and released on GitHub [14]. This interface provides a fast and attractive interface (see Figure 1) to the data and in addition, allows the data to be browsed as linked data using the RDF interface as provided by (McCrae et al., 2014). In addition, clear links are provided to the GitHub to encourage contributions and to the Global WordNet Association.

8 Conclusion

In this paper, we have presented a new version of WordNet for English that has been developed as a fork of the Princeton Wordnet and in particular we describe the first release of this resource as a 'drop-in' replacement for the Princeton WordNet. As a main contribution, we have moved the development of English WordNet to an open-source framework, ensuring that the development of WordNet is not constrained by the funding situation at a single institute. Instead, we commit to a yearly update cycle and welcome contributions from many directions. We believe that one of the most important challenges with this will be ensuring that WordNet can remain a gold standard resource for NLP applications. Moreover, we note that as this resource has fixed over 3,500 errors in WordNet, the English WordNet 2019 release is naturally of higher quality than any previous Princeton WordNet release.

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Unions Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure and grant agreement No 825182, Prêt-à-LLOD.

References

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2014. Lexicon Model for Ontologies: Community Report. W3C community group final report, World Wide Web Consortium.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like resource for English slang. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, pages 4329–4332.

Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. Supervised distributional hypernym discovery via domain adaptation. In *Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1-5; Austin, TX. Red Hook (NY): ACL; 2016. p. 424-35.* ACL (Association for Computational Linguistics).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. UBY: A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.

Juraj Juraska, Panagiotis Karagiannis, Kevin K Bowden, and Marilyn A Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. *arXiv preprint arXiv:1805.06553*.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of Euralex*.

Simon Krek, John McCrae, Iztok Kosem, Tanja Wissek, Carole Tiberius, Roberto Navigli, and Bolette Sandford Pedersen. 2018. European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Pawe Kdzia. 2016. PlWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL.

[12] https://angularjs.org/

[13] https://rocket.rs/

[14] https://github.com/jmccrae/wordnet-angular

John P. McCrae. 2018. Mapping WordNet Instances to Wikipedia. In *Proceedings of the 9th Global WordNet Conference*.

John P. McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and Linking WordNet using lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

John P. McCrae and Narumol Prangnawarat. 2016. Identifying Poorly-Defined Concepts in WordNet with Graph Metrics. In *Proceedings of the First Workshop on Knowledge Extraction and Knowledge Integration (KEKI-2016)*.

John P. McCrae, Piek Vossen, Luis Morgado da Costa, and Francis Bond. 2019. The Global WordNet Association Schemas. In *Submitted to LILT Special Edition on WordNets*.

John P. McCrae, Ian Wood, and Amanda Hicks. 2017. The Colloquial WordNet: Extending Princeton WordNet with Neologisms. In *Proceedings of the First Conference on Language, Data and Knowledge (LDK2017)*.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Ian Niles and Adam Pease. 2003. Linking lixicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Ike*, pages 412–416.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian wordnet for reasoning. Technical report, COLING 2012.

Bolette Pedersen, John McCrae, Carole Tiberius, and Simon Krek. 2018. ELEXIS - a European infrastructure fostering cooperation and infor-mation exchange among lexicographical research communities. In *Proceedings of the 9th Global WordNet Conference*.

Ewa Rudnicka, Wojciech Witkowski, and Michał Kaliński. 2015. Towards the Methodology for Extending Princeton WordNet. *Cognitive Studies*, 15(15):335–351.

Selja Seppälä, Alan Ruttenberg, and Barry Smith. 2017. Guidelines for writing definitions in ontologies. *Ciência da Informação*, 46(1).

Vivian S Silva, Siegfried Handschuh, and André Freitas. 2018. Recognizing and justifying text entailment through distributional navigation on definition graphs. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 international workshop on Intercultural collaboration*, pages 139–146. ACM.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet Affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Citeseer.

WM Wang, Z Li, ZG Tian, JW Wang, and MN Cheng. 2018. Extracting and summarizing affective features and responses from online product descriptions and reviews: A kansei text mining approach. *Engineering Applications of Artificial Intelligence*, 73:149–162.

Assessing Wordnets with WordNet Embeddings

Ruben Branco[1] and **João Rodrigues**[1] and **Chakaveh Saedi**[1,2] and **António Branco**[1]

[1]University of Lisbon
NLX-Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal

[2]Macquarie University
Department of Computing
Sydney, Australia
{ruben.branco,jrodrigues,chakaveh.saedi,ahb}@di.fc.ul.pt

Abstract

An effective conversion method was proposed in the literature to obtain a lexical semantic space from a lexical semantic graph, thus permitting to obtain WordNet embeddings from WordNets. In this paper, we propose the exploitation of this conversion methodology as the basis for the comparative assessment of WordNets: given two WordNets, their relative quality in terms of capturing the lexical semantics of a given language, can be assessed by (i) converting each WordNet into the corresponding semantic space (i.e. into WordNet embeddings), (ii) evaluating the resulting WordNet embeddings under the typical semantic similarity prediction task used to evaluate word embeddings in general; and (iii) comparing the performance in that task of the two word embeddings, extracted from the two WordNets. A better performance in that evaluation task results from the word embeddings that are better at capturing the semantic similarity of words, which, in turn, result from the WordNet that is of higher quality at capturing the semantics of words.

1 Introduction

Lexical semantics studies the semantic properties of lexical units, and is often defined as the study of word meaning. Given its importance, the computational representation of lexical meaning is a core challenge in natural language processing (NLP).

Since the meaning of a word is strongly related to the meaning of other words, the relations between words are a key ingredient for the representation of their meaning. There have been different types of representations proposed for lexical semantics, which, in general, can be viewed as pertaining to one of three main family of representations, namely semantic networks (Quillan, 1966), feature-based models (Minsky, 1975; Bobrow and Norman, 1975), and semantic spaces (Harris, 1954; Osgood et al., 1957).

Semantic networks are a type of approach for lexical semantics that is based on graphs. In a nutshell, a lexical unit, typically a word, is recorded as a node in a graph while the semantic relations among words, such as hyponymy or synonymy, etc., are recorded as labeled edges among the nodes of the graph. One of the most popular semantic networks is WordNet (Fellbaum, 1998). It stands out as being a lexical semantics network based on non trivial linguistic intuitions of human experts.

Feature-based models representing lexical semantics, in turn, resort to a hash table that stores the lexical units as keys, and the semantically related units as the respective values. Small World of Words (De Deyne et al., 2013) is an example of such a model. In its development, the semantic features (related words) of a lexical entry can be obtained straightforwardly from laypersons by using the lexical entry as a cue to evoke possible words associated to it.

Finally, in semantic spaces, the meaning of a lexical unit is represented as a vector in a high dimension space — also known as word embedding —, typically obtained on the basis of the frequency of its co-occurrence with other lexical units, resorting to a large collection of documents. Word2vec (Mikolov et al., 2013) is an example of a method to obtain semantic spaces.

Bridging between these different types of lexical meaning representations is instrumental for a wider use of all the existing lexical semantics resources. Unifying this knowledge in one lexical semantic representation would carry an immediate impact across a range of NLP tasks.

An existing form of (partial) bridging is ob-

tained with the conversion of one type of representation to another as in (Saedi et al., 2018), with the wnet2vec methodology. Wnet2vec permits the conversion from lexical semantic networks to lexical semantic spaces, termed as WordNet embeddings.

The success of this type conversion can be measured by using the typical semantic space evaluation process. That is obtained by comparing the semantic similarity scores between the vectors of words arranged in pairs against the gold scores of semantic similarity among the words in the pairs, which were obtained from human subjects.

The evaluation of the semantic similarity task based on the semantic space wnet2vec used the SimLex-999 (Hill et al., 2016), a mainstream semantic similarity data set composed of 999 pairs of words with a correspondent similarity strength value. Semantic similarity detection with wnet2vec (Saedi et al., 2018) shows an almost 20% superior result against a strong baseline, namely Google's word2vec semantic space, which is trained on a very large collection of 100 Billion token texts.

Our goal in the present paper is to propose the exploitation of this conversion methodology as the basis for the comparative assessment of WordNets: Given two WordNets, for the same language, their relative quality in terms of capturing the lexical semantics of that language, can be assessed by (i) converting each WordNet into the corresponding semantic space (i.e. WordNet embeddings), (ii) evaluating the resulting embeddings in the semantic similarity prediction task; and (iii) comparing the performance in that task of the two word embeddings, extracted from the two WordNets. A better performance results from the word embeddings that better capture the semantic similarity of words, which, in turn, results from the WordNet that is of higher quality at capturing the semantics of words.

In order to illustrate this proposed methodology for the comparative assessment of WordNets with a first exercise with its application, we resort to two WordNets of the same language, Portuguese, developed under two distinct methodologies, namely MWN.PT — hand-crafted — and OWN-PT — built (semi-)automatically.

The next Section 2 reports on the conversion of the hand-crafted WordNet to the respective WordNet embeddings and on the performance of the latter in the semantic similarity prediction task. The following Section the same exercise is undertaken but now with the WordNet built (semi-)automatically. Sections 4 and 5 present, respectively, the discussion of the results and the related work. The conclusions are presented in Section 6.

2 Embeddings from hand-crafted WordNet

The MultiWordnet of Portuguese (MWN.PT) is developed under the same methodological principles as the seminal Princeton English Wordnet — including the resorting to manually validated representations. Its synsets are aligned with the translationally equivalent synsets in Princeton WordNet. It is available from ELRA-European Language Resources Association.[1] Besides the difference in the language covered, MWN.PT differentiates to Princeton WordNet by being smaller, encompassing 17k concepts/synsets (against over 120k of Princeton), by encoding only synonymy and hyponymy/hypernymy (against some 25 semantics relations in Princeton WordNet), and by including only nouns (against all open categories), and includes mostly the sub-ontologies of Person, Organization, Event, Location and Art works. Hence, it offered interesting contrasting conditions to proceed with an empirical study of the strength of the wnet2vec methodology when applied to quite different and more challenging empirical settings than the one originally resorted to in (Saedi et al., 2018) to convert the Princeton WordNet into its WordNet embeddings.

To obtain word embeddings, the mainstream methods have used the frequency of co-occurrence in large corpora between the target word and its neighboring words to construct the respective vector. Instead of texts and the frequency of co-occurrence between words, wnet2vec resorts to lexical semantics graphs and the knowledge encoded in them, using the semantic networks as the empirical source to obtain the vectors of the corresponding semantic space. The key insight in the conversion process is that a stronger semantic affinity between two lexical units is found between nodes that are closer and have a higher number of connecting paths.

[1] MWNT.PT was obtained from http://catalogue-old.elra.info/product_info.php?cPath=42_45&products_id=1101&language=en

In a nutshell, the wnet2vec methodology starts by creating a matrix with all of the possible semantic relations between all the words, resulting in an adjacency matrix M. Then it populates each cell M_{ij} of the matrix resorting to a WordNet, in the present experiment MWN.PT, as the semantic graph G. Each cell M_{ij} is set to 1 if and only if there is a direct edge between synsets including the two words $word_i$ and $word_j$ the cell encodes/represents. Words present in the same synset have a synonym relation and thus are assigned a value of 1. If there is no edge between the two words that cell is set to 0.

For all nodes not directly connected, that is connected through other nodes in between, the representation of their affinity strength is obtained by following the cumulative iteration:

$$M_G^n = I + \alpha M + \alpha^2 M^2 + \ldots + \alpha^n M^n \quad (1)$$

M^n is the matrix where every two words, $word_i$ and $word_j$, are transitively related by n edges. I represents the identity matrix and α is used as a decay factor for longer paths.

The iteration converges into the matrix M_G, obtained by an inverse matrix operation:

$$M_G = \sum_{e=0}^{\infty} (\alpha M)^e = (I - \alpha M)^{-1} \quad (2)$$

After the convergence, a Positive Point-wise Mutual Information transformation (PMI+) is applied to reduce the frequency bias, followed by an L2-norm to normalize each line of M_G, and finally, a Principal Component Analysis (PCA) is applied to reduce the dimension of the vectors. Further details on this conversion can be found in (Saedi et al., 2018).

2.1 From the semantic graph to a corresponding semantic space

When converted to a semantic space and the resulting semantic space is evaluated on the semantic similarity task with SimLex-999, Princeton WordNet supports a wnet2vec whose performance has an accuracy score of 0.50 in terms of Spearman's coefficient (Saedi et al., 2018). On the same task and testing dataset, Google's word2vec semantic space, used as the baseline, obtains 0.44 accuracy score.

While the semantic space obtained from the English WordNet was evaluated with the original SimLex-999 dataset, given we are handling here Portuguese instead, we resort to LX-SimLex-999 (Querido et al., 2017), which resulted from the translation of SimLex-999 into Portuguese.[2]

And while the corpus-based baseline for English was the Google's word2vec semantic space, for the corpus-based baseline here, we resort LX-DSemVectors 2.2b (Rodrigues and Branco, 2018) for Portuguese that also uses word2vec learning tools.[3] This semantic space was trained over a collection of text with more than 2 Billion tokens and obtains state-of-the-art results in a wide range of test datasets, including the LX-SimLex-999. Its best-reported accuracy score with this testing dataset is 0.35, in terms of Spearman's coefficient. All evaluations use the cosine distance measure between the vectors.

We use the same settings as in the experiment with the English WordNet, using here a decay factor of 0.75 and all available semantic relations being taken into account. The dimensions of the embeddings were kept at 850, the best-reported size. In the experiment with English, only 60k of the over 120k synsets in Princeton WordNet were used due to memory footprint limitations. No such reduction was necessary for the MWN.PT conversion due to the smaller size (17k synsets) of this semantic graph.

Our experiments were performed with an Intel Xeon E5-2640 V2 with 2 CPUs, each CPU has 8 cores. The training resorted to an upper bound of 120GB of memory and took 2 days.

2.2 Results

The result obtained with the WordNet embeddings obtained from MWN.PT using wnet2vec methodology can be found in Table 1, together with the score of the baseline. The graph-based semantic space obtained from 15886 words with wnet2vec is 11 percentage points better than the corpus-based baseline obtained from 2.2B words with word2vec.

Given the difference in the size of their vocabularies, the number of similarity pairs with unknown words differs among the two semantic spaces. The LX-DsemVectors 2.2b, trained on more than 2 Billion tokens, covers almost all of the words of the 999 pairs, with only 3.5% pairs with unknown words. The semantic space obtained

[2] Obtained from https://github.com/nlx-group/LX-DSemVectors

[3] Obtained from http://lxcenter.di.fc.ul.pt/datasets/models/2.2b/

with the MWN.PT has a coverage with 74.9% pairs with unknown words.

Lexical Semantic Model	Similarity
MWN.PT (wnet2vec)	0.4643
LX-DSemVectors 2.2b (word2vec)	0.3502

Table 1: Performance in the semantic similarity task over the LX-SimLex-999, given by Spearman's coefficient (higher score is better).

3 Embeddings from (semi-)automatic WordNet

Given the lessons learned with the creation of a semantic space from the MWN.PT, in the second phase of our experiments we applied the same conversion methodology to another Portuguese WordNet, the OpenWordnet-PT (OWN-PT) (de Paiva et al., 2012).

While MWN.PT was built manually by resorting to human experts labor, OWN-PT is different in that it resorts to (semi-)automatic and machine learning methodologies, and has a dimension that is over three times larger — over 54k words (against over 15k in MWN.PT) —, thus offering an interesting case for empirical study.

We resorted to OpenWordnet-PT in the LMF format (Vossen et al., 2013), whose last release in this format we found is from October 2018. To reuse the scripts ready for the conversion to wordnet2vec, we converted this LMF format into a Princeton WNDB format, having retained 54390 words.[4] This conversion was done by iterating over the lexicon and keeping track of lexical entries and their lemmas and senses, according to a unique id to differentiate between them, and also keeping a log of the semantic relations between synsets. Only two semantic relations present in OWN-PT are not represented in the final converted network, due to them not being present in the Princeton WNDB format[5]. Those two semantic relations are "exemplifies" and "is_exemplified_by".[6]

For the sake of comparability, and given the different sizes of the two WordNets for Portuguese, three experiments were performed with OWN-PT.

In a first experiment, the 54390 words of OWN-PT in the WNDB format were used.

In a second experiment, a subset of OWN-PT was selected with the same number of words of MWN.PT (15886). The words that are common to both WordNets were selected. Given that not all of the MWN.PT words exist in the OWN-PT, further words were selected from OWN-PT to attain the aimed dimension. Remaining synsets were ordered from the ones with more outgoing edges to less outgoing edges and the words from the more connected synsets were selected until the intended dimension was reached. In previous experiments with English (Saedi et al., 2018), it became apparent that selecting words from synsets with more outgoing edges leads to semantic spaces with better performance in the semantic similarity task.

In a third experiment, a subset of equal dimension to the MWN.PT set was again extracted, this time with the simpler methodology of the second part of the selection undertaken in the second experiment: synsets were ordered from the ones with more to less outgoing edges and the words from the more connected synsets were selected until the intended dimension was reached.

Table 2 presents the scores obtained in these experiments.

4 Discussion

The result of these experiments with MWN.PT is in line with the results of the experiments with English (Saedi et al., 2018), even though now the experiment was with another language and with WordNets that are quite different in dimension and coverage than the English one. When evaluated in the semantic similarity task with a mainstream test dataset, the semantic space obtained from a concept-based semantic network with wnet2vec methodology outperforms the strong baseline consisting of a semantic space obtained from mainstream corpus-based methods, namely with word2vec trained with a very large collection of text, with 2.2B tokens in the present case.

The results of the subsequent experiments with OWN-PT are also in line with those findings. Even though it was built with a methodology resorting to heuristics and (semi-)automatics methods, the semantic space obtained from a second concept-based semantic network of Portuguese with wnet2vec methodology also outperforms the same strong baseline.

[4] https://wordnet.princeton.edu/documentation/wndb5wn

[5] All semantic relations from Princeton WNDB (https://wordnet.princeton.edu/documentation/wninput5wn) were resorted to

[6] This script is available from https://github.com/nlx-group/WordNet-Format-Conversion

WordNet	Similarity	Words
MWN.PT	0.4643	15886
OWN-PT All words (1st experiment)	0.3124	54390
OWN-PT Same size, common words w/ MWN.PT (2nd exp.)	0.4060	15886
OWN-PT Same size, synsets w/ more relations (3rd exp.)	0.4020	15886

Table 2: Performance of the models obtained from the conversion of MWN.PT and OWN-PT WordNets over LX-SimLex-999 given by Spearman's coefficient (higher score is better).

In this connection, we offer the observation that when a subset of the English WordNet was experimented using a number of synsets (25k) that is closer to our experiments reported here, a 0.45 score was obtained (against 0.53 with 60k synsets) (Saedi et al., 2018). This may indicate that improving the existing WordNets of Portuguese with a larger number of lexical units and relations may bring even better performance.

Additionally, the results of the experiments reported above suggest that when using the wnet2vec methodology to obtain a semantic space from a semantic graph, under comparable experimental circumstances (i.e. over 15k words that hold higher number of relations), 15% better semantic similarity performance scores are obtained with a manually crafted WordNet — 0.46 with MWN.PT — than with a WordNet obtained (semi-) automatically — 0.40 with OWN-PT.

This is in line with what is expected given the noise introduced by the (semi-)automatic methods used in the construction of WordNets. What is new with respect to the methodology proposed here is that there is now a quantitative way to assess the difference between WordNets in what concerns their different quality at capturing the semantics of words.

5 Related work

A proposal for the conversion from the Princeton WordNet to a semantic space different from the one used here can be found in (Goikoetxea et al., 2015). That is different in that in this other proposal the conversion from semantic graph to semantic space is not direct. First, a synthetic corpus is generated by a random walk in the WordNet. Then on the basis of that artificial text, common corpus-based techniques are used to obtain the word embeddings.

In (Gonçalo Oliveira, 2018), in turn, another approach was used to obtain a semantic space from Portuguese semantic networks also resorting to a random walk but, differently from the approach mentioned above, via direct conversion. Instead of using a concept-based semantic network (WordNet) as in our study reported here, semantic networks based on words only (no synsets) were used and converted to semantic spaces. Also, a different method than ours was used, a random walk with 30 iterations.

Although these differences render the results not comparable, it may be still interesting to draft some observations with the necessary caution and grains of salt. The best accuracy score reported in (Gonçalo Oliveira, 2018) with LX-SimLex-999 is 0.61 in terms of Spearman's coefficient. This score is obtained with a network with more than 200k words, more than ten times larger than the network used in our study reported here, with approximately 17k synsets.

The system that, in turn, is reported there as having a performance score of 0.45, in line with the score of 0.46 we found here for a 17k network, was trained over a network five times larger than the one used here. This may be another sign of the higher quality of (hand-crafted) WordNets at recording the lexical semantics of words.

In future work, it will be interesting to undertake further experiments to try to understand to what extent the strength of the findings reported here are due to intrinsic strength of the conversion algorithm adopted here or to the intrinsic quality of the semantic networks used, or just of a bit of both factors and of their combination.

6 Conclusions

In a previous study in the literature (Saedi et al., 2018), a conversion method (wnet2vec) was explored to obtain a semantic space (aka word embeddings) from a semantic graph, by applying it to the English Princeton WordNet. The WordNet embeddings wnet2vec thus generated, on the basis of 60k synsets, outperforms a strong baseline which is a corpus-based word embedding word2vec,

based on 100B words. It outperforms in the semantic similarity detection task over the mainstream SimLex-999 test dataset, with an accuracy score of 0.50 against 0.44 in terms of Spearman's coefficient, for wnet2vec and word2vec respectively .

In the present paper, we experimented with this conversion method under further empirical conditions. We applied it over a WordNet manually built under the same construction principles as Princeton WordNet (over 120k synsets) only that it is more than seven times smaller (17k synsets) and is for another language, namely Portuguese. We experimented also with another WordNet for Portuguese but constructed under an alternative approach that resorts to (semi-) automatic methods.

The WordNet embeddings obtained were tested under the semantic similarity task over the Portuguese translation of the mainstream SimLex-999 test dataset (Querido et al., 2017). The baseline was the word embeddings obtained with the corpus-based word2vec procedure over a 2.2B words corpus of Portuguese (Rodrigues and Branco, 2018).

The results obtained are in line with earlier findings. The wnet2vec conversion method to obtain a semantic space from a semantic network is very effective.

The semantic similarity detectors based on word embeddings wnet2vec — obtained from any of the WordNets experimented with in this paper — outperform the strong baseline detector based on the corpus-based word embeddings word2vec.

The semantic similarity detector based on the manually built WordNet, in turn, outperformed the detector based on the WordNet that was built (semi-) automatically. These results suggest that, when using the wnet2vec methodology to obtain a semantic space from a semantic graph, under comparable experimental circumstances, better semantic similarity performance scores are obtained with a manually crafted WordNet rather than with a WordNet obtained (semi-) automatically. This is as expected given the noise introduced by automatic methods. What is new with respect to the assessment methodology proposed here is that it offers a new quantitative way to evaluate the difference between WordNets in what concerns their different quality at capturing the meaning of words.

7 Acknowledgments

The present research was partly supported by the Infrastructure for the Science and Technology of Language (PORTULAN CLARIN) by the National Infrastructure for Distributed Computing (INCD) of Portugal, and by the ANI/3279/2016 grant.

References

[Bobrow and Norman1975] Daniel G. Bobrow and Donald Arthur Norman. 1975. Some principles of memory schemata. In *Representation and Understanding: Studies in Cognitive Science*, page 131–149. Elsevier.

[De Deyne et al.2013] Simon De Deyne, Daniel J Navarro, and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, 45(2):480–498.

[de Paiva et al.2012] Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India, December. The COLING 2012 Organizing Committee. Published also as Techreport http://hdl.handle.net/10438/10274.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

[Goikoetxea et al.2015] Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies Conference (NAACL-HLT25)*, pages 1434–1439. Association for Computational Linguistics.

[Gonçalo Oliveira2018] Hugo Gonçalo Oliveira. 2018. Distributional and knowledge-based approaches for computing portuguese word similarity. *Information*, 9(2):35.

[Harris1954] Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

[Hill et al.2016] Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Googlenews-vectors-negative300.bin.gz - efficient estimation of word representations in vector space. *arXiv preprint*

arXiv:1301.3781. https://code.google.com/archive/p/word2vec/.

[Minsky1975] Marvin Minsky. 1975. A framework for representing knowledge. In *Psychology of Computer Vision*. McGraw-Hill.

[Osgood et al.1957] Charles E Osgood, George J Suci, and Percy H Tannenbaum. 1957. The measurement of meaning. *Urbana: University of Illinois Press*.

[Querido et al.2017] Andreia Querido, Rita de Carvalho, João Rodrigues, Marcos Garcia, Catarina Correia, Nuno Rendeiro, Rita Valadas Pereira, Marisa Campos, João Silva, and António Branco. 2017. LX-LR4DistSemEval: A collection of language resources for the evaluation of distributional semantic models of Portuguese. *Revista da Associação Portuguesa de Linguística*, 3.

[Quillan1966] M Ross Quillan. 1966. Semantic memory. Technical report, Bolt Beranek and Newman Inc., Cambridge MA.

[Rodrigues and Branco2018] João Rodrigues and António Branco. 2018. Finely tuned, 2billion token based word embeddings for portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

[Saedi et al.2018] Chakaveh Saedi, António Branco, João António Rodrigues, and João Silva. 2018. Wordnet embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 122–131. Association for Computational Linguistics.

[Vossen et al.2013] Piek Vossen, Claudia Soria, and Monica Monachini. 2013. Wordnet-lmf: A standard representation for multilingual wordnets. *LMF Lexical Markup Framework*, pages 51–66.

Spoken WordNet

Kishore Kashyap
Department of Information Technology
Gauhati University, India
kb.guwahati@gmail.com

Shikhar Kr Sarma
Department of Information Technology
Gauhati University, India
sks001@gmail.com

Kumari Sweta
Department of Information Technology
Gauhati University, India
swetagupta647@gmail.com

Abstract

WordNets have been used in a wide variety of applications, including in design and development of intelligent and human assisting systems. Although WordNet was initially developed as an online lexical database, (Miller, 1995 and Fellbaum, 1998) later developments have inspired using WordNet database as resources in NLP applications, Language Technology developments, and as sources of structured learned materials. This paper proposes, conceptualizes, designs, and develops a voice enabled information retrieval system, facilitating WordNet knowledge presentation in a spoken format, based on a spoken query. In practice, the work converts the WordNet resource into a structured voiced based knowledge extraction system, where a spoken query is processed in a pipeline, and then extracting the relevant WordNet resources, structuring through another process pipeline, and then presented in spoken format. Thus the system facilitates a speech interface to the existing WordNet and we named the system as "Spoken WordNet". The system interacts with two interfaces, one designed and developed for Web, and the other as an App interface for smartphone. This is also a kind of restructuring the WordNet as a friendly version for visually challenged users. User can input query string in the form of spoken English sentence or word. Jaccard Similarity is calculated between the input sentence and the synset definitions. The one with highest similarity score is taken as the synset of interest among multiple available synsets. User is also prompted to choose a contextual synset, in case of ambiguities.

1. Introduction

WordNets have become resources for many NLP applications, language technology developments, as well as a knowledge database (Morato, 2004). This is different from any other form of information storage as in WordNet words are stored in a way where different word forms (sysnsets) semantically linked. Many semantic relations are embedded in WordNet, and lexical units are defined and described with examples, concepts, and synonyms. Thus making the database a resourceful lexico-semantic knowledge base.

In this paper. we have developed a new voice based system integrating a new system pipeline for processing spoken query, and presenting retrieved knowledge also in spoken format. This has enabled usage of the textual lexico-semantic knowledge base as a spoken knowledge base, thus creating the new concept of Spoken WordNet. The intelligent spoken language interfacing technique is already been explored (Inagaki, 2013). The relevance of voiced enabled interface for information extraction has become more intense in recent years, and user base of voice interfaces is growing. Also, for visually challenged persons, voice based interface, and spoken knowledge presentation signifies a lot. For our current work, we used the Princeton English WordNet.

Query in the form of short sentence, or discrete word is inputted through the voice based interface, either the Web version, or the App version. For experimental and demonstration purpose, we limited the scope of work to Nouns only. Spoken query is processed for extracting/ identifying words. Then it undergoes a series of computational steps, ultimately defining the token of interest in the form of text. This token of

interest is then subjected to the retrieval process into the main WordNet database and required information is identified and extracted from WordNet. At this moment, we also mine to depth 1 of hypernym, extracting upper layer knowledge. This is now put in another process pipeline for structuring the presentable knowledge. Predefined format is used to embed the extracted segments, formatting new sentence level presentations. And ultimately such structured and formatted textual sentences are presented through the interface in spoken format.

2. The Algorithm and the Core Engine

Algorithm 1: Algorithm

Input: sentence **S**
Output: Word description and hypernym of depth 1

```
[S_t] = Tokenized S
[W] = POS tagged words of [S_t]
retrieve W_i from wordnet database where W_i ∈ [W], W_i = Noun
begin do ∀W_i ∈ [W]
        if N = 0, where N is number of synsets found for W_i
            speak "No synset found for W_i"
            goto step 4
        else
            if 1 < N
                get highest Jaccard Similarity from N W_i definition and S
                if highest Jaccard Similarity found
                    speak W_i definition, W_i hypernym of depth 1
                    goto step 4
                else
                        list multiple synsets and prompt user to select one synset
                        get selected synset and goto step 12
            else
                goto step 12
end do
```

FIGURE 1

In this section we will discuss the overall system model. The speech signal is taken as input from the user and send over the Internet in real time to convert it into text using open source Speech-To-Text. Upon receiving the text data, the core engine of Spoken WordNet works as the algorithm (figure 1) we have developed.

The Jaccard similarity is calculated as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

(If A and B are both empty, we define $J(A,B) = 1$.)

$$0 \leq J(A,B) \leq 1.$$

where, A is the set of words in the input sentence and B is the set of words in the synset definition. Both A and B are free of English stop words.

Here, Jaccard Similarity is used for word sense disambiguation (WSD).

The core engine is entirely written in Python3 and NLTK (Bird, 2004) is used for WordNet information retrieval. For POS tagging TextBlob (https://textblob.readthedocs.io/en/dev/) package is used. English stop words are removed using NLTK.

3. User Interface Design and Implementation

It is essentially very important in today's technological use case scenario that a great deal of usage of any successful software product should include a good and user friendly UI. Now, users are more Internet centric as compared to users a decade ago. So, the authors decided to develop two important areas of Human Computer Interaction UI. Namely Smartphone App and Web Interface.

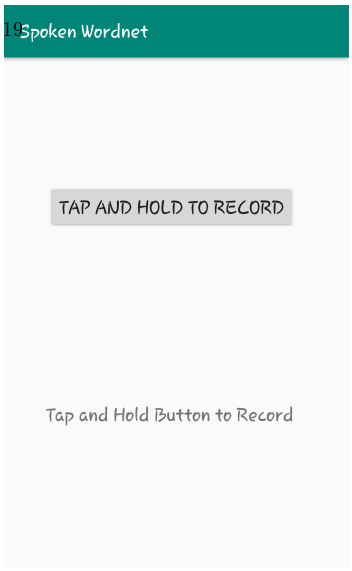


FIGURE 2

3.1. Smartphone App

People are becoming more interested in App based information retrieval, because of user friendliness and readymade linking to the host servers, and cloud based services. But spoken interfaces are very limited and there are no substantial evidence of such an existing system. Although access to WordNets are free, and many APIs including APIs for smartphones have been developed, but we have not seen any smartphone App for voiced based connection to WordNet. As WordNet contain knowledge, and also the WordNets are scalable, an App based Voice Interface with linking and structured retrieving of knowledge will facilitate a wide range of users. This is more user friendly, as query is in the form of voice prompts, and also significantly important to visually challenged persons. App is designed and developed using Android Studio and Firebase. Speech signal from user is taped using the Smartphone microphone, and then the App module temporarily stores it and then sends to the Firebase cloud server. The voice segment is stored as a .wav file in the Firebase server. This file is used for processing in our core system pipeline. The structured and formatted output from the core processing is then send back to the server, and the App accesses the server, takes the file, and reproduces the file into spoken format. Thus, for user, it's a voice-input-voice-output system, creating the App for Spoken WordNet.

3.2. Web Interface Design and Implementation

We have developed a Web Interface for the system which currently runs in the Google chrome browser. User can click on the main interface and then start speaking. The Speech To Text conversion and Text To Speech Conversion is done using the JavaScript Web Speech API.

Flask is used for web server creation. Flask is a micro framework for python web development

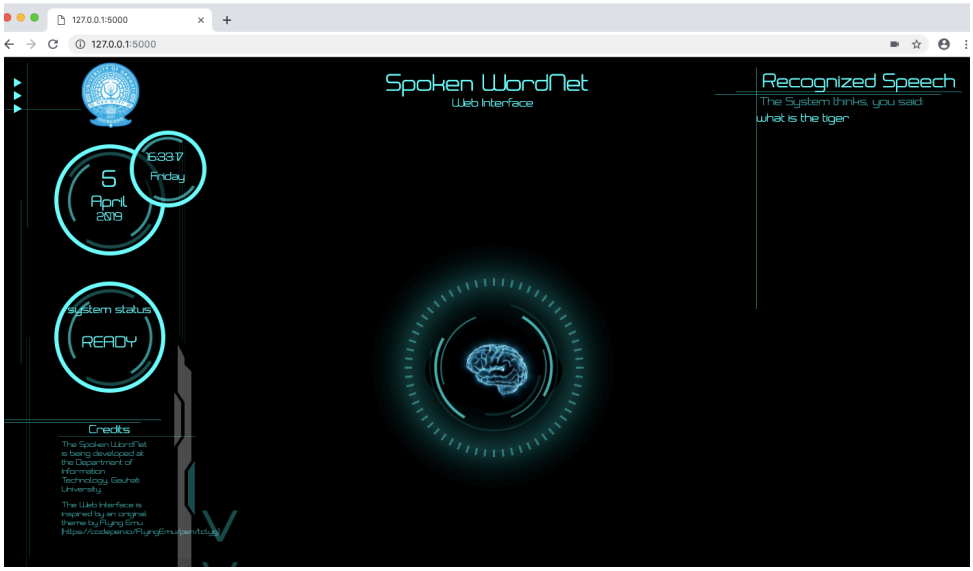


FIGURE 3

(Ronacher http://mitsuhiko.pocoo.org/flask-pycon-2011.pdf).
For message passing to and from the web interface to the core engine is done through SocketIO. Socket.IO enables real-time, bidirectional and event-based communication.

4. Result

The system is working well for nouns as we have limited our concentration to the nouns POS. We have also experiments with the Lesk WSD algorithm available in NLTK. But, Jaccard Similarity is found to be working well for nouns and having more than one words in the query sentence, e.g., knowing what words are to be given as input for extracting the correct sense of the word.

5. Future Work

We have presented a system of voiced based interface for the WordNet in both Smartphone App and Web environment. In this prototyping work, we have considered only nouns for their semantic extraction from WordNet. In future, this can be extended for other parts of speech. Furthermore, it has scope for experimenting with other sentence similarity measures for WSD. Exploring the word expansion for simplifying spoken text may be another future direction of research through this system.

References

Armin Ronacher, Opening the Falsk, URL:http://mitsuhiko.pocoo.org/flask-pycon-2011.pdf

Christiane Fellbaum 1998, ed. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

George A. Miller 1995, WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

Hirohito Inagaki, Takaaki Hasegawa, Satoshi Takahashi, and Yoshihiro Matsuo 2013, Toward Intelligent Spoken Language Interface Technology, NTT Technical Review, Vol. 11 No. 7

Jorge Morato, Miguel Angel Marzal, Juan Lloréns, and José Moreiro, 2004, WordNet Applications, Proceedings of GWC2004, pp. 270–278.

Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (ACLdemo '04). Association for Computational Linguistics, Stroudsburg, PA, USA, , Article 31 . DOI=http://dx.doi.org/10.3115/1219044.1219075

OntoLex-Lemon as a Possible Bridge between WordNets and Full Lexical Descriptions

Thierry Declerck
DFKI GmbH
Saarbrücken, Germany
& ACDH-OEAW
Vienna, Austria
declerck@dfki.de

Melanie Siegel
Hochschule Darmstadt
University of Applied Sciences
Darmstadt, Germany
melanie.siegel@h-da.de

Dagmar Gromann
University of Vienna
Vienna, Austria
dagmar.gromann@univie.ac.at

Abstract

In this paper we describe our current work on representing a recently created German lexical semantics resource in OntoLex-Lemon and in conformance with WordNet specifications. Besides presenting the representation effort, we show the utilization of OntoLex-Lemon to bridge from WordNet-like resources to full lexical descriptions and extend the coverage of WordNets to other types of lexical data, such as decomposition results, exemplified for German data, and inflectional phenomena, here outlined for English data.

1 Introduction

We aim at publishing German WordNet conformant data in the Linguistic Linked Open Data (LLOD) cloud.[1] We selected the OntoLex-Lemon model (Cimiano et al., 2016), a successor and World Wide Web Consortium (W3C) standardization of the *lemon* model (McCrae et al., 2012b), in order to represent different kinds of lexical semantics data, since it has established itself as the defacto community standard for representing lexical data in the Linked Data framework. Guidelines for mapping Global WordNet formats to a *lemon*-based Resource Description Framework (RDF) representation have been published[2] and already some WordNets have been mapped to *lemon*, as described for example in (McCrae et al., 2014).

A candidate for representing German lexical semantics data in OntoLex-Lemon is GermaNet, which is a manually designed WordNet resource for German (Hamp et al., 1997). Developed more than 20 years ago, it represents a very stable, well-tested, and precise lexical semantics resource. However, its access is restricted by its current license. Without such open data access, reuse of GermaNet in global initiatives, such as Open Multilingual WordNet (OMW) (Bond and Paik, 2012), is inhibited. One of the objectives of this paper is to represent lexical semantics data openly linked to other OMW datasets in the LLOD.

Two alternatives compliant with WordNet specifications and available under an open-source license are the *lemonUby* set of resources (Eckle-Kohler et al., 2015) and the Open-de-WordNet (OdeNet) effort.[3] *lemonUby* is an export of lexical data from the large-scale linked UBY (Gurevych et al., 2012)[4], which unites collaboratively and expert-developed resources (e.g. FrameNet and Wiktionary) in English and German, to *lemon*. *lemonUby* contains the German version of OmegaWiki[5], which encodes WordNet compliant descriptions of German words. OdeNet provides a German resource to the OMW initiative.

A mapping from *lemonUby* to OntoLex-Lemon can be expected to be straightforward due to a high compliance between both models. Thus, this publication concentrates on mapping the WordNet-compliant XML code of OdeNet to OntoLex-Lemon, while in the long run a cross-linking or, where possible, a merging of *lemonUby* and OdeNet in the LLOD is foreseen. We exemplify the richness of lexical descriptions offered by OntoLex-Lemon with the case of components of compounds, German in this submission, and inflectional morphological variations, here in the case of sense variations across English nominal plural inflections.

[1] See http://linguistic-lod.org/ and (Chiarcos et al., 2012), which describes the first instantiation of the LLOD, while (McCrae et al., 2016) details the further developments of LLOD.

[2] https://globalwordnet.github.io/schemas/#rdf

[3] https://github.com/hdaSprachtechnologie/odenet

[4] http://www.ukp.tu-darmstadt.de/uby/

[5] https://lemon-model.net/lexica/uby/ow_deu/

In the following sections we first describe OdeNet, before presenting the main characteristics of OntoLex-Lemon. In Section 4 we present the current state of the mapping from OdeNet to OntoLex-Lemon, before finally discussing the potential added-value of having WordNets represented in OntoLex-Lemon.

2 OdeNet

OdeNet combines two existing resources: The OpenThesaurus German synonym lexicon[6] and the Open Multilingual WordNet (OMW)[7] (Bond and Foster, 2013). In terms of English resources, it includes the Princeton WordNet of English (PWN) (Fellbaum, 1998). Integrating OpenThesaurus in OdeNet means making use of a large resource for German that is generated and updated by the crowd. A consequence of this approach is that OdeNet needs to be curated, as the authors of the resource mention.

We downloaded the most recent version[8] and first analyzed its content. OdeNet is in an XML format and shares its Document Type Definition (DTD)[9] with other WordNets in the OMW initiative. Lexical entries provide information on different senses of a lexeme, such as "Kernspaltung" or "Kernfission" (*nuclear fission*) in the same synset:

```
<LexicalEntry id="w1">
    <Lemma writtenForm="Kernspaltung" partOfSpeech="n"/>
    <Sense id="w1_1-n" synset="odenet-1-n"/>
</LexicalEntry>
<LexicalEntry id="w2">
    <Lemma writtenForm="Kernfission" partOfSpeech="n"/>
    <Sense id="w2_1-n" synset="odenet-1-n"/>
</LexicalEntry>
```

Lexical senses are grouped to synsets, i.e., groups of word senses with the same meaning. Hierarchical relations are introduced as synset relations, such as here a hypernymy relation:

```
<Synset id="odenet-1-n" ili="i107577"
  partOfSpeech="n" dc:description="a nuclear reaction in
  which a massive nucleus splits into smaller nuclei with
  the simultaneous release of energy">
  <SynsetRelation target='odenet-5437-n'    relType='hypernym'/>
</Synset>
```

Another example is the entry for "Stuhl" (*chair*):

```
<LexicalEntry id="w224" confidenceScore="1.0">
  <Lemma writtenForm="Stuhl" partOfSpeech="n"/>
  <Sense id="w224_49-n" synset="odenet-49-n"/>
  <Sense id="w224_1172-n" synset="odenet-1172-n"/>
</LexicalEntry>

<Synset id="odenet-49-n" ili="i51746"
  partOfSpeech="n" confidenceScore="1.0">
 <Definition>
   Eine Sitzgelegenheit für eine Person, mit einer Lehne
   im Rücken.
</Definition>
  <SynsetRelation target='odenet-11251-n' relType='hypernym'/>
  <SynsetRelation target='odenet-8518-n' relType='hyponym'/>
  <SynsetRelation target='odenet-20127-n' relType='hyponym'/>
  <SynsetRelation target='odenet-34983-n' relType='hyponym'/>
  <Example>
    Sie sitzt auf dem Stuhl.
  </Example>
</Synset>
```

Access to the lemma information for hypernyms and hyponyms is also possible, for instance for the `odenet-49-n` synset for "Stuhl" it would be:

```
>>> hypernyms("odenet-49-n")
  odenet-11251-n:
  ['Sitz', 'Platz', 'Sitzplatz', 'Sitzgelegenheit']

>>> hyponyms("odenet-49-n")
  odenet-8518-n:
  ['Rolli', 'Krankenfahrstuhl', 'Rollstuhl']),
  odenet-20127-n:
  ['Lehnsessel', 'Fauteuil']),
  odenet-34983-n:
  ['Lehnstuhl', 'Bergère', 'Sessel',
  'Polsterstuhl', 'Polstersessel'])]
```

3 OntoLex-Lemon

The OntoLex-Lemon model was originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the description of ontology elements are equipped with an extensive linguistic description (McCrae et al., 2012a; Cimiano et al., 2016). This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e., the meaning of these lexical entries with respect to an ontology or to specialized vocabularies. The main organizing unit for those linguistic descriptions is the lexical entry, which enables the representation of morphological patterns for each word and/or affix. The connection of a lexical entry to an ontological entity is marked mainly by the `denotes` property or is mediated by the `Lexical Sense` or the `Lexical Concept` properties, as this is represented in Figure 1, which displays the core module of the model.

OntoLex-Lemon, as well as its predecessor *lemon*, have also been deployed for the representation of WordNets, as described for example in (McCrae et al., 2014) and guidelines are available for mapping WordNets to an RDF code compliant to OntoLex-Lemon.[10] A main difference between *lemon* and OntoLex-Lemon is that the latter model includes an explicit way to encode conceptual hierarchies, using the SKOS standard.[11]

[6] `https://www.openthesaurus.de/`

[7] `http://compling.hss.ntu.edu.sg/omw/`

[8] `https://github.com/hdaSprachtechnologie/odenet`

[9] `https://github.com/globalwordnet/schemas/blob/master/WN-LMF.dtd`

[10] `https://globalwordnet.github.io/schemas/#rdf`

[11] SKOS stands for "Simple Knowledge Organization System". SKOS provides "a model for expressing the basic struc-

As can be seen in Figure 1, lexical entries (lemmas) can be linked, via the `ontolex:evokes` property, to such SKOS concepts, which can represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the `ontolex:reference` property or mediated by the instances of the `ontolex:LexicalSense` class.[12] The `ontolex:LexicalConcept` class seems to be best appropriated to model the "sets of cognitive synonyms (synsets)"[13] that (PWN describes, while the `ontolex:LexicalSense` class is meant to represent the bridge between lexical entries and ontological entities (which do not necessarily have semantic relations between them).

More recently the OntoLex-Lemon model has been more and more considered also for modeling lexical data as such, in the context of projects and studies related to the development of digital lexicography, like for example in the past COST action "ENeL" (European Network of e-Lexicography).[14] This development towards a more generic representation model for lexicographic purposes is documented among others in (McCrae et al., 2017).

4 Mapping OdeNet to OntoLex-Lemon

One main issue that occurred due to partly crowdsourced data in OdeNet was that additional textual information or special characters were added by the crowd to the headwords. A second issue was the improper use of Part-of-Speech (PoS) tags if word classes were different from noun, verb, or adjective or could not be clearly assigned to one of these. These entries are marked with PoS "p", which we filter and link to well-established German lexical data in the LLOD cloud in order to extract the correct PoS information. To clean the data, we wrote a Python script, which not only filters out noisy data, but also maps certain GWN codes (like PoS) to the vocabularies used in OntoLex-Lemon, for example the LexInfo vocabulary for PoS and semantic relations.[15]

As for now, we have an OntoLex-Lemon encoding of OdeNet 120,012 lexical entries, the same number of lexical senses and 36,192 synsets, which are encoded as `ontolex:LexicalConcepts` and included in a SKOS-based conceptual hierarchy, supporting also the description of lexical semantic relations between synsets, like synonymy, hyponomy etc.

The following listings provide details on the OntoLex-Lemon encoding of the first OdeNet entry, which is "Kernspaltung" (*nuclear fission*).

Listing 1: The lexical entry for *Kernspaltung*

```
:entry_w1
  rdf:type ontolex:MultiWordExpression ;
  decomp:constituent :Kern_comp,
                     :spaltung_comp ;
  rdf:_1 :Kern_comp ;
  decomp:subterm :entry_w3542 ;
  rdf:_2 :spaltung_comp ;
  decomp:subterm :entry_w23527 ;
  lexinfo:hypernym :synset_odenet-5437-n ;
  wn:partOfSpeech wn:noun ;
  ontolex:canonicalForm :form_w1 ;
  ontolex:sense :sense_w1_1-n ;
  ontolex:evokes :synset_odenet-1-n ;
.
```

In Listing 1 we display the full OntoLex-Lemon entry, which allows us to represent the components of compound words by encoding information as a `ontolex:MultiWordExpression` instance. This class marks any type of entries that can be segmented, thus, including compounds. The term "Kernspaltung" is associated with its two components "Kern" and "Spaltung". Each component represents a full lexical entry with all of its semantic relations. Reuse of components across OntoLex-Lemon entries reveals relations between different instances of `ontolex:MultiWordExpression` based on their component entries. This possibility demonstrates one of the added-values of linking synsets to the (complex) representation of lexical entries, as we can state (see below) semantic relations be-

ture and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary" (https://www.w3.org/TR/skos-primer/)

[12] Quoting from Section 3.6 "Lexical Concept" `https://www.w3.org/2016/05/ontolex/`: "We [...] capture the fact that a certain lexical entry can be used to denote a certain ontological predicate. We capture this by saying that the lexical entry denotes the class or ontology element in question. However, sometimes we would like to express the fact that a certain lexical entry evokes a certain mental concept rather than that it refers to a class with a formal interpretation in some model. Thus, in lemon we introduce the class Lexical Concept that represents a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses. A lexical concept is thus a subclass of skos:Concept."

[13] Quoted from `https://wordnet.princeton.edu/`.

[14] `https://www.cost.eu/actions/IS1305/#tabs|Name:overview`

[15] See `https://www.lexinfo.net/ontology/2.0/lexinfo` and also (Cimiano et al., 2011).

tween synsets associated to the components of a compound word and its synsets.

Listing 2 below displays the form information associated with entry `:entry_w1` in Listing 1.

Listing 2: The ontolex:Form "Kernspaltung"

```
:form_w1
  rdf:type ontolex:Form ;
  ontolex:writtenRep "Kernspaltung"@de ;
.
```

Listing 3 shows the conversion of the original OdeNet sense information to an instance of the `ontolex:LexicalSense` class.

Listing 3: The lexicalSense associated to the entry for "Kernspaltung"

```
:sense_w1_1-n
  rdf:type ontolex:LexicalSense ;
  ontolex:isLexicalizedSenseOf
    :synset_odenet-1-n ;
  ontolex:isSenseOf :entry_w1 ;
  ontolex:reference
    https://www.wikidata.org/wiki/Q11429 ;
.
```

A sense can be linked to a synset via the property `ontolex:isLexicalizedSenseOf`, which relates a lexical sense to that lexical concept it lexicalizes, here a synset. The entry can be linked to the synset via the property `ontolex:evokes`, as displayed in Listing 1, which is defined as relating a lexical entry to one of the abstract lexical concepts that a speaker of the language would associate with the words in the lexical entry. In contrast to evokes that links to a lexical concept, `ontolex:reference` links to an ontological concept that represents a denotation of the lexical entry, here in the form of a Wikidata entry.

Listing 4 displays the representation of the synset associated with both the lexical entry `entry_w1` and the `sense_w1_1-n`. There we can also see that this lexical concept (synset) is also "evoked" by other entries/senses. For example by the entries for "Kernfission" or "Atomspaltung", which are synonyms of "Kernspaltung". The `lexinfo:hypernym` property provides information on the semantic relation this synset has to another synset.

Listing 4: The `LexicalConcept` (synset) associated with the entry for "Kernspaltung"

```
:synset_odenet-1-n
  rdf:type ontolex:LexicalConcept ;
  skos:inScheme :ODEnet ;
  skos:definition "a nuclear reaction
  in which a massive nucleus splits
  into smaller nuclei with the
  simultaneous release of energy" ;
  wn:ili ili:i107577 ;
  ontolex:isEvokedBy :entry_w1 ;
  ontolex:isEvokedBy :entry_w2 ;
  ontolex:isEvokedBy :entry_w3 ;
  ontolex:isEvokedBy :entry_w4 ;
  ontolex:lexicalizedSense :sense_w1_1-n ;
  ontolex:lexicalizedSense :sense_w2_1-n ;
  ontolex:lexicalizedSense :sense_w3_1-n ;
  ontolex:lexicalizedSense :sense_w4_1-n ;
  lexinfo:hypernym :synset_odenet-5437-n ;
.
```

Finally, in Listing 5 we display the entries for the components of the compound word "Kernspaltung". Those components point to the lexical entries they are related to (the entry `:entry_w23527` is for example the one corresponding to the noun "Spaltung" (*split*, *fission*, *separation*, *cleavage*, etc.), which has again its own senses and associated synsets. We can here disambiguate the meaning of "Spaltung" as used in the compound, as being the one of "fission". And the whole compound can then be considered as a hyponym of the synset for "fission".

Listing 5: The two components of entry for "Kernspaltung"

```
:Kern_comp
  rdf:type decomp:Component ;
  decomp:correspondsTo :entry_w3542 ;
.
:spaltung_comp
  rdf:type decomp:Component ;
  decomp:correspondsTo :entry_w23527 ;
.
```

In Listing 1, we can see the information on the sequence those components have in this entry. For sure, those component entries can be re-used separately for other compounds, such as for "Atomspaltung". Thereby, we can collect all the corresponding meanings of a word, also when they are used in compounds and in dependency on their relative position in the compounds. A detailed representation of the decomposition module of OntoLex-Lemon is shown in Figure 2.

In this section we described the current state of the OntoLex-Lemon representation of filtered or cleaned data we can find in OdeNet. Furthermore, we touched upon the possible use of OntoLex-Lemon as a bridge between WordNet-like resources and full lexical descriptions, here exemplified with the case of German compound nouns. In the next section we address the issue on representing sense variants in dependency of the singular or plural inflection of an entry.

5 Added-Values of the Use of Lemon-OntoLex for Representing WordNets

As stated in the preceding section, we see the use of OntoLex-Lemon for representing WordNets as a chance to not only port information from one format to another, but also as an opportunity to extend the coverage of WordNet descriptions to more complex lexical phenomena, beyond lemma and PoS considerations. One case we have been investigating concerns the different synsets that are attributed in PWN to the singular and to the plural forms of one word.

When searching for a word in the PWN interface[16], all potential synsets for this word are returned. While it is possible to actively search for plural forms of a noun, in a vast majority of cases the interface returns results for its uninflected counterpart because it lemmatizes the queried word. In cases of complementary plural entries, WordNet displays augmented lists of synsets: those associated with the singular, e.g. *people*, and those associated with the plural, e.g. *peoples*. All senses for this example are displayed in Listing 6.

Listing 6: The Synsets for "people" vs. "peoples"

```
people.n.01
  ((plural) any group of human
  beings ... collectively)
  citizenry.n.01
   (the body of citizens of a state
   or country)
people.n.03
  (members of a family line)
  multitude.n.03
   (the common people generally)
peoples.n.01
  (the human beings of a particular
  nation or community or ethnic group)
```

This differentiation of grammatical number in the representation of synsets and associated meanings intuitively suggests that plural and singular forms do not share all meanings. Regular cases, such as *car* returns no additional synsets and senses for its inflected form *cars*. Thus, it can be assumed that the change of grammatical number does not cause any sense variant in those cases. This means, in turn, that it can be assumed that the availability of additional senses indicates that semantic differences exist between the inflectional forms.

[16] http://wordnetweb.princeton.edu/perl/webwn

We also observe that querying a plural in WordNet always results in the listing of all singular senses of a word and, where available, senses specific to the plural. However, this rigorous listing of singular senses also applies to plural nouns that share no sense with their singular counterpart. For instance, querying the pants *khakis* would result in a listing of all senses related to *khaki* and that of the plural. In case a sense exists only for a plural form, it would be desirable for the system to return only the corresponding synset.

Mixed cases exist for this phenomena, the ones where singular and plural share senses and those where senses are specific the singular or plural form. We showcase this behavior with the word pair *letter-letters*. While several senses can be associated with both the singular and the plural form of the lexical entry *letter*, the literary culture sense can be associated only with the plural form. On the other hand, the sense of literal interpretation (e.g. in the case of law texts that are interpreted by the *letter*) is generally assigned to the singular form. In the following listings, we show, in a simplified manner, the way this complex information can be encoded in OntoLex-Lemon.

Listing 7 displays the lexical entry for *letter*. It is stated that two forms are associated with this noun: a singular (the `canonicalForm`) and a plural (the `otherForm`) form. In this simplified entry, we link only to one sense: the one of an exchange between two parties (see Listing 10).

Listing 7: The lexical entry for *letter*

```
:letter
  rdf:type ontolex:Word ;
  lexinfo:partOfSpeech
                  lexinfo:noun ;
  ontolex:canonicalForm
                  :Form_letter ;
  ontolex:otherForm
                  :Form_letters ;
  ontolex:sense
          :LexicalSense_letter_1 ;
.
```

Listings 8 and 9 display the basic encoding for the two possible word forms for the entry *letter*, the singular and the plural forms.

Listing 8: The form for *letter* in singular

```
:Form_letter
  rdf:type ontolex:Form ;
  lexinfo:number
              lexinfo:singular ;
  ontolex:writtenRep
                  "letter"@en ;
.
```

Listing 9: The form for *letters* in plural

```
:Form_letters
  rdf:type ontolex:Form ;
  lexinfo:number
              lexinfo:plural ;
  ontolex:writtenRep
                "letters"@en ;
.
```

The next listing is about the shared sense associated with the lexical entry. As there is a Wikidata entry for the type of entity this sense can refer to, we make use of the `ontolex:reference` property in order to link to this data source.

Listing 10: The lexical sense for the entry *letter* (which can have singular and plural forms)

```
:LexicalSense_letter_1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "letter as a missive from
       one party to another (taken from
      Wikidata)" ;
  ontolex:isSenseOf :letter ;
  ontolex:reference <https://www.
      wikidata.org/wiki/Q133492> ;
.
```

Listing 11 introduces the additional lexical entry for the plural form of *letter* that has a specific meaning that cannot be associated with its singular form. Therefore we link this entry only to the plural instance of the class `Form` and to the specific sense encoded in Listing 12, where we additionally formulate the constraint that the usage of this sense is restricted to the plural form *letters*.

Listing 11: The special lexical entry for *letters*

```
:letters
  rdf:type ontolex:Word ;
  lexinfo:partOfSpeech
                    lexinfo:noun ;
  rdfs:comment "encoding singular
             and plural entries" ;
  ontolex:canonicalForm
                   :Form_letters ;
  ontolex:sense
          :LexicalSense_letters_1 ;
.
```

Listing 12: The sense for *letters* in plural

```
:LexicalSense_letters_1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "letters"
       as "literary culture" ;
  ontolex:usage :Form_letters ;
.
```

In fact the use of the `ontolex:usage` property could suffice in order to mark that a sense is restricted to a particular inflectional form of an entry, as exemplified below in Listing 13 for the sense of the literal interpretation, without the need to introduce a new lexical entry.

Listing 13: The literal interpretation sense for *letter* in singular

```
:LexicalSense_letter_2
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "letter"
       as "strictly literal
           interpretation" ;
  ontolex:usage :Form_letter ;
.
```

OntoLex-Lemon in this case seems to be able to provide for a representation that would support morpho-semantic phenomena. As part of our future work, a possibility to associate senses to forms as well as lexical entries in the OntoLex-Lemon model is investigated.

6 Conclusion

We described our current work consisting in porting a recently developed German WordNet compliant lexical resource, OdeNet, to OntoLex-Lemon, in order to support its publication in the Linguistic Linked Open Data cloud. While processing those data, we noticed that OntoLex-Lemon can be used for bridging the WordNet type of lexical resources to a full description of lexical entries, leading possibly to an extension of the coverage of WordNets beyond the consideration of lemmas and PoS information. We documented this with the example of the representation of components of German compounds and the distinct senses that can exist between certain singular and plural forms of English words.

In terms of future work, other types of full lexical descriptions will be modeled in OntoLex-Lemon and associated with the presented resources. Furthermore, this type of modeling allows for cross-linking to other German WordNets in the LLOD, such as *lemonUby*. This cross-linking effort intends to finally interlink multilingual WordNets in a Linked Data-based format and its rich potential for full lexical descriptions.

Acknowledgments

Contributions by Thierry Declerck have been supported in part by by the H2020 project “Prêt-à-LLOD” with Grant Agreement number 825182 and by the H2020 project “ELEXIS” with Grant Agreement number 731015.

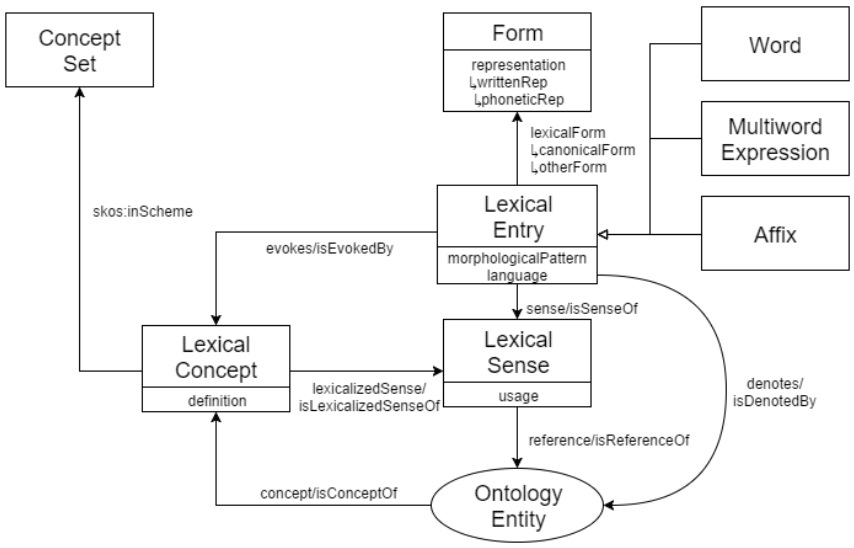


Figure 1: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

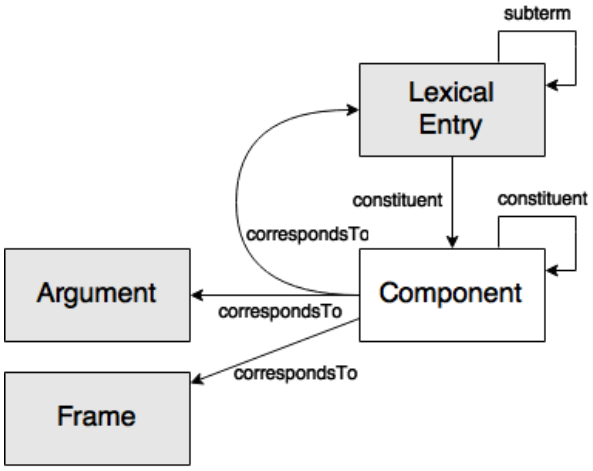


Figure 2: The Decomposition module of OntoLex-Lemon. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362, Sofia.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. *Small*, 8(4):5.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012. *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*. Springer.

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. Lexinfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51, 2.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report.

Judith Eckle-Kohler, John Philip McCrae, and Christian Chiarcos. 2015. lemonuby - a large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web Journal*, 6(4):371–378.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - a large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France, April. Association for Computational Linguistics.

Birgit Hamp, Helmut Feldweg, et al. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asuncion Gomez-Perez, Jorge Garcia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012a. Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, D ennis Spohr, and Tobias Wunner. 2012b. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.

John P. McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3 rd Workshop on Linked Data in Linguistics*.

John P. McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. 2016. The open linguistics working group: Developing the linguistic linked open data cloud. In *The 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Slovenia, Portorož*.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*, pages 587–597.

Semi-automatic Annotation of Event Structure, Argument Structure, and Opposition Structure to WordNet by Using Event Structure Frame

Seohyun Im
Automation and System Research Institute, Seoul National University
Seoul, South Korea
Seohyunim71@gmail.com

Abstract

In this paper, we present semi-automatic annotation of the Event Structure Frames to synsets of English verbs in WordNet. The Event Structure Frame is a sub-eventual structure frame which combines event structure (lexical aspect) with argument structure represented by semantic roles and opposition structure which represents the presupposed and entailed sub-events of a matrix event. Our annotation work is done semi-automatically by GESL-based automatic annotation and manual error-correction. GESL is an automatic annotation tool of the Event Structure Frame to verbs in a sentence. We apply GESL to the example sentence given for each synset of a verb in WordNet. We expect that our work will make WordNet much more useful for any NLP and its applications which require lexical semantic information of English verbs.

1 Introduction

This paper aims to present our work of linking the Event Structure Frame (henceforth, ESF) to WordNet to improve its usability for NLP applications such as multimodal (and textual) inference tasks which require the lexical semantic information of words.

WordNet represents the distinct senses of verbs very delicately and organizes the semantic relations such as synonymy and hypernymy of the verbs. The semantic relation is one of the major strengths of WordNet. However, WordNet lacks the following two factors which consist of the lexical meaning of verbs. First, the lexical aspect of verbs, which is represented as event structure, is essential lexical semantic information (Pustejovsky, 1995). Different lexical aspects have different event structure frames. Secondly, argument structure with semantic roles also is a necessary factor to represent the meaning of verbs.

We argue in this paper that the ESF, originally developed by Im & Pustejovsky (2009, 2010) and Im (2013), enriches WordNet. Linking ESF to WordNet makes it possible to provide information about sub-eventual structure and argument structure of English verbs together with original information about the semantic relation of verbs WordNet gives.

The ESF of a verb with its specific sense divides its sub-events into pre-state, process, and post-state. This will be a big help to any kind of inferencing or reasoning tasks which use the word meaning of verbs. For instance, the ESF of the English verb *arrive* in (1) gives the information required to derive the lexically entailed result state after the arriving event and the presupposed state before it.

(1) The Event Structure Frame of *arrive*
(arrive.v.01)
se1: pre-state: not_be_at (theme, goal)
se2: process: arriving (theme)
se3: post-state: be_at (theme, goal)

Given the sentence *John arrived at school at 9 am today,* we get the inferred statements from the ESF of arrive.v.01 by Word Sense Disambiguation (linking *arrive* to an appropriate WordNet synset arrive.v.01): 'John was not at school before 9 am today' and 'John was at school after 9 am today'.

We began the WordNet-ESF linking project around the end of last year (2018). The tagging work goes through the two steps: automatic annotation of the ESF for each verb sysnset in

WordNet by GESL and manual error correction. GESL is an automatic annotation tool of the ESF for verbs in a sentence developed by Im (2013). Since WordNet synsets have their example sentences, GESL is applied to the sentences for automatic ESF annotation.

In this paper, we present our main idea regarding the task and small annotated data focused on English motion verbs. The structure of this paper is as follows: in the next section, we briefly introduce the theoretical background of the Event Structure Frame and show the list of pre-defined ESFs in Im (2013). Section 3 describes our main task. First, we introduce GESL, the automatic ESF annotating system to verbs in text. Second, we explain how to assign ESFs to WordNet synsets. In section 4, we explain ESF-based verb classification and the extended list of ESFs for WordNet-ESF linking. In section 5, we show small size of data in which we annotated ESFs to WordNet synsets for a part of motion verbs. After that, we mention FrameNet and VerbNet and explain why we chose linking ESF to WordNet in the next section. Finally, we summarize our main idea and future work in section 6.

2 Event Structure Frame

In this section, we explain the theoretical background of the ESF. The idea is originated from Im and Pustejovsky (2009, 2010) and fully developed in Im (2013). The ESF is based on event structure and argument structure in Generative Lexicon Theory (Pustejovsky, 1995) and opposition structure (Pustejovsky, 2000). As shown in (1), the ESF is a merger of event structure, argument structure, and opposition structure.

A complex event has its sub-eventual structure which consists of temporally ordered sub-events. In (1), se1 precedes se2 and se3. The event structure of a complex event is composed of pre-state, process, and post-state. Pre-state is a presupposed sub-event. That is, it is a presupposition of the verb which denotes the main process (event). For instance, our common sense requires the presupposition that Kennedy was alive before killing him in order to use the word *kill*. On the other hand, post-state is temporally later than the killing process. The post-state is a lexical entailment of the verb *kill*. When Osswald killed Kennedy, it normally entails that Kennedy died and Kennedy is dead.

To sum up, the combination of pre-state, process, and post-state is a temporally ordered structure of lexical presuppositions, main process, and lexical entailments.

Based on the theoretical viewpoint about ESF, Im (2013) suggests 23 pre-defined ESF-dependent verb classes. As shown in Table 1, verb classification in GESL consists of three steps of classification.

aspectual	semantic	event type
state	state	state
process	process	process
	motion	motion
transition	change-of-location	leave, arrive, pass, transfer
	change-of-possession	lose, get, give
	change-of-state	come-into-existence, go-out-of-existence, become, begin, continue, end positive-causation, negative-causation, cos-leave, cos-arrive, cos-transfer, scalar-change change-state

Table 1. Verb classification in Im (2013)

The first step is to classify verbs according to the lexical aspect of verbs - state, process, and transition, based on Generative Lexicon Theory. State and process are simple events and transition is a complex event. Therefore, transition verbs have sub-eventual structure with more than one sub-event.

The next step is semantic classification of verbs. Im (2013) classifies process verbs into two groups – process and motion. It is because motion verbs have their own special lexical semantic properties. Their lexical aspect is heavily dependent on their contextual meaning. For instance, the motion verb *run* belongs to motion process but it changes into change-of-location class when it co-occurs with the prepositional phrases which denote goal, source, duration, etc. (e.g. *run to the store, run from the store, run for 30 minutes*). Transition verbs are classified into change-of-location, change-of-possession, or change-of-state verbs semantically.

The last step is to divide each semantic class into more specific ESF-dependent classes. Each verb class we finally get has its own ESF. Specifically, the change-of-location verb class has arrive, leave, pass, and transfer classes. The change-of-possession verbs are classified into lose, get, or give. Change-of-state verbs in-

clude aspectual classes (begin, continue, end), positive-/negative-causation (e.g. *cause_to / prevent_from*), become (e.g. *turn_red*), come_into_existence (e.g. *be_born*), go_out_of_existence (e.g. *die*), scalar_change (e.g. *increase, broaden*, etc.). COS-leave, COS-arrive, COS-transfer groups are for metaphorical or metonymical expressions of change-of-location which belong to change-of-state verb class semantically (e.g. *the water came to a boil*).

3 GESL-based Semi-Automatic Annotation of Event Structure Frame to WordNet

Our main task in WordNet-ESF linking is to assign a proper ESF to each synset of a verb in WordNet. We do the task semi-automatically via the two steps: automatic annotation of ESF with GESL and manual error correction. In section 3.1, we first introduce the automatic event structure tagging tool, GESL. Second, section 3.2 describes the procedure of WordNet-ESF linking.

3.1 The Generator of the Event Structure Lexicon (GESL)

GESL is the automatic event structure annotation tool developed by Im (2013) and Im and Pustejovsky (2009, 2010), which generates an appropriate event structure for each English event-denoting verb in text. Figure 1 shows the input and output of GESL.

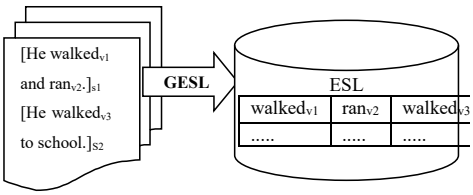


Figure 1. The input and output of GESL

As shown in Figure 1, the input of GESL is a text document. GESL gets English text data and generates the event structure of each event-denoting verb together with its lexical semantic information including its grammatical tense, aspect, and dependencies. For example, if GESL gets the sentence *Osswald killed Kennedy November 22, 1965*, the tool gives the ESL of the event-denoting verb *kill* as its output (Table 2).

verb	KILLED
vid	V1
tense	past
aspect	none
dependency	nsubj (killed, Osswald), dobj (killed, Kennedy), time (killed, November-4)
aspectual class	Transition
semantic class	change-of-state
event type	go_out_of_existence
event structure	se1: pre-state: not_be_killed (Kennedy) se2: pre-state: there_be (Kennedy) se3: process: killing (Osswald, Kennedy) se4: post-state: be_killed (Kennedy) se5: post-state: there_not_be (Kennedy)
sid	S1
sentence	*Osswald killed Kennedy November 22, 1965.*

Table 2. The Event Structure Lexicon of *kill*

Table 2 shows the GESL annotation result of the event-denoting verb *kill* in the special context the sentence generates. GESL classifies the contextual meaning of an English verb into one of the pre-defined event structure types via the three steps of classification – aspectual, semantic, and event type classification. The verb *kill* in the sentence above belongs to transition class aspectually and its semantic class is change-of-state (COS). Finally, the event type of the verb is go-out-of-existence.

GESL goes through several steps to derive the event structure of an event-denoting verb. We show the architecture of GESL in Figure 2.

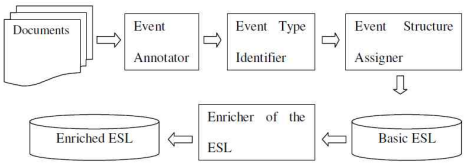


Figure 2. The architecture of GESL

GESL first determines whether a verb in text denotes an event or not. If it denotes an event, it classifies the verb into one of the pre-defined event types via the three classification steps and assigns the proper ESF to the verb. In addition, it links arguments to the semantic roles in the ESF by using the information from the given sentence. The last step is to enrich the event structure by adding synonyms, hypernyms, and antonyms[1].

[1] Refer to Im (2013) if you want to know in more detail about the enriching procedure of the ESL. We can infer additional information like 'Kennedy is dead', 'Kennedy died', 'Kennedy was alive', etc. by the enrichment.

3.2 WordNet-ESF Linking

Because WordNet synsets have their corresponding example sentences, we apply GESL to them in order to annotate the ESF to each synset in WordNet. After automatic annotation of ESF by GESL, we correct errors manually (Figure 3).

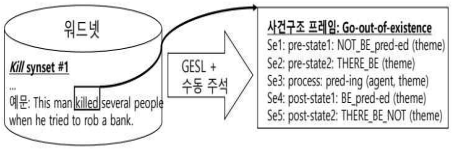


Figure 3. Annotation of ESF to WordNet synset

We have two reasons that we need manual error correction. First, many examples in WordNet synsets are not complete and thus GESL's performance is worse than its ordinary application to text documents. Second, quite many WordNet synsets do not have examples. In those cases, GESL is not applicable. Therefore, we need manual annotation of ESFs.

4 Verb Classes and Pre-defined Event Structure Frames

The ESFs and verb classes in GESL are designed as simple as possible, because it is an automatic annotation system. For instance, GESL does not distinguish between a verb class and its causative counterparts in terms of their ESFs. Instead, the issue is solved by the argument linking algorithm in GESL.

However, the ESFs linked to WordNet need to be more specific than the ESFs in GESL, since WordNet-ESF linking aims to make NLP applications like a textual inference system get the event structure-related inferences only by Word Sense Disambiguation with no other special NLP work.

First, we add its causative counterpart to each verb class (e.g. arrive – cause_arrive). This makes it easier to use the ESF of each synset of English verbs in WordNet without special difficulty in linking arguments to semantic roles in ESFs. Secondly, we separate semelfactive verb class from process class, although Im (2013) did not distinguish the two. The ESFs of the two verb groups are not different. However, we need to consider semelfactive verbs independently. The third change is to divide motion verbs into more specific groups considering motion_direction, motion, self_motion, move_backward, move_down, move_up, pull, push. self_motion verbs do not result in change-of-location. Fourth, the change-of-location verb class originally consists of arrive, leave, transfer but we added move_toward_speaker, move_from_speaker, bring, take, and carry. Fifth, scalar_change verb group is divided into: scale_up, scale_down, and scale_move . The sixth change is to add change_direction and change_posture. Finally, we added precede/follow, happen, maintain, skip, spread, info_transfer, performative (speech act verbs). Appendix A shows the list of verb classes for WordNet-ESF linking and their ESFs. ESFs and verb classes are not limited to the list but can be extended or modified. WordNet has more than 2100 verbs. Our final goal is to assign proper ESFs to all synsets of the verbs. In the next section, we show the examples of annotated ESF.

5 Data: Annotated WordNet Synsets

As of now, we have the ESFs for all synsets of verbs in WordNet by applying GESL to the example sentences in synsets of WordNet. We are working on manual error correction.

In this section, we present the result of experiment with the motion verbs which occur in the season 1 episodes of the drama named "Friends", which will be used in the Video Turing Test (VTT) Project we have been working on since 2017. We use the WordNet version 2.1 embedded in NLTK, Natural Language ToolKit developed at Stanford NLP Lab. The total number of verbs is 91 and they have 952 synsets. We assigned a proper ESF to each synset through automatic annotation by GESL and manual correction of the annotated ESF. We note that one verb can have several different ESFs since different synsets can have different ESFs. For instance, the 41 synsets of the verb *run* has 12 different types of ESF: motion, cause-motion, state, process, follow, leave, spread, change_state, cause-change_state, continue, become .

The scalar_change verbs need more consideration of the kinds of scales. We leave it as a future work.

motion [run.v.1, 6, 11, 28, 33, 34; play.v.18; ply.v.03], **cause-motion** [run.v.26], **change_state** [run.v.24, 41; melt.v.01; ladder.v.01], **cause-change_state** [run.v.31], **continue** [prevail.v.03], **follow** [hunt.v.01], **leave** [scat.v.01], **pass** [run.v.29], **process** [campaign.v.01; carry.v.15; move.v.13; operate.v.01; function.v.01; guide.v.05; race.v.02; run.v.13, 15, 16, 19, 21, 23, 25, 30, 32], **spread**

The target motion verbs are listed in Appendix B. Because the verbs used in the experiment are motion verbs, many synsets belong to motion or change-of-location-related classes. 30.6 % of the synsets (291 out of total 952 synsets) belong to motion or change-of-location-related verb classes. About 40% of the synsets are one of state, process, and change-of-state classes. It is a natural result because those groups have much more verbs than the others.

We additionally assigned the ESFs to the synsets of total 207 verbs including the 85 verbs used in the sentences which describe the scenes of Friends season 1 and their related phrasal verbs and idioms (Appendix C). The scene descriptions were automatically derived by the action recognition algorithm our co-workers developed in the field of Computer Vision. You can see the annotated data in GitHub.[4]

6 Related Work

Since lexical knowledge of words is crucial for various NLP applications including textual inference, computational lexical semanticists have been trying to build lexical resources which annotate many kinds of lexical knowledge. FrameNet, VerbNet, and WordNet, out of the built resources, are well-known and used in the field of NLP and its applications.

FrameNet is a lexical database of English that is both human- and machine-readable with manually annotated sentences, which is based on Frame Semantics (Fillmore, 1976). The basic idea is that the meaning of most words can be understood on the basis of a semantic frame: a description of a type of event, relation, or entity and the participant in it. The FrameNet project is still in progress. However, FrameNet's frames do not annotate the sub-eventual structure of verbs systematically, since it concentrates on semantic roles rather than event structure (Osswald and Van Valin, 2012).

Although VerbNet (Kipper, 2005), a hierarchical verb lexicon based on Levin's classes, also represents sub-eventual structure of verbs, its event structure annotation is neither complete nor consistent (Zaenen et al., 2008). More importantly, neither of the resources has much knowledge about semantic relations of verbs.

WordNet does not include the knowledge about the event structure of verbs but it has the other important factors of lexical semantic knowledge of verbs – semantic relations like synonym, antonym, hypernym, hyponym, etc. Therefore, adding event structure to WordNet will make the resource much more helpful to any NLP applications which need lexical knowledge of verbs. Especially, WordNet-ESF linking would allow us to derive event structure of a verb in text only by Word Sense Disambiguation which maps it to its proper synset, because the synset would have its ESF. In conclusion, WordNet-ESF linking is a good attempt of combining crucial lexical knowledge of verbs.

7 Conclusion

In this paper, we briefly described our semi-automatic annotation task of Event Structure Frames to WordNet synsets via the following two steps. GESL, an automatic event structure annotation tool, assigns a proper ESF to each WN synset of English verbs in WordNet and we correct errors manually. Since each WordNet synset has its own example sentence, GESL, which annotates event structure to verbs in a full sentence, can be applied to the target verb in the sentence so that it annotates an ESF to the verb. If a synset has no example sentence, GESL cannot annotate an ESF to the sysnset. It is one of the reasons that we need manual error correction.

Although WordNet is very useful to develop NLP application tools which require word meaning, it lacks event structure, argument structure, semantic role, and opposition structure. We expect that the enriched WordNet by WordNet-ESF linking will be a big help to NLP applications such as textual or multimodal inference tasks.

For WordNet-ESF linking, we extended ESF-dependent verb classes in GESL in order to represent the event structural meaning of each synset of verbs more specifically. GESL has 23 verb classes and each of them has its own event structure frame. We suggest 44 classes and their causative counterparts in this paper. The classes are not fixed. Since we still work on the WordNet-ESF linking task, verb classes can undergo change.

Acknowledgments
This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-01780, The

[run.v.27, 30], **state** [run.v.05, range.v.01, tend.v.01], **become** [run.v.14]

[4] https://github.com/ish97/VTT/blob/master/

technology development for event recognition/relational reasoning and learning knowledge based system for video understanding).

References

Fillmore, Charles J. 1976. Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech* 280: 20-32.

Im, Seohyun. 2013. *The Generator of the Event Structure Lexicon (GESL): Automatic Annotation of Event Structure for Textual Inference Tasks*. PhD Dissertation, Brandeis University, MA, USA.

Im, Seohyun and James Pustejovsky. 2010. Annotating Lexically Entailed Subevents for Textual Inference Tasks. In the *Proceedings of FLAIRS 23*, Daytona Beach, Florida, USA, 2010.

Im, Seohyun and James Pustejovsky. 2009. Annotating Event Implicatures for Textual Inference Tasks. In the *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon*, Pisa, Italy, 2009.

Kipper, Karin Schuler. 2005. VerbNet: A Broad-coverage, Comprehensive Verb Lexicon. *PhD Dissertation*. University of Pennsilvenia.

Miller, George A. 1995. Wordnet: A Lexical Database for English. *Communications of the ACM* 38, no. 11.

Osswald, Rainer and Jr. Robert D. Van Valin. 2012. FrameNet, Fame Structure, and the Syntax-Semantics Interface (draft).

Pustejovsky, James. 2000. Events and the Semantics of Opposition, Pustejovsky and Tenny (eds.) *Events as Grammatical Objects*. CSLI Publications.

Pustejovsky. James. 1995. *The Generative Lexicon*. The MIT Press.

Zaenen, Annie, Cleo Condoravi, and Danny. Bobrow. 2008. The Encoding of Lexical Implications in VerbNet. *Proceedings of LREC 2008*. Morocco, 2008.

Appendix A. Verb Classes and Event Structure Frames

* CAUSATIVE counterparts: causer-argument added

STATE
se1: state: pred-ing_(prep) (theme)
PROCESS [cause_process]
se1: process: pred-ing_(prep) (agent)
SEMELFACTIVE [cause_semelfactive]
se1: process: pred-ing_(prep) (theme)
MOTION [cause_motion]
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing (theme)
d-se2: post-state: be_loc-prep (theme, goal)
MOVE_BACK [cause_move_back]
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing_back (theme)
d-se2: post-state: be_loc-prep (theme, goal)
d-se3: post-state: be_behind (goal, source)
d-se2 = d-se3
MOVE_UP [cause_move_up]
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing_up (theme)
d-se2: post-state: be_loc-prep (theme, goal)
d-se3: post-state: be_higher_than (goal, source)
d-se2 = d-se3
MOVE_DOWN [cause_move_down]
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing_downward (theme)
d-se2: post-state: be_loc-prep (theme, goal)
d-se3: post-state: be_lower_than (goal, source)
d-se2 = d-se3
MOVE_TOWARD_SPEAKER
[cause_move_toward_speaker]
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing (theme)
d-se2: post-state: be_loc-prep (theme, goal)
d-se3: post-state: be_near (goal, speaker's location)
d-se2 = d-se3
MOVE_FROM_SPEAKER
[cause_move_from_speaker]
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing (theme)
d-se2: post-state: be_loc-prep (theme, goal)
d-se3: post-state: not_be_near (goal, speaker's location)
PULL
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing (agent, theme)
d-se2: post-state: be_loc-prep (theme, goal)
PUSH
d-se1: pre-state: be_loc-prep (theme, source)
se1: process: pred-ing (agent, theme)
d-se2: post-state: be_loc-prep (theme, goal)
CARRY
se1: process: pred-ing (agent, theme)
se2: state: having (agent, theme)
se1 = se2
LEAVE [cause_leave]
se1: pre-state: be_loc-prep (theme, source)
se2: process: pred-ing (theme)
se3: post-state: not_be_loc-prep (theme, source)
PASS [cause_pass]
se1: pre-state: be_loc-prep (theme, source)
se2: process: pred-ing (theme)
se3: state: be_loc-prep (theme, path)
se4: post-state: be_loc-prep (theme, goal)
se2 = se3
ARRIVE [cause_arrive]
se1: pre-state: not_be_loc-prep (theme, goal)
se2: process: pred-ing (theme)
se3: post-state: be_loc-prep (theme, goal)

TRANSFER [cause_transfer]
se1: pre-state: be_loc-prep (theme, source)
se2: process: pred-ing (theme)
se3: post-state: be_loc-prep (theme, goal)
SPREAD [cause_spread]
se1: pre-state: not_be_over (theme, ground)
se2: process: pred-ing (agent, theme, ground)
se3: post-state: be_over (theme, ground)
BRING
se1: pre-state: not_be_loc-prep (agent & theme, goal)
se2: process: pred-ing_goal-prep (agent, theme, goal)
se3: post-state: be_loc-prep (agent & theme, goal)
TAKE
se1: pre-state: be_loc-prep (agent & theme, source)
se2: process: pred-ing_source-prep (agent, theme, source)
se3: post-state: not_be_loc-prep (agent & theme, source)
LOSE [cause_lose]
se1: pre-state: have (possessor, theme)
se2: process: pred-ing (possessor, theme)
se3: post-state: not_have (possessor, theme)
GET [cause_get]
se1: pre-state: have (recipient, theme)
se2: process: pred-ing (recipient, theme)
se3: post-state: not_have (recipient, theme)
GIVE
se1: pre-state: have (possessor, theme)
se2: process: pred-ing (possessor, recipient, theme)
se3: post-state: have (recipient, theme)
EXCHANGE
se1: pre-state: have (possessor, theme1)
se2: pre-state: have (recipient, theme2)
se3: process: pred-ing (possessor, recipient, theme1, theme2)
se4: post-state: have (possessor, theme2)
se5: post-state: have (recipient, theme1)
INFO_TRANSFER
se1: pre-state: have (possessor, theme:info)
se2: process: pred-ing (possessor, theme:info)
se3: post-state: have (possessor & recipient, theme:info)
COME_INTO_EXISTENCE [cause_come_into_existence]
se1: pre-state: not_be_pred-ed (theme)
se2: pre-state: there_be_not (theme)
se3: process: pred-ing (theme)
se4: post-state: be_pred-ed (theme)
se5: post-state: there_be (theme)
GO_OUT_OF_EXISTENCE [cause_go_out_of_existence]
se1: pre-state: not_be_pred-ed (theme)
se2: pre-state: there_be (theme)
se3: process: pred-ing (theme)
se4: post-state: be_pred-ed (theme)
se5: post-state: there_be_not (theme)
BECOME [cause_become]
se1: pre-state: not_be_pred-ed (theme, state)
se2: pre-state: not_be (theme, state)
se3: process: pred-ing (theme, state)
se4: post-state: be_pred-ed (theme, state)
se5: post-state: be (theme, state)
BEGIN [cause_begin]
se1: pre-state: not_in_progress (event)
se2: process: pred-ing (event)
se3: post-state: in_progress (event)
CONTINUE [cause_continue]
se1: pre-state: in_progress (event)
se2: process: pred-ing (event)
se3: post-state: in_progress (event)
END [cause_end]
se1: pre-state: in_progress (event)
se2: process: pred-ing (event)
se3: post-state: not_in_progress (event)
POSITIVE_CAUSATION
se1: pred-ing (causer, event)
se2: happen (event)
NEGATIVE_CAUSATION
se1: pred-ing (causer, event)
se2: not_happen (event)
SCALE_UP [cause-scale_up]
d-se1: pre-state: be_loc-prep (theme, source_scale)
se1: process: pred-ing (theme)
d-se2: post-state: be_loc-prep (theme, goal_scale)
d-se3: post-state: be_higher_than (goal, source_scale)
d-se2 = d-se3
SCALE_DOWN [cause-scale_down]
d-se1: pre-state: be_loc-prep (theme, source_scale)
se1: process: pred-ing (theme)
d-se2: post-state: be_loc-prep (theme, goal_scale)
d-se3: post-state: be_lower_than (goal, source_scale)
d-se2 = d-se3
SCALE_MOVE [cause-scale_move]
se1: process: pred-ing (theme, scale)
CHANGE_DIRECTION [cause-change_direction]
se1: pre-state: not_be_pred-ed (theme)
se2: pre-state: be (theme, source_direction)
se3: process: pred-ing (theme)
se4: post-state: be_pred-ed (theme)
se5 = post-state: be (theme, goal_direction)
CHANGE_POSTURE [cause-change_posture]
se1: pre-state: not_be_pred-ed (theme)
se2: pre-state: be (theme, source_posture)
se3: process: pred-ing (theme)
se4: post-state: be_pred-ed (theme)
se5: post-state: be (theme, goal_posture)
CHANGE_STATE [cause_change_state]
se1: pre-state: not_be_pred-ed (theme)
se2: pre-state: be (theme, source_state)
se3: process: pred-ing (theme)
se4: post-state: be_pred-ed (theme)
se5: post-state: be (theme, goal_state)
COS_LEAVE [cause_cos_leave]
same as the ESF of LEAVE
COS_ARRIVE [cause_cos_arrive]
same as the ESF of ARRIVE
COS_TRANSFER [cause_cos_transfer]
same as the ESF of TRANSFER
PERFORMATIVE (speech act)

se1: pre-state: not_be_pred-ed_to_by (theme, addressee, speaker)
se2: process: pred-ing (speaker, addressee, theme)
se3: post-state: be_pred-ed_to_by (theme, addressee, speaker)
HAPPEN [cause_happen]
se1: state: there_be (event)
MAINTAIN
se1: pre-state: be (state)
se2: process: pred-ing (agent, state)
se3: state: be (state)
se2 = se3
PRECEDE
se1: state: pred-ing (theme1, theme2)
se2: state: be_before (theme1, theme2)
se1 = se2
FOLLOW
se1: state: pred-ing (theme1, theme2)
se2: state: be_after (theme1, theme2)
se1 = se2

Appendix B. The list of motion verbs in Friends Season 1 episodes

arrive, back, bail, barge, base, board, bring, brush, bury, camp, carry, chase, clean, come, conduct, creep, dance, dip, drag, draw, drift, drive, drop, dump, enter, erase, fall, fax, fling, float, flush, fly, follow, go, head, hike, hop, inch, invade, jump, kick, land, lay, lead, leave, load, move, park, pass, plunge, pop, pour, pull, push, put, raise, reach, remove, return, ride, roll, run, rush, send, ship, shove, shuffle, sit, ski, skip, slather, slide, slip, stand, step, stomp, sweep, swoop, take, throw, travel, tremble, turn, twist, usher, vacuum, walk, wave, wind, wipe, wobble

Appendix C. The list of verbs in the scene description sentences provided by a Computer Vision Action Recognition algorithm

apply, assemble, attack, bark, beat, box, burn, celebrate, cheer, clean, comb, cook, crash, cry, cut, decorate, demonstrate, drink, dunk, eat, explain, explode, fight, film, fish, fix, floor, fold, give, have, hit, hold, hug, hunt, install, interact, interview, involve, kiss, lick, lie, make, mix, paint, perform, pet, ping, place, play, pose, preform, prepare, punch, race, read, record, rub, scoop, score, scream, sew, shoot, show, sing, skate, ski, sleep, slice, smash, smile, solve, speak, spray, stretch, surf, swim, talk, teach, use, wash, watch, weave, work, wrestle, write

Enhancing Conceptual Description through Resource Linking and Exploration of Semantic Relations

Svetlozara Leseva
DCL – IBL, BAS
Sofia, Bulgaria
zarka@dcl.bas.bg

Ivelina Stoyanova
DCL – IBL, BAS
Sofia, Bulgaria
iva@dcl.bas.bg

Abstract

The paper presents current efforts towards linking two large lexical semantic resources – WordNet and FrameNet – to the end of their mutual enrichment and the facilitation of the access, extraction and analysis of various types of semantic and syntactic information. In the second part of the paper, we go on to examine the relation of inheritance and other semantic relations as represented in WordNet and FrameNet and how they correspond to each other when the resources are aligned. We discuss the implications with respect to the enhancement of the two resources through the definition of new relations and the detailisation of conceptual frames.

1 Introduction

The first part of the paper outlines the principles and procedures of aligning WordNet and FrameNet. The focus is on WordNet as the main lexical-semantic structure (the verbal domain, in particular), which we aim at enhancing with richer linguistic description from FrameNet and VerbNet. The second part of the paper proposes an analysis of the correspondences between the frame-to-frame relations in FrameNet and the synset-to-synset relations in WordNet.

The aim is two-fold: (a) from a theoretical perspective, to provide insights into the scope and definition of overlapping or corresponding relations and the relational structure of the two resources, to establish similarities and discrepancies that may come from different semantic construal or from errors; (b) from an applied perspective, to provide directions for the mutual enhancement and improvement of (i) the relational structure of the two resources; (ii) the accuracy of the frame assignment based on the theoretical observations.

The contribution of the paper consists in:

• An implementation of a mapping between WordNet synsets and FrameNet frames by extending existing mappings using the hierarchical structure of WordNet and the concept of inheritance. In addition, considerable improvements on the data are made including disambiguation of FrameNet frame assignment (selecting a single frame for a given synset, where the mapping has yielded more than one), correction of errors, consistency checks.

• A theoretical study of frame relations and their correspondences in WordNet and discovery of existing but inexplicit relations in one of the resources that are mappable to the other to the end of enhancing the relational structure of both resources and proposing procedures for a more reliable frame assignment using semantic inheritance.

This work is a key part of ongoing research on defining a conceptual framework for encoding semantic relations between verb and noun synsets based on a detailed conceptual representation of verbs and the identification of semantic classes of nouns satisfying the selectional restrictions imposed on frame elements in the verb's frame.

After a brief discussion of related work (Section 2), we outline the alignment between WordNet and FrameNet (Section 3) based on existing mappings and procedures for their enhancement and expanding. Section 4 focuses on the theoretical and practical aspects of semantic relations in FrameNet and how they are reflected (with respective semantic relations) within WordNet. Section 5 sketches the implications from these observations, while (Section 6) focuses on the role of this research in the context of other ongoing research.

2 Related work

One of the main directions of development of semantic resources is finding ways of uniting their strengths through integrating them and exploiting their features in a complementary way. Mapping

of existing semantic resources has been undertaken in a number of works (cf. section 3.1).

Another line of research in the development and enhancement of the interconnected resources is explicitly linking and generalising existing, but unrelated information in them. A poorly studied direction of research has been the exploration and use of the internal structure of these resources towards their mutual enhancement. One area of research along these lines has been the extension of frame relations by using information from WordNet. (Virk et al., 2016) propose a supervised model for enriching FrameNet's relational structure through predicting new frame-to-frame relations using structural features from the existing FrameNet network, information from the WordNet relations between synsets, and corpus-collected lexical associations. Leseva et al. (2018) have employed features of both relational structures to develop an algorithm for assigning FrameNet frames to WordNet synsets by transferring the relational knowledge for pairs of related synsets to matching lexical units and frames in FrameNet.

An interesting theoretical and practical issue arising from the mapping of the 'building blocks' of the two resources is how the underlying relational structures relate and correspond to each other, how they can be mapped to each other, and further explored. In the second part of this paper, we have attempted to tackle this issue.

3 Aligning WordNet and FrameNet

Our work relies on two main resources – WordNet (WN) and FrameNet (FN), and employ VerbNet (VN) as a complementary resource in some tasks related to alignment and verification. We use WordNet (Miller, 1995; Fellbaum, 1999) as the basic lexical resource. FN (Baker et al., 1998) represents conceptual structures (frames) which describe particular types of objects, situations, etc. along with their participants, or frame elements (Ruppenhofer et al., 2016). Frames are then assigned to lexical units (LUs), e.g. the verb *mature* is assigned the frame *Aging* with the description 'An Entity is undergoing a change in age typically associated with some deterioration or change in state'. FrameNet is internally structured using a set of relations, which are discussed in Section 4. The VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008) classes represent formations of verbs with shared semantic and syntactic properties and behaviour organised in a shallow hierarchy.

3.1 Existing mappings

Previous efforts at linking these resources include Shi and Mihalcea (2005), Baker and Fellbaum (2009), WordFrameNet[1] (Laparra and Rigau, 2009; Laparra and Rigau, 2010), MapNet[2] (Tonelli and Pighin, 2009), and more enhanced proposals, such as the system Semlink[3] (Palmer, 2009) which brings together WN, FN and VN with PropBank, and its follow-up Semlink+ that brings in mapping to Ontonotes (Palmer et al., 2014). Analysis of the available resources for linking WN, FN and VN, as well as procedures for automatically extending the mapping, are presented by Leseva et al. (2018).

These efforts generally suffer from limited coverage and compatibility issues due to multiple release versions of the original resources. Moreover, to the best of our knowledge, no further checks and verification have been performed on the results. This reduces considerably their applicability and further development.

A complementary approach is to exploit the relational structure of the two resources through assigning frames to synsets not only on the basis of direct correspondence between FN LUs and WN literals, but also on the basis of the inheritance of conceptual features in hypernym trees and the assignment of frames by inheritance from hypernyms to hyponyms. The main drawback of this approach is that for deeper level WN synsets the inherited frames may be underspecified. Our current and prospective work builds upon this paradigm, notably by looking for ways of refining previous proposals (Leseva et al., 2018) through validation which results in enriching the frame structure with systematic relations (e.g. causative, inchoative, etc. frame correspondences). Further, we envisage to define new, more detailed frames on the basis of more rigid selectional restrictions on frame elements.

3.2 Linking procedures

Linking FN to WN is not straightforward. There are two principal types of mappings that have already been applied on the lexical resources discussed in section 3.1: (a) lexical mapping – lexical units (from one resource) have been assigned

[1] http://adimen.si.ehu.es/web/WordFrameNet
[2] https://hlt-nlp.fbk.eu/technologies/mapnet
[3] https://verbs.colorado.edu/semlink/

categories from another, e.g. a FN lexical unit is mapped to a WN literal and hence its FN frame is also assigned to the literal (and the synset); and (b) structural mapping – classification categories from one resource have been aligned to categories from another, e.g. a VN class assigned to a synset is linked to a FN frame, so the FN frame is transferred onto the synset. In this way we are able to verify individual mappings by examining the result in terms of the overall structure.

Initially, our mapping is based on three sources of existing lexical mappings: 2,817 direct mappings provided within FN (Baker and Fellbaum, 2009), 3,134 from eXtendedWordFrameNet (Laparra and Rigau, 2010), and 1,833 from MapNet (Tonelli and Pighin, 2009). Structural mapping using VN contributed 1,335 mappings. Overall, there are 4,306 unique WN synset to FN frame mappings. The main procedure we apply to improve and extend mapping coverage is based on the relations of inheritance within WordNet. First, we manually verified the frames assigned to 250 out of the 566 root verb synsets: we corrected 75 mappings and assigned valid frames to additional selected 27 root synsets with a large number of hyponyms. We then transferred the hypernym's frame to its hyponyms in the cases where the hyponyms are not directly mapped to FN frames. As a result, we obtained an extended coverage of 12,880 synsets (with an assigned FN frame). With the further defined procedures we aim at improving the quality of this assignment.

The procedures for validation of frame assignments to verb synsets include: (i) manual checks of the assigned frame; (ii) checks for existing but unmapped correspondences between literals and LUs (e.g., by reapplying lexical mapping); (iii) automatic or semiautomatic consistency checks based on correspondences between VN classes (or superclasses) and FN frames; (iv) automatic or semiautomatic consistency checks based on systematic relations within the resources, e.g. causativity. If no appropriate frame exists, we propose to posit a new category (and a frame) provided that it is predictable and complying with FN's frame structure. For instance, while *Motion* is linked to *Cause_motion*, *Self_motion* (e.g. *jump:1, leap:1* 'move forward by leaps and bounds') does not have a causative counterpart to which verbs such as *jump:11, leap:4* 'cause to jump or leap' can be mapped, so we formulate one.

An envisaged direction for refining the inheritance assignment is by employing relational information based on the exploration of FN-to-WN relations discussed below, as well as through identifying meaningful information in the WN glosses that may point to a more appropriate frame.

4 Theoretical and practical aspects of semantic relations within FrameNet reflected in WordNet

FN and WN each have its own relational structure which is based on conceptual relations between language units (WN) or conceptual representations (FN). The WN structure is by far the richer in types and instances of relations; in addition to the conceptual relations it comprises lexical relations, derivational relations and some other relations. Although the relations in the two resources have different number and scope, at least part of them are grounded in similar universal assumptions which leads to partial overlap, depending on their definition and the specific information in the resources. For instance, there is a clear correspondence between the *Inheritance* relation in FrameNet and the *hypernymy* relation in WordNet, to the extent that both represent a modelling of the is-a relation (Ruppenhofer et al., 2016), or between the *Causativity* relation (FN) and the *causes* relation (WN). Figure 1 presents the process of linking WN and FN. In what follows, we are going to explore how the FN frame-to-frame relations translate into WN relations (when they do) and to outline the main trends in the correspondence between relations in the two resources.

The core part of the data to be examined are pairs or longer chains of WN synsets such that: (a) are related through a given WN relation, and (b) are assigned FN frames, which are (c) related through a particular FN relation.

The main WN relation to be considered is hypernymy, which is the principal tree structure organising relation in the resource. We take into account both direct hypernymy (direct relation between a parent and a child node) and indirect hypernymy (where the hypernym is not a parent of the hyponym but there are intermediate parents between them). Other relations that emerge from the studied data are: antonymy, also see, causes, verb group, as well as some distant shared hypernyms (i.e. the synsets are in the same tree). Below we present the definition and theoretical grounding

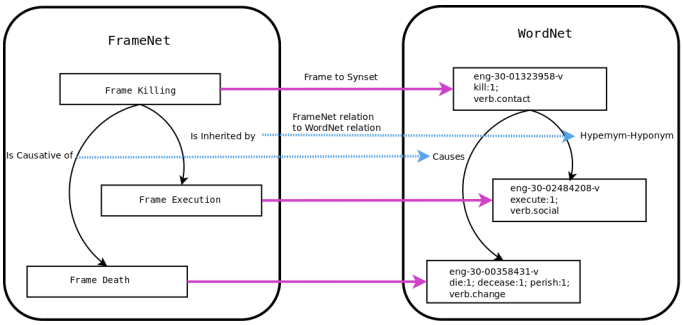


Figure 1: Representation of WordNet to FrameNet linking.

of FN relations (Ruppenhofer et al., 2016), along with the observations about their correspondence with WN relations.

4.1 Inheritance (Is Inherited by ↔ Inherits from)

Inheritance is defined as the strongest relation in FN; it denotes a relationship between a more general (parent) frame, and a more specific (child) frame in such a way that the child frame elaborates the parent frame. The basic idea, although not always straightforwardly applicable, is that each semantic fact about the parent must correspond to an equally specific or more specific fact about the child (Ruppenhofer et al., 2016, p. 81-82). This means that, generally, there should be a correspondence between entities, frame elements, frame relations and semantic characteristics in the parent and the child frame (Petruck, 2015).

Example 1. Frame *Killing Is Inherited by* frame *Execution*

Frame: *Killing*
Core frame elements: Killer; Victim:Sentient; Cause; Means:State_of_affairs; Instrument:Physical_entity
FN definition: A Killer or Cause causes the death of the Victim.
Example synset: kill:1

Frame: *Execution*
Core frame elements: Executioner:Sentient; Executed:Sentient
FN Definition: An Executioner punishes an individual (Executed) with death as a consequence of some action of the Evaluee (the Reason).
Example synsets: execute:1 (direct hyponym of kill:1); hang:3 (indirect hyponym)

As per the definition of *Inheritance*, the configurations of the two frames are similar and the frame elements in the parent frame have correspondences in the child frame, which may be the same or more specific: e.g. Killer has no selectional restrictions, unlike its more specific descendant Executioner (which is specified as Sentient).

Based on this definition, one should expect a considerable overlap between *Inheritance* and *hypernymy*: that is, when a pair of WN synsets is related through hypernymy and their corresponding frames are related through a frame-to-frame relation in FN, this relation should be *Inheritance*.

What the data show (Table 1) diverges from this expectation in two ways: (a) there is another frame-to-frame relation which is very strongly favoured for a counterpart of the hypernymy relation, i.e. *Using* (compare results in Table 1); (b) in a substantial number (20%) of the cases we find out an inverse relationship, i.e. for a hypernym–hyponym pair, the hyponym is assigned the more general (parent) frame, and the hypernym – the child frame in an existing *Inheritance* relation (the last two rows in Table 1). This is illustrated in Example 2 where the hyponym is assigned the frame *Respond to a proposal*, while the hypernym receives the child frame *Agree or refuse to act*.

Example 2.

Hypernym: refuse:1, decline:3; **Gloss**: show unwillingness towards; **Frame**:*Agree or refuse to act*
Hyponym: reject:4, spurn:1; **Gloss**: reject with contempt; **Frame**: *Respond to proposal*

When looking closely at the data, we find out that in a substantial number of the cases of reversed relation, this is not so much the result of incorrect automatic assignment of frames, as the result of different construal of the conceptual

	Is Inherited by		Is Used by		Is Perspectivized in		Has Subframe(s)		Causative of	
WN relation	total#	#diff.	total#	#diff.	total#	#diff.	total#	#diff.	total#	#diff.
Direct hypernymy	84	43	67	33	3	2	6	2	13	7
Indirect hypernymy	454	66	576	70	37	2	129	2	41	8
Direct hyponymy	35	22	39	13	0	0	0	0	11	6
Indirect hyponymy	108	21	51	18	0	0	0	0	36	6

Table 1: WN relations hypernymy/hyponymy for different FN relations.

and the lexical domain as the parent and child frames show a high level of similarity. This is the case, though not in all instances, with frame pairs such as *Referring_by_name* and *Labeling*, *Ingest_substance* and *Ingestion*, *Statement* and *Telling*, *Statement* and *Affirm_or_deny*, *Assistance* and *Supporting*, *Change_position_on_a_scale* and *Proliferating_in_number*, among others.

4.2 Using (Is Used by ↔ Uses)

Another hierarchical relation in FN is *Using*. It is defined as a relationship between two frames where the first one makes reference in a very general kind of way to the structure of a more abstract, schematic frame (Ruppenhofer et al., 2016). The definition has been further specified as a relation between a child frame and parent frame in which only some of the FEs in the parent have a corresponding entity in the child, and if such exist, they are more specific (Petruck and de Melo, 2012); hence, the relation may be viewed as a kind of weak Inheritance (Petruck, 2015).

The data confirm that the majority of synsets mapped to FN frames with the *Using* relation are hypernym-hyponym pairs; also, the numbers for *Using* are similar to the respective numbers for the *Inheritance* relation, as shown in Table 1.

Example 3. Frame *Placing Is Used by* frame *Arranging*
Frame: *Placing*
Core frame elements: Agent:Sentient; Cause; Theme:Physical_object; Goal:Goal
FN definition: An Agent places a Theme at a location, the Goal, which is profiled.
Example synset: put:1, set:1, place:1, pose:5

Frame: *Arranging*
Core frame elements: Agent:Sentient; Theme:Physical_object; Configuration
FN Definition: An Agent puts a complex Theme into a particular Configuration.
Example synsets: arrange:1, set up:5

The child frame and the parent frame to which it refers have similar configurations of elements, with the more specific Configuration (of things) corresponding to Goal (principally a location).

Similarly to *Inheritance*, cases of inverse assignment of the *Using* relation, where a hypernym is assigned a child frame, and a hyponym – a parent frame, are also found on a regular basis (12% of the cases) although not as often as with the *Inheritance* relation. Examples like (4) show that synset members and language units may be mapped to descriptions with different level of specification: in this case *garage:1* is construed as more specific in WordNet, but is assigned the more general *Placing* frame than its hypernym, which receives the frame *Storing*.

Example 4.
Hypernym: store:2; **Gloss**: find a place for and put away for storage; **Frame**: *Storing*
Hyponym: garage:1 **Gloss**: keep or store in a garage; **Frame**: *Placing*

The inverse assignment in many of the cases concerns frame pairs which display higher level of similarity and a weaker hierarchical relation. Such frame pairs, though not exclusively, include: *Placing–Storing*, *Abounding_with–Mass_motion*, *Attempt_suasion–Suasion*, *Evidence–Explaining_the_facts*.

The inverse frame assignment with both *Inheritance* and *Using* represents an interesting theoretical issue with respect to the analysis of lexical units (verbs) in terms of their lexical definitions and their conceptual properties.

4.3 Perspective (Is Perspectivized in ↔ Perspective on)

Perspective is defined as similar to, but more specific and restrictive than *Using* (Ruppenhofer et al., 2016, p. 82). It indicates that a situation viewed as neutral may be specified by means of perspectivised frames that represent different possible points-of-view on the neutral state-of-affairs.

It follows from this definition that the neutral frame is more abstract than the perspectivised frames and that there should be a great extent of

correspondence between the conceptual description and frame elements of the neutral and the perspectivised frames; these features *Perspective on* shares to a degree with both *Inheritance* and *Using*. It is not surprising, then, that this relation may translate as the *hypernymy-hyponymy* relation (Table 1), and in fact, this is the only WN relation that corresponds to it, even though in a very limited way: only 2 pairs of frames are found to be represented by related synsets: *Transfer* – which is perspectivised in *Giving* (cf. Example 5) and *Hostile_encounter* – which is perspectivised in *Attack*:

Example 5.

Hypernym: give:3; **Gloss**: transfer possession of something concrete or abstract to somebody; **Frame**: *Transfer*

Hyponym: contribute:2, give:25, chip in:1; **Gloss**: contribute to some cause; **Frame**: *Giving*

Apart from the actual WN relations, we find *Perspective on* between synsets having a common direct or indirect hypernym, where the same pairs *Giving–Transfer* and *Hostile encounter–Attack* are the only two discovered. Only among more structurally distant pairs of synsets do we find other pairs of neutral–perspectivised frames: *Transfer–Receiving*, *Import export scenario–Importing*, *Import_export_scenario–Exporting*.

This observation shows that the kind of semantic generalisation underlying the *Perspective* relation does not correlate well with the WN conceptual and lexical relations. In fact, looking more in depth into the data, we find out that synsets related through a WN relation may be perspectivised frames of a non-lexical neutral frame. Such example is provided by the antonym pair *import:1* (bring in from abroad') – *export:1* (sell or transfer abroad'): the two synsets are assigned the frames *Importing* and *Exporting*, respectively, which perspectivise the neutral *Import_export_scenario*, and although they have a common hypernym *trade:1, merchandise:1*, there is no suitable lexicalisation of the neutral frame. A similar case is presented by other converse (antonym) pairs.

4.4 Subframe (Has Subframe(s) ↔ Subframe of)

Subframe is a relation between a complex frame referring to sequences of states and transitions, each of which can itself be separately described as a frame, and the frames denoting these states or transitions (Ruppenhofer et al., 2016, p. 83–84). It is also noted that the frame elements of the complex frame may be connected to the frame elements of the subparts, although not all frame elements of one need have any relation to the other. Another feature of this relation is that the ordering and other temporal relationships of the subframes can be specified by the binary *Precedence* relation.

The definition of *Subframe* allows for it to correspond to hypernymy, which, apart from 2 instances of *also see*, is the only WN corresponding relation (Table 1), even though it is represented in a very limited way – only 2 pairs of frames are found, *Cause_motion–Placing* and *Cause_motion–Removing* (Example 6), and the predominant trend is for non-direct, rather than for direct hypernymy.

Example 6.

Hypernym: raise:2, lift:1, elevate:2, get up:3; **Gloss**: raise from a lower to a higher position; **Frame**: *Cause_motion*

Hyponym: shoulder:1; **Gloss**: lift onto one's shoulders; **Frame**: *Placing*

In more distant structural relations between WN synsets with common non-direct, distant hypernyms, other pairs of frame-to-frame relations are found as well, such as *Traversing–Departing*, *Traversing–Arriving*, *Intentional_traversing–Quitting_a_place*, *Self_motion–Quitting_ a_place*.

Although *Subframe* is much better represented through (indirect) hypernymy than *Perspective*, it shares with it the feature that much like the neutral frame, the complex frame may represent a conceptual structure that does not have a lexicalised correspondence and that it is feasible to look for WN relations between subframes of a complex frame (rather than between a complex frame and a subframe). Another supporting example comes from the domain of antonymy – two synsets related by means of the *antonymy* relation may be assigned subframes of a complex frame, e.g. *fall asleep:1, dope off:1*... (*Fall_asleep*) <antonym> *wake up:2* (*Waking_up*) with respect to *Sleep_wake_cycle*.

4.5 Precedence (Precedes ↔ Is Preceded by)

This relation holds between component subframes of a single complex frame and provides additional information by specifying the chronological ordering of the states and events (subevents) within a complex event (Ruppenhofer et al., 2016; Petruck, 2015). A small number of *Precedence* instances are found among antonyms (12 pairs) and the majority of the instances are between synsets having

a common (direct or indirect) hypernym. The following pairs of frame-to-frame relations are found with antonyms: *Placing–Removing*, *Arriving–Departing*, *Activity_stop–Activity_ongoing*:

Example 7.

Antonym: file in:1; **Gloss**: enter by marching in a file; **Frame**: *Arriving*

Antonym: file out:1; **Gloss**: march out, in a file; **Frame**: *Departing*

This relation may result in complex structures involving a number of subframes such as the notable example of the *Sleep_wake_cycle* (Petruck, 2015). It does not have a counterpart in the WN structure, but it may be transferred, thus bringing an additional dimension of semantic description through linking otherwise unrelated subevents and through specifying their temporal ordering.

4.6 Causation (Causative of) and Inchoativity (Inchoative of)

Causation and *Inchoativity* are systematic non-inheritance relationships between stative frames and the inchoative and causative frames that refer to them (Ruppenhofer et al., 2016, p. 85). Obviously, *Causation* should correspond straightforwardly to the WN relation *causes*. In fact, it does in a small number of cases (30 pairs), which is due to the fact that this relation has not been implemented consistently in FN (Ruppenhofer et al., 2016, p. 85). It may well be argued that its implementation needs to be enhanced in WordNet as well, as a lot of pairs for which this relation holds have not been linked in the resource. For instance, while the causative and the inchoative sense of *freeze* (see Example 8.) are connected through the *causes* relation, the respective antonym senses have been collapsed in a single synset: *dissolve:9, thaw:1, unfreeze:1, unthaw:1, dethaw:1, melt:2* (become or cause to become soft or liquid').

Example 8.

Synset (causes): freeze:4; **Gloss**: cause to freeze; **Frame**: *Cause_change_of_phase*

Synset (is caused by): freeze:2; **Gloss**: change to ice; **Frame**: *Change_of_phase*

The lack of the *causes* relation between causative and inchoative senses is well observed, for instance, in the hypernym trees whose roots are *change:1, alter:1, modify:3* ('cause to change; make different; cause a transformation') *causes* > *change:2* ('undergo a change; become different in essence; losing one's or its original nature').

There are a considerable number of hypernym-hyponym pairs (see Table 1) that have been assigned the *Causation* relation. A look at the data shows that these are cases of wrong frame assignment as exemplified in the following case where the causative *boost:2* ('give a boost to; be beneficial to') has been assigned the inchoative frame *Change_position_on_a_scale* instead of the causative frame of the parent synset *increase:2* ('make bigger or more'), i.e. *Cause_change_of_position_on_a_scale*. Such errors in the assignment are commonly found due to the similarity of the formulation of meanings and the common morphological roots of the causative and the inchoative members.

There are 39 correspondences between FN *Causative of* and WN *verb group*, most of which refer to true causative–inchoative pairs which have not been identified as members of the *causes* relation in WN, as in the following example: *corrode:1, eat:6, rust:2* ('cause to deteriorate due to the action of water, air, or an acid'), with the frame *Corroding_cause – corrode:2, rust:1* ('become destroyed by water, air, or a corrosive such as an acid'), with the frame *Corroding*. In these cases, we propose the addition of the more informative *causes* relation between the respective pairs.

The *Inchoativity* relation is very poorly represented in the data so we do not consider it herein.

4.7 See also

See also is a relation that has no direct semantic meaning but rather serves to differentiate frames which are similar and confusable (Ruppenhofer et al., 2016, p. 85, 82). It may be construed in quite different ways, which is reflected in the data, through its mapping to a greater variety of WN relations: *also see* (16 pairs), *antonymy* (8 pairs), *verb group* (22 pairs), *causes* (3 pairs), *hypernymy* (582 pairs). Example 9 illustrates a *See also* relation that corresponds to the WN *also see* relation and denotes an unspecified relation of similarity between the *Placing* and the *Filling* frame, which represent different profilings of a situation.

Example 9.

Also see synset: put:1, set:1, place:1, pose:5; **Gloss**: put into a certain place or abstract location; **Frame**: *Placing*

Also see synset: put on:7, apply:4; **Gloss**: apply to a surface; **Frame**: *Filling*

The greatest part of the synsets with an actual WN relation whose frames are linked by means of *See also* are related through *hypernymy*. A typical case is presented in Example 10.

Example 10.

Hypernym: search:4; **Gloss**: subject to a search; **Frame**: *Scrutiny*

Hyponym: frisk:2; **Gloss**: search as for concealed weapons by running the hands rapidly over the clothing and through the pockets; **Frame**: *Seeking*

The difference between the two frames is stated as one of different primary focus (to the Sought entity or to the Ground)[4]. While this semantic difference is captured by the distinct conceptual structures, it seems to be too fine and does not create a problem in construing *search:4* as the hypernym of *frisk:2*. Judging from the examples of the hypernym–hyponym pairs and the definition of the frames, the same conclusion is valid for many other pairs of frames, such as: *Sound_movement–Make_noise*, *Exchanging–Replacing*, *Cause_motion–Manipulation*, *Worry–Experiencer_focused_emotion*, *Placing–Filling*, *Motion–Ride_vehicle* among others.

In addition, when examining the *See also* pairs we find out that many of them are in fact linked through another, more informative relation, e.g. Using: *Cause_motion–Bringing*, *Motion–Operate_vehicle*; Inheritance: *Motion–Self_motion*, *Deciding–Choosing*; Subframe: *Cause_motion–Placing*, *Cause_motion–Removing*.

5 Implications from the observations

The main conclusions that we can make based on the observations so far are:

(1) The internal structure of FrameNet and WordNet is determined primarily by the notion of inheritance (and several non-inheritance relations). In FrameNet this notion is represented by the relations of *Inheritance* (strong inheritance), *Using* (weak inheritance) and *See also* (an unspecified relation of similarity often construable as inheritance), as well as by relations such as *Subframe*, and *Perspective on*, although in a limited way. WN inheritance is implemented through the *hypernymy-hyponymy* relation. The comparison between the two structures sheds light on the nature of inheritance and hypernymy, especially in the ways it may diverge from the notion of subsumption. Especially interesting are the cases of inverted relations as they may point to errors in assignment or to a variability in semantic construal.

(2) A practical implication from the comparison refers to the insights into the possible ways of perfecting or enhancing the two resources. We have paid special attention to the way FN relations are translated into WN relations. Particularly interesting are cases where relations showing significant similarity in their scope do not correspond in the two resources. Such cases point to peculiarities in the relational structure of the two resources or assignment errors. Inverted relations are also a productive source of information as they point to greater hypernym–hyponym similarity than in straightforward cases and may give clues as to possible collapsing of hierarchical information.

(3) Validation procedures for discovering incorrect assignments of FN frames to WN synsets have been proposed on the basis of discrepancies between the two structures through: (i) identifying incompatible relations in the two resources, e.g. FN *Causative_of* and corresponding hypernym-hyponym pairs; (ii) adding relations based on observations, e.g. adding the *causes* relation between synsets related through *verb_group*; (iii) finding out inaccurately assigned frames by considering pairs of frames not related in FrameNet, but assigned to synsets related through a particular WN relation, e.g. *Cause_motion–Self_motion*, *Cause_to_be_dry–Express_publicly*, etc.

(4) Suggestion of additional groupings (relations) between synsets on the basis of existing relations. The purpose is to make explicit certain relationships that are not captured (systematically) in WordNet, such as the ones between synsets marked as being subframes of a non-lexicalised complex frame or perspectivised frames of a non-lexicalised neutral frame. The suggestion takes a cue from the way in which temporal relationships between subframes are made explicit through the *Precedence* relation. For instance, *fall asleep:1* and *wake up:2, awake:1* are mapped to the *Fall_asleep* and *Waking_up* FN frames and are both subframes of the *Sleep_wake_cycle*. While

[4]Seeking: A Cognizer_agent attempts to find some Sought_entity by examining some Ground. The success or failure of this activity (the Outcome) may be indicated. NB: This frame should be compared to the Scrutiny frame, in which the primary focus is on the Ground; https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Seeking

they are linked through the WN *antonymy* relation, their relationship with synsets representing other subframes of the same scenario remains unaccounted for: *get up:2, turn out:12* (*Getting_up*) and *sleep:1, kip:1, slumber:1* (*Sleep*).

Towards the consistent representation of causativity, we suggest: (a) linking pairs of senses in corresponding causative and inchoative or stative trees, such as the causative and inchoative change trees (the roots synsets are themselves related through the *causes* relation); (b) transferring the *causes* relation to relevant LUs and frames.

(6) The study of the relational structure of the two resources, their overlap and possible improvement has more far-reaching impact with a view to the elaboration of the conceptual structure of verbs undertaken by our team. Based on the properties of the semantic relations in FN and their correlation with hypernymy, we attempt at formulating principles for transferring conceptual information based on the inheritance of features: in particular, configurations of frame elements and imposed selectional restrictions. The observations on *Inheritance* and *Using* are especially useful as they shed light on the specialisation that takes place from parent to child: reducing core frame elements by incorporating one of them in the verb meaning – e.g. *whip:4* incorporates the Instrument of *strike:1*; reducing the scope of the frame – e.g. *drive:1* as a hyponym of *operate:3* applies only to land vehicles; profiling a different frame element – e.g. *rob:1* profiles the Victim, while its hypernym *steal:1; rip off:2, rip:4* profiles the stolen Goods. Among the non-hierarchical relations *Causative_of* and the underrepresented *Inchoative_of* bear importance to the conceptual description as they determine the relations between similar structures with common major frame elements and selectional restrictions. The *See also* relation denotes similarity between conceptual structures that may very well translate as distinctions between similar configurations of frame elements (as in Example 10) or differences between similar (but not identical) sets of frame elements with similar semantic restrictions.

6 Conclusion and future work

The alignment between WordNet and FrameNet at the lexical level (literals within WordNet synsets – lexical units within frames) offers limited coverage and shows some inconsistencies in the representation of semantic relations. The expansion of the coverage relies on: the understanding of the relational structure of both resources and exploring the possibility of identifying the frames relevant to certain synsets (based on inheritance and other semantic relations); defining new frames and synsets in order to provide consistency in the representation of relations, etc. The verification of the resources as well as their alignment and mutual enhancement can be based on automatic consistency checks of inheritance (both strong, e.g. Inheritance, and weak, e.g. relations such as Using, Perspective, Subframe) and on paying special attention to cases with inverted inheritance (frame F_1, assigned to synset S_1, is inherited by frame F_2, assigned to synsets S_2, but within WordNet S_2 is more general than S_1). Further exploration of inheritance can yield more (ir)regularities which may facilitate the enhancement of the resources.

This work is an integral part of our research on defining a conceptual framework for encoding semantic relations between verbs (as represented in synsets) and relevant sets of noun synsets to the end of creating a relationally densely populated semantic network. In particular, the study of inheritance and the remaining relations in FN and how they translate into WN relations enables us: (i) to formulate procedures for exploring the relational structure of the resources towards increasing the coverage of the mapping between FN frames and WN synsets based on these relations; (ii) to define more rigid and clear-cut conceptual classes of verbs on the basis of the enhanced mapping of conceptual frames; (iii) to undertake the building of a rich relational structure through defining relations between verbs belonging to particular frames and sets of nouns with particular semantic properties (as reflected in WN subtrees, ontological categories, etc.) corresponding to key frame elements in the verb's frame. The last task is sensitive to the precision and scope of the conceptual description and is thus dependent on the validation, extension and enhancement of the assignments.

Acknowledgments

This study has been carried out as part of the project *Towards a Semantic Network Enriched with a Variety of Semantic Relations* funded by the National Science Fund of the Republic of Bulgaria under the Fundamental Scientific Research Programme (Grant Agreement 10/3 of 14.12.2016).

References

C. F. Baker and C. Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09), Association for Computational Linguistics, Stroudsburg, PA, USA*, pages 125–129.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.

Christiane Fellbaum, editor. 1999. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. Language Resources and Evaluation. *Commun. ACM*, 42(1):21–40.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon. PhD Thesis*. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

E. Laparra and G. Rigau. 2009. Integrating WordNet and FrameNet using a knowledge-based Word Sense Disambiguation algorithm. In *Proceedings of Recent Advances in Natural Language Processing (RANLP09), Borovets, Bulgaria*, pages 208–213.

E. Laparra and G. Rigau. 2010. eXtended WordFrameNet. In *Proceedings of LREC 2010*, pages 1214–1219.

Svetlozara Leseva, Ivelina Stoyanova, and Maria Todorova. 2018. Classifying Verbs in WordNet by Harnessing Semantic Resources. In *Proceedings of CLIB 2018, Sofia, Bulgaria*.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

M. Palmer, C. Bonial, and D. McCarthy. 2014. SemLink+: FrameNet, VerbNet and Event Ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014), Baltimore, Maryland USA, June 27, 2014*, pages 13–17. Association for Computational Linguistics.

M. Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*. 9–15.

M. R. Petruck and G. de Melo. 2012. Precedes: A semantic relation in FrameNet. In *Proceedings of the Workshop on Language Resources for Public Security Applications*, pages 45–49.

M. R. Petruck. 2015. The Components of FrameNet. `http://naacl.org/naacl-hlt-2015/tutorial-framenet-data/FNComponentsMRLP.pdf`.

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, C. F. Baker, and J. Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springe, Berlin, Heidelbergr.

S. Tonelli and D. Pighin. 2009. New Features for Framenet – Wordnet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09), Boulder, USA*.

S. M. Virk, P. Muller, and J. Conrath. 2016. A Supervised Approach for Enriching the Relational Structure of Frame Semantics in FrameNet . In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan*, pages 3542–3552.

Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia

Kiril Simov, Petya Osenova, Laska Laskova, Ivajlo Radev, Zara Kancheva
IICT-BAS, Sofia, Bulgaria
{kivs|petya|laska|radev|zara}@bultreebank.org

Abstract

The paper reports on an ongoing work that manually maps the Bulgarian WordNet BTB-WN with Bulgarian Wikipedia. The preparatory work of extracting the Wikipedia articles and provisionally relating them to the WordNet lemmas was done automatically. The manual work includes checking of the corresponding senses in both resources as well as the missing ones. The main cases of mapping are considered. The first experiments of mapping about 1000 synsets show the establishement of more than 78 % of exact correspondences and nearly 15 % of new synsets.

1 Introduction

There is still lack of sufficient knowledge for solving many important NLP tasks, such as word sense disambiguation (WSD), relation extraction, named entity linking, event detection, etc. Up to now a number of attempts have been provided in the community that integrate various linguistic and semantic resources in smart ways. These are, among others, SemLink (Palmer, 2009), Predicate Matrix (de Lacalle et al., 2014), UBY (Gurevych et al., 2012), BabelNet (Navigli and Ponzetto, 2012). SemLink combines PropBank (Kingsbury and Palmer, 2002), VerbNet (Kipper-Schuler, 2005), and FrameNet (Baker, 2008). Predicate Matrix extends SemLink with a mapping from its lexical units to WordNet synsets (Fellbaum, 1998). UBY was created for two languages — English and German. It combines WordNet and GermaNet with Wiktionary, Wikipedia, FrameNet and VerbNet for English, and Wiktionary and Wikipedia for German. BabelNet also combines many multilingual resources including WordNet and Wikipedia. All these examples demonstrate two facts: (1) a single knowledge resource is not sufficient for the most of the NLP tasks; and (2) the automatic integration of the various distinct resources is error prone. This is especially true for low-resource languages that totally miss such resources or their existing resources are rather small in size.

Here we report on an effort to integrate Bulgarian WordNet (BTB-WN) (Osenova and Simov, 2018) with the Bulgarian Wikipedia. We are considering mapping of two semantic objects — *concepts* (meaning expressed by common words) and instances of such concepts called *named entities*. The integration is meant to be performed manually in order to ensure high quality of the result. The integrated knowledge graph will include the current version of BTB-WN extended with: a) new senses and new synonyms for the existing synsets — all extracted from the articles in the Bulgarian Wikipedia; b) a controlled number of named entities that are specific to Bulgaria and c) increasing the number of terminological concepts in various domains. Thus the integrated resource will combine general lexica with encyclopedic knowledge (terminology).

The expected result would be twofold: a) the mutual enrichment and improvement of both resources and b) handling of WSD in a more effective way by integrating the encyclopedic knowledge from Wikipedia and the lexical information from WordNet.

The structure of the paper is as follows: in the next section related work is presented. Section 3 outlines the approach to the mapping as well as the results. The last section concludes the paper.

2 Related Work

Needless to say, one of the most notable resources that link WordNet and Wikipedia is BabelNet — an automatically created very large, wide-coverage multilingual semantic network (Navigli and Ponzetto, 2012). BabelNet encodes knowl-

edge as a labeled directed graph. It is created by linking the largest multilingual Web encyclopedia – Wikipedia, to the most popular computational lexicon — WordNet. BabelNet has been built in 3 steps. The first step was to automatically combine WordNet and Wikipedia by mapping the WordNet senses to Wikipedia articles. The second step was to collect the multilingual lexicalizations of the BabelNet synsets by using human-generated translations. These translations were provided by Wikipedia as well as by a machine translation system for translating the occurrences of the concepts within sense-tagged corpora. The third step was to establish relations between the Babel synsets through collecting all the relations found in WordNet together with all Wikipedias in the languages of interest. The integration was performed by an automatic mapping and by filling the lexical gaps in resource-poor languages with the aid of Machine Translation. The result is an "encyclopedic dictionary" that provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations.

In spite of having at disposal such a resource as BabelNet, our motivation to invest efforts in mapping the WordNet to the Wikipedia was as follows: a) adding more locally important content into the existing mappings and b) enriching the resource that was constructed automatically with validated data. The Babelfy service is very good at detecting concepts and names (given the availability of relevant data per language), but it still has problems with disambiguation among local people or places with the same name, or between a concept and a name. For example, the verb литва (litva, "start to fly") is identified only as the country Литва (Litva, "Lithuania") whose graphical form coincides with the verb; similar for the adjective русия (rusiya, "blond") and the name of the country Русия (Rusiya, "Russia").

In (Osenova and Simov, 2018) Osenova and Simov mention the initial attempt for annotating of named entities (NE) in Bulgarian Treebank (BulTreeBank) with URIs from DBpedia. This process was done with the same goal, namely to extend BTB-WN in two directions: (1) the number of senses for lemmas that are already in BTB-WN; and (2) the instances of the concepts. However, the BulTreeBank appeared to contain only a small number of named entities. Thus the extension was insufficient and it required the use of the Wikipedia URIs and DBpedia classes for the missing NEs. The authors also report on the automatic extension of BTB-WN with automatically derived Bulgarian synsets on the basis of the English ones through the usage of the English Wiktionary. After manual checking of around 11000 suggestions, BTB-WN was enriched by around 5000 synsets.

(McCrae, 2018) reports on the manual mapping of the Princeton WordNet (PWN) instances to the English Wikipedia. He proposes that a subset of PWN instance concept synsets is automatically linked and manually evaluated on Wikipedia articles in order to "provide a gold standard for link discovery". This is done by matching PWN lemmas to all Wikipedia titles containing the lemma. Then by using a special tool, human annotators evaluate the links. This tool shows the PWN definitions and the first paragraph from the Wikipedia article so the annotators are able to confirm or reject the mapping. The same paper also suggest 5 types of links between PWN and Wikipedia: exact — one synset to one article; broad — several synsets to one article; narrow — one synset to several articles; related — one-to-one relation, but not the same concept; unmapped — not possible to map. This method proved to be highly successful and even yielded a report with 8 errors which aimed to improve PWN. We follow very closely the approach of this work except that we are interested in mapping not only the instances, but all possible lemmas in BTB-WN.

(Rudnicka et al., 2017) present another attempt at linking two large lexico-semantic databases, namely the Princeton WordNet of English and the plWordnet of Polish language. The approach considers models and ideas originating from the bilingual lexicography and translation studies. For the creation of the PlWordnet language data from contexts of use attested in large language corpora was used rather than from dictionaries and the approach focused on word uses, not concepts.

A synset in PWN is viewed as a representation of a lexicalised concept, while in plWordNet it is a set of lexical units sharing constitutive lexico-semantic relations and features. The synset includes such lexical units that share a set of lexico-semantic relations, called constitutive relations (hyper/hyponymy, holo/meronymy, type/instance, etc.). In some cases the constitutive relations might be irrelevant, so constitutive features are also used – stylistic register, aspect,

semantic classes of verbs and semantic classes of adjectives. Glosses, examples and substitution tests are also applied in the plWordnet. The mapping strategy refers to the synset level and includes looking for pairs of plWordnet and PWN synsets that are close in meaning. The stages of the mapping are as follows: an analysis of the sense and relation structure of a source synset, the selection of candidates for a target synset, the choice of a target synset and an inter-lingual relation that links the source and target synsets. Having in mind the complex schema of mapping between the two WordNets we doubt that such a mapping could be successfully established between the WordNet and the Wikipedia even for the one and the same language. Expectedly, when named entities are highly predominant in the mapping, we might envisage also a high number of exact mappings, but for common words this is not so straightforward. For that reason, we decided to perform the mapping manually. For the first step our goal was to extend BTB-WN with new synsets, synonyms and mappings to Bulgarian Wikipedia.

Another approach that could be taken into account when aiming to extend the WordNet is its alignment with a FrameNet (if such a resource has been constructed for the language). A recent and rather innovative example of the development of a FrameNet based on a corpus of written Dutch, and annotated with PropBank predicates and roles is the project of (Vossen et al., 2018). In this project the creation of the FrameNet also exploits already manually classified data about real world events which specify frame constraints on the described situations. This data is manually related to texts describing the events. In future work we will consider this approach to extend the coverage of BTB-WN as well as to add new constraints on the combinations of the senses within texts.

3 BTB-WN to Wikipedia Mapping

In this section we present the correspondences between the synsets within BTB-WN and the pages from the Bulgarian Wikipedia.

3.1 Wikipedia Page to Synset Correspondence

The first step was to establish a correspondence between lexical entries in BTB-WN and the Bulgarian Wikipedia. For each lemma within BTB-WN we automatically selected all the articles in Wikipedia that match that lemma. In order to do this, the article titles were cleaned from the modifiers given in brackets like in the following example: the lemma маса (masa) corresponding to "table" (a piece of furniture), "mass" (a body of matter), and "mob" (a disorderly crowd of people) is mapped to Wikipedia articles with titles like: Маса "Mass" (physical term); Маса (мебел) "Table (furniture)", etc. The special disambiguation articles play an important role like Маса (пояснение) in this example. Their importance comes from the fact that they provide additional information about the potential synonyms. Such an example in this case is the connection from Маса to Заземяване "Ground (electricity)" which was a missing sense within the current version of BTB-WN. For each Bulgarian Wikipedia article we also extracted the title of the corresponding English article in order to facilitate the process of selecting the right meaning and the process of mapping between BTB-WN and the English WordNet.

After the extraction of the relevant Wikipedia pages we grouped together the pages corresponding to a given lemma and all the BTB-WN synsets that contain the lemma. These groups have been represented in XML and loaded into CLaRK System[1] for inspection and mapping. A screen shot of the data loaded in the system is presented in Fig. 1. Each group is represented via `<eq>` element. In the representation we use the tree layout settings of the system in order to present not only the structure elements but also their content. Each group contains one or more pages, thus one or more entries for the same lemma. If an entry contains more than one lemma, this entry will be added to several groups if there are appropriate Wikipedia pages. In the figure we can observe two expanded groups — one for the lemma "Iceberg" and one for the lemma "Aquarium". For each page the layout shows the Bulgarian title of the page, then the English title (if there is a link to an English Wikipedia page). Thus, the annotator[2] could understand the sense described by the Wikipedia article without expanding the structure of the page. Of course, if necessary, the annotator could read more from the content of the page. For each entry

[1]For a description see (Simov et al., 2004b). The system could be downloaded from `http://bultreebank.org/en/clark/`.

[2]We call the people that manually establish the mapping between the two resources *annotators*, but a more appropriate term is necessary such as *mappers* or *knowledge relaters*.

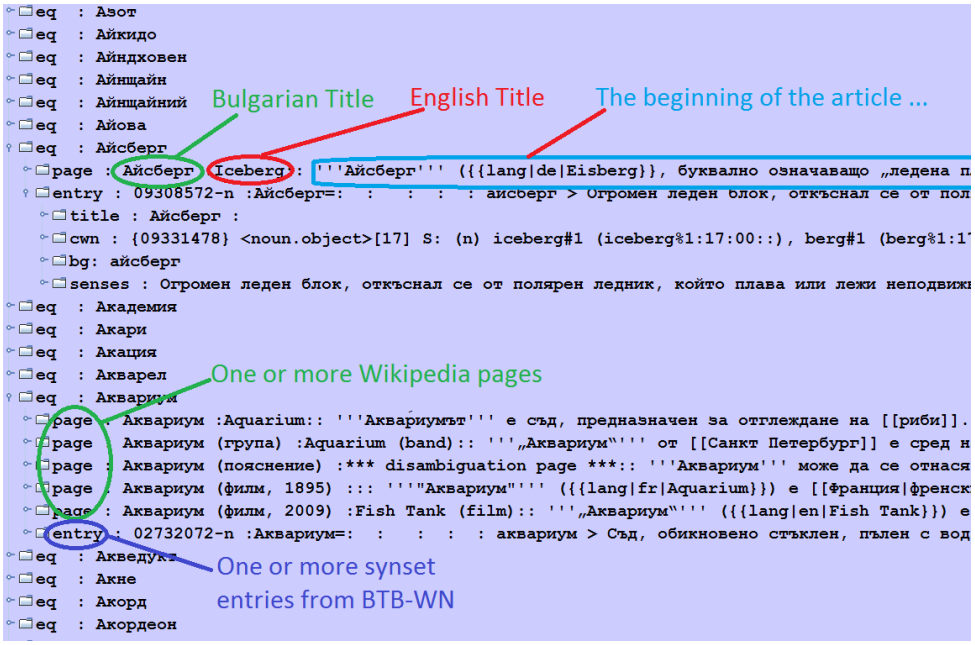


Figure 1: Representation of the groups of matched Wikipedia pages and BTB-WN synsets (represented via <entry> element.)

the layout shows the PWN identifier; the mapping to Wikipedia page (if such has been selected); the list of lemmas for the synset; and finally the definition related to the synset. Again, the annotator might read the important information without expanding the structure of the entry. In the example of the group for "Iceberg" the structure of an entry is as follows. The element <cwn> contains the mapping information to PWN. The element <bg> contains the list of lemmas of the synset. The element <senses> contains one or more definitions (if selected from different sources) and zero or more examples of uses of the lemmas in the corresponding sense.

The group for "Iceberg" represents the simplest case of one-to-one mapping. The actual connection is established by copying the title of the appropriate Wikipedia page as a first element of the entry. The group for "Aquarium" demonstrates the case when more than one Wikipedia page corresponds to a given lemma. Here we have a page corresponding exactly to an entry in BTB-WN. Several pages exist for named entities like a band, two movies – one French and one British. Also there is a disambiguation page, marked with "*** disambiguation page ***". In cases of disambiguation page we also added the pages that are mentioned within the disambiguation list. Similarly, we add the redirect pages pointing to some of the other pages within the group. In some cases such redirect pages provide synonyms or derivative lemmas. In this way we try to provide as much information as possible from the Bulgarian Wikipedia to the annotator.

Following the mapping strategy, mentioned above, for about 22 000 synsets in BTB-WN we extracted a little more than 13 000 Wikipedia articles. For each sense (sense in BTB-WN is defined as a lemma in some of the synsets) in BTB-WN the annotators received a list of the corresponding Wikipedia articles. Thus they were able to check whether the selected sense is presented within Wikipedia and to establish correspondence if it is the case. After consulting the individual senses in BTB-WN, the annotators checked whether new meanings had to be added to it. The new meaning could be a sense for the common word or a named entity. In both cases the annotator created a new lexical entry in BTB-WN.

3.2 Named Entities Processing

Because of the high productivity in the case of named entities, many common words are presented as named entities in Wikipedia. Since our main goal was to introduce more locally centered names, these respectively were considered as important. Thus, the annotator first filtered the candidates in order to introduce only the important names. More specifically, we defined names of importance in the following way:

- As a first step, only names of persons, organizations and locations are considered;
- For location names we select names of Bulgarian places or of well-known foreign places;
- For the rest of the names only well-known names are considered.

Although this definition is not very precise, it helped us to filter quite a lot of location named entities. Here we additionally introduced a restriction to include larger cities in Europe (larger than 100 000 citizens if they are not well-known). In this way for example, Шенген ("Schengen") is included in BTB-WN although it has less than 4000 citizens, but Буден ("Boden", a city in Sweden) is not included although its transliteration in Bulgarian coincides with an adjective. In our future work we need to make the definition more precise in order to cover all the names in Wikipedia, but without overloading the WordNet with the ambiguity coming from very rare named entities.

The above selection criteria are to some extent arbitrary[3]. For example, for some countries the limit of 100 000 citizens is too restrictive. Especially for small countries or countries in Europe. For other countries this might allow many not well known cities. In order to provide an additional evaluation of the importance of the named entities, we use a gazetteer created during the development of the BulTreeBank Pipeline for Bulgarian — see (Simov et al., 2004a) and (Savkov et al., 2012) and during the compilation of the Bulgarian treebank (2001-2004). The names in it were collected from the following sources: (1) Bulgarian law documents containing the names of all villages, towns, cities, municipalities in Bulgaria; (2) Names from touristic advertisements; and (3) list of names manually selected from a ranking list of potential named entities from a large corpus of Bulgarian. We consider the names in the gazetteer as representative for Bulgarian texts. They also contains all Bulgarian location names. The gazetteer contains more than 26 000 records, but some of them are not basic forms (lemmas) because during the preparation of the gazetteer we selected non-basic forms like vocatives, plurals and definite forms.

All the Wikipedia pages were extracted that correspond to the names in the gazetteer. We extracted 10 899 pages altogether. From them 1 515 pages were already extracted on the basis of the lemmas within BTB-WN. Thus we marked there 1 515 as important, but still the annotators could select names that are marked in this way. The rest 9 384 pages were classified as Bulgarian locations, other locations, people, organizations and other. They will be checked for inclusion in BTB-WN at a later stage. In this way we selected also some important names that are not considered at the beginning of this work.

3.3 Mapping Cases

Here we consider different cases of correspondence among pages and entries, grouped together on the basis of the lemmas from BTB-WN. Each annotator was instructed to check the aligned WordNet synsets and the Wikipedia articles for the following cases:

- Exact mapping of senses represented in both resources;
- A concept represented in Wikipedia, but not in WordNet. In such a case they had to create a new synset and to establish a mapping;
- An admissible named entity in Wikipedia, missing in WordNet. In such a case they had to create a new synset and to establish a mapping.

Whenever a new synset was created, it was also mapped to the corresponding PWN synset when possible (for more details see (Osenova and Simov, 2018)). The annotation was performed by 5 people that considered nearly 1000 WordNet lemmas, automatically mapped to more than 1300 Wikipedia articles. Table 1 presents the distribution of the different cases.

The first category (first line — None) contains the number of no correspondences between the

[3] As it was pointed to us by one of the reviewers.

Correspondence	Number Total: 1309	%
None	276	21.08
Equality	688	52.57
Many to One	128	09.78
New Concept	128	09.78
New Named Entity	68	05.19
New Synonyms	21	01.60

Table 1: Percentage of the different cases.

two resources. In this case none of the Wikipedia articles describes a synset in BTB-WN. The reason for this usually is the named-entity-centered nature of Wikipedia. For example, under the title Плейбой "Playboy" Wikipedia has only one article on the Playboy journal. In BTB-WN there is an entry corresponding to PWN synset with a gloss "a man devoted to the pursuit of pleasure." The closest page in English Wikipedia is "Playboy lifestyle" which requires a more complex mapping. Such a page is missing in the Bulgarian Wikipedia. Similarly, the word Стожер (stozher) in the Wikipedia is only a name of a village and a newspaper, while WordNet records only the concept стожер (stozher) as pillar. Thus, the WordNet entity cannot be mapped to Wikipedia. This case corresponds to McCrae's *Unmapped links*.

The second category (Equality) describes the equality relation, where both resources describe the same concept. For example, Столица (stolitsa), "capital" is defined in the same way in both resources. These cases are the majority of all mappings. It corresponds to McCrae's *Exact links*.

The third category (Many-to-One) presents the case where different parts of the same Wikipedia article are dedicated to different concepts. Often, but not always, this is the case for the disambiguation pages. Among the concepts, one usually corresponds to the mapped WordNet synset. For example, in Wikipedia, Стойка (Stoyka) has several representations as a given name or a surname, but it also refers to the concept of (body) posture and the concept of stand. BTB-WN contains only one concept — that of the posture. Another problem in this case is that the two pages for these general concepts do not exist, but they are defined only in the disambiguation page. Thus, the annotator has to use a special relation to the disambiguation page. The Индекс (indeks), "index" page illustrates another example that is treated in a similar manner. In this case, the authors of the Wikipedia page point out that the word индекс might refer to several things, among which a list of items, a superscript or subscript character, a hierarchical classifier, and a value on a measurement scale. Two of these concepts are lexicalized as индекс in the WordNet and they are mapped to the article with a Many-to-One relation, which corresponds to McCrae's *Broad links*.

In both cases, the annotator has to perform one more operation before moving on, that is, to check whether the BTB-WN does not already contain the seemingly missing concepts; it might be the case that they are lexicalized in a different way, i.e. in other terms. Here, the annotators rely on information from Wikipedia, and, of course, on their own linguistic competence. Whenever deemed necessary, and especially when dealing with terminological units, they consult a synonym dictionary or a thesaurus. Needless to say, there would be two possibilities: a) the right match is found, or b) not found, because it is missing. In the Стойка example, the concept for "stand" was already present in the WordNet, so the annotator established a Many-to-One correspondence between the article and the synset, and added the term стойка to the set of synonyms. In the Индекс example, the new concepts found in Wikipedia were indeed missing from the BTB-WN and thus the annotator created two new synsets mapping them to the article with a Many-to-One relation.

In some cases, the new concept introduced in the Wikipedia article, is given only a short definition and the term is linked to an empty page. Given the dynamic nature of Wikipedia, we decided to map this type of pages to the corresponding BTB-WN synsets with an additional *empty* relation; from here we can expect one of the two positive outcomes — on the one hand, the annotators are free to contribute to the Bulgarian Wikipedia by providing new content (a time-consuming task which at this point is given a low priority), an on the other hand, we keep the possibility of future resource enrichment by not excluding a potentially useful mapping.

The fourth and the fifth categories (New Concept and New Named Entity) correspond to the case in which the Wikipedia article introduces one or more new concepts — both types or instances. We can distinguish several cases here.

The Wikipedia article lists some or all of the hy-

ponyms of the concept named in the title. For example, Абак (abak), "abacus", contains information about the different types of abacuses. Each of these types prompts the creation of a new synset. In this cases we reuse the definitions from the Wikipedia article. We also select examples from the article. This allows for BTB-WN to be used independently from Wikipedia.

The Wikipedia article is dedicated to different concepts which are not linked by a hypernym — hyponymy relation. This type of relatedness corresponds to McCrea's understanding of *Related links*. The nature of the relatedness remains unspecified but the new concept is always linked to some existing one in the WordNet: through homonymy, derivation, systematic polysemy, semantic expansion, etc. Let us give some examples.

- The article Авария (avariya) describes a technogenic disaster. It is related to the synset авария, катастрофа (breakdown, equipment failure) by a causal link.

- The article Инвалидност (invalidnost), disability is related to the synset инвалид (invalid), "disabled person" derivationally. Here we annotate the mapping as derivational, but in future we will add more specific relations depending on the semantic relation.

- As for the systematic polysemy, two are the most common types.
  The first one regards the relation between a title understood as "an identifying appellation signifying status or function", and the person who is given this title because they have the corresponding status or function. As a rule, the Wikipedia article describes the title while the existing WordNet concept is related to the person. The annotators create a new synset linked to the page with an Equality relation and also indicate the specific type of relatedness between the preexisting synset and the page.
  The second type of systematic polysemy is characteristic of some geographical named entities, such as Бахамски острови (Bahamski ostrovi, Bahamas). This multiword expression has two meanings. It can refer to the country, the Commonwealth of the Bahamas, or to a geographical region, in this case the island group known as Lucayan Archipelago. The annotators apply the same strategy as the one described above.

The Wikipedia article introduces a hypernym. For example, Камион (kamion), "truck" in Wikipedia is a hyperonym of the two synsets for truck and van, presented within the current version of BTB-WN.

The sixth category (New Synonyms), features the case when the corresponding synset is part of the WordNet, but there are some missing synonyms that come from the Wikipedia. For example, the multiword expression Кралство Камбоджа, "Kingdom of Cambodia" is missing in the synset that contains the name of Cambodia.

As it can be seen, in more than 78 % of the cases we establish a correspondence between synsets in BTB-WN and the Bulgarian Wikipedia. In our view this is a good coverage. Also we have added about 15 % new concepts and named entities.

4 Conclusion

The paper presents our initial attempts in enriching BTB-WN with mappings to the Bulgarian Wikipedia. The first annotation results are promising in showing that WordNet profits well from this mapping — especially in adding synonyms, new senses and new instances.

The importance of such a resource is envisaged at least in the following directions: enhancing named entity linking, relation extraction and word sense disambiguation of high quality for tasks, involving Bulgarian data. The mapping also provides access to the whole Wikipedia articles which could contribute valuable information for the usage of the corresponding concepts and named entities.

The main source of enriching BTB-WN appeared to be the named entities and the domain terms. We also noticed that Wikipedia is a valuable resource for including MWEs — predominantly terminological units, but not only. Since the named entities are too many, as mentioned above, we focused on local ones because they are important for processing Bulgarian data, and also — they can be viewed as a valuable localized supplementary contribution to BabelNet.

Another issue is that Wikipedia contains mainly nouns. Thus, the mappings enriched the noun network and the instances of names. For the verbs, adjectives and adverbs other enriching sources

should be considered. Through the derivation relations in WordNet, however, we still could incorporate the presented in Wikipedia deverbal and adjectival nouns.

In future work we envisage to map BTB-WN also to other semantic resources such as Wikidata. We have started with Wikipedia because it provides more human oriented information which facilitates the mapping. In addition, Wikidata is heavily extracted from Wikipedia and we hope this to allow for an easy mapping.

In the long run, we envisage also incorporating more Bulgarian concepts and named entities with the idea to construct a Bulgarian knowledge graph aligned to linguistic knowledge — senses and grammatical features.

Acknowledgements

This research was funded by the Bulgarian National Science Fund grant number 02/12/2016 — *Deep Models of Semantic Knowledge (DemoSem)*. The contribution of Ivajlo Radev and Zara Kancheva has been partially supported by the Bulgarian Ministry of Education and Science under the National Research Programme “Young scientists and postdoctoral students" approved by DCM # 577 / 17.08.2018. We are grateful to the anonymous reviewers for their valuable remarks, comments, and suggestions. All errors remain our own responsibility.

References

Collin Baker. 2008. FrameNet, present and future. In Jonathan Webster, Nancy Ide, and Alex Chengyu Fang, editors, *The First International Conference on Global Interoperability for Language Resources*, Hong Kong. City University, City University.

Maddalen Lopez de Lacalle, Egoitz Laparra, and German Rigau. 2014. Predicate matrix: extending semlink through wordnet mappings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 903–909, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France, April. Association for Computational Linguistics.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

John P. McCrae. 2018. Mapping WordNet Instances to Wikipedia. In *In: Proceedings of Ninth Global WordNet Conference*, pages 62–69. The Global WordNet Association.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and applicationof a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Petya Osenova and Kiril Simov. 2018. The Data-driven Bulgarian WordNet: BTBWN. *Cognitive Studies | Études cognitives*, 18(1713).

Martha Palmer. 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, page 9–15.

Ewa Katarzyna Rudnicka, Maciej Tomasz Piasecki, Tadeusz Piotrowski, Łukasz Grabowski, and Francis Bond. 2017. Mapping WordNets from the perspective of inter-lingual equivalence. *Cognitive Studies | Études cognitives*, 17(1373):1–17.

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova, and Kiril Simov. 2012. Linguistic Processing Pipeline for Bulgarian. In *Proceedings of LREC 2012*, pages 2959–2964.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, and Dimitar Doikoff. 2004a. A Language Resources Infrastructure for Bulgarian. In *Proceedings of LREC 2004*, pages 1685–1688.

Kiril Simov, Alexander Simov, Hristo Ganev, Krasimira Ivanova, and Ilko Grigorov. 2004b. The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004*, pages 235–238.

Piek Vossen, Antske Fokkens, Isa Maks, and Chantal Van Son. 2018. Open Dutch Framenet. In Tiago Timponi Torrent, Lars Borin, and Collin F. Baker, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).

English-Turkish Parallel Semantic Annotation of Penn-Treebank

Bilge Nas Arıcan
Starlang Yazılım Danışmanlık, Turkey
bnarican@gmail.com

Özge Bakay
Boğaziçi University, Turkey
ozge.bakay@boun.edu.tr

Begüm Avar
Boğaziçi University, Turkey
begum.avar@boun.edu.tr

Olcay Taner Yıldız
Işık University, Turkey
olcaytaner@isikun.edu.tr

Özlem Ergelen
Boğaziçi University, Turkey
ozlem.ergelen@boun.edu.tr

Abstract

This paper reports our efforts in constructing a sense-labeled English-Turkish parallel corpus using the traditional method of manual tagging. We tagged a pre-built parallel treebank which was translated from the Penn Treebank corpus. This approach allowed us to generate a resource combining syntactic and semantic information. We provide statistics about the corpus itself as well as information regarding its development process.

1 Introduction

Parallel corpora, which are a collection of texts in one language and their translations in at least one other, can be used in a variety of fields, such as translation studies and contrastive linguistics. They are used for many different purposes including creating new linguistic resources such as lexicons and WordNet (Petrolito and Bond, 2014). As for the relationship between parallel corpora and natural language processing (NLP) studies, in addition to the fact NLP studies use parallel corpora as material bases or testing arenas, NLP studies also contribute to the development of corpora in many areas, especially in corpus annotation.

In this paper, we present a sense-tagged English-Turkish parallel corpus, which is the only corpus for the English-Turkish combination having both semantic and syntactic information. It has been built on the preceding parallel treebank construction and morphological analysis efforts reported in (Yildiz et al., 2014) and (Gorgun et al., 2016). The aim of this study is to investigate the possibility of a parallel semantic annotation for an English-Turkish corpus. The motivation behind the study is the potential contribution of this parallel semantic annotation to several NLP tasks such as automatic annotation, statistical machine translation and word sense disambiguation.

This paper is organized as follows: We give background information about lexical semantics in Section 2 and present the related work in Section 3. The details of our corpus and how it is constructed are given in Section 4. We provide the annotation statistics about the corpus in Section 5 and conclude in Section 6.

2 Lexical Semantics

In linguistics, lexical semantics is the study of word meaning. The main challenge in this field is generated from 'polysemy', which is the term used for the phenomenon of a single orthographic word having multiple, interrelated senses. In classical dictionaries, these senses are listed under a single lexical entry and, as stated in (Firth, 1957), "You shall know a word by the company it keeps", that is, only with the help of the context one can pin down the particular sense in which a word is used. A further challenge in the field stems from collations, i.e. groups of words having "a unitary meaning which does not correspond to the compositional meaning of their parts" (Saeed, 1997).

Hence, as far as compositionality is considered to be crucial to semantic analysis, there are two central concerns for the semanticist: (i) At the lexical level, choosing the correct sense of a given word within a context, and (ii) at the sentence level, determining how a particular combination of words should be interpreted.

Languages also differ in terms of how lexical items are combined, which is directly related to how compositionality is to be interpreted. Therefore, the success and adequacy of a multi-

lingual semantic analysis not only requires taking "into consideration the multitude of different senses of words across languages", but also "effective mechanisms that allow for the linking of extended word senses in diverging polysemy patterns" (Boas, 2005).

When it comes to interlingual semantics studies, even further complications arise. For one, there is a huge discrepancy between languages in terms of which semantic components they lexicalize. For instance, in analytic languages like English, functional morphemes are free forms, such as determiners and appositions, whereas in agglutinative languages, such as Turkish, syntactic relations are expressed mainly via affixation. Hence, a single orthographic word in Turkish may correspond to a phrase consisting of a combination of multiple free morphemes in English.

3 Related Work

In this section, we present previous work and provide a comparison of our corpus with other corpora mainly with reference to their sense annotation process and the number of annotated words.

3.1 English Semantically-Annotated Corpora

Among many corpora concentrated on English is SemCor (Miller et al., 1993), which is the most widely-used and largest sense-tagged English corpus with 192,639 instances. SemCor's input comes from the novel of The Red Badge of Courage and the Brown corpus, which presents one million words in contemporary American English obtained from various sources. As for the word-sense mappings, they were done based on WordNet entries.

Another significant study in this area is the line-hard-serve corpus (Leacock et al., 1993). Having extracted its data from three different resources, it is comprised of 4,000 sense-tagged examples of each of the words line (noun), hard (adjective), and serve (verb), which are also mapped with their WordNet senses.

Table 1 shows the English partition of our corpus in comparison with the other English sense-tagged corpora. Our English corpus can be considered as a noteworthy example in terms of its target, the number of annotated words and the version of WordNet used. Having all words annotated by using the latest version of WordNet (WN 3.1), our corpus annotates 41,986 words in total.

3.2 Multilingual Semantically-Annotated Corpora

Among interlingual studies aligned with SemCor, there is the English/Italian parallel corpus called MultiSemCor (Bentivogli et al., 2005), which is aligned at the word level and annotated with PoS, lemma and word sense. Their corpus contains around 120,000 English words annotated, approximately 93,000 of which are transferred to Italian and annotated with Italian word senses. Another important project is by (Lupu et al., 2005). Targeting all words to be annotated, their corpus, SemCor-En/Ro, contains around 48,000 tagged words in Romanian.

The comparison of our multilingual corpus with the other multilingual sense-tagged corpora is given in Table 2. Our corpus is notable when compared to the other corpora for three main reasons; first, it uses the latest version of WordNet (WN 3.1) unlike many other multilingual corpora; second, the total number of words annotated for both languages in our corpus is substantial for a preliminary work; third, it is the first parallel semantically annotated corpus for English-Turkish language pair.

3.3 Turkish Semantically-Annotated Corpora

METU-Sabanci Turkish Treebank (Oflazer et al., 2003), which is a parsed, morphologically-analyzed and disambiguated treebank of 6,930 sentences, is a substantial corpus for Turkish. The sentences were extracted from the METU Turkish corpus, which is a compilation of 2 million words from written Turkish samples gathered from several resources (Say et al., 2002). In these sentences, 5,356 lemmas are annotated, with 627 of them having at least 15 occurences.

Another exemplary corpus for Turkish is the Turkish Lexical Sample Dataset (TLSD) (İlgen et al., 2012). It includes noun and verb sets and both sets have 15 words each with high polysemy degree. An important strength of this corpus is that each word has at least 100 samples which were gathered from various Turkish websites and encoded with the senses of TDK (the Turkish Language Institution's dictionary) by human interpreters.

Our Turkish corpus, on the other hand, is prominent among the current Turkish corpora. As Table 3 suggests, it is the only Turkish corpus both an-

Table 1: Comparison of English sense-annotated corpora

Corpus	# Words Tagged	WordNet	Target
SemCor3.0-all (Miller et al., 1993)	192,639	WN 3.0	all
SemCor3.0-verbs (Miller et al., 1993)	41,497	WN 3.0	verbs
Gloss Corpus (Miller et al., 1993)	449,355	WN 3.0	some
Line-hard-serve (Leacock et al., 1993)	4,000	WN 1.5	some
DSO corpus (Ng and Lee, 1996)	192,800	WN 1.5	nouns, verbs
Senseval 3 (Snyder and Palmer, 2005)	2,212	WN 1.7.1	all
MASC (Ide, 2012)	100,000	WN 3.0	verbs
SemEval-2013 Task 13 (Jurgens and Klapaftis, 2013)	5,000	WN 3.1	nouns
Our corpus	41,986	WN 3.1	all

Table 2: Comparison of multilingual sense-annotated corpora

Corpus	# Words Tagged	Languages	WordNet	Target
MultiSemCor	92,420	Italian	MultiWN	all
(Bentivogli et al., 2005)	119,802	English	WN 1.6	
SemCor-En/Ro	48,392	Romanian	BalkaNet	all
(Lupu et al., 2005)	n/a	English	WN 2.0	
NTU-MC	36,173; 27,796	Chinese; Indonesian	COW; WN Bahasa	all
(Tan and Bond, 2012)	15,395; 51,147	Japanese; English	Jpn WN; PWN	
SemEval-2013 Task 12	3,000; 3,000	French; Spanish	BabelNet	all
(Navigli et al., 2013)	3,000; 4,000	German; Italian		
Our corpus	61,127; 41,986	Turkish; English	KeNet 1.0; WN 3.1	all

Table 3: Comparison of Turkish sense-annotated corpora

Corpus	# Words Tagged	# Lemma	Target	Syntactic Parse
SemEval-2007 (Orhan et al., 2007)	5,385	26	noun; verbs	Available
TLSD (İlgen et al., 2012)	3,616	35	noun; verbs	Unavailable
Our corpus	61,127	7,017	all	Available

notating all words and providing their syntactic information and it annotates by far the largest number of words in total, 61,127. Second, it is also the only Turkish corpus which is parallel annotated.

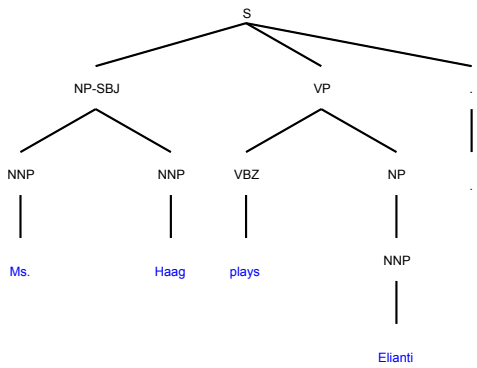
4 Corpus

In this section, we describe how the data in our corpus were extracted and organized, give details of our annotation tool, explain how the data in both Turkish and English partitions were annotated, give an account of our data format, and finally, evaluate our annotation.

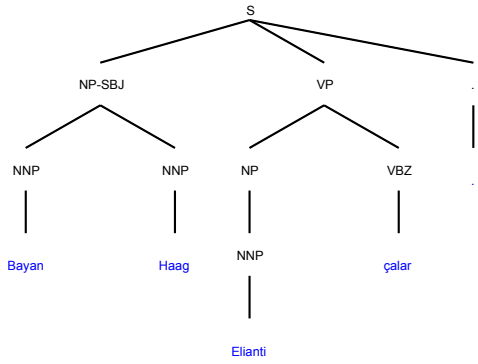
4.1 Preliminary Corpus

As a preliminary work for our corpus, we disambiguated the Turkish-English parallel Treebank (Yildiz et al., 2014) where the English parse trees were converted into their equivalent Turkish parse trees with the application of several transformation heuristics. First, the subtrees were permuted with reference to the Turkish sentence structure rules. Then, leaf tokens were replaced with the most synonymous Turkish counterparts. Finally, an output which was both translated and syntactically-parsed was formed.

Regarding the differences related to syntax, one should note that the majority of Turkish sentences have the Subject-Object-Verb word order whereas most English sentences have Subject-Verb-Object order. When translating English trees, they permute its subtrees to reflect the change of constituent order in Turkish. For example, when translating the sentence in Figure 1(a), VBZ and NP subtrees are exchanged so that the correct con-



(a)



(b)

Figure 1: An example English sentence from Penn-Treebank corpus (a) and its translated form (b)

stituent order in Turkish is constructed in the translated form (Figure 1(b)).

They also use the *NONE* tag when they cannot use any direct gloss for an English token. The semantic aspects expressed by prepositions, modals, particles and verb tenses in English in general correspond to specific morphemes attached to the corresponding word stem in Turkish. By using *NONE* tag, permuting the nodes and choosing the full inflected forms of the glosses in the Turkish tree, they have a working method to convert subtrees to an inflected word.

Following the translation phase, the corpus has been improved with morphological annotations to use in tree-based statistical machine translation (Gorgun et al., 2016). In that work, human annotators selected the correct morphological parse from multiple possible analyses returned from the

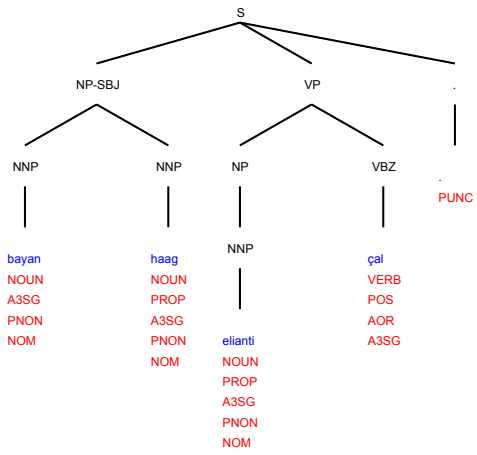


Figure 2: Morphologically-disambiguated form of the sentence in Figure 1(a)

automatic parser. The tag set and morphological representation were quoted from the study reported in (Oflazer et al., 2003). Each output of the parser comprises the root of the word, its part-of-speech tag and a set of its morphemes, each separated with a "+" sign. Figure 2 illustrates the morphologically disambiguated form of the sentence in Figure 1(a).

4.2 Annotation Tool

The annotators use a custom application (written in Java) for browsing sentences and annotating them with senses. The toolkit is freely available [1]. The current implementation of the application is designed for the import of text files that adhere to the Penn Treebank data format (that is, translated and morphologically analyzed).

Once a pre-processed sentence has been imported into the semantic editor, human annotators are presented with the visualized syntactic parse tree of that sentence. Annotators can click on leaf nodes, which correspond to the words. When a word is selected, a drop-down list is displayed, in which all the available WordNet entries of the selected lemma are listed. Figure 3 shows a screenshot from the system interface, depicting the screen presented to the annotators when annotating the verb "çalar" in the Turkish sentence "Bayan Haag Elianti çalar." Right after the selection of the most appropriate sense, the drop-down

[1] https://github.com/olcaytaner/DataCollector

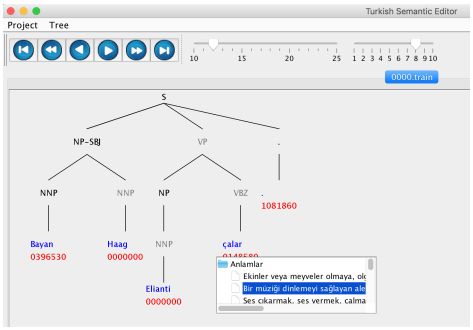


Figure 3: A screenshot from the system interface

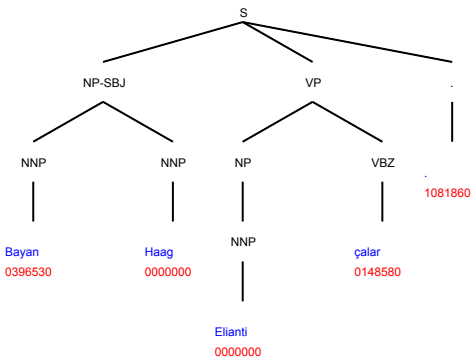


Figure 4: Sense-annotated form of the Turkish sentence in Figure 1(a)

list is hidden and the ID of the submitted synset is displayed under the word. Figure 4 shows the sense-annotated form of the Turkish sentence in Figure 1(a).

4.3 Turkish Sense Annotation

4.3.1 Extracting Preliminary WordNet from Turkish Dictionary

For the Turkish sense annotation, the Turkish WordNet KeNet 1.0 (Ehsani et al., 2018) was used. KeNet was stored in an XML format that is quite similar to BalkaNet's (Stamou et al., 2002). The structure of a sample synset is as follows:

```
<SYNSET>
<ID>0066140</ID>
<SYNONYM>
<LITERAL>baba<SENSE>1</SENSE>
</LITERAL>
<LITERAL>peder<SENSE>1</SENSE>
</LITERAL>
<SYNONYM>
<POS>n</POS>
<DEF>Çocuğu olmuş erkek</DEF>
<EXAMPLE>Babasını çok sever.
</EXAMPLE>
</SYNSET>
```

Table 4: Unambiguous entities in the Turkish WordNet

Id	Entity
0000000	Proper noun
0000003	Time
0000004	Date
0000006	Hash tag
0000007	E-mail
0000010	Integer
0000011	Ordinal number
0000013	Percentage
0000015	Rational number
0000018	Interval
0000020	Real number

Each entry in the dictionary is enclosed by <SYNSET> and </SYNSET> tags. Synset members are represented as literals and with their sense numbers. Similar to BalkaNet, synonym literals are joined within a synset. <ID> shows the unique identifier given to the synset. <POS> and <DEF> tags denote the part of speech and the definition, respectively. As for the <EXAMPLE> tag, it gives a sample sentence for the synset.

For the Turkish side of the corpus, unambiguous entities, such as proper nouns, numbers or dates, are also included in the task where they are assigned with the IDs for their specific synsets (See Table 4). For instance, in Figure 4, the words "Bayan" and "Elianti" are assigned the ID of "0000000", which is the synset ID for proper nouns.

4.3.2 Extracting Candidate Sense List

The available senses of a word are obtained by querying its root word in this new WordNet. For example, in the converted sentence shown in Figure 2, the Turkish verb "çalar" can be morphologically decomposed in three different ways as illustrated below.

çal + VERB + POS + AOR + A3SG (plays)
çal + VERB + POS + AOR∧DB + ADJ + ZERO (playing X)

çalar + NOUN + A3SG + PNON + NOM (player)

As mentioned before, morphological disambiguation has been done by human annotators in the past study reported in (Gorgun et al., 2016). In the course of annotation, our system queries the dictionary with "çal" (play) or "çalar" (player) according to the selected morphological analysis. This morphological disambiguation prior to the annotation process is crucial especially in agglutinative languages such as Turkish. Thanks to this morphological disambiguation, the annotation process has been accelerated since the annotators have been provided with shorter lists of possible senses depending on the part of speech (POS) of the word being annotated in the given sentence. For example, when the annotator is to annotate the word "çalar" (play) in Figure 4, the software lists its senses as a verb and excludes the other senses provided by other POSs such as the sense(s) of "çalar" (player) as a noun.

Another issue that must be handled by the sense disambiguation tool is collocations. Many English words have a multi-word translation into Turkish and they need special attention to obtain a sense list. As a solution, we take cartesian product of derived forms of each word and search the WordNet for each combination. If any sense is found, we add it into the sense lists of the words that are included in the collocation. For instance, consider the following parallel sentences:

Hisse senedini sattı.

He sold the stock.

In the Turkish sentence, there is one collocation, namely "hisse senedi" which corresponds to "stock" in the English partition. After taking all the possible productions of the two words, "hisse" and "senedini" ("hisse senet", "hisse senedi", "hisse senedini"), the available senses displayed in the droplist for the word "hisse" contain both the possible senses of the simplex "hisse" and the ones corresponding to the collation of "hisse senedi".

4.4 English Sense Annotation

4.4.1 Sense Inventory

For the English sense annotation, we use Princeton WordNet (PWN) version 3.1. Although PWN does not provide a web page for obtaining synsets and/or their relations, the data files are present. After retrieving the synset data files from the site, we constructed a WordNet XML file similar to the Turkish one as given in Section 4.3.1:

```
<SYNSET>
<ID>10100638</ID>
<LITERAL>father<SENSE>1</SENSE>
</LITERAL>
<LITERAL>begetter<SENSE>1</SENSE>
</LITERAL>
<POS>n</POS>
<DEF>a male parent</DEF>
<EXAMPLE>...</EXAMPLE>
</SYNSET>
```

4.4.2 Extracting Candidate Sense List

For the English partition, extracting simple senses is much easier. We only ask for the available senses of the English word in PWN. Complexities arise for verbs marked for third person (-s), gerund (-ing), past participle (-ed); and for adjectives in comparative (-er) or superlative (-est) forms. For those cases, we strip down the affixes and search for the root form in PWN. For irregular forms (such as irregular verbs), we use the exception list of PWN to get the root forms.

Whereas function words are left unannotated, their copular or lexical counterparts are annotated. For instance, while the auxiliary verbs "be" and "have" are not annotated with a sense, their copular or lexical counterparts, such as "have" in the example of "The company had a loss", have been assigned a sense by the annotators. 868 of all the occurrences of "be" and "have" are lexical; and thus, were annotated with a sense.

For collocations, the situation is again easy for the English partition. We search for 2 or 3 word collocations in PWN with respect to the adjacent words of the current word. For instance, consider the sentence "They get up early". While showing the sense list of "get", we do not only show the sense list of "get" in isolation, but also add the senses of "get up" to that list. There are also collocations written with a hyphen in-between. For the ones listed as a single entry in the dictionary, such as "way-out", we add the senses under each word included in the collocation. The number of that kind of collocations with senses annotated is 219. However, the ones that cannot be treated as single lexical items, such as "three-months", were left unannotated. In total, 998 collocations with a hyphen could not be assigned a sense.

For the sake of consistency, since the corpus has a number of recurring words, annotators have compiled a list of the most frequently occurring 82 polysemous words, with multiple sense definitions

differing only slightly from each other. They have then decided on what sense is to be chosen and assigned to these words, and in which contexts. In addition, the annotators have agreed on certain conventions in annotating quantificational expressions, including numerals. The preparation of such a convention-guide, which is used as a sense-annotation-lexicon, helped each annotator to consistently select the same sense for a given word occurring in the same context and increased the inter-annotator agreement rate.

4.5 Data Format

In order to be able to process further, we remain highly faithful to the standard Penn Treebank notation of syntactic bracketing in the backend. We extend the original format with the relevant information, given between curly braces. For example, the word "plays" in the sentence shown in Figure 1 in the standard Penn Treebank notation, may be represented in the data format provided below:

```
(VBZ plays)
```

After all levels of processing are finished, the data structure stored for the same word has the following form in our system:

```
(VBZ {turkish=çalar}{english=plays}
{turkishSemantics=0703650}
{englishSemantics=15161405-n})
```

If there are multiple words on the Turkish side, the senses of each word is separated via a dollar sign:

```
(JJR {turkish=daha nazik}
{english=gentler}
{turkishSemantics=0178860$0572140}
{englishSemantics=01458191-s})
```

except collocations, for which a single sense ID is sufficient:

```
(NN {turkish=hisse senedi}
{english=stock}
{turkishSemantics=0348790}
{englishSemantics=13438244-n})
```

4.6 Annotation Evaluation

In this current work, all Turkish and English words in the input sentences have been disambiguated by human annotators, who are graduate students in language departments. They are native speakers of Turkish and advanced users of English.

For the evaluation of the annotated dataset, we used an inter-annotator agreement measure. Two different groups of annotators annotated the same sentences. Due to time limitations, we could re-annotate only 500 sentences from both Turkish and English partitions. We got %77.0 and %77.4 of inter-annotator agreement for Turkish and English, respectively.

Table 5: Distribution of sense annotations per synset

(a) Turkish		(b) English	
# of sense annotations	# of synsets	# of sense annotations	# of synsets
(500-1200)	6	(500-665)	2
(300-499)	11	(300-499)	3
(200-299)	15	(200-299)	4
(100-199)	42	(100-199)	22
(50-99)	128	(50-99)	72
(40-49)	53	(40-49)	34
(30-39)	108	(30-39)	79
(20-29)	200	(20-29)	141
(10-19)	521	(10-19)	478
(5-9)	898	(5-9)	921
4	491	4	494
3	529	3	694
2	1524	2	1678
1	2443	1	4037

5 Statistics About the Corpus

5.1 Distribution of Sense Annotations

Except the unambiguous entities, the current status of the Turkish side of the corpus contains 59,847 sense annotations. There are 6,969 distinct sense annotations and the average number of samples per sense is 8.59. The distribution of sense annotations per synset is given in Table 5(a).

For the English partition of the corpus, only entities residing in PWN are annotated, which include nouns, verbs, adjectives and adverbs. The current status of the English partition of the corpus contains 41,986 sense annotations. There are 8,629 distinct sense annotations and the average number of samples per sense is 4.87. The distribution of sense annotations per synset is given in Table 5(b).

5.2 Missing Annotations

When we compare annotations on the English partition with the annotations on the Turkish side, we see that, for some words in English, there is no corresponding semantic annotation in Turkish.

In total, there are 1,323 such words in English, composed of mostly modals (a total of 534: 100 “were”, 209 “was”, 7 “have”, 9 “has”, 6 “had”, 32 “been”, 16 “be”, 155 “are”) and prepositions (a total of 457: 13 “a”, 10 “about”, 2 “around”, 17 “as”, 36 “at”, 12 “back”, 2 “before”, 21 “down”, 5 “even”, 20 “for”, 30 “in”, 12 “into”, 15 “no”, 61 “not”, 22 “of”, 17 “off”, 14 “on”, 53 “out”, 11 “over”, 6 “through”, 4 “to”, 65 “up”, 9 “well”).

5.3 Multiword Expressions

Not only some words on the English partition may have multiword expression counterparts on the Turkish side, but also there are multiword expressions on the English partition whose counterparts are also multiword expressions on the Turkish side. The annotation framework can detect multiword expressions consisting of two and three word expressions (See Section 4.3.2). In total, there are 3,911 two-word (1,215 distinct) and 29 three-word (18 distinct) annotated multiword expressions.

6 Conclusion

In this paper, we reported our experience on manual tagging of English and Turkish senses in an English-Turkish parallel treebank, which had been parsed and enhanced with morphological features before the semantic annotation process. Our study has shown that it is possible to perform a parallel semantic annotation for an English-Turkish corpus and that the pre-processing stage for the parsing and morphological enhancement has been useful as it has accelerated the sense annotation process by providing the annotators with shorter lists of senses of a word in a given sentence.

As a future work, we plan to expand the size of the corpus by following the same manner of procedure, perform word sense disambiguation experiments on it with various classifiers and feature sets and make use of our parallel corpora in various NLP tasks including automatic annotation, statistical machine translation or word sense disambiguation.

References

L. Bentivogli, E. Pianta, and M. Ranieri. 2005. Multisemcor: an English Italian aligned corpus with a shared inventory of senses. In *Proceedings of the Meaning Workshop*, page 90, Trento, Italy, February.

H. Boas. 2005. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18.

Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz. 2018. Constructing a wordnet for turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3):24.

J. R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis, Philological Society*, pages 1–32.

O. Gorgun, O. T. Yildiz, E. Solak, and R. Ehsani. 2016. English-Turkish parallel treebank with morphological annotations and its use in tree-based smt. In *International Conference on Pattern Recognition and Methods*, pages 510–516, Rome, Italy.

N. Ide. 2012. Multimasc: An open linguistic infrastructure for language research. In *Fifth Workshop on Building and Using Comparable Corpora*, Istanbul.

D. Jurgens and I. Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *7th International Workshop on Semantic Evaluation*, Atlanta, Georgia.

C. Leacock, G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 260–265, Princeton, NJ.

M. Lupu, D. Trandabat, and M. Husarciuc. 2005. A Romanian semcor aligned to the English and Italian multisemcor. In *1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School*, pages 20–27, EUROLAN, Cluj-Napoca, Romania.

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308, Stroudsburg, PA, USA.

R. Navigli, D. Jurgens, and D. Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (SEM 2013)*, pages 14–15, Atlanta, Georgia.

H. T. Ng and H. B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An examplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47, Santa Cruz, CA, USA.

K. Oflazer, B. Say, and N. B. Atalay. 2003. The annotation process in the turkish treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, Budapest, Hungary.

Z. Orhan, E. Çelik, and N. Demirgüç. 2007. Turkish lexical sample task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 59–63, Prague, Czech Republic.

T. Petrolito and F. Bond. 2014. A survey of wordnet annotated corpora. pages 236–245, 01.

J. I. Saeed. 1997. *Semantics*. Blackwell.

B. Say, D. Zeyrek, K. Oflazer, and U. Özge. 2002. Development of a corpus and a treebank for present-day written turkish. In *Proceedings of the Eleventh International Conference of Turkish Linguistics*, pages 183–192, Eastern Mediterranean University, Cyprus, August.

B. Snyder and M. Palmer. 2005. The English all-words task. In *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41–43.

S. Stamou, K. Oflazer, K. Pala, D. Christodoulakis, D. Cristea, D. Tufis, S. Koeva, S. Totkov, D. Dutoit, and M. Grigoriadou. 2002. Balkanet: A multilingual semantic network for balkan languages. In *Proceedings of the First International WordNet Conference*, pages 21–25, Mysore, India.

L. Tan and F. Bond. 2012. Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). *International Journal of Asian Language Processing*, 22:161–174.

O. T. Yildiz, E. Solak, O. Gorgun, and R. Ehsani. 2014. Constructing a Turkish-English parallel treebank. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 112–117, Baltimore, Maryland, June. Association for Computational Linguistics.

B. İlgen, E. Adalı, and A. C. Tantuğ. 2012. Building up lexical sample dataset for turkish word sense disambiguation. In *IEEE International Symposium on Innovations in Intelligent Systems and Applications*, pages 1–5, Trabzon, Turkey, July.

Comparing Sense Categorization Between English PropBank and English WordNet

Özge Bakay
Boğaziçi University, Turkey
ozge.bakay@boun.edu.tr

Begüm Avar
Boğaziçi University, Turkey
begum.avar@boun.edu.tr

Olcay Taner Yıldız
Işık University, Turkey
olcaytaner@isikun.edu.tr

Abstract

Given the fact that verbs play a crucial role in language comprehension, this paper presents a study which compares the verb senses in English PropBank with the ones in English WordNet through manual tagging. After analyzing 1554 senses in 1453 distinct verbs, we have found out that while the majority of the senses in PropBank have their one-to-one correspondents in WordNet, a substantial amount of them are differentiated. Furthermore, by analysing the differences between our manually-tagged and an automatically-tagged resource, we claim that manual tagging can help provide better results in sense annotation.

1 Introduction

The main challenge in lexical semantics is generated from 'polysemy', which refers to the phenomenon of a single orthographic word having multiple, interrelated senses. Only with the help of the context one can pin down the particular sense in which a word is used. A further challenge in the field stems from multi-word expressions, i.e. groups of words having "a unitary meaning which does not correspond to the compositional meaning of their parts" (Saeed, 1997). Hence, there are two central concerns for semantic analysis, centered around compositionality: (i) At the lexical level, choosing the correct sense of a given word within a context, and (ii) at the sentence level, determining how a particular combination of words should be interpreted.

Having semantic analysis of annotated corpora along with the syntactic architecture enhances Natural Language Processing (NLP) applications such as information retrieval, machine translation, information extraction, and question answering. Using the added semantic layer, syntactic parser refinements can be achieved, which not only increases the efficiency but also improves application performance. PropBank (Kingsbury and Palmer, 2002; Kingsbury and Palmer, 2003; Palmer et al., 2005) is one of the studies on this concept, widely accepted by computational linguistics communities.

In this paper, we present a sense category evaluation between English PropBank and WordNet. In order to compare the sense categories, we first manually disambiguate English verbs in the input sentences with sense tags from English WordNet. Then, these annotations are compared with sense annotations in English PropBank.

This paper is organized as follows: Since we compare senses in English PropBank and WordNet, we provide information about these resources in Section 2 and touch upon the related work about combinations of these resources in Section 3. The details of our sense-annotated corpus and how it is constructed are given in Section 4. We give the comparison details and statistics in Section 5, and differences between automatic vs. manual tagging in Section 6. Lastly, we conclude in Section 7.

2 Resources

PropBank is a corpus where predicate-argument information is annotated and semantic roles or arguments each verb can take are posited (Babko-Malaya, 2005). PropBank uses conceptual labels for arguments from Arg0 to Arg5. Only Arg0 and Arg1 indicate the same roles across different verbs, standing for Agent/Causer and Patient/Theme, respectively. The rest of the argument roles can vary across different verbs. For instance, the roles of the predicate "attack" from PropBank are as follows: Arg0 is "attacker", Arg1 is "entity attacked", and Arg2 is "attribute".

WordNet is a graph data structure where the nodes are word senses with their associated word forms and edges are semantic relations between

the sense pairs. The first WordNet project was Princeton WordNet (PWN) which was initiated in 1995 by George Miller, (Miller, 1995). Over time, PWN evolved to become a comprehensive relational representation of the word senses of English (Fellbaum, 1998). WordNet includes relations between synsets such as hypernym, instance hypernym, hyponym, instance hyponym, meronym, holonym, antonym, entailment, etc.

3 Related Work

Among many previous studies similar to ours is the one by (Pazienza et al., 2006), which aims to extract frame pairs by combining the lexical database of WordNet with the syntactic and semantic information given in VerbNet and semantically-annotated corpus of verbs in PropBank. Having inferred 989 frame pairs with troponymy, entailment, causation, and antinomy; they conclude that NLP applications can benefit from such repositories by making use of automatic or semi-automatic techniques to map arguments across the frames.

In another study, Kwon and Hovy (2006) first assigned the frame for each verb sense in WordNet from FrameNet and then, aligned roles among FrameNet, WordNet, and LCS depending on their mappings. In total, 4240 senses are linked with FrameNet frames, 674 of which are also linked with LCS, 1250 with PropBank, and 1757 with both.

SemLink (Palmer, 2009) is another project which aims to combine different information provided by various lexical resources (VerbNet (Kipper-Schuler, 2005), FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005) and WordNet (Fellbaum, 1998)). With mappings among these resources, the project aims to develop an NLP resource with extended overall coverage.

Aiming for interoperability among the same resources used in SemLink, López de Lacalle et al. ((de Lacalle et al., 2014a; de Lacalle et al., 2014b; de Lacalle et al., 2016a; de Lacalle et al., 2016b)) focus on predicates and try to develop a common semantic infrastructure, which is called the Predicate Matrix (PM). They define a set of methods, such as advanced graph-based word sense disambiguation algorithms and various corpus alignment methods to automatically achieve this integration. While they base their work on the central motivation of SemLink, the authors criticize the limitations of the manual methods used in developing the SemLink project and argue that "building large and rich enough predicate models for broad-coverage semantic processing takes a great deal of expensive manual effort" (de Lacalle et al., 2016b).

López de Lacalle et al.'s (2016b) work is definitely a major progress for NLP studies centering around predicate structure. Their approach, however, does not seem to put enough emphasis on the cognitive nature of language. Undoubtedly, manual tagging requires significant human effort, hence is more costly, and manually tagged corpora may be more limited in terms of systematicity and coverage. However, whether any analysis of language can be fully automatized is still a very skeptical issue. According to the approach adopted in the present study, the use of human annotators is considered worthwhile for developing corpora by focusing on semantic information. There may well be 'human errors', but overall, we believe that manual tagging still gives us better results in qualitative terms, though maybe not in quantitative terms.

4 Sense Annotation

4.1 Annotation Tool

The annotators in the present study use a custom application (written in Java) for browsing sentences and annotating them with senses. The toolkit is publicly available[1]. The current implementation of the application is designed to import the text files that adhere to the Penn Treebank data format. Once a sentence has been imported into the semantic editor, the human annotator is presented with the visualized syntactic parse tree of that sentence. Annotators can click on the leaf nodes corresponding to words. When a word is selected, a drop-down list is displayed, in which all the available WordNet entries of the selected word's lemma are listed.

Moreover, sense options whose POS (parts of speech) do not agree with the given word's POS, are disabled to optimize the task/help the annotators. Upon the selection of the most appropriate sense, the drop-down list is hidden and the ID of the submitted synset is displayed under the word. Figure 1 shows a screenshot from the system interface, depicting the screen presented to the an-

[1] https://github.com/olcaytaner/DataCollector

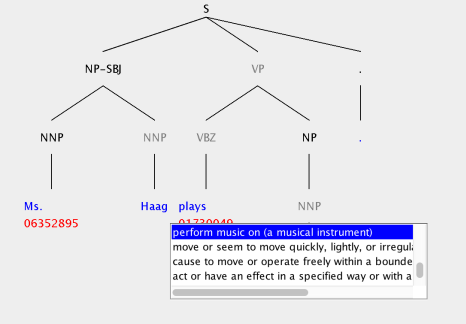


Figure 1: A screenshot from the system interface

notators when annotating the verb "plays" in the English sentence "Ms. Haag plays Elianti."

4.2 Sense Inventory

For the sense annotation, we use PWN version 3.1. Although PWN does not provide a web page for obtaining synsets and/or their relations, the data files are present. After retrieving the synset data files from the site, we constructed a WordNet XML file similar to the BalkaNet's (Stamou et al., 2002).

4.3 Extracting Candidate Sense List

For extracting senses, we only ask for the available senses of the English word in PWN. Complexities arise for verbs marked for third person (-s), gerund (-ing), past participle (-ed); and for adjectives in comparative (-er) or superlative (-est) forms. For those cases, we strip down the affixes and search for the root form in PWN. For irregular forms (such as irregular verbs) we use the exception list of PWN to get the root forms.

For collocated verbs, we just search for 2 or 3 word collocations in PWN with respect to the adjacent words of the current word. For instance, consider the sentence "They get up early". While showing the sense list of "get", we do not only show the sense list of "get" in isolation, but also add the senses of "get up" to that list.

4.4 The Comparison Process

For the comparison of sense categorization of verbs between English PropBank and English WordNet, a list of sentences (7576 sentences with 1554 senses of 1453 distinct verbs), all of which had been annotated by human annotators, was extracted. Instead of single-word annotations, we preferred to have the annotations of all the words appearing in those sentences. Two human annotators who are both graduate students in language-related departments were then provided with the list which displayed all the verbs alphabetically with all of the sentences they were used in. Seeing all the exemplary sentences of the verbs together is believed to have helped the annotators analyze the meaning differentiations within the verbs more closely (See Table 1). Before moving onto the comparison between the two sense categorizations, annotators also checked the accuracy of the annotated meanings of the verbs. This second step is considered to have strengthened the accuracy of sense annotation as well as the comparison of PropBank and WordNet. During the comparison stage, human annotators analyzed the two datasets to find out how similarly or differently the senses in English PropBank were reflected in English WordNet.

5 Comparison Details

In the comparison of the senses in English PropBank and English WordNet, what has been found out is that whereas most of the senses in WordNet seem to match the ones in PropBank, some of them do not. We came across these mismatches for three reasons; because of (i) the senses going under differentiation, (ii) the senses getting combined in English WordNet or (iii) the overlaps of senses in a given verb in that one sense of a verb in WordNet corresponds to more than one single sense of the same verb in PropBank.

In Sections 5.1 and 5.2, we will review one-to-one and one-to-many sense matchings between English PropBank and English WordNet, whereas in Sections 5.3 and 5.4 we will review many-to-one and many-to-many sense matchings between English PropBank and English WordNet.

5.1 One-to-one Sense Matches between English PropBank and English WordNet

The majority of the senses in English PropBank (1184 senses of 1118 different verbs) is found to match the ones in English WordNet. In other words, the sense categorizations in English PropBank seem to be retained in English WordNet once the senses are replaced with their WordNet equivalents. For example, as it is shown in Case 1, the sense of "abate" in PropBank, "to decrease, become less strong", is observed to be equal to "be-

Table 1: Example sense categorizations for English PropBank verbs

Case	Verb	PropBank Sense	WordNet Sense	Example
1	abate_01	to decrease, become less strong	become less in amount or intensity	The dollar posted gains in quiet trading as concerns about equities abated.
2	strengthen_01	(cause to) become stronger	gain strength	As Wall Street strengthened, the London trading room went wild.
	strengthen_01	(cause to) become stronger	make strong or stronger	In 1986, Congress strengthened the penalty by making it a felony.
3	absorb_01	suck up	assimilate or take in	Most dealers can not continue to absorb this supply.
	absorb_01	suck up	take up, as of debts or payments	Deal stocks led the market down as they absorbed the heaviest losses.
4	buy_01	purchase	accept as true	U.S. officials, however, said they are n't buying the Israeli argument.
	buy_01	purchase	obtain by purchase	Everybody was out buying Monets.
	buy_01	purchase	buy what had previously been sold, lost, or given away	So far, the company had bought back 1.6 million shares.
5	celebrate_01	honor, show respect to	have a celebration	The ads celebrate the achievements of some of Lake View 's residents.
	celebrate_02	have a party, occasion to mark an event	have a celebration	They don't even give a nod to human sensibilities by celebrating Halloween.
6	build_01	construct	make by combining materials and parts	A Taiwanese steelmaker recently announced plans to build a Nucor-like plant.
	build_01	construct	develop and grow	You built your career on prejudice and hate.
	build_02	grow	bolster or strengthen	Seagram says the promotion is designed to build brand loyalty rather than promote heavy drinking.
	build_02	grow	develop and grow	The great silver clouds on the horizon build themselves on the pale water.

come less in amount or intensity" in WordNet (See Table 1).

5.2 Sense Differentiations in English WordNet

A significant difference between the two sense categorizations is that in 352 senses of 329 different verbs, the senses given in English PropBank branch up to 12 distinct, and hence more specific, senses in English WordNet. Those differentiations may be meaning- or syntax-related. Regarding the syntax-related differentiation as indicated in Case 2, for example, in English PropBank, for the verb "strengthen", there is one sense for both transitive and intransitive forms, "(cause to) become stronger". However, in English WordNet, this sense is differentiated into two; "gain strength" for the intransitive and "make strong or stronger" for the causative, i.e. transitive form (See Table 1).

As an example for meaning-related differentiations, when we look at the verb "absorb" in Case 3, we see that whereas the only sense provided for it by Propbank is "suck up", two different senses for that same sense are given in WordNet; (i) assimilate or take in and (ii) take up, as of debts or payments. Although the former can be considered as the equivalent of "suck up", the latter indicates a different sense, which seems to be miss-

Table 2: Verbs with the highest number of senses

Verbs	Senses Annotated	Senses in WordNet
take_01	12	42
give_01, have_03, see_01	11	44, 19, 24
break_01, move_01	9	59, 16
come_01, know_01, turn_01	8	21, 11, 26
do_02, draw_02, find_01, lead_01, look_01, place_01	7	13, 36, 16, 14, 10, 16

ing in PropBank (See Table 1). It is also important to note that the number of the added senses in WordNet may vary depending on the verb. Table 2 shows the list of the 15 verbs with the highest number of senses. For instance, while PropBank lists only one sense for the verb "take", which is "take, acquire, come to have, choose, bring with you from somewhere, internalize, ingest", WordNet lists 42 different ones and in the current dataset, 12 of them were assigned.

Apart from the higher number of senses provided by English WordNet, another factor playing a role in presenting new senses is collocations. For example, as Case 4 shows, for the verb "buy", three senses were assigned in total (See Table 1). While two of them are among the senses given for "buy" in English WordNet, the third one is the sense given for the collocation of "buy back". Thus, in addition to the senses of the verbs as individual forms, the senses of collocations including the verbs were also annotated. In total, for 18 senses of 15 verbs, senses of their collocations were also assigned.

5.3 Sense Combinations in English WordNet

Another difference between the two categorizations is that some senses in English PropBank are combined into a single sense in English WordNet. For instance, in Case 5, the two senses used in PropBank for "celebrate" are combined into one (See Table 1). In total, 6 verbs senses of which are combined in WordNet are "celebrate, cite, clear, explode, scuttle, and prepare".

5.4 Overlapping Categorizations

In some of the verbs, there is no one-to-one sense match between the two categorizations. In other words, a single sense in WordNet is annotated for at least two different senses of the same verb in PropBank. These overlapping categorizations are different from the sense combinations explained in 5.3 since in these verbs, a particular sense given in WordNet may replace more than one sense of a verb in PropBank in some of the sentences, whereas in the rest of the sentences, different senses, other than the overlapping one, are still annotated. For example, for the verb "build" given in Case 6, the sense of "develop and grow" was annotated for two different senses of "build" in English PropBank: "construct" and "grow" (See Table 1). However, in both of these sense categories of "build", other senses were also annotated: "make by combining materials and parts" in "build_01" and "bolster and strengthen" in "build_02". So, instead of combining these two senses, the sense of "develop and grow" seems to occur across different senses of the same verb. The number of verbs with sense overlaps are 29.

6 Automatic vs. Manual Tagging

Although automatic corpus alignment methods are preferred over manual tagging because of their systematicity and lower cost in many lexical resource integration studies, we argue that the manual tagging method is still highly needed. In an attempt to investigate the effectiveness of automatic and manual taggings, we compared our manually-tagged lexical resource integration with the automatically-tagged PM created by López de Lacalle et al.'s ((de Lacalle et al., 2014a; de Lacalle et al., 2014b; de Lacalle et al., 2016a; de Lacalle et al., 2016b)) based on PropBank senses. As a result of this comparison, we found that while the matchings of 418 WordNet senses in 413 verbs in PropBank and WordNet are the same in our and their integrations (See Case 1 in Table 3), the higher number of items with differences (1721 senses in 1387 verbs) in their matchings are worth attention. To mention some of those differences, for 307 PropBank senses in 281 verbs, our work and the PM do not match in any of the assigned senses. Also, for 199 PropBank senses in 178 verbs, there are both matching and mismatching senses.

Table 3: Example sense categorizations for English PropBank verbs by manual and automatic annotation

Case	Verb	Propbank Sense	WordNet Sense	
			Manual	Automatic
1	abdicate_01	to relinquish (power or responsibility)	give up, such as power, as of monarchs and emperors, or duties and obligations	give up, such as power, as of monarchs and emperors, or duties and obligations
	accelerate_01	make be faster, the act of speeding up	move faster	move faster
	accelerate_01	make be faster, the act of speeding up	cause to move faster	cause to move faster
2	zap_01	destroy	-	strike at with firepower or bombs
	zigzag_01	(cause to) move in zigzag fashion	-	travel along a zigzag path
3	emote_01	express emotion	give expression or emotion to, in a stage or movie role	-
	encrypt_01	encode, scramble digital information	convert ordinary language into code	-
4	accept_01	take willingly	consider or hold as true	consider or hold as true
	accept_01	take willingly	give an affirmative reply to	give an affirmative reply to
	accept_01	take willingly	receive willingly something given or offered	receive willingly something given or offered
	accept_01	take willingly	-	tolerate or accommodate oneself
	accept_01	take willingly	-	react favorably to
5	answer_01	give an answer, reply	react verbally	react verbally
	answer_01	give an answer, reply	give the correct answer or solution to	-
6	appeal_01	legal transaction	take a court case to a higher court for review	be attractive to
	appeal_03	be attractive	be attractive to	be attractive to

First of all, as we cover only a small part of PropBank data in our incomplete and still ongoing study, we lack 593 senses of 323 verbs included in the PM. In other words, those senses (such as "zap_01" or "zigzag_01" as shown in Case 2 in Table 3) are not included in our comparison at all. However, 8 verbs that are annotated in our limited integration, namely "emote, encrypt, franchise, indemnify, jell, motorize, outsell and squeegee" in Case 3 in Table 3, do not seem to be automatically annotated in the PM, which could be taken as the first evidence to suggest that automatic tagging may not be sufficient.

Secondly, when we look at the number of the matches assigned for each sense, we observe that the PM has matches that do not currently exist in our integration. For example, in Case 4 given in Table 3, while our integration provides only three senses for the item "accept_01" ((i) consider or hold as true, (ii) give an affirmative reply to and (iii) receive willingly something given or offered), the PM has two additional ones ((iv) tolerate or accommodate oneself and (v) react favorably to), adding up to five in total. The number of items with additional sense annotations is 497 in 477 verbs. The reason we suggest for those missing senses in our integration is that our work captures a portion of the whole Penn Treebank and as larger portions get annotated, those senses will be added to our integration, as well. On the other hand, in

addition to finding missing senses in our integration, we also came across senses that are included in our corpus but not in their PM. In total, 125 senses of 123 verbs in our integration do not exist in theirs. For instance, for the item "answer_01" in Case 5 in Table 3, we annotated two senses, which are (i) react verbally and (ii) give the correct answer or solution to. In the PM, only the first sense is annotated. So, we take the lack of those unassigned senses in the PM as our second evidence.

Thirdly, when we analyze the senses within the same verbs in the PM, we see that while some of them were annotated by taking into account their differences, some of them were merged, which resulted in the loss of some meanings. As an example, the same sense is assigned to "appeal_01" and "appeal_03" in the PM while different senses are annotated in our work as shown in Case 6 in Table 3. Due to this merger, PM fails to capture the sense that is needed, for example, for the sentence "Minpeco attorneys said they would appeal the decision to a federal district court.". Although not all the verbs with multiple senses were subject to that kind of wrong merging, we still consider this as an issue that needs to be resolved and take it as our third evidence to show the importance of manual tagging. Given that those errors cannot be noticed in automatic tagging, we suggest that manual tagging still has a crucial role in detecting those systematic errors resulting from automatic tagging.

7 Conclusion

In this paper, we reported our comparison results of English verb sense annotations in PropBank with senses in English WordNet for the sentences from Penn Treebank. In opposition to the idea that automatic tagging is good enough to eliminate the necessity for manual tagging, based on our comparison of our work with the PM, we contend that manual tagging is still needed to have qualitatively-better results and that it would be quite useful to apply it, at least in combination with automatic tagging.

Another issue that makes our work promising is its extendibility to a larger Turkish dataset. Related to that, Ak et al. (Ak et al., 2018) have recently constructed a Turkish Proposition Bank using translated sentences of English PropBank. So far, 9560 translated sentences are annotated with semantic roles and framesets are created for 1914 verb senses. In spite of its limited size, their study constitutes a base for Turkish Proposition bank. Therefore, we hope that our English Propbank and English Wordnet parallelization study can be used to extend many larger datasets in other languages, starting with the Turkish Proposition bank.

References

K. Ak, C. Toprak, V. Esgel, and O. T. Yildiz. 2018. Construction of Turkish proposition bank. *Turkish Journal of Electrical Engineering & Computer Sciences*.

O. Babko-Malaya, 2005. *Guidelines for Propbank Framers*.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

M. Lopez de Lacalle, E. Laparra, and G. Rigau. 2014a. First steps towards a predicate matrix. In *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, Jan 25-29.

M. Lopez de Lacalle, E. Laparra, and G. Rigau. 2014b. Predicate matrix: extending semlink through wordnet mappings. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

M. Lopez de Lacalle, E. Laparra, I. Aldabe, and G. Rigau. 2016a. A multilingual predicate matrix. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Paris, France, may.

M. Lopez de Lacalle, E. Laparra, I. Aldabe, and G. Rigau. 2016b. Predicate matrix: automatically extending the semantic interoperability between predicate resources. *Language Resources and Evaluation*, 50(2):263–289, Jun.

C. Fellbaum. 1998. ed. wordnet: an electronic lexical database. *MIT Press, Cambridge MA*, 1:998.

P. Kingsbury and M. Palmer. 2002. From treebank to propbank. In *LREC*. European Language Resources Association.

P. Kingsbury and M. Palmer. 2003. Propbank: The next level of treebank. In *Proceedings of Treebanks and Lexical Theories*, Växjö, Sweden.

K. Kipper-Schuler. 2005. *VerbNet: a broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March.

M. Palmer. 2009. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the Generative Lexicon Conference*, Pisa, Italy.

M. T. Pazienza, M. Pennacchiotti, and F. M. Zanzotto. 2006. Mixing wordnet, verbnet and propbank for studying verb relations. 01.

J. I. Saeed. 1997. *Semantics*. Blackwell.

S. Stamou, K. Oflazer, K. Pala, D. Christodoulakis, D. Cristea, D. Tufis, S. Koeva, S. Totkov, D. Dutoit, and M. Grigoriadou. 2002. Balkanet: A multilingual semantic network for balkan languages. In *Proceedings of the First International WordNet Conference*, pages 21–25, Mysore, India.

Building ASLNet, a Wordnet for American Sign Language

Colin Lualdi[1,2] **Jack Hudson**[1] **Christiane Fellbaum**[3,4] **Noah Buchholz**[4]

[1]SignSchool Inc., Madison, Wisconsin, USA
[2]Department of Physics, University of Illinois, Urbana-Champaign, Illinois, USA
[3]Department of Computer Science, Princeton University, Princeton, New Jersey, USA
[4]Program in Linguistics, Princeton University, Princeton, New Jersey, USA
colin@signschool.com, jack@signschool.com
fellbaum@princeton.edu, noah.buchholz@princeton.edu

Abstract

We discuss the creation of ASLNet by aligning the Princeton WordNet (PWN) with SignStudy, an online database of American Sign Language (ASL) signs. This alignment will have many immediate benefits for first- and second- sign language learners as well as ASL researchers by highlighting semantic relations among signs. We begin to address the interesting theoretical question of to what extent the wordnet-style organization of the English lexicon (and those of wordnets in other spoken languages) is applicable to ASL, and whether ASL requires positing additional, language- or modality-specific relations among signs. Significantly, the mapping of SignStudy and PWN provides a bridge between ASL and the worldwide wordnet community, which comprises speakers of dozens of languages working in academic and language technology settings.

1 Background and Motivation

We discuss plans for developing ASLNet, the large-scale alignment of the Princeton WordNet (Miller, 1995; Fellbaum, 2010) and SignStudy (www.signstudy.org), a database of American Sign Language (ASL) signs. The popularity of the Princeton WordNet (PWN) has spawned wordnets in dozens of other spoken languages (Bond and Foster, 2013; Vossen, 2004), including those outside the Indo-European language family. Crossing modalities, ImageNet (Deng et al., 2009), a database created to support image recognition, contains thousands of images linked to PWN's synsets. Sign languages fall squarely within the family of human languages but communicate meaning in the visual-kinesthetic modality. Aligning the synsets of PWN (and by extension those of the wordnets in other spoken languages) with ASL signs is both a logical and challenging next step.

1.1 SignStudy

SignStudy (SS) is an online ASL lexical resource created and supported by SignSchool (www.signschool.com), an online ASL learning platform. SS is freely available to any registered user who wants to learn, explore or conduct research on the ASL lexicon.

Users can search for signs by typing an English word into a search window, which returns a video showing the corresponding ASL sign. SS will return multiple signs if there are several variants (synonyms) of a sign that share the same meaning; multiple signs will also be returned for a polysemous word form whose different English meanings correspond to distinct signs in ASL. Signs are demonstrated via videos with user-controllable pausing and playback speeds. Additionally, signs are accompanied by four annotated parameters: the dominant hand starting and ending handshapes, and the non-dominant hand starting and ending handshapes. The database is structured in terms of semantic categories (e.g., Nature) and subcategories such as Nature:Animals and Nature:Landforms. In the current version of SS, the depth of the semantic hierarchies is limited to two levels.

1.2 Benefits of SS

As a large repository of ASL signs, SS has the potential to offer a centralized platform for the ASL research community to study various aspects of that language. Supporting the study and comparison of signs along with their properties will enable the expansion of theoretical linguistics research on sign language. SS would also benefit ASL lexicography efforts by enabling the analysis of which

signs (among their variants) are considered more prevalent and standard than others by various subgroups of the ASL community (e.g., native vs. non-native signers), perhaps through the implementation of a sign rating mechanism, whether via a crowd-sourced or a controlled polling process.

A deeper understanding of ASL gained from such research has the potential to improve ASL teaching resources such as those available via SignSchool, offering tools that enable users of other languages to become familiar with ASL. ASL is taught in some high schools and universities in the U.S., but older individuals and those who do not have access to adult education facilities offering ASL classes would clearly benefit from an online resource that can be accessed with a computer anywhere and anytime.

Increasing the accessibility of ASL learning resources is critical for raising the general public's awareness of ASL as it would enable improved communication accessibility among deaf ASL users and the hearing. It would also address the fact that many hearing speakers are unaware of ASL as a full language with all the complexity and expressiveness of a spoken language, including a rich lexicon and a grammar that differs considerably from English but falls well within the parameters of Universal Grammar.

1.3 Limitations of SS

SS is well equipped as a resource to assist with these broader goals with its respectable coverage of the ASL lexicon. The database currently contains 4,500+ sign videos (demonstrated by over 10 deaf and hard of hearing models) associated with 6,000+ English equivalents. Signs are annotated by 67 handshapes, 38 semantic categories, and 238 semantic subcategories. Nonetheless, SS will benefit from further additional coverage. As is the case for spoken languages, the size of the lexicon cannot be conclusively determined, in part because the notion of *word*, as familiar from a spoken language, does not map straightforwardly to *sign*. In the context of our work, we define 'word' as a unique mapping of meaning and form, regardless of modality. For example, the sign DOG[1] and the spoken form [dog], both referring to canines ("dog" in written English), can both be considered 'words', and thus part of the lexicons of English and ASL.

SS aims to be more than a flat list of signs with their English equivalents. The meanings of signs can be more clearly represented in a thesaurus-like fashion, where signs with intuitively similar meanings are interconnected. While SS has already manually grouped its current vocabulary into semantic categories and subcategories, much more structure will be added.

2 Enhancing SS with PWN Relations

A promising method for the semantic organization of SS's signs is to map them to PWN, creating ASLNet. However, it is critical to avoid the misconception that ASL is simply a signed version of English, which leads to the incorrect impression that one may develop ASLNet by the simple mapping of signs to their corresponding English words in PWN. As is the case with creating wordnets for languages other than English, ASLNet-internal additions are required to accommodate links among signs, some of which are not (and cannot) be encoded for wordnets representing the lexicon of a spoken language. With this in mind, we emphasize that our objective is to utilize the semantic structure offered by PWN to assist with the semantic organization of ASL signs and their linking to corresponding senses in other languages with wordnets.

There are multiple benefits resulting from a mapping of ASL signs to PWN entries. Deaf and hearing learners of ASL can explore the ASL lexicon by following the links in multiple, intuitive ways. For example, if the signs HAND and FINGER are linked to PWN synsets containing (the corresponding senses of) *hand* and *finger*, the semantic relation between these two words (meronymy, the part-whole relation) that is encoded in PWN will be transferred to the signs. SS does not need to independently encode such relations among signs, so long as signs and PWN words are semantically equivalent.

A structured lexical resource for ASL will offer major pedagogical benefits and enable semantically-driven learning of ASL (Miller and Fellbaum, 1992). It will support language acquisition by Deaf children by enabling them to quickly acquire the meanings of new signs. For example, the signs LEGISLATURE and JUDICIARY could be linked to the sign GOVERNMENT by the meronymy relation; the signs EX-

[1] In this paper we use the convention of writing ASL signs in all capital letters.

PENSIVE and CHEAP by the antonym relation, etc. Children's books designed to foster word learning commonly present words in such semantically related groups. Second-language students of ASL will be able to expand their ASL lexicon by diving into a semantic rabbit hole as they discover a sign that leads them to semantically related signs, and so on. Entire lessons could be organized semantically so that critical areas of the lexicon are quickly filled out with meaningfully related signs.

Perhaps the most important benefit of linking ASL to PWN is the immediate connection to dozens of wordnets in other languages. PWN can be thought of as the hub to which wordnets in many languages are linked. Departing from a given signed word will allow one to go from the corresponding English word to its equivalents in Spanish, Basque and Hindi, for example. Additionally, the link between PWN and ImageNet raises interesting possibilities for exploring questions of iconicity in ASL (Perniss et al., 2010).

3 Related Work

We are aware of only one effort to link wordnets to a database of signs. (Prinetto et al., 2011; Shoaib et al., 2012) describe plans for developing a Sign Bank for Italian Sign Language (LIS) and its alignment to MultiWordNet (Pianta et al., 2002), a lexical database for Italian, Romanian, Spanish, Portuguese, Latin and Hebrew modeled on, and linked to an early, smaller version (1.6) of PWN that is no longer the standard for natural language processing applications. However, while the concept of LIS was developed and described in (Shoaib et al., 2012), to the best of our knowledge, the LIS Sign Bank was not in fact created.

4 Units of Meaning

Both SS and PWN are databases whose atomic units are form-meaning mappings. However, there are inherent technical challenges to aligning such pairings across different modalities. Spoken and written words are discrete units of form-meaning mappings. By contrast, signed words are expressed through movement in continuous space. This allows signers to modify a given meaning in a continuous manner in many cases; signers are not necessarily limited to the words of a spoken language that divides a range or scale into discrete steps. Additionally, while there also exist signs with quasi-discrete differences in meaning, boundaries between these meanings may not always be clear-cut due to the continuous nature of signing. Consequently, a successful mapping between discrete spoken words and continuous signed words requires a careful analysis of the meaning(s) of ASL signs.

4.1 Gradability

Residing in continuous space, ASL signs have a vast parameter space that results in many signs having highly variable senses. One consequence of this is a large number of signs being gradable. For example, basic lexical signs, such as SNOWING, may be endlessly modified to assume slightly modified senses (e.g., "snowing" vs. "snowing heavily"). The lack of discrete steps in such modifications makes it difficult to map such signs to discrete synsets. In fact, such mapping efforts may reveal the rich, fluid way in which troponyms and hyponyms are expressed in ASL.

One possible solution, at least for gradable signs, is to distinguish between gradable and complementary signs. For gradable concepts, we propose creating a special construction in ASLNet that associates groups of signs with numerical ratings to indicate their location on their shared scale. Discrete synsets can then be linked to this grouping. For NLP applications, one could add threshold ranges to ASLNet queries that will return the signs that, for example, fall between intensities 5 and 7 on a 10-point scale. In fact, ASL instructional material (e.g., textbooks) typically simplify the scale to three degrees: less, normal, more. This three-way classification may be sufficient for ASLNet purposes.

Note that a one-dimensional gradability scale is assumed here; further analysis may reveal the need for higher-dimensional scales to characterize signs with more than one gradable aspect. We anticipate this to be the case for verbs, where the troponymy relation distinguishes among verbs that can elaborate the common event along different dimensions. For instance, the sign WALK may be modified along a scale corresponding to walking speed (cf. English "run" and "amble") and along another scale corresponding to step length (cf. English "mince" and "stride").

4.2 Classifier Constructions

The continuous-space nature of ASL also manifests itself via classifier constructions. Essentially, they are certain handshapes that are asso-

ciated with different semantic classes (e.g., size, shape, action, etc.). Thus, when these handshapes are paired with specific sign parameters or used in certain sentence constructions, they can be used to communicate nuances in meaning and provide highly detailed descriptions (of objects, actions, etc.). As a simple example, classifiers may be used to elaborate the meanings of basic lexical nouns. For instance, one possible way to sign "tome" is to first produce the lexical sign meaning "book" followed by a classifier construction that indicates the referent (the book) possesses the property of substantial thickness. Additionally, classifier modifications are often gradable; the production of the "thickness" classifier may be adjusted on a thickness scale to change the description of the book from that of a tome to that of a pamphlet.

Thus, signs that include classifiers would benefit from being encoded into ASLNet in a manner that indicate the classifier(s) used and their position(s) on the relevant scale(s). This will allow for a more complete documentation of the semantic meaning of a particular sign and how its components contribute to the meaning of the sign. Doing so will assist with the semantic linking of signs within ASLNet as discussed below.

4.3 Non-Manual Signals

Expressions of gradability and classifiers often involve the use of non-manual signals (NMS), which consist of various facial and body movements that accompany signs. They serve many purposes, including modifying individual signs and indicating sentence structure. In the context of ASLNet, we are primarily interested in NMS that are an integral component of a sign and NMS that modify the meaning of signs. For the former, certain signs require a particular NMS (namely, a mouth morpheme) to assume a particular meaning. Thus, for STRUGGLE the handshapes and their movements are almost always accompanied by a STA-STA mouth movement. For the latter, NMS can be used to indicate the precise meaning of a sign; the classifier construction used in TOME may be paired with different NMS to indicate whether a particular tome is an average tome (neutral face), or a very thick tome (incorporating a CHA mouth movement).

In developing ASLNet, it is important to distinguish between non-grammatical and grammatical NMS. Non-grammatical NMS are analogous to voice inflections; an ASL speaker may assume a happy facial expression when conveying happy news. While such expressions do convey meaning at the conversational level, they do not directly affect the meaning of individual signs, and thus are not of interest in the context of ASLNet. This is in contrast to grammatical NMS (as discussed above), which are more structured to the extent that they modify the meaning and function of signs.

4.4 Lexical Gaps

While certain ASL signs may have a clear English correspondence in PWN, there are lexical gaps in both languages, as is the case for any language pair.

As an example, the sign TRUE BUSINESS, which may be literally translated as "true business", is often used to emphasize the authenticity of a statement or to introduce a surprising twist to a previous statement. There is no obvious lexical equivalent in English, though various (highly context-dependent) translations exist. Such gaps are also also prevalent in certain usage cases involving classifiers. For instance, classifier-based signs such as CL-"peeling a banana" and CL-"close a refrigator" do not have lexeme status in English, where they are freely composed. While such signs obviously are to be included in ASLNet, they will not map onto a single synset in PWN. There are also many examples where the reverse is true, i.e., a simplex word in English requires multiple signs in ASL that speakers do not consider a lexical unit. For instance, the English word "concierge" requires signing out a full phrase analogous to "the hotel employee who assists guests".

A solution for handling crosslingual lexical gaps is to add a placeholder, without a word, for signs and words in the network that show the gap, either PWN or ASLNet.

5 ASLNet-Specific Links

As ASL possesses linguistic properties distinct from those of spoken languages such as English, ASLNet-internal additions are required to accommodate links among signs. Without those modifications, ASLNet runs the risk of projecting ASL into an incompatible framework, preventing its study without biases towards spoken languages.

5.1 Phonological and Lexical Links

ASLNet requires the encoding of information pertaining to the five generally recognized parameters of ASL (handshape, location, movement, orientation (palm), and non-manual signals) for each sign. Links may then be established between signs that share common phonological properties.

The five parameters are analogous to phones in a spoken language. In isolation, phones like [k], [a] or [t] carry no meaning, but when composed into a morpheme or word, they assume a meaning ([kat] = 'cat'). The same is true for sign language parameters: individual parameters of signs combine to give the whole sign its meaning. Thus, modifying one parameter of a given sign will result in a different sign that may or may not have a related meaning. While this seems analogous to substitution of a different phoneme of a given word (*hat-cat-car*), the meanings of such words are not usually similar for spoken languages.

The framework for such a phonological categorization has been demonstrated by Caselli et al. (2017) with the development of ASL-LEX, a broad lexical database of approximately 1,000 signs. Each sign in ASL-LEX is coded with six phonological properties (sign type, selected fingers, flexion, major and minor location, and movement). An ASLNet-specific lexical encoding is also likely beneficial, as demonstrated by ASL-LEX, where signs are additionally coded for four lexical properties: initialization, lexical class, compounding, and fingerspelling.

By utilizing aspects of ASL-LEX's design and structure as a model, ASLNet will be able to incorporate ASL-specific phonological and lexical properties necessary to develop some of the ASLNet-internal links. ASLNet can then build upon the work done by Caselli et al. by introducing the additional dimension of semantic links by virtue of its integration with PWN.

5.2 Other ASLNet-Specific Links

There are additional ASLNet-specific links that ASLNet would likely benefit from including. An example is the close relation between the members of ASL noun-verb pairs. There are many signs that can simply switch between noun and verb forms by, for instance, changing the number of movements, such as chair (CHAIR [2x]) and sit (CHAIR [1x]). This parallels the many noun-verb pairs in English related by zero morphology (love, drive, travel, Google, etc.). Encoding such pairs in ASLNet would allow for the comparison of the relations among the different part-of-speech forms of ASL signs with other languages and whether certain relations are more prevalent for certain semantic categories, e.g., teleologically related noun-verb pairs for artifact nouns.

Additionally, signs containing classifier constructions should be linked if they have similar classifiers since such signs may have similar meanings. For instance, recalling our TOME example from earlier, the classifier used to indicate the thick nature of a book would be applicable to signs relating to a "beam" as in "wooden beam"; they are both thick objects. This is analogous to sound symbolism in English, as in many words beginning with [gl] ("gleam", "glitter", and "glossy") that all seem to have a meaning related to light. Combining these classifier links, along with the phonological and lexical links discussed above, with purely semantic links from PWN will allow for the exploration of phonesthesia in ASL (i.e., the non-accidental relation between form and meaning).

6 Implementation

We discuss methods and steps required for building ASLNet.

6.1 Crosslingual Wordnets

After PWN gained widespread popularity, wordnets were built in a number of different languages. EuroWordNet (Vossen, 2004) comprises eight European languages, including Estonian and Basque, which are genetically and typologically unrelated to Indo-European languages.

An important goal was to connect all wordnets to one another, so that equivalent words and meanings could easily be identified. EuroWordNet took PWN as its hub to which each new wordnet was mapped. In some cases, the wordnet developers simply translated the English synsets into their language; in other cases, wordnets were initially built up independently and later merged with the English version. Not all languages lexicalize the same concepts, and for words that have no English equivalent a simple record was added to PWN pointing to and from the language-specific words. In this way, PWN became the union of all concepts lexicalized in all wordnets, but not shared by all. Consequently, the structure of EuroWordNet per-

mits one to find equivalent words and meanings in all eight languages by going via PWN, making it a valuable tool for crosslingual study and applications.

Since the techniques described by (Vossen, 2004) proved successful for connecting PWN to non-Indo-European languages, it is reasonable to believe that they are applicable to the case of ASL, which is genetically unrelated to English. The only unknown is whether and to what extent such techniques also work across modality differences (spoken vs. signed).

6.2 Proof-of-Concept Demonstration

Following the methods described by Vossen, we propose to use a hybrid approach in building ASLNet, starting first by directly mapping straightforward cases such as many common nouns. We then encode words existing only in ASL and ASLNet-specific semantic relations among signs, such as noun-verb mappings and gradable groups, within ASLNet. Once we have accommodated ASLNet-specific entries and links, the next step is to merge these with PWN.

Thus, as our first step we plan to develop a proof-of-concept demonstration of ASLNet (ASLNet V1.0) by starting with lexical ASL nouns. Advantages of working with lexical nouns include their relative ease of encoding and more straightforward PWN mappings by virtue of their lexicalized nature. This is in contrast to verbs whose expression differs significantly in ASL. To guide ASLNet V1.0 development, we intend to draw lexical nouns from existing PWN noun categories that have a rich hierarchical structure, such as vehicles, artifacts, and food. Starting with these categories will allow for the increased likelihood of observing novel semantic relationships between ASL signs early in ASLNet development, providing opportunities for evaluating the success of our development technique.

6.2.1 Technological Infrastructure

Steps have already been taken to develop a preliminary system for mapping SS signs to PWN. The structure of the SS database now allows for the association of ASL signs with their corresponding English equivalents. These equivalents are in turn linkable to PWN synsets. Thus, the links associated with those equivalents will be automatically inherited to SS from PWN.

SS has developed a simple web application to allow for computer-assisted manual linking of SS signs to PWN synsets and the assignment of POS labels. This "WordnetMapper" tool[2] utilizes the existing English equivalents of SS signs to query PWN for the purpose of suggesting possible additional English equivalents as well as PWN synsets to map to. Manual mapping is also possible via this application.

6.2.2 Development Procedure

A relatively complete and functional prototype will be developed by following PWN groupings and systematically filling out all or most of the terms in PWN (e.g., "eyelid" and "nostril" are parts of "eye" and "nose", respectively).

A team of contributors will supply any additional signs necessary for filling out of specific corners of the ASL lexicon within SS. Each additional sign will be accompanied by filmed demonstration by a Deaf and native signer of ASL. Fluent ASL signers experienced with ASL-English translation will then assist with the phonological and lexicographical encoding (including ASLNet-specific links) of those signs into the SS database, including their mapping to their corresponding PWN synsets.

6.2.3 Lexicography Considerations

As much is not yet known regarding developing a wordnet for a sign language, the development process for ASLNet V1.0 will be carefully monitored and documented by the development team to generate data that will guide and inform the subsequent development of ASLNet.

ASL nouns that resist straightforward mapping to PWN, such as those without direct equivalents in English will be recorded with placeholders to mark lexical gaps in English. The same procedure will be applied to English nouns with no obvious ASL lexical equivalent.

As ASL signs are subject to significant regional and articulation variations, special attention will be given to adding as many synonymous signs as possible in order to make ASLNet representative of ASL. The equivalent of such signs in English are words like "hoagie", "submarine", "po'boy", "hero", and "grinder". Such regional variations are usually encoded in PWN as synonyms (members of one synset); sometimes, the "gloss" names the region where a specific term is used. Thus, in

[2]WordnetMapper is still in the evaluation stages and is currently not publicly available.

ASLNet such signs will be manually grouped together as variants of the same sign (and sense).

Special attention will be given to polysemy, especially where it is interesting, as in the case of metaphors, when signs for inclusion in ASLNet are recorded. For example, filming the sign for "line" as a queue, will accompany filming a sign for "line" as a long, narrow bar. Incorporating such data may make ASLNet V1.0 more versatile in identifying interesting directions to pursue in the study of systematic polysemy in ASL during subsequent development cycles.

The encoding of non-manual signals (NMS) will likely be challenging. In particular, it is not always clear what constitutes grammatical or non-grammatical NMS. Thus, for the purposes of ASLNet V1.0 the lexicographers will concentrate on signs that are not usually subject to modifications by NMS (i.e., many lexical nouns).

7 Future Work

Once ASLNet V1.0 is completed, the resulting wordnet will be analyzed for any novel characteristics and properties. These results will be reported along with an analysis of our proposed development procedure and its effectiveness.

Along with remedying any difficulties encountered during V1.0 development, building V1.1 will include identifying the mapped nouns that belong to noun-verb pairs along with the consideration of modified lexical nouns (e.g., with the use of classifier constructions). Such mappings are likely to be challenging for reasons discussed in Section 4. V1.1 will be followed by subsequent versions that incorporate mappings for increasingly complicated aspects of ASL, particularly those that differ significantly from spoken languages. Once mapping techniques have been developed and proved viable for the core aspects of ASL, large-scale lexical expansion of ASLNet may then commence. Such work may lead to ASLNet becoming a part of the Collaborative Interlingual Index (CILI), a means of linking wordnets without depending on PWN's semantic structure (Bond et al., 2016; Vossen et al., 2016). CILI integration may be beneficial in the face of lexical gaps as well as differences in word encoding and linking between ASL and English.

It is also important to be mindful of the fact that sign languages are not universal; there exist many other sign languages distinct from ASL. As the linguistic properties of other sign languages may not be entirely identical to those of ASL, it is rewarding to develop the structure of ASLNet such that it is as general as possible with regard to sign languages so that this work may give rise to similar research opportunities with other sign languages without unintentionally introducing a bias towards ASL. This is comparable to how PWN led to the development of wordnets for additional languages.

8 Conclusion

This work opens many interesting avenues for research. As discussed previously, developing ASLNet will provide insight into lexical gaps between ASL and English. With the inclusion of ASL verbs, ASLNet will permit the exploration of verb troponymy within ASL. By highlighting semantic relationships between signs, ASLNet may also offer insights into many properties of ASL, including systematic metaphoricity, compounds, idiomatic expressions, compositionality and similarities, and iconicity. Furthermore, since ASLNet will be linked to PWN, and in extension, wordnets for many other languages, comparisons of these linguistic properties may be made between ASL and other languages.

Acknowledgements

The authors would like to thank Evan Corden and the full SignSchool team for developing SS as a resource to enable this work as well as valuable insights from Elaine Wright and James Waller. Comments and feedback from the 2019 Global Wordnet Conference attendees as well as three anonymous reviewers are also greatly appreciated.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1352–1362.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In *Proceedings of the Eighth Global WordNet Conference*, pages 50–57.

Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2017. ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods*, 49(2):784–801.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Christiane Fellbaum. 2010. WordNet. In *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer.

George A. Miller and Christiane Fellbaum. 1992. WordNet and the Organization of Lexical memory. In Merryanna L. Swartz and Masoud Yazdani, editors, *Intelligent Tutoring Systems for Foreign Language Learning*, pages 89–102, Berlin, Heidelberg. Springer Berlin Heidelberg.

George A Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Pamela Perniss, Robin Thompson, and Gabriella Vigliocco. 2010. Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1:227.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *First International Conference on Global WordNet*, pages 293–302.

Paolo Prinetto, Umar Shoaib, and Gabriele Tiotto. 2011. The Italian Sign Language sign bank: Using WordNet for sign language corpus creation. In *2011 International Conference on Communications and Information Technology (ICCIT)*, pages 134–137. IEEE.

Umar Shoaib, Nadeem Ahmad, Paolo Prinetto, and Gabriele Tiotto. 2012. A platform-independent user-friendly dictionary from Italian to LIS. In *LREC*, volume 12, pages 2435–2438.

Piek Vossen, Francis Bond, and John P. McCrae. 2016. Toward a truly multilingual Global Wordnet Grid. In *Proceedings of the Eighth Global WordNet Conference*, pages 419–426.

Piek Vossen. 2004. EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *International Journal of Lexicography*, 17(2):161–173.

A collaborative system for building and maintaining wordnets

Tomasz Naskręt
G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Science and Technology, Wrocław, Poland
tomasz.naskret@pwr.edu.pl

Abstract

A collaborative system for wordnet construction and maintenance is presented. Its key modules include WordnetLoom editor, Wordnet Tracker and JavaScript Graph. They offer a number of functionalities that allow solving problems on every stage of building, editing and aligning wordnets by teams of lexicographers working in parallel. The experience collected in recent years has allowed us to refine applications and add new modules to provide the best user experience in a reliable and easily maintainable way.

1 Introduction

Wordnet is not yet another electronic dictionary. It is a complex lexico-semantic network. Its construction, especially when done manually by a team of lexicographers, and its further editing and/or aligning with other resources requires very advanced and flexible tools. Such tools should offer the possibility of simultaneous work of many team members on the same lexicon (a wordnet for a particular language), simultaneous work of different teams on different lexicons, and the subsequent manual or semi-automated linking of the constructed resources.

Dictionary compiling tools are mostly designed as complex XML editors such as, for instance, Lexonomy (Měchura, 2017). This approach may not be beneficial in working with graph-like structures. Therefore, several dedicated tools have been designed and are currently used by different wordnet teams e.g DEBVisDic (Horák et al., 2006), sloWTool (Fišer and Novak, 2011). Visualisation of wordnet graphs in most tools follows a radial pattern: a synset in focus is presented in the middle and all links, irrespectively of their types are placed radially around the central element, e.g. *sloWTool* or *WordTies* (Pedersen et al., 2012). *GernEdiT* (Henrich and Hinrichs, 2010) offers visualisation of the wordnet structure in the range selected by the user, but it is hierarchical and focused mainly on hypernymy. Moreover, the visual presentation does not allow for direct editing of the structures. WordnetLoom stands out of the remaining tools, because it offers a graph-based visualisation of wordnet data and provides entirely different workflow based on the direct interaction with graph nodes. In this paper, we will present the most recent development of WordnetLoom and progress in relation to earlier releases. We have improved the graphic design for better user experience and implemented the lexical unit graph visualisation.

Both dictionary making and wordnet building are usually carried out by teams of lexicographers and/or developers. Collaborative work, especially in distributed teams working from , requires control tools to provide quality assurance and development progress. In-built auditing/change backlog feature is often absent in these systems and data versioning is handled by external VCS[1] software or done manually. The newest version of WordnetLoom is interconnected with the Wordnet Tracker module which provides additional feedback channel for lexicographers to enrich their workflow. Every activity of each lexicographer is registered and can be monitored by a senior lexicographer. This paper will showcase how auditing and monitoring can be handled.

We will also present a new web-related module, namely JavaScript Graph. JavaScript Graph module is an answer to user needs and provides the possibility of embedding graph visualization to existing websites or applications.

[1] VCS - Version Control System e.g. Git, Subversion

In this work, we will present the key modules that are part of a collaborative system for wordnet construction and maintenance including WordnetLoom editor, Wordnet Tracker and JavaScript Graph.

2 WordnetLoom Demo

Up to version 1.68, WordnetLoom was a standalone java fat client application directly connecting to its database with all logic contained on the client side. Such approach ensures that scaling of the application could only be possible by scaling the database server, in this case MySQL[2]. In order to meet the growing numbers of users and challenges in providing dedicated endpoints not only for the client editor application, but also for other external applications, web pages or mobile applications, all business logic was extracted to a separate application built on top of JEE8[3] framework. The application is responsible for data validation, data auditing, user activity monitoring, user management and data processing. It provides a communication channel via REST API (Fielding, 2000) in the form of Siren[4]-like hypermedia specification. Scaling of the application itself is done by docker-compose[5] replicas, while database scaling can be achieved by replication configurations where at least two databases are available. Master database configuration is optimized for writing and slave databases have configuration optimized for high performance reading. Further scaling can be ensured by introducing new slave database nodes for each distinct consumer such as a mobile application or a web page.

The main consumer of the API is a thick client in the form of WordnetLoom Editor java application (main application workspace presented at Fig. 1a) which has been slimmed down and does not contain essential business logic which reduces it to the role of a simple REST client. It enables advanced search functionalities and basic CRUD[6] operations on typical core objects being part of the semantic structure such as synset, sense (see sense editing properties Fig. 1b), sense relation, synset relation, and relation type. From the editor level, the user with administrator privileges can modify and add elements to dictionary entities such as: part of speech, domain, register (see editing dictionaries Fig. 1c) and adding or editing types of semantic relations (see editing relation types Fig. 1d). The main advantage of the application is the possibility of working with visualization in the form of a graph, which provides quick and easy navigation and simplifies the creative process. Due to the fact that the Editor has recently undergone a major architectural transformation, it has allowed for even simpler modifications and easier addition of new components, such as in the case of implementing the extended semantic description panel for Dictionary of Polish Borrowings in Yiddish[7] (see Fig. 1e). Also within this project we have created a graph visualization of lexical units which has become the part of a core application.

3 Wordnet Tracker Demo

An important aspect of the process of building and maintaining wordnet is the ability to monitor changes made by team members. It is made possible by the Wordnet Tracker module which provides tracking of user activity (see Wordnet Tracker dashboard Fig. 2) in terms of the number of lexical units, synsets and semantic relations entered (see Fig. 3 for synset relation changes). Through this application, the lexicographer has also access to the full history of changes that have been made within a given lexical unit (see Fig. 4 for current changes of lexical units). All changes in the synset structure are presented in Fig. 5 where the left side column displays the current synset state, while the right side column shows all changes in the synset elements. The user as well as the coordinator have access to current changes in real time for constant monitoring. This functionality turned out to be particularly valuable when working with new, inexperienced lexicographers. The application administrator has the possibility to create diagnostic queries within lexicons or even within the entire dataset, as well as to create statistic queries. In both cases the generated query results are available for download in the form of files. Wordnet Tracker also provides basic user management panel where the privileged user can add new users, reset passwords or restrict user access to chosen lexicons.

[2] https://www.mysql.com/

[3] Java Enterprise Edition 8 specification https://javaee.github.io/javaee-spec/

[4] https://github.com/kevinswiber/siren

[5] https://docs.docker.com/compose/overview/

[6] CRUD are four basic functions of persistent storage (such as create, read, update and delete)

[7] https://polonjid.wn.uw.edu.pl/?lang=en

4 JavaScript Graph Module Demo

Presenting work results in the form of a graph visualization outside WordNet Loom editor environment is possible now by a created javascript module. The module tries to faithfully preserve the navigation functions as in the WordNet editor, but at the same time gives the possibility to adjust the color scheme and nodes style to the host application/page design. The presentation data model is fetched from the WordnetLoom server via the REST endpoint and the D3.js[8] library with custom modifications handles graph visualization and user interaction. The module is constructed in such a way so as to allow easy embedding in other applications, such as a mobile application or a website. A very good example can be the main page of plWordNet[9] where the module is used in the form of a pop-up window or as a full scale central element of the website presented at the online Dictionary of Polish Borrowings in Yiddish[10](see Fig. 6). Simple library import and basic configuration will allow to present wordnet lexicon as graph visualization on every platform where JavaScript is supported.

5 Conclusions and Further Works

This concludes our brief description of each module. We have seen that the combination of presented tools offers solutions to common tasks and problems encountered while building wordnets particularly by distributed teams. We will continue to be open-source software licensed under GNU LGPL 3.0. The source code is hosted in GitHub repository(https://github.com/CLARIN-PL/wordnetloom).

We will continue to actively develop presented tools over the next years focused on adding new functionalities based on the needs of users. We will also direct our development towards a reliable, fully-flagged web-based system and we will strive to continue to simplify system deployment by an extensive use of docker[11] containers.

Acknowledgment

The work co-financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and the project funded by the National Science Centre, Poland under the grant agreement No UMO-2015/18/M/HS2/00100.

[8] D3.js is a JavaScript library for manipulating documents based on data

[9] http://plwordnet.pwr.edu.pl/wordnet/

[10] http://polonjid-dictionary.clarin-pl.eu

[11] https://www.docker.com/

References

Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, 2000. AAI9980887.

Darja Fišer and Jernej Novak. Visualizing sloWNet. In *Proceedings of eLex*, pages 76–82, 2011. URL http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-9.pdf.

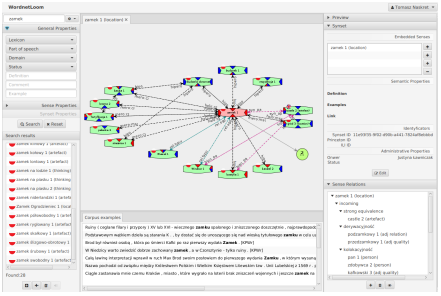
Verena Henrich and Erhard Hinrichs. GernEdiT – the GermaNet editing tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Aleš Horák, Karel Pala, Adam Rambousek, and Martin Povolný. DEBVisDic — first version of new client-server wordnet browsing and editing tool. In *Proceedings of the Third International WordNet Conference — GWC 2006*, pages 325–328. Masaryk University, 2006.

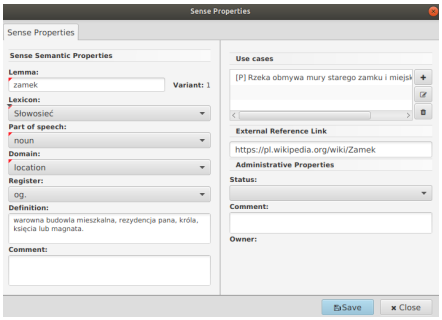
M. B Měchura. Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, The Netherlands.*, 2017. URL https://www.lexonomy.eu/docs/elex2017.pdf.

B.S. Pedersen, L. Borin, M. Forsberg, K. Lindén, H. Orav, and E. Rögnvalssson. Linking and validating nordic and baltic wordnets – a multilingual action in META-NORD. In *Proceedings of 6th International Global Wordnet Conference*, pages 254–260., Matsue, Japan., 2012.

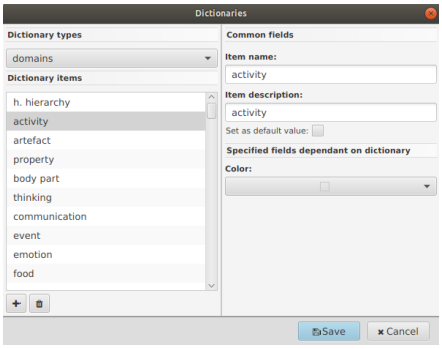
Figure 1: Key windows in WordnetLoom



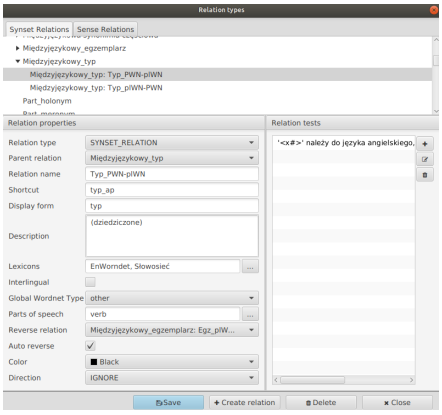
(a) Application main workspace.



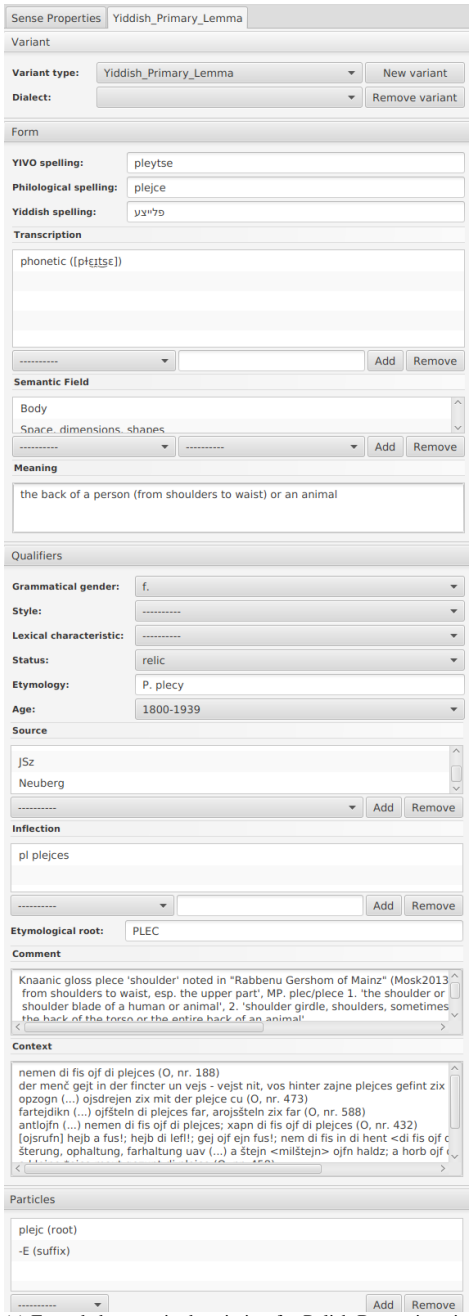
(b) Sense properties window.



(c) Dictionaries window.



(d) Relation types window.



(e) Extended semantic description for Polish Borrowings in Yiddish dictionary.

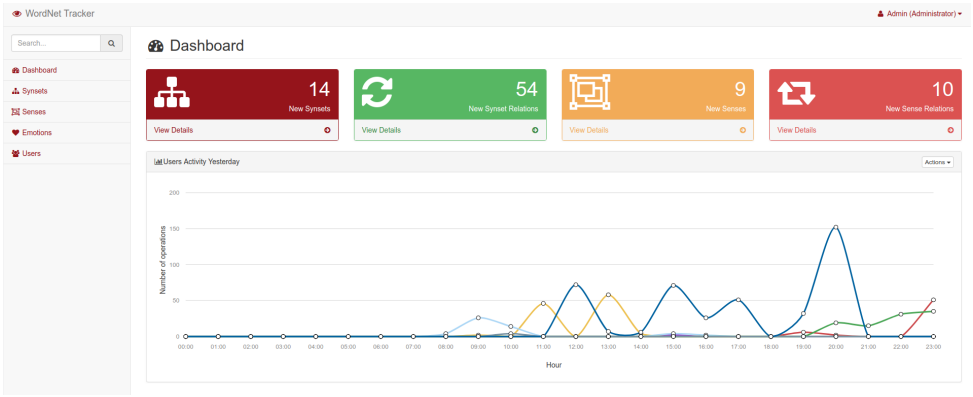


Figure 2: Tracker dashboard.

Figure 3: Synset relations changes history view.

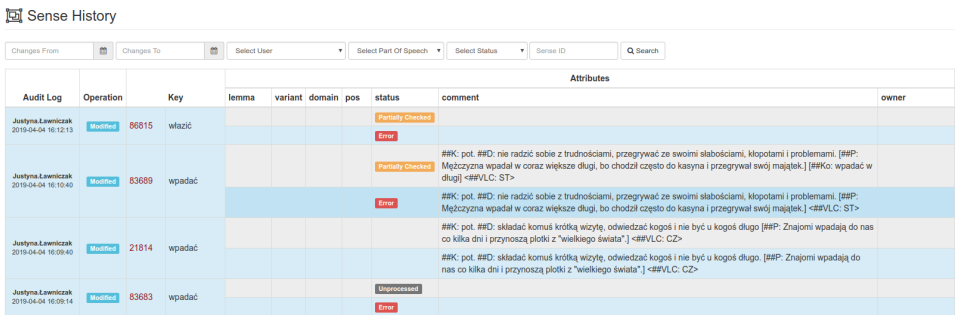


Figure 4: Senses changes view.

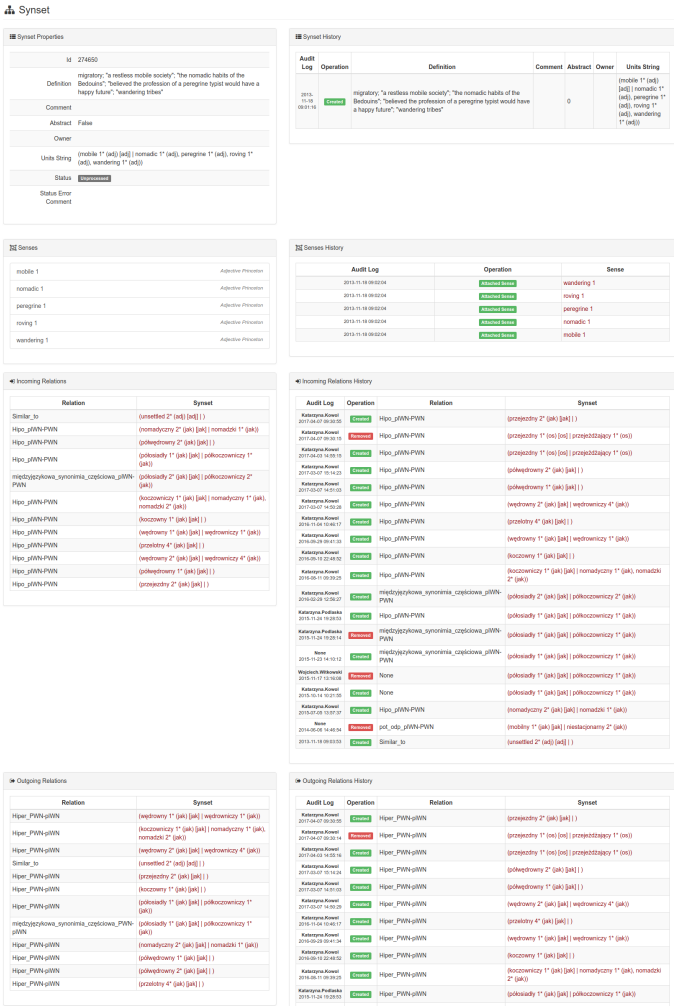


Figure 5: Selected synset history view.

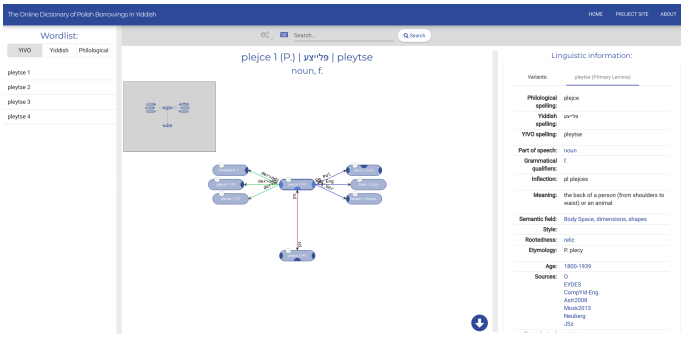


Figure 6: Example of embedded java script visualization module.

Enriching a Keywords Database Using Wordnets – a Case Study

Tomasz Jastrząb
Silesian University of Technology
Gliwice, Poland
Technicenter Sp. z o.o.
Bytom, Poland
Tomasz.Jastrzab@polsl.pl

Grzegorz Kwiatkowski
Silesian University of Technology
Gliwice, Poland
Grzegorz.Wojciech.Kwiatkowski
@polsl.pl

Abstract

In the paper, we study the case of building a keywords database related to the Polish Classification of Activities (PKD 2007). The database enables automatic classification of the companies to the industry branches. The classification is performed based on the company's activity description. We present the initial design of the keywords database and the ways in which wordnets were used to enrich it. Finally, we present the preliminary statistical evaluation of the produced resource.

1 Introduction

The Polish Classification of Activities (PKD 2007) (Council of Ministers, 2007), based on the European Classification of Economic Activities (NACE) (EUROSTAT European Commission, 2006), defines a hierarchical structure of industry branches and activities conducted by Polish companies. It is divided into five levels comprising sections (industries), divisions, groups, classes, and subclasses. There are 21 sections, 88 divisions and 654 subclasses, denoted by symbols consisting of letters (sections), numbers (divisions, groups, classes) or letters and numbers (subclasses).

The Polish Classification of Activities serves as a guideline for governmental institutions such as the Central Statistical Office of Poland. It acts as a source of information for services such as wskaznikibranzowe.pl[1], which publishes the quarterly and yearly financial ratios for the respective industry branches distinguished in PKD 2007. Furthermore, it can be used as a text corpus for different natural language processing tasks. In this paper, we follow the latter possibility. In particular, we use the descriptions of sections, divisions, and subclasses to build a keywords database defining each PKD 2007 section. The keywords database is then used to classify the companies to their industry branches, based on the company's activity descriptions.

Our motivations are as follows. Firstly, we want to help company owners to better describe their activities. Secondly, we want to provide an automatic tool for classifying the company to its industry. Such a tool may support search engines and allow company managers to find their competition easier. Finally, we would like to allow for simpler integration with services such as wskaznikibranzowe.pl, which given the company description, can provide the appropriate financial ratios.

The contributions of the paper are two-fold. First, we present the designed keywords database, which adds new value to the existing lexical and semantic resources. Second, we discuss the ways in which wordnet enriched the database. This way we also evaluate the wordnet in terms of data availability and completeness. We use plWordNet (Maziarz et al., 2016) as the one containing more data than the other Polish wordnet, PolNet (Vetulani, 2014).

The remainder of this paper is divided into four sections. In Section 2 we review the literature pertaining to the applications of wordnets, e.g., for building lexical resources. In Section 3 we describe the process of building and enriching the keywords database. In Section 4 we present the results of the statistical evaluation of the keywords database, while in Section 5 we summarize the paper and provide future research perspectives.

2 Related Works

Wordnets constitute lexico-semantic resources, whose basic building blocks are usually synonym sets (synsets) (Miller et al., 1990; Miller, 1990) or less frequently, lexical units (Maziarz et al., 2016). These blocks are interconnected by means of var-

[1]Available at https://wskaznikibranzowe.pl.

ious relations, such as hypernymy, hyponymy, meronymy, and others.

Wordnets support various natural language processing tasks, which we divide into the following categories:

1. Creation, extension, and enrichment of lexical and semantic resources of different types, including e.g., other wordnets, thesauri, and taxonomies.

2. Text processing tasks, such as word-sense disambiguation, entity linking, sentiment/polarity analysis, and semantic features mapping.

Within the first category of wordnet applications, the primary source of information is the Princeton WordNet (Fellbaum, 1998), which was used to construct various national wordnets. It also plays an important role in the development of multilingual resources such as the EuroWordNet or the MultiWordNet projects (Vossen, 1998; Magnini et al., 1994). Furthermore, thanks to the mapping of Princeton WordNet to the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2003) or its use in the creation of the Yago ontology (Suchanek et al., 2007), the Princeton WordNet is used as a reliable link between these ontologies and other wordnets, e.g., plWordNet (Kędzia and Piasecki, 2014). For other projects and resources based on Princeton WordNet, created with the aim of supporting research or providing entertainment, the reader is referred to (Princeton University, 2010).

Wordnets, and in particular the Princeton WordNet, are often combined with other resources such as Wikipedia or Wiktionary to produce new or to improve existing resources. As an example of such a resource, the semantic network BabelNet could be mentioned (Navigli and Ponzetto, 2012). It combines the knowledge (synsets, relations) included in the Princeton WordNet with the data collected from Wikipedia. A similar approach, but using additional resources, was also taken when creating the ConceptNet (Speer and Havasi, 2012). A key feature of the ConceptNet and its relation to Princeton WordNet is that it aligns the Princeton WordNet concepts with other resources making it a part of the Linked Data movement. The idea of linking the Princeton WordNet within the framework of the Linguistic Linked Open Data cloud is also mentioned in (McCrae, 2018). The author focuses on the interconnection between proper nouns included in the Princeton WordNet and Wikipedia articles.

An example of a resource that is based on the Polish wordnets is the integrated wordnet discussed in (Krasnokucki et al., 2017). The resource combines the information included in PolNet and plWordNet by merging the common elements and extending the amount of information available in one of the wordnets with the contents of the other one and vice versa. The use of plWordNet in relation to ontologies is also mentioned in (Postanogov and Jastrząb, 2017), where it is considered as a source of reusable information for building new ontologies.

It is worth to mention that, although usually successful, the use of wordnets as sources of additional knowledge can also end up with a failure. An example of such a case is reported in (Poprat et al., 2008). The authors aimed at using the existing software infrastructure and data formats of Princeton WordNet to create the links between the wordnet and an Open Biomedical Ontology. It turned out that neither the data format nor the software was suitable for biomedical data representation. It mainly suffered from the limited number of relations supported by Princeton WordNet or restrictions regarding the number and format of the created concepts. Finally, the authors claimed that the Princeton WordNet provides a limited coverage of biomedical-specific terms. The limited coverage of required information in wordnets was also mentioned in (Liebeskind et al., 2018). The authors tried to create a thesaurus for Hebrew, based on the Hebrew WordNet, but due to its limited coverage they had to supplement the process by manual labour.

The second category of wordnet applications mentioned before is related to the support of natural language processing tasks. One of the key applications is the use of wordnets for opinion mining as well as sentiment and polarity analysis. Examples of semantic resources created with this purpose in mind include the SentiWordNet (Esuli and Sebastiani, 2006), Q-WordNet (Agerri and Garcia-Serrano, 2010), and plWordNet emo (Janz et al., 2017). The first two resources are based on Princeton WordNet, while the last one is based on plWordNet. The Princeton WordNet was also used e.g., for word-sense disambiguation in text clustering (Wei et al., 2015) as well as for

the document exapansion in information retrieval systems (Agirre et al., 2010).

The use cases of Polish wordnets, and especially the plWordNet, include e.g., the analysis of the amount of emotions-related information covered by plWordNet, which was investiged in (Kwiatkowski and Jastrząb, 2016a; Kwiatkowski and Jastrząb, 2016b). As shown in (Jastrząb et al., 2016; Jastrząb et al., 2017) wordnets can be also used for the semantic features mapping, which in turn can support the valence schema matching.

3 Keywords Database Design

The keywords database construction was based on the XML version of the PKD 2007 document (Główny Urząd Statystyczny, 2007). The keywords database was constructed according to the following steps:

1. Information selection and extraction,
2. Information processing,
3. Keywords extraction,
4. Keywords enrichment.

Of the above steps, the first three were performed based on the source document only, while the last step was performed with the use of wordnets.

The *information selection and extraction* step consisted in choosing the most relevant elements of the XML document. We decided to parse the contents of the following XML elements (the translations in parentheses are added for clarity, since the original names are in Polish):

- `poziom` ("level") – this is the basic element grouping the information on various levels of the PKD 2007 hierarchy;
- `numerPoziomu` ("level number"), `nazwaPoziomu` ("level name") – these two sub-elements of the `poziom` element allowed us to gain the knowledge about document structure and also to filter out the information we considered irrelevant. We decided to use only the elements corresponding to levels 1, 2 and 5, i.e., sections, divisions, and subclasses;
- `element` ("element") – this is the basic element grouping the descriptions of respective sections, divisions, and subclasses;
- `nazwa` ("name"), `symbol` ("symbol") – these two sub-elements of the `element` tag uniquely identify the members of the PKD 2007 hierarchy and can be also used for tracking the relationships between the different levels of the hierachy;
- `opisObejmujeNieobejmuje` ("description includes/excludes") – this element is the most crucial from the perspective of the keywords database construction. It contains the descriptions of the industry branches and company activities included in or excluded from the given level of the hierarchy.

Let us consider the following examples of the document contents. On level 1, we can find a section with symbol *A* and name *Rolnictwo, leśnictwo, łowiectwo i rybactwo* ("Agriculture, forestry and fishing"). On level 2, we can find a division with symbol *01* and name *Uprawy rolne, chów i hodowla zwierząt, łowiectwo, włączając działalność usługową* ("Crop and animal production, hunting and related service activities"). Finally, on level 5, we can find a subclass with symbol *01.41.Z* and name *Chów i hodowla bydła mlecznego* ("Raising of dairy cattle"). The following excerpt of the *description includes/excludes* element for subclass 01.41.Z (note the HTML tags) is an example of the source text used for the keywords database: "<h2>01.41.Z</h2> Podklasa ta obejmuje: chów i hodowlę bydła mlecznego,
produkcję surowego mleka krowiego lub z bawołów.
" ("01.41.Z This subclass includes: raising and breeding of dairy cattle, production of raw milk from cows or buffaloes.") (Główny Urząd Statystyczny, 2007).

Since the keywords database aims to support the assignment of companies to sections only, we used the *name* and *symbol* elements to combine the descriptions of divisions and subclasses with the description of the section. This way we obtained a more detailed description of each section, which constituted the input for the second step of the database construction. Note that from now on, when we speak about section description, we consider the combined descriptions mentioned above.

In the *information processing* step we first removed from the descriptions all the elements that were not words, such as HTML tags, punctuation marks, and digits (we used a set of simple regular expressions to do so). Then, based on the white

signs (spaces, tabulations, new line characters) we divided the text into words. Next, we removed those words that certainly could not become the keywords, such as conjuctions, pronouns, and prepositions. We did it semi-automatically, by removing words of length not greater than three. The process was also complemented by manual verification of the words that remained. Considering the excerpt presented above, the resulting set of words would be {Podklasa, obejmuje, chów, hodowlę, bydła, mlecznego, produkcję, surowego, mleka, krowiego, bawołów}. The number of words was finally reduced by creating a set of unique words describing each section.

The *keywords extraction* step was initialized with the calculation of the edit (Levenshtein) distance between the words describing each section. The aim of this process was to merge similar words together to further reduce their number e.g., the following words could be merged bawół, bawołów, bawoły, bawolę. While calculating the edit distance we temporarily ignored the Polish diacritics, in the sense that characters such as e.g., 'ł' and 'l', were considered to be the same. The reason for omitting the differences resulting from the use of Polish diacritics was again to limit the number of keyword candidates. Based on the obtained Levenshtein distances we merged together the words for which the distance was not greater than three. Additionally, we performed manual verification of the outcomes, to make sure that no undesired merges were made. As a result, for each section i we obtained a set of keyword candidates $K_i = \{\text{word}\}$. For each keyword candidate k, we calculated the following metric:

$$W_k = \sum_{i=1}^{n} x_i \tag{1}$$

where n is the number of sections and x_i is a binary variable such that $x_i = 1$, when $k \in K_i$, and $x_i = 0$, otherwise. Hence, for any keyword candidate k, W_k is an integer from the interval $[1, n]$. The initial set of keywords was established by removing those keyword candidates k for which $W_k \geq 2$. Since the devised set of keywords contained various forms of the same word (resulting from flection), we have manually revised all the keywords producing the set of common word forms. The final set of keywords was constructed after repeating the calculation of the W_k metric, denoted by W'_k, for the set of common word forms and rejecting the words for which the condition $W'_k \geq 2$ was satisfied.

Given the sets of keywords, we performed the *keywords enrichment* step, which involved the use of wordnets and the APT_PL tagger (Pęzik and Laskowski, 2017) used for obtaining lemmas of the keywords[2]. In this step we decided to include synonymy, hypernymy, hyponymy, and cohyponymy relations to expand the sets of keywords for each section. The reason for choosing these relations were as follows. The synonyms represent words which can be used interchangeably in the company's descriptions, so the more synonyms we can gather the better the classification quality should be. Besides, the synonyms are available straightforwardly in the wordnets, since the basic building blocks are synsets. The hypernyms allowed us to gain some knowledge on more general terms describing the concepts represented by keywords, while hyponyms allowed us to get a more detailed view on them. The cohyponyms, although usually incompatible, were chosen to enable a broader view on the given concept. Note that, before introducing the words resulting from any of the relations mentioned above, we verified whether they will not change the value of W'_k to become greater than the assumed threshold value. Words that did not satisfy this condition were rejected.

4 Statistical Evaluation

To assess the database quantitatively we measured the sizes of the resource at the various design stages. In particular, we measured the initial size of the database, calculated at the end of *information processing* step, the sizes after the application of the W_k and W'_k metrics in the *keywords extraction* step and the final size of the database after *keywords enrichment* step. The observed size changes are reflected in Figure 1. As can be observed the use of W_k and W'_k metrics reduced the initial database size almost three times. On the other hand, using the wordnet we managed to increase the number of keywords significantly, since the number of unique synsets added was approximately equal to 50 500, which means over threefold increase in the number of keywords.

The distribution of the number of keyword can-

[2]The tagger enabled us to improve the coverage of keywords by plWordNet, providing the base word forms used also in the wordnet.

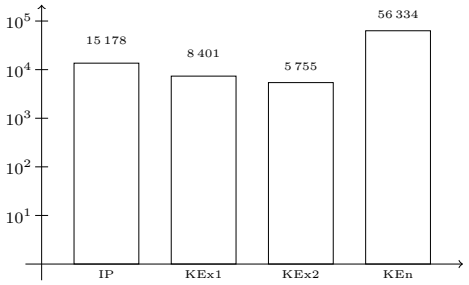


Figure 1: Changes in size of the keywords database after the *information processing* (IP) step, the W_k and W'_k metrics application in the *keyword extraction* (KEx1 and KEx2) step, and the *keyword enrichment* (KEn) step

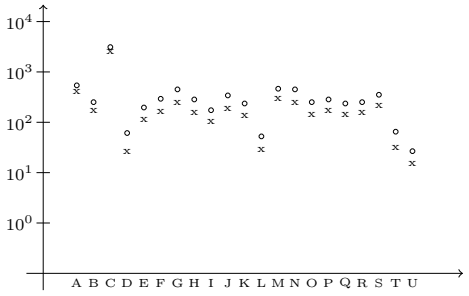


Figure 2: Keyword candidates number distribution in sections *A*–*U* of the PKD 2007. The circles represent the number of keyword candidates after W_k metric application, while the 'x' symbols denote the numbers after W'_k metric was used.

didates in the different sections is shown in Figure 2. The figure presents the information on the number of words remaining after the application of W_k (circles) and W'_k metrics ('x' symbols). It can be observed that the sections containing the fewest keyword candidates were D (Electricity, gas, steam, hot water and air conditioning manufacturing and supply), L (Real Estate Activities), T (Households as employers; goods-and-services-producing activities of households for own use), and U (Extraterritorial organisations and bodies). On the other hand, the section described by the largest number of keywords was section C (Manufacturing), which had over 3000 keyword candidates.

We also collected the information on the contribution of plWordNet towards the extension of the keywords database (Table 1). From the table it follows that the hyponymy and cohyponymy (Hyp and CoHyp columns) relations brought the largest number of keywords. Let us also note that the values presented in Table 1 are actually synsets, so the real number of words added to the database is even larger. The value given in the last column (Total) denotes the total number of unique synsets resulting from all four relations considered.

	Syn	**Hpr**	**Hyp**	**CoHyp**	**Total**
A	1187	1907	47 586	16 290	59 279
B	500	967	6131	7783	14 342
C	6695	5416	93 660	39 587	111 731
D	93	295	2683	1756	4595
E	369	840	7582	6874	14711
F	429	831	5854	5921	12 256
G	750	1404	11 674	10 542	22 533
H	472	945	6020	6845	13 364
I	321	680	5296	3929	9720
J	655	1061	59 855	7852	65 784
K	429	871	6815	5322	12 409
L	90	232	633	1623	2553
M	867	1418	47 887	10 784	55 646
N	785	1279	29 135	10 757	38 092
O	420	825	21 760	5529	26 926
P	603	1239	8844	7270	16 621
Q	382	811	5206	7033	12 929
R	478	1042	4078	6590	11 442
S	633	1181	51 828	8763	57 668
T	67	245	1169	886	2287
U	51	156	1184	1525	2825

Table 1: The number of synsets contributed by the synonymy (Syn), hypernymy (Hpr), hyponymy (Hyp), and cohyponymy (CoHyp) relations, and the total number of synsets added to the keywords database

We have observed that around 95% of initial keywords were found in plWordNet, which is a very good result. To further compare the respective sections, we have analyzed the keywords coverage percentage shown in Fig. 3. We noted that sections I, M and S were covered to the least extent. In case of sections I and S the missed keywords were usually quite specific, e.g., they were different hotel types (section I) or abbreviations (section S). In case of section M we noted the problems with the coverage of biomedical terms (see also (Poprat et al., 2008)). On the other end we observed the full coverage of sections T and

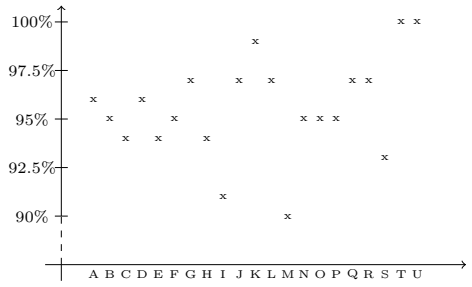


Figure 3: Keywords coverage in plWordNet (expressed as a percentage of the initial number of keywords)

U, which had relatively small number of not-so-specific keywords.

5 Summary

In the paper, we discussed the creation of a new resource related to the Polish Classification of Activities. The designed keywords database has been constructed on the basis of the official documentation related to the PKD 2007 hierarchy. The database was enriched with the use of plWordNet, the largest Polish wordnet. We used the synonymy, hypernymy, hyponymy and cohyponymy relations available in the wordnet. The results of our preliminary evaluation show that plWordNet can be a good source of information related to the activities of Polish companies.

In the future we plan to use the keywords database for the classification of companies to the respective industries given by PKD sections. We want to perform an analysis of multi-word expressions and a word-sense disambiguation step to include only the most relevant terms. Note however, that with the current design the database serves its purpose, because the not-related meanings will not appear in the company's description.

Acknowledgements

The research has been supported by the European Union under the Regional Operational Program of the Śląskie Voivodeship 2014-2020 within the project *Opracowanie zaawansowanych algorytmów automatycznego wspomagania procesów decyzyjnych w przedsiębiorstwach* ("Development of advanced algorithms for automatic decision support in enterprises") awarded to Technicenter Sp. z o.o.

References

Rodrigo Agerri and Ana Garcia-Serrano. 2010. Q-WordNet: Extracting polarity from WordNet senses. In *Proceedings of the 7th conference on Language Resources and Evaluation (LREC'10)*, pages 2300–2305.

Eneko Agirre, Xabier Arregi, and Arantxa Otegi. 2010. Document expansion based on WordNet for robust IR. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 9–17.

Council of Ministers. 2007. Regulation of the Council of Ministers of december 24th, 2007. JL No. 251, item 1885.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422.

EUROSTAT European Commission. 2006. Statistical classification of economic activities in the European Community NACE Rev. 2. WE 1893/2006.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Główny Urząd Statystyczny. 2007. Polska Klasyfikacja Działalności PKD 2007. Available at: `https://www.dane.gov.pl/media/resources/20151019/pkd2007.xml`, last accessed: 07/04/2019.

Arkadiusz Janz, Jan Kocoń, Maciej Piasecki, and Monika Zaśko-Zielińska. 2017. plWordNet as a basis for large emotive lexicons of Polish. In *Proceedings of the 8th Language & Technology Conference (LTC'17)*, pages 189–193.

Tomasz Jastrząb, Grzegorz Kwiatkowski, and Paweł Sadowski. 2016. Mapping of selected synsets to semantic features. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*, volume 613 of *CCIS*, pages 357–367, Cham. Springer.

Tomasz Jastrząb, Grzegorz Kwiatkowski, Paweł Sadowski, and Adam Dyrek. 2017. A comparison of Polish wordnets in the view of semantic features mapping. In *Man-Machine Interactions 5*, volume 659 of *AISC*, pages 375–386, Cham. Springer.

Paweł Kędzia and Maciej Piasecki. 2014. Ruled-based, interlingual motivated mapping of plWordNet onto SUMO ontology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pages 4351–4358.

Daniel Krasnokucki, Grzegorz Kwiatkowski, and Tomasz Jastrząb. 2017. A new method of XML-based wordnets' data integration. In *Beyond Databases, Architectures and Structures. Towards*

Efficient Solutions for Data Analysis and Knowledge Representation, volume 716 of *CCIS*, pages 302–315, Cham. Springer.

Grzegorz Kwiatkowski and Tomasz Jastrząb. 2016a. An experimental comparison of Polish wordnets in the context of emotions analysis. In *Badania i Rozwój Młodych Naukowców w Polsce – Nauki Techniczne i Inżynieryjne*, pages 60–66.

Grzegorz Kwiatkowski and Tomasz Jastrząb. 2016b. A survey of wordnets' applications to sentiment analysis and related problems. In *Badania i Rozwój Młodych Naukowców w Polsce – Nauki Techniczne i Inżynieryjne*, pages 54–60.

Chaya Liebeskind, Ido Dagan, and Jonathan Schler. 2018. Automatic thesaurus construction for modern Hebrew. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1446–1451.

Bernardo Magnini, Carlo Strapparava, Fabio Ciravegna, and Emanuele Pianta. 1994. A project for the construction of an Italian lexical knowledge base in the framework of WordNet. Technical Report 9406-15, IRST.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268.

John McCrae. 2018. Mapping WordNet instances to Wikipedia. In *Proceedings of the 9th Global WordNet Conference*.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

George A. Miller. 1990. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, pages 412–416.

Piotr Pęzik and Sebastian Laskowski. 2017. Evaluating an averaged perceptron morphosyntactic tagger for polish. In *Proceedings of the 8th Language & Technology Conference (LTC'17)*, pages 372–376.

Michael Poprat, Elena Beisswanger, and Udo Hahn. 2008. Building a BioWordNet by using WordNet's data formats and WordNet's software infrastructure – a failure story. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 31–39. ACL.

Igor Postanogov and Tomasz Jastrząb. 2017. Ontology reuse as a means for fast prototyping of new concepts. In *Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation*, volume 716 of *CCIS*, pages 273–287, Cham. Springer.

Princeton University. 2010. About WordNet. Related projects. `https://wordnet.princeton.edu/related-projects`.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3679–3686. European Language Resources Association (ELRA).

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A core of semantic knowledge unifying WordNet and Wikipedia. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web*, pages 697–706. ACM.

Zygmunt Vetulani. 2014. *Komunikacja człowieka z maszyną*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Netwoks*. Kluwer Academic Publisher.

Tingting Wei, Yonghe Lu, Huiyou Chang, QiangZhou, and Xianyu Bao. 2015. A semantic approach for text clustering using WordNet and lexical chains. *Expert Systems with Applications*, 42(4):2264–2275.

Propagation of emotions, arousal and polarity in WordNet using Heterogeneous Structured Synset Embeddings

Jan Kocoń*, **Arkadiusz Janz***, **Monika Riegel**†, **Małgorzata Wierzba**†,
Artur Marchewka†, Agnieszka Czoska‡, Damian Grimling ‡,
Barbara Konat‡¡, Konrad Juszczyk‡¡, Katarzyna Klessa¡, Maciej Piasecki*
*Wroclaw University of Science and Technology
† Laboratory of Brain Imaging, Nencki Institute of Experimental Biology of Polish Academy of Sciences
‡W3A.PL Sp. z o.o., ¡Adam Mickiewicz University
{jan.kocon,arkadiusz.janz,maciej.piasecki}@pwr.edu.pl,
{a.marchewka,m.riegel,m.wierzba}@nencki.gov.pl,
{agnieszka,damian,barbara,konrad}@sentimenti.pl, klessa@amu.edu.pl

Abstract

In this paper we present a novel method for emotive propagation in a wordnet based on a large emotive seed. We introduce a sense-level emotive lexicon annotated with polarity, arousal and emotions. The data were annotated as a part of a large study involving over 20,000 participants. A total of 30,000 lexical units in Polish WordNet were described with metadata, each unit received about 50 annotations concerning polarity, arousal and 8 basic emotions, marked on a multilevel scale. We present a preliminary approach to propagating emotive metadata to unlabeled lexical units based on the distribution of manual annotations using logistic regression and description of mixed synset embeddings based on our Heterogeneous Structured Synset Embeddings.

1 Introduction

Rapid growth of interest in sentiment analysis comes from its vast potential in automatic detection of subjectivity (whether the text expresses a subjective opinion rather than an objective fact) and polarity (whether the expressed opinion is positive, negative, or neutral) in large amount of textual data. For instance, sentiment analysis systems proved useful in automatic analysis of many different kinds of textual data, such as emails, tweets, blogs, reviews, newspaper headlines or novels (Dodds et al., 2015; Mohammad, 2016). Whereas introduction of advanced computational methods (e.g. machine learning) to natural language processing resulted in the development of sentiment analysis methodologies, the scarcity of high quality and large scale data sources greatly constrains their usage.

Numerous attempts were made to annotate words in terms of emotions for various languages (Riegel et al., 2015). However, such datasets are typically limited in size and consist of several thousands of words, while natural lexicons are known to be much bigger. Since annotations are provided manually by either qualified experts (usually 2-3 independent annotators) or a group of naive participants, data collection is typically very expensive in terms of time and money. Therefore, most of the available resources describe word meanings in terms of polarity, without further distinguishing various emotion categories attributed to them.

In emotion research, words are usually characterized according to two dominant theoretical accounts on the nature of emotion: dimensional account and categorical account. According to the first one proposed in (Russell and Mehrabian, 1977), each emotional state can be represented by its location in a multidimensional space, where valence or polarity (ranging from negativity to positivity) and arousal (from low to high) explain most of the observed variance. A competing account distinguishes several basic categories of emotional states, with more complex, subtle emotion states emerging as their combination. There have been various interpretations of the basic emotions concept, and different numbers of emotion categories were proposed by different theories, with (Ekman, 1992) and (Plutchik, 1982) gaining most recognition in the scientific community.

In this work, we used a large dataset described in (Kocoń et al., 2019a), containing metadata for a total of over 30000 word meanings from Polish WordNet (Piasecki et al., 2009), annotated

in terms of polarity, arousal, as well as 8 basic emotion categories (i.e. joy, sadness, trust, disgust, fear, anger, surprise, anticipation). Here, we present a novel propagation approach to automatically extend the original dataset by deriving emotion metadata for lexical units and synsets that are not present in this dataset. If effective, our approach could alleviate the problem of data scarcity and facilitate the widespread use of sentiment analysis in various applications including but not restricted to artificial intelligence, computational linguistics, psychology or business.

2 Dataset description

In the Sentimenti database (Kocoń et al., 2019a), a total of over 20,000 unique respondents (with approximately equal number of male and female participants) was sampled from Polish population. Multiple demographical characteristics such as: sex, age, native language, place of residence, education level, marital status, employment status, political beliefs and income were controlled. The annotation schema was based on the procedures most widely used in previous studies, aiming to create the first standardized datasets of Polish words characterized in terms of emotion (NAWL, (Riegel et al., 2015); NAWL BE, (Wierzba et al., 2015); plWordNet-emo (Zaśko-Zielińska et al., 2015; Janz et al., 2017)). Thus, we collected annotations of valence (polarity), arousal, as well as eight emotion categories: joy, sadness, trust, disgust, fear, anger, surprise and anticipation. By combining simple annotation schema with crowd annotation, we were able to effectively acquire large amount of data and preserve its high quality at the same time.

The total number of over 30,000 word meanings was annotated, with each meaning ranked at least 50 times on each scale. Moreover, in a follow-up study a total number of over 7,000 texts (short phrases or paragraphs of text) were annotated in the same way, with each text assessed at least 25 times on each scale. Before attempting the assessment task, subjects were instructed to rank word meanings rather than words, as well as encouraged to indicate their immediate, spontaneous reactions. Participants had unlimited time to complete the task and they were able to quit the assessment session at any time and resume their work later on. The source of texts were reviews from two domains: *medicine*[1] (2000 reviews) and *hotels*[2] (2000 reviews). Due to difficulties in observing neutral reviews in the selected sources, we have chosen them from websites describing medical information[3] (500 paragraphs) and the hotel industry[4] (500 paragraphs). We also selected phrases using *lexico-semantic-syntactic patterns* (LSS) manually created by linguists to capture one of the four effects affecting sentiment: *increase, decrease, transition, drift*. Most of these phrases belong to the previously mentioned subject areas. The source for the remaining phrases were Polish WordNet glosses (Piasecki et al., 2009).

2.1 Data conversion

We decided to treat the problem of emotive propagation as a multilabel classification task, where the individual lexical units are classified considering all emotive categories. Eight basic emotions and arousal were annotated on a scale of integers from range $[0, 4]$. The valence was annotated using $[-3, 3]$ scale. To perform the classification task a proper conversion schema should be applied. For most of emotive dimensions we used a simple averaging strategy, where the final score is an average value of all assigned scores, normalized to the range $[0, 1]$. In the case of valence scores we divided the annotations into two separate groups: positive scores (Valence$_p$) and negative scores (Valence$_n$). This division results from the fact that some texts have mixed annotations, both positive and negative. To keep the original distribution of valence scores we decided to use a separate approach (see Algorithm 1). The positive scores were separated from negative ones to measure the degree of positive valence (valence$_p$) and negative valence (valence$_n$). With this approach a single lexical unit obtains two normalized valence scores.

We decided to partition all scores for each dimension into two clusters using k-means clustering (Hartigan and Wong, 1979). We assign a label representing a membership of an individual lexical unit to specific emotive category if the final score of a lexical unit in the dimension representing that category is greater than the threshold determined by k-means. Each lexical unit can be described by multiple categories, thus we might obtain multiple

[1] www.znanylekarz.pl
[2] pl.tripadvisor.com
[3] naukawpolsce.pap.pl/zdrowie
[4] hotelarstwo.net, www.e-hotelarstwo.com

Algorithm 1 Estimating the average value of positive and negative valence for a single lexical unit.

Require: V: list of all valence scores; $m = 3$: maximum absolute value of polarity

Ensure: Pair (p, n) where p is average positive valence, and n is average negative valence

```
(p, n) = (0, 0)
for v ∈ V do
    if v < 0 then n = n + |v| else p = p + v
return (p ÷ (|V| · m), n ÷ (|V| · m))
```

labels assigned to a single lexical unit.

2.2 Polarity transfer from units to synsets

On the basis of the previous work where we analysed the contemporary annotation of plWordNet to see how diverse synsets are in terms of lexical units valence, we assumed that we can average all dimensions of lexical units (which belong to synset A) separately to obtain the metadata description of synset A. Previously acquired statistics show that synsets are strongly homogeneous in terms of the units valence, so we decided to move annotations from unit-level to synset-level in that way (Kocoń et al., 2018a; Kocoń et al., 2018b).

3 Emotive Propagation

In this study, we decided to follow the idea presented in (Kocoń et al., 2018a; Kocoń et al., 2018b). The idea is to apply a semi-supervised learning on a large seed of emotively annotated synsets. The seed is used to train a classifier and then to predict the polarity categories for unlabeled synsets in a close vicinity of labeled ones. Starting from the labeled synsets we visit their neighbors and annotate them iteratively by applying our classifier. The propagation process ends when all synsets become annotated.

In (Kocoń et al., 2018b) the authors proposed a rich set of wordnet-based features to describe the synsets representing an initial seed for emotive propagation. For every synset existing in a seed they extracted the features capturing the structure of its neighborhood by taking into account the neighboring synsets (with their relative location) and assigned polarities. The features were generated on a basis of a template being defined in terms of 4 feature variables. A single feature is generated by initializing the variables in the template with a specific combination of possible values. The variables in the template are defined as follows:

- *Relation* – one of the 13 most common WordNet relations,
- *Direction* – the direction of the relation,
- *Element* – a type of an element used to construct a *bag-of-words* model; two types of *elements* were used: *synset_ID* (any number) and *synset_polarity* (one of the following numbers: $-1, 0, 1$; it represents 3 polarity classes: *negative, neutral, positive*),
- *Level* – a distance (number of hops) between the initial synset being described and its neighbors, e.g. the synsets of *second level* means neighboring synsets in a distance of two hops (excluding the synsets of *first level*).

Extracted features were converted to a *bag-of-words* model, where the elements of a bag were representing the synsets or their polarities. Then the authors used these features as a signal for a classifier to decide whether a given synset should be positive, negative, neutral or ambiguous.

Such an approach generated vectors of very large dimensions, thus it increases the overall propagation time. The classification procedure was time-consuming because the process of feature generation produced high-dimensional data.

3.1 Heterogeneous Structured Synset Embeddings

To reduce the dimensionality of the input feature space, we decided to build upon the methods designed for embedding the lexical knowledge bases. The main aim is to produce a meaningful vector space representation of concepts existing in a knowledge base by capturing their lexico-semantic properties and embedding the structure of their neighborhood. In our case the concepts are represented by synsets and we utilize the structure of a wordnet to construct synset embeddings. Our approach is based on the skip-gram model (Mikolov et al., 2013) which takes as its input a large textual corpus and produces a distributional representation of words by capturing the neighboring words appearing in a small context window. The main assumption is that the words sharing the similar context should have similar vector space representations. The neural network based on skip-gram architecture learns the

vector space representations of words (existing in a corpus) by minimizing the loss in the task of *predicting the context words given the input word*.

To produce synset embeddings we adapted the solution presented in (Goikoetxea et al., 2015). The authors generated an artificial wordnet-based corpus to train a skip-gram model by performing multiple random walks on a wordnet. The input corpus consisted of synset identifiers generated during random walking process. Previous solutions were limited only to synset links and did not include the information about relation types. We decided to expand this idea by including the lexical units and their links while generating the corpus with a random walk. Thus, we might obtain a corpus with artificial words (elements) represented by the identifiers of synsets, lexical units, and the identifiers of relation types. In the case of relation types we decided to differentiate the identifiers depending on the type of linked concepts, e.g. the identifiers starting with the rSS prefix represent the relations between synsets while rSL (or rLS) represent the links between synsets and lexical units. The elements with the rSS prefix have an additional identifier representing a specific type of wordnet relation (e.g. 10 represents $hyponymy$). The additional information about the types of links should lead to obtaining a more heterogeneous and accurate embeddings of synsets. A short random walk of length 13 can be represented as the following sequence of elements:

```
s_7078349 -> rSS_136 -> s_60485 -> rSL
-> l_85957 -> rLS -> s_60485 -> rSS_10
-> s_22456 -> rSS_11 -> s_55576 -> rSS_11
-> s_55575 -> rSS_11 -> s_7077974
-> rSS_10 -> s_55575 -> rSL -> l_79892
-> rLL_3425 -> l_10483 -> rLS -> s_3974
-> rSS_11 -> s_7077977
```

To generate the embeddings, we used a popular FastText method (Bojanowski et al., 2017; Joulin et al., 2017). This method was used in many different NLP tasks, especially with applications to sentiment analysis e.g. hate speech detection (Badjatiya et al., 2017), sentiment polarity recognition, emotion and sarcasm identification (Felbo et al., 2017), aspect-based sentiment analysis in social media (Wojatzki et al., 2017), text classification on multiple sentiment datasets (Joulin et al., 2017).

3.2 Emotive classification

The knowledge base embeddings alone might be insufficient to successfully solve downstream tasks due to the lack of contextual information connecting these embeddings with real world data. To prepare a more meaningful and contextual representation for our emotive classifier we decided to augment our model with plain word embeddings. In (Kocoń et al., 2018a) the authors showed that the size and the quality of training corpora might affect the overall performance in downstream tasks. They also tested several parameter settings of word embeddings for Polish language using the implementation of CBOW and skip-gram methods provided with FastText (Bojanowski et al., 2017). These models are available under an open license in the CLARIN-PL project DSpace repository[5]. With these embeddings, the best results were obtained in two NLP tasks: recognition of temporal expressions (Kocoń and Gawor, 2018; Kocoń et al., 2019b) and recognition of named entities (Marcińczuk et al., 2018).

We used the same model to produce a complementary feature space for the task of polarity prediction. To prepare a complementary embedding space for HSSE we averaged the embeddings of lemmas linked with the synsets through their lexical units. For each synset in a seed we computed the averaged embedding of its lemmas and concatenated it with HSSE embedding.

Our emotive classifier is represented as an ensemble of binary classifiers, each one predicting one of 11 emotive categories. Each classifier in this ensemble was trained on a seed of synsets representing a specific emotive category with its positive and negative examples.

4 Results and Discussion

Comparing the F1 scores of the model in a task of valence, arousal and emotion propagation, we can observe that the valence propagation performs better than the propagation of most emotive dimensions (Table 1). The best results are obtained for *trust* (F1-score: 77.48%) and anticipation (F1-score: 74.94%). F-score for all dimensions except *disgust* are above 63%.

5 Conclusions

The results of the propagation task suggest that the novel method presented here can be useful for both valence and emotions datasets. Heterogeneous Structured Synset Embeddings allow for effective scaling up of annotated datasets.

[5] https://clarin-pl.eu/dspace/handle/11321/606

Dim.	P[%]	R[%]	F[%]
Valence$_p$	70.18	81.17	75.27
Valence$_n$	65.57	85.26	74.12
Arousal	66.74	75.76	70.97
Joy	59.27	80.56	68.29
Surprise	59.55	69.37	64.07
Anticip.	70.58	79.12	74.94
Trust	74.86	80.46	77.48
Sadness	59.18	83.16	69.13
Anger	56.99	84.14	67.93
Fear	51.98	81.40	63.43
Disgust	45.50	82.65	58.71

Table 1: Precision, recall and F1-score for Valence, Arousal and Emotion propagation.

The paper also describes a large dataset: "Sentimenti" which covers more than 30,000 word-meanings in Polish annotated with 8 basic emotions as well as polarity and arousal.

The data gathered in the psycholinguistic study were used to enhance affective annotation of Polish WordNet (Piasecki et al., 2009). Tthe emotive metadata were propagated to unlabeled lexical units, enabling emotive categorization of the whole WordNet. Categorization results proved that the ML methods trained on the data enriched with detailed description of synset features and relations between lexical units and synsets resulted in effective models for emotional metadata propagation.

Such metadata propagation methods are successful only when based on large data sets and multilevel annotations of a given WordNet. In the current study, we showed that their effectiveness can be high also for very subjective emotional features, especially valence propagation. Emotional metadata propagation for Polish proved to have a high accuracy for most of the emotion values. However, a question remains why some of the emotion values were attributed less effectively - whether it was the lack of input data or it is a property of those emotions expressed verbally to be less evident and more dispersed.

Acknowledgments

Funded by National Centre for Research and Development, Poland, under grant “Sentimenti – emotions analyzer in the written word” no POIR.01.01.01-00-0472/16.

References

[Badjatiya et al.2017] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

[Bojanowski et al.2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[Dodds et al.2015] Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, Karine Megerdoomian, Matthew T McMahon, Brian F Tivnan, and Christopher M Danforth. 2015. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences of the United States of America*, 112(8):2389–94.

[Ekman1992] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

[Felbo et al.2017] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

[Goikoetxea et al.2015] Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2015. Random walks and neural network language models on knowledge bases. In *Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 1434–1439.

[Hartigan and Wong1979] John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

[Janz et al.2017] Arkadiusz Janz, Jan Kocoń, Maciej Piasecki, and Zaśko-Zielińska Monika. 2017. plWordNet as a Basis for Large Emotive Lexicons of Polish. In *LTC'17 8th Language and Technology Conference*, Poznań, Poland, November. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

[Joulin et al.2017] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Va-

lencia, Spain, April. Association for Computational Linguistics.

[Kocoń and Gawor2018] Jan Kocoń and Michał Gawor. 2018. Evaluating KGR10 Polish Word Embeddings in the Recognition of Temporal Expressions Using BiLSTM-CRF. *Schedae Informaticae*, 2018(Volume 27).

[Kocoń et al.2018a] Jan Kocoń, Arkadiusz Janz, and Maciej Piasecki. 2018a. Classifier-based Polarity Propagation in a Wordnet. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*.

[Kocoń et al.2018b] Jan Kocoń, Arkadiusz Janz, and Maciej Piasecki. 2018b. Context-sensitive Sentiment Propagation in WordNet. In *Proceedings of the 9th International Global Wordnet Conference (GWC'18)*.

[Kocoń et al.2019a] Jan Kocoń, Arkadiusz Janz, Piotr Miłkowski, Monika Riegel, Małgorzata Wierzba, Artur Marchewka, Agnieszka Czoska, Damian Grimling, Barbara Konat, Konrad Juszczyk, Katarzyna Klessa, and Maciej Piasecki. 2019a. Recognition of emotions, valence and arousal in large-scale multi-domain text reviews. In Zygmunt Vetulani and Patrick Paroubek, editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics*. Wydawnictwo Nauka i Innowacje, Poznań, Poland, May.

[Kocoń et al.2019b] Jan Kocoń, Marcin Oleksy, Tomasz Bernaś, and Michał Marcińczuk. 2019b. Results of the PolEval 2019 Shared Task 1: Recognition and Normalization of Temporal Expressions. *Proceedings of the PolEval2019 Workshop*.

[Marcińczuk et al.2018] Michał Marcińczuk, Jan Kocoń, and Michał Gawor. 2018. Recognition of Named Entities for Polish-Comparison of Deep Learning and Conditional Random Fields Approaches. In Maciej Ogrodniczuk and Łukasz Kobyliński, editors, *Proceedings of the PolEval 2018 Workshop*, pages 77–92. Institute of Computer Science, Polish Academy of Science.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mohammad2016] Saif M. Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herb Meiselman, editor, *Emotion Measurement*. Elsevier.

[Piasecki et al.2009] Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

[Plutchik1982] Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

[Riegel et al.2015] Monika Riegel, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka. 2015. Nencki Affective Word List (NAWL): the cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

[Russell and Mehrabian1977] James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273 – 294.

[Wierzba et al.2015] Małgorzata Wierzba, Monika Riegel, Marek Wypych, Katarzyna Jednoróg, Paweł Turnau, Anna Grabowska, and Artur Marchewka. 2015. Basic emotions in the Nencki Affective Word List (NAWL BE): New method of classifying emotional stimuli. *PLOS ONE*, 10(7):1–16.

[Wojatzki et al.2017] Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, pages 1–12.

[Zaśko-Zielińska et al.2015] Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for Polish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 721–730.

Testing Zipf's meaning-frequency law with wordnets as sense inventories

Francis Bond,♠ Arkadiusz Janz,♢ Marek Maziarz♢ and Ewa Rudnicka♢
♠ Nanyang Technological University, Singapore
♢ Wrocław University of Science and Technology, Poland
bond@ieee.org, {arkadius.janz|marek.maziarz|ewa.rudnicka}@pwr.edu.pl

Abstract

According to George K. Zipf, more frequent words have more senses. We have tested this law using corpora and wordnets of English, Spanish, Portuguese, French, Polish, Japanese, Indonesian and Chinese. We have proved that the law works pretty well for all of these languages if we take - as Zipf did - mean values of meaning count and averaged ranks. On the other hand, the law disastrously fails in predicting the number of senses for a single lemma. We have also provided the evidence that slope coefficients of Zipfian log-log linear model may vary from language to language.

1 Introduction

The dependency between meaning and frequency is undisputable. Since Zipf's discovery of the high correlation between mean sense count and mean rank (Zipf, 1945), the law was confirmed by several research teams. Among many Zipfian laws, the modelling of the law of meaning-frequency dependency is probably the most fascinating one, because it directly concerns semantics. Meaning strongly influences word frequency (Piantadosi, 2014) and it is clear that semantics precedes language form in text generation (Ferrer-i-Cancho, 2018).[1]

Originally, Zipf tested the law on Thorndike's list of 20k most frequent words of standard English[2] and meanings taken from the *Thorndike-Century Senior Dictionary*[3] (Zipf, 1945). The dictionary meaning account was based on the actual usage in English newspapers, so there were no obsolete or rare senses. The corpus itself was 10^7 running words large, the lemmas on the frequency list were divided into bins of 500 and 1,000 words. Zipf proved a very strong correlation between the *average* number of word senses and rank of lemmas (Zipf, 1945, p. 253), formulating the following statistical law (Zipf, 1949, ch. 3):

$$m_i \propto {f_i}^{\delta} \tag{1}$$

where i is a given word's rank, f_i is its frequency, m_i represents the number of lemma meanings, Zipf also claimed that the coefficient $\delta \approx \frac{1}{2}$. Taking the logarithm of both sides leads to the equation in 2:

$$log_{10}(m_i) \propto \delta \cdot log_{10}(f_i) \tag{2}$$

The corresponding equation for the meaning-rank law was formulated as follows:

$$log_{10}(m_i) \propto -\gamma \cdot log_{10}(i) \tag{3}$$

where i is a word rank. Zipf thought that $\gamma = \delta$.

Zipf justified the straight meaning-rank line in log-log scale with the "conflicting Forces of Unification and Diversification". While a lazy speaker would always tend to use only a few highly frequent and strongly polysemous words, a demanding hearer would prefer numerous unequivocal/monosemous words. Since these balancing forces act within each frequency bin, language equips more frequent words with more senses to maintain a constant ('compromise') polysemy ratio (Zipf, 1949).

He argued that the slope coefficient was close to 0.5, which is now called the *strong* Zipf's law

[1] Consider, for instance, a simple model of a random walk on an undirected graph of lexico-semantic relations. The stationary probabilities of lending in each vertex are proportional to the degree of the vertices (cf. Avrachenkov et al. (2015), Lovász and Winkler (1995, p. 5)), that is to the number of sense relations, including the number of interconnected polysemous senses of the same lemma. As a result, one gains more polysemous words being chosen more frequently.

[2] *A Teacher's Word Book of 20,000 Words*, New York: Teachers College, 1932.

[3] New York: Appleton-Century, 1941.

(Ferrer-i-Cancho, 2016), although the exact value was in fact 0.466 (Zipf, 1945).[4]

Although he only proved the dependency between the mean number of senses $\bar{m}$ and the mean rank $\bar{i}$ within each frequency bin,[5] Zipf was sure that the law (1) was applicable to every single lemma:

> (…) if we had a rank-frequency distribution of the 20,000 most frequent words of the Thorndike analysis, it would probably be rectilinear (…), at least for the first 10 to 12 thousand most frequent words. (Zipf, 1949, ch.3)

A more recent verification of Zipf's meaning-frequency law has revealed that the relationship is more complex than could have been foreseen in the middle of the 20^{th} century.

The aim of this paper is to provide new broad empirical evidence for the *weakened* version of Zipf's meaning-frequency law[6] based on corpora and wordnets as sense inventories. Five European languages (English, Spanish, Portuguese, French and Polish) and three Asian languages (Mandarin, Indonesian and Japanese) representing four distinct language families (Indo-European, Sino-Tibetan, Japonic and Austronesian) were inspected.

All data sets and source code are available at `https://github.com/MarekMaziarz/Zipf-s-Meaning-Frequency-Law`.

2 Related Work

Despite the fact that most of Zipf's laws, like the law of frequency-rank distribution or the law of abbreviation, were studied thoroughly, the meaning-frequency law itself gained relatively less attention (Casas et al., 2019). Still, some attempts were made by several research teams.

Edmonds (2004) repeated Zipf's experiment on the British National Corpus with the use of Princeton WordNet 2.0. He gathered lemmas in bins of 100 words each and estimated the γ coefficient at 0.40 (cf. Table 1).

experiment	**lang.**	γ
Zipf, 1945, 1949	en	.47
Edmonds, 2004	en	.40
Ilgen & Karaoglan, 2007	tr	.42, .39
Casas et al. 2019	en	.38
Casas et al. 2019	es	.27
Casas et al. 2019	nl	.25

Table 1: The power γ of Zipf's law exponent of Eq. (3) in hitherto experiments, for bins of 500 (with exception of Zipf's paper and Edmonds' article, details in text).

Ilgen and Karaoglan (2007) tested the law on two Turkish corpora (newspapers, novels, short-stories), one of which was manually tagged with word senses, while the second one was compared to an electronic Turkish dictionary. The authors tested different frequency bin sizes ranging from 50 words up to 1000 words, showing gradual predictive power loss while moving from larger to smaller bins. They obtained the slope coefficient slightly lower than that of Zipf's (0.42 and 0.39; Table 1).

Hernández-Fernández et al. (2016) tested the robustness of Zipf's meaning law on two different corpora (child and child-directed speech corpus – CHILDES[7], and the SemCor corpus) and two sense inventories (WordNet and WordNet senses that appear in SemCor).[8] The authors merged the resources in different combinations which, surprisingly, in all cases led to non-zero correlation coefficients. Unlike previous parametric research, the authors did not focus on mean values of sense count and frequency, but estimated direct correlations between row values of the two. They focused on those parts of speech that were present in WordNet (nouns, adjectives, verbs and adverbs). The authors concluded that positive and statistically significant correlation between sense count and frequency seemed to be corpus-independent.

In Casas et al. (2019) the above non-parametric approach was expanded to two other European languages: Dutch and Spanish; English is analysed again. The sources of frequency were the CHILDES corpus and Wikipidia, while the sources of sense inventories were wordnets: Word-

[4] Provided the first 500 words were omitted, which Zipf tended to treat as function words.

[5] Thorndike's frequency list divided words into bins of 500 and 1,000 words without giving any precise information about the exact number of occurrences of each lemma.

[6] *Strong* Zipf's law of meaning-frequency relationship forces the slope coefficient γ to be equal to 0.5, while *weak* version of the law simply states that $\gamma > 0$ (Ferrer-i-Cancho, 2016).

[7] `https://childes.talkbank.org/`

[8] `http://web.eecs.umich` They also took frequency counts from the English part of the CELEX corpus `edu/mihalcea/downloads.html#semcor`

Net, Open Dutch WordNet and the Multilingual Central Repository for Spanish. Each wordnet is "a proxy for the number of meanings of a word" (*ibidem*).

Casas and colleagues (Casas et al., 2019) also did some experiments with parametric modelling with wordnets as sense inventories and CHILDES corpus as a source of frequencies. They calculated slope coefficients and R-squared values for bins of 100 and 500 words (cf. Tab. 1). They also observed that smaller bins gave worse R-squared statistics.

The criticism of the rank-frequency models was addressed by Piantadosi (2014). Piantadosi raised an important question of the explanatory validity of Zipf's law derivation in various theoretical models. Since there are dozens of different ways of deriving the Zipf's equations (such as random-typing, stochastic models, semantic accounts, communicative accounts etc.), the derivation lacks its explanatory power and "[t]he key will be (...) to generate novel predictions and to test their underlying assumptions with more data than the law itself" (*ibidem*).

Altmann and Gerlach (2016) argue that linguistic statistical models should be validated not only by measures of fit like R-squared determination coefficient, but also with additional measures of randomness of model residuals (they propose significance level set at 1%): "A low p-value is a strong indication that the null model is violated and may be used to refute the law (e.g., if p-value < 0.01)." According to them, ordinary Zipfian rank-frequency linear models unfortunately lack this randomness property (p-values $\ll 0.01$). Piantadosi (2014) similarly points that rank-frequency models based on corpus data when analysed in a standard way (i.e., on the same sample), suffer from correlated errors, since the ranking is constructed out of the very same frequency distribution as frequency estimation itself. Luckily, this argument cannot be applied to the same extent to meaning-frequency and meaning-rank distributions, since they are prepared with either frequency, or rank at once. Especially, if the Zipf's meaning-frequency law (or meaning-rank law) is modelled on the basis of different language resources (like a wordnet and a corpus) the problem vanishes.

3 Method

We checked the validity of Zipf's meaning-rank law by collating frequency counts and *corresponding* meaning counts. We did this by comparing general corpora, representing language in usage, and sense numbers taken from wordnets, which represent each language lexical system, cf. Fellbaum (1998). Another way to see these two sides of language reality is to compare frequencies and polysemy count in the very same text (in a widely sense tagged corpus). We did this on the richly annotated Sherlock Holmes subcorpus of Nanyang Technological University Multilingual Corpus, *NTU-MC* (Bond and Tan, 2012).

3.1 Data sets: Wordnets

We treat wordnet as a useful model of human mental lexicon, and wordnet sense numbers as the approximation of real polysemy of a lemma. The choice of wordnets is motivated by their shared properties (e.g. similar relational description models, existence of synsets, glosses) which allow us to directly compare Zipfian curves for different languages. For the purposes of our study, we have chosen eight wordnets. The wordnets include: Princeton WordNet (henceforth, PWN) (Fellbaum, 1998), Polish WordNet (henceforth, plWN) (Maziarz et al., 2016), Wordnet Libre du Français (henceforth, WOLF), Multilingual Central Repository (henceforth, MCR) (Gonzalez-Agirre et al., 2012), Japanese Wordnet (henceforth, WNJA) (Bond et al., 2008), Wordnet Bahasa (WNB) (Bond et al., 2014), and Chinese Open Wordnet (henceforth, COW) (Wang and Bond, 2013). The wordnets are listed in Table 2 together with languages they represent, number of lemmas from wordnets and corpus coverage. They all appear in the Open Multilingual WordNet 1.0[9] (Bond and Foster, 2013) and are thus inter-linked via PWN. The numbers are given with the exclusion of multi-word lexical units and synsets not linked to Princeton WordNet and, hence, not linked to CILI.

3.2 Data sets: Corpora

To test Zipf's meaning-frequency law, we have inspected two types of text data sets: (i) general corpora for English, Spanish, French, Portuguese, Chinese, Japanese and Polish built at Centre for

[9] http://compling.hss.ntu.edu.sg/omw/

wordnet	lang.	#S [10^3]	#L [10^3]	poly. #S/#L
COW$^+$	zh	8.1	3.2	2.53
WNJA	jp	158.1	92.0	1.72
MCR	es	57.8	36.7	1.58
OpenWN-PT	pt	74.0	54.0	1.37
plWN$^+$	pl	288.4	191.8	1.50
PWN$^+$	en	218.6	159.4	1.37
WNB	id	95.3	26.9	3.54
WOLF	fr	102.7	55.4	1.85

Table 2: Data sets: OMW wordnets. Symbols: COW - Chinese Open Wordnet, WNJA - Japanese Wordnet, MCR - Multilingual Central Repository, OpenWN-PT - Open Portuguese Wordnet, plWN - Polish WordNet, PWN - Princeton WordNet, WNB - Wordnet Bahasa, WOLF - Wordnet Libre du Français; #S - number of senses, #L - number of lemmas, poly. - average polysemy; $^+$ – wordnet taken in whole. Please note that for most wordnets we have taken only PWN equivalents (connected via (C)ILI). All numbers are given for one-word lexical units only.

corpus	size [10^9]	min f	L [10^3]	cov. [%]
en-IC	.18	218	14.5	72
en-RC	.10	1,100	3.8	70
pl-IC	1.80	10,972	8.8	88
es-IC	.14	248	7.2	48
fr-IC	.18	2,080	4.3	86
pt-IC	.19	2,400	4.0	80
zh-IC*	.28	183	11.0	22
zh-GC*	.24	377	7.0	28
jp-IC	.25	567	10.0	67
en-SH	.02	1	2.8	56
id-SH	.01	1	1.8	48
jp-SH	.03	1	3.1	37
zh-SH	.02	1	3.8	67

Table 3: Data sets: corpora. Symbols: en - English, es - Spanish, id - Indonesian (Bahasa), jp - Japanese, fr - French, pl - Polish, pt - Portuguese, zh - Chinese (Mandarin); IC - internet corpus, RC - Reuters Corpus, GC - Gigaword Corpus, SH - NTU-MC subcorpus of Sherlock Holmes stories; * - word frequency list; f – corpus frequency, L – number of lemmas given for frequency lists united with each wordnet list, cov – coverage of the original frequency list as covered by a particular wordnet.

Translation Studies, University of Leeds[10], and at Wroclaw University of Science and Technology, Poland,[11] (Broda et al., 2010), as well as (ii) a part of the NTU-MC, containing two Sherlock Holmes stories and their translations into Indonesian, Chinese and Japanese, henceforth: *SH* (Bond and Tan, 2012). All used frequency lists are available under open licences.

Corpus statistics are presented in Table 3. Most general corpora are collections of Web documents (marked as *IC*) gathered by Web crawling within the WaCky project (Baroni et al., 2009), covering 100–300 million running words each. The Web as a Corpus approach was also used to make a corpus of Polish, the largest one, comprising almost 2 billion words, the source of lemma frequency list in the case of Polish (Maziarz et al., 2016). To analyse the impact of the used corpora on our results we made use of frequency lists for Reuters Corpus (a collection of news from Reuters, *RC*) and Gigaword Corpus for Chinese (henceforth: *GC*)[12] Frequency lists for Chinese were word form based,[13] whereas all the rest of the lists contained lemma frequencies.

In order to make sure that our lists contain only content words we threw out all words of rank 100 and above (rank $i \leq 100$). On the other hand, all frequency lists were shortened at different cut-off points. For instance, the Reuters Corpus was clipped down to the rank i = 5,000 (1,100 occurrences in the corpus), while Chinese Gigaword corpus was cut at the rank i = 25,000 (377 occurrences). Intersections with wordnets' lemma lists gave as a result a similar order of magnitude of the resulting lists for all languages (5–15$\times 10^3$). These lists had quite good coverage of the original corpora frequency lists (on the average 70-80%, with the exception of Chinese corpora, which had poor coverage of 20-30%).

The final analysis was conducted on a relatively small multilingual SH corpus. The Sherlock Holmes subcorpus of the NTU-MC consists of two of Conan Doyle's short stories (*The Adventure of the Speckled Band*, 1892, and *The Adventure of the Dancing Men*, 1903/1904) annotated with wordnet

[10] http://corpus.leeds.ac.uk/list.html, CC-BY.

[11] In the case of Polish, http://nlp.pwr.wroc.pl/en/tools-and-resources/resources/frequency-list CC BY-NC-SA 3.0

[12] The selection contains only news which makes it comparable to the Reuters Corpus.

[13] Chinese has practically no inflection

senses. The coverage appears low, but this is an undercount, some concepts cover multiple words (especially in Japanese, where the segmenter segments to morphemes).

Apart from nouns, adjectives, verbs and adverbs, the annotation also included pronouns, so to make the lists more comparable to those made out of general corpora, the top of all lemma frequency lists, comprising mostly function words,[14] was cut saving words less frequent than 100 occurrences ($f < 100$).[15] The intersection with wordnets' lemma lists was also comparable to that of general corpora (2–4$\times 10^3$, cf. Tab. 3, SH rows).

3.3 Constructing rank bins

The original Zipf's work on meaning-frequency dependency was in fact the research on averaged values of sense number and ranks. Ilgen and Karaoglan (2007) and Casas et al. (2019) proved that the relationship is strong for larger bins, but becomes more and more relaxed for smaller frequency bins.

Since our frequency lists are intersected with wordnet lemma sets, some ranks from the original corpus lists occasionally fall out, so we receive gaps within continuous stream of ranks. If a corpus coverage by each intersection set (see Table 3) is close to 80%, on the average every fifth rank darts out. This is the reason why we cannot take a final intersected corpus-wordnet list and simply divide it into bins of particular size. Instead, we ought to deal with specific rank ranges.

The process of varying word bin sizes was slightly different for general and Sherlock Holmes corpora, thus we give their descriptions separately.

General corpora. We explored only nouns, adjectives, verbs and adverbs, but omitted words of ranks 1–100 in order to avoid introducing non-content words into frequency counts.[16] The bins were collated for specific rank ranges (λ = 1, 50, 100, ..., 500). Since ranks 1–100 were intentionally omitted, we started our rankings in the best case from $i = 101$. Similarly to Casas et al. (2019), we constructed bins of the range λ such that a word with i^{th} rank fitted j^{th} bin if and only if the following inequalities were fulfilled:

$$100 + \lambda \cdot (j-1) + 1 \leq i \leq 100 + \lambda \cdot j, \quad (4)$$

where j = 1, 2, 3, ..., round($\frac{n}{\lambda}$), λ is a rank range, I is a set of ranks i ($i \in \mathbf{N}$, $i > 100$), and n is a maximum rank.

Figure 1 illustrates the process of making the frequency bins smaller and smaller and shows regression lines for some successive rank bins in the Reuters Corpus.

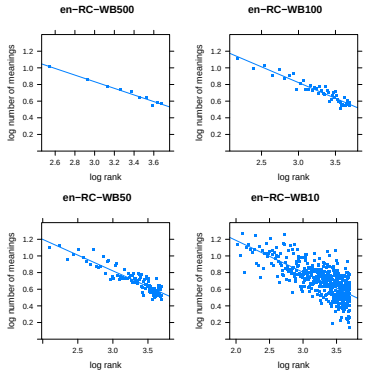


Figure 1: Meaning-rank dependency for the Reuters Corpus and PWN 3.1, with regard to word bins of different rank range. Symbols: WB - word bin λ = 10, 50, 100, 500 ranks. The slope coefficient $-\gamma$ equals -0.42 for λ = 500.

Sherlock Holmes stories. We explored again only nouns, adjective, verbs and adverbs, but threw out words of frequencies greater than a hundred occurrences in a corpus. The bins were collated for the following rank ranges: λ = 1, 3, 5, ... 99. We construct such bins of the range λ that a word with i^{th} rank fitted j^{th} bin if and only if the following inequality was fulfilled:

$$(\lambda \cdot (j-1) + 1 \leq i \leq \lambda \cdot j) \;\&\; (f_i < 100), \quad (5)$$

where j = 1, 2, 3, ..., round($\frac{n}{\lambda}$), λ was a rank range, I was a set of ranks i ($i \in \mathbf{N}$), and n was a maximum rank, f_i was a frequency count for the i^{th} word.

3.4 The log-log model

We investigated the weak version of Zipf's meaning-frequency law in the form of Eq. 3 by changing values of the rank range λ from large bins

[14] In fact they are not strongly polysemous.

[15] For English, e.g., there were words *I*, *be*, *you*, *he*, *we*, *say*, *she*, *this*, *not*; while for Indonesian lemmas – *-nya* (as a pronoun and an article), *itu* (a pronoun/article), *saya* 'I, me, mine', *dia* 'he, she, it', *kami* 'we, our, us'.

[16] In his original paper, Zipf cut off the first 500 words, claiming they were function words (Zipf, 1945). Limiting our analysis to nouns, adjectives, verbs and adverbs would result in a slightly different number of words in each bin.

to small. We aimed at discovering the determination coefficient R^2, as well as the slope coefficient $-\gamma$ for largest bins. R^2 values were used previously as a measure of model fit (Zipf, 1945, 1949; Edmonds, 2004; Ilgen and Karaoglan, 2007; Casas et al., 2019). We checked also the slope coefficient non-zeroness with the t-Student test.

To avoid any possibility of infecting our model with correlated errors, we also inspected residuals with the Shapiro-Wilk statistics, as suggested by Altmann and Gerlach (2016). The Shapiro-Wilk test is the most powerful normality test available now to researchers. Originally designed for small samples, now it is applicable also to samples up to 5,000 observations (Razali and Wah, 2011). Hence, if a model was constructed on a larger sample,[17] we applied sampling 5,000 instances from the original set of observations *without* replacement.

As far as we know, this is the first time when the residuals of the linear Zipfian log-log model for meaning distribution are inspected for non-normality.

4 Results

4.1 Predictive power

General corpora. Seven languages (five Indo-European, Chinese and Japanese) and nine corpora were checked for Zipf's meaning-rank law (Eq. 3) efficiency. Table 4 shows the results for λ = 500, 100, 50 and 1. Clearly the very same pattern that was observed earlier in Ilgen and Karaoglan (2007) and in Casas et al. (2019) is also visible in our data. The bigger rank range λ is, the more efficient is Zipf's law. Making word bins smaller and smaller leads to smaller R^2 values, with a collapse at $\lambda = 1$ (no bins).

Figure 2 presents a more detailed picture of what is happening (λ = 1, 50, 100, 150, ..., 500). The determination coefficient R^2 maintains its values down to quite small bin sizes (λ equals 50-100) and then rapidly lowers to poor percentages of 10-20% of variance explained. The process is accompanied by a non-normal behaviour of residuals (p-value drops below the significance level of 1% at λ in the range of 50–150).

In the case of Chinese corpora (the Internet corpus, *IC*, and the news corpus, *GC*) R-squared values are smaller as compared to other languages. It becomes clearer why it is so when one compares the coverage of both corpora by the Chinese Open Wordnet (Tab. 3), which is relatively small (coverage is between 20-30%). For most languages, the coverage is much higher resulting in small difference between real bin size and the face value λ (they differ by one-fifth). For Chinese, the proportion is much worse and the real bin size might be on the average only one-third of the nominal value. Simply when looking at Chinese data we look at much smaller bins.

[17] It was possible for some general corpora in the case of no-bins, see Table 3, the column *L* and rows *en-IC*, *pl-IC*, *es-IC*, *zh-IC*, *zh-GC* and *jp-IC*.

			rank range λ			
data set	γ_{500}	**n**	**500**	**100**	**50**	**1**
en-IC$^+$	.42	40	.98	.94	.90	.22
en-RC$^+$	.40	10	.98	.90	.81	.11
pl-IC$^+$	.22	20	.96	.83	.69	.06
es-IC$^+$	.47	30	.94	.86	.81	.26
fr-IC$^+$	.51	10	.98	.94	.90	.04
pt-IC$^+$	.32	10	.96	.88	.77	.09
zh-IC*	.22	100	.86	.61	.45	.06
zh-GC*	.21	50	.86	.56	.40	.06
jp-IC$^+$	.26	30	.94	.83	.72	.08
tr-B	.42	27	.97	.94	.89	—
tr-G	.39	45	.89	.70	.66	—
en-CH	.38	19	.98	.86	—	—
nl-CH	.25	5	.99	.78	—	—
es-CH	.27	7	.95	.59	—	—

Table 4: Loss of Zipf's meaning-rank law predictive power in terms of determination coefficient R^2 with regard to different rank bin sizes (λ = 500, 100, 50, 1). Symbols: γ_{500} marks the slope coefficient of the regression line for $\lambda = 500$, 'n' is number of rank bins used for calculating γ; 'tr-B' and 'tr-G' denotes BilTD and GozD Turkish corpora, respectively, in Ilgen and Karaoglan (2007), '*-CH' marks the CHILDES corpus in 3 language versions: English (*en*), Dutch (*nl*) and Spanish (*es*), taken from Casas et al. (2019, Tab. 1, 2); we have chosen only values for child language.

Sherlock Holmes stories. In Table 5 we provide the actual R^2 values for the NTU-MC subcorpus of Holmesian stories. The gradual loss of Zipf's law predictive power is clear – the smaller a bin is the lower the correlation coefficient becomes. Contrary to the results for general corpora/wordnets coupling, the final variance amount explained by Zipf's model is not very low.

The magnitude of the effect itself might be hid-

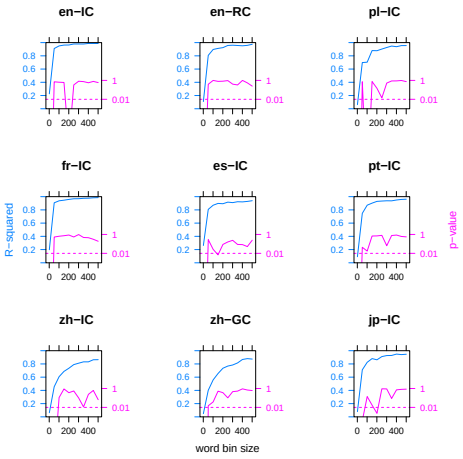


Figure 2: Loss of Zipf's meaning-rank law predictive power in terms of determination coefficient R^2 (blue line) with regard to different frequency bin sizes (λ = 1, 50, 100, ..., 500). With pink line we mark p-values of Shapiro-Wilk normality test for residuals of the model.

den just by these relatively large correlation values. It remains in agreement with our expectation about how the dependency should act in real texts. The meaning number space is much lower than in the case of comparing general corpora and wordnets, and there is an upper limit imposed on the number of meanings equal to the frequency itself.[18]

The real problem with taking full frequency lists becomes obvious if we inspect the meaning-*frequency* relation for $f > 0$ (Figure 3). Despite the fact that R-squared values are very high, residuals of each model are not normal (p-values $< 1\%$), leading to the presumption that the long tail of words occurring in usage only once per a corpus forces residuals to be correlated.[19] In the case of meaning-*rank* dependencies, this issue is hidden with the common rank ordering practice that we have followed.

All words that occur in a corpus once receive consecutive ranks, rescuing model residuals from total disaster. To be clear, this proves that for *SH* corpora containing *hapax legomena* (like in Table 5) Zipf's law does not function properly, even for

[18]We cannot get more senses of a lemma than the number of its occurrences in a text.

[19]Since we have a huge amount of points with co-ordinates $f_i = 1$ and $m_i = 1$.

data set	γ_{100}	n	rank range λ			
			100	**50**	**10**	**1**
en-SH	.43	28	.96	.94	.86	.44
id-SH	.36	18	.94	.90	.81	.34
jp-SH	.31	31	.96	.94	.83	.37
zh-SH	.33	38	.94	.92	.85	.42

Table 5: Loss of Zipf's meaning-rank law predictive power in terms of the determination coefficient R^2 with regard to different rank bin sizes (λ = 100, 50, 10, 1, $f > 0$ in all cases). The symbol γ_{100} marks the absolute value of the negative slope coefficient of the regression model for λ = 100.

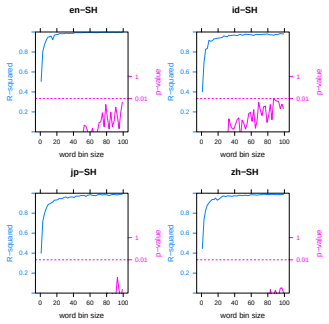


Figure 3: R-squared values of Zipf's meaning-frequency linear model with regard to different word bin sizes (λ = 1, 3, 5, ..., 99) for whole SH corpora (with hapaxes, $f > 0$).

mean values.[20]

It is justified to test Zipf's law for words occurring in a corpus more than once. This compromise gives us also an opportunity to compare such shortened lists for Holmes stories with largely abridged lists from general corpora. Figure 4 presents the data. After removing hapaxes, the Zipf's model starts to behave properly: p-values soar above 1%, R-squared values become large when rank ranges are bigger than 20.

4.2 The slope

General corpora. Table 4 provides slope coefficients γ for the largest bins ($\lambda = 500$) in Internet and news corpora. In more detail, we illustrate it with Figure 5 (for different bin sizes). All γ values occurred to be statistically significant in

[20]This extraordinary property of Zipf's law does not contradict the results obtainable from general corpora, since they are always shortened with the least frequent lemmas, possibly having hidden this phenomenon out of sight of researchers.

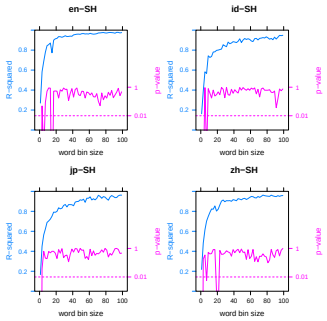


Figure 4: R-squared values of Zipf's meaning-rank linear model with regard to different word bin sizes (λ = 1, 3, 5, ..., 99) for $f > 1$.

t-Student test (p-values are much smaller than the significance level of 1%). It is obvious from the data that the coefficients are mostly less than 0.5 – the value hypothesized by Zipf himself. The values seem also quite stable concerning the vast range of rank bins, however it is not obvious whether slope coefficients are independent of the corpora and frequency lists used.

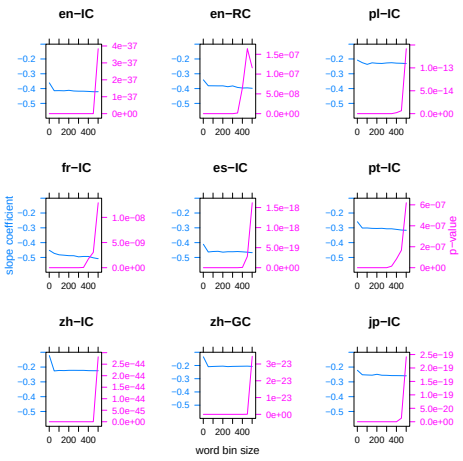


Figure 5: Stability of the slope coefficient $-\gamma_\lambda$ values (blue line) with regard to different frequency bin sizes (λ = 1, 50, 100, ..., 500). Frequency lists are taken in as a whole. With the pink line we mark p-values of t-Student test for non-zeroness of the slope values.

Our frequency lists vary with respect to length, they also come from differently sized corpora (see Table 3). To overcome this problem, we decided to confront languages using relative frequencies. We have chosen the frequency of 12.5 occurrences per million in a corpus as a maximum rank for the need of comparison. The abridged lists cover the most frequent vocabulary of each language, i.e., the top 4000–6000 most frequent lemmas. Table 6 provides Zipfian curve coefficients for $\lambda = 200$ and the shortened frequency lists. The coverage of frequency lists for most languages is very good (80–90%).

Languages differ in terms of regression coefficients. The clear dependency links the slope value and the intercept. The more steep a regression line is, the bigger an intercept becomes. This cross-lingual pattern finds its counterpart in each language data.

corpus	**max. rank**	**cov.** [%]	**poly.** $med(m)$	γ_{200}	I
en-IC$^+$	4856	86	4(5.5)	0.40	2.0
en-RC$^+$	4501	77	4(5.8)	0.38	2.0
pl-IC$^+$	6038	89	3(3.9)	0.22	1.3
es-IC	4575	78	4(4.7)	0.29	1.6
fr-IC	4672	87	5(7.1)	0.48	2.4
pt-IC	4987	79	3(3.4)	0.30	1.5
zh-IC$^+$	6521	48	2(2.6)	0.08	0.7
zh-GC$^+$	6486	45	2(2.7)	0.19	1.1
jp-IC	4681	76	3(4.3)	0.19	1.2

Table 6: Comparison of Zipfian curve coefficients: the slope $-\gamma_{200}$ and the intercept I. The cut-off point is the relative frequency of 12.5 occurrences per million in a corpus. For different languages and corpora the cut-off maximum rank differ. We have chosen $\lambda = 200$ to ensure normality of residuals. Symbols: $^+$ - wordnet taken in whole (not only ILI part), $cov.$ - the coverage of a frequency list by wordnet lemmas, $poly.$ – median of (med) / mean (m) polysemy (senses per lemma).

Figure 6 shows the pattern more thoroughly by presenting regression slope and intercept for different cut-off ranks (maximum ranks) in each language/corpus. Dashed vertical line represents the maximum rank corresponding to relative frequency of 12.5 occurrences per million (chosen as a basis of comparison in Table 6). Coefficients may change their values a lot, as Spanish or French data proves. Yet again, both regression coefficients react inversely to elongating frequency list. While intercepts grow, simultaneously slope coefficients

$-\gamma_{200}$ drop. This reproduces the fact that lengthening frequency lists is the same as adding less and less polysemous lemmas.

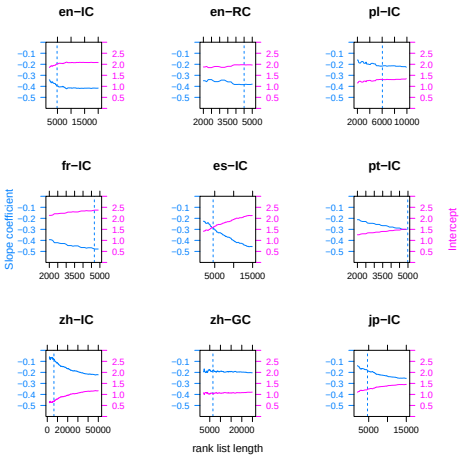


Figure 6: Variation of the slope coefficient $-\gamma_{200}$ values (blue lines) and intercepts (pink lines) with regard to different frequency list lengths (for one particular bin size, λ = 200). With dashed blue lines we mark ranks corresponding to the relative frequency 12.50 occurrences per million.

It is rather unlikely that each language will finally reach its own Zipfian $\gamma = -0.5$ magical zone, even for very long frequency lists. Sherlock Holmes stories give us a unique opportunity to check slope coefficient values under controlled conditions.

Sherlock Holmes stories. As in the case of general corpora, slope coefficients present stable behaviour while changing frequency bin sizes for corpora abridged by *hapaxes* (Fig. 7). Comparing them to values obtained for general corpora and described in the literature shows that although they do change, the change rate is rather moderate (close to $\pm.10$).

Consider the γ values for English. In Zipf's experiment it was .47, Edmonds (2004) estimated it at .40 (Tab. 1), in CHILDES corpus Casas et al. (2019) found it to be close to .38, in Leeds corpora it equals .40/.42 (Internet) and .38/.40 (news), while in Sherlock Holmes stories it is .43. In the case of Japanese, we got .19/.26 in the Leeds corpus and .31 in Sherlock Holmes, not so distant. For Chinese the values change bit more (*zh-IC*: .08/.21, *zh-GC*: .19/.22, *zh-SH*: .33), but this might

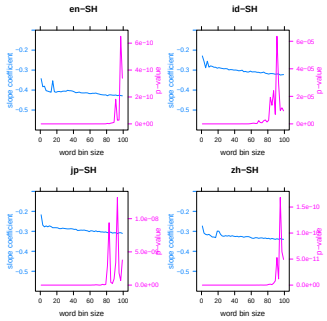


Figure 7: Stability of the slope coefficient $-\gamma_\lambda$ values (blue line) with regard to different frequency bin sizes (λ = 1, 3, 5, ..., 99) for $f > 1$ (no *hapax legomena*). With the pink line we mark p-values of t-Student test for non-zeroness of the slope.

be caused by the fact that the coverage for Leeds corpora are too low. For Spanish we found the slope coefficient close to .29/.47 in our data, while (Casas et al., 2019) obtained value of .27. The difference might be explained with the poor coverage of the Spanish child speech corpus with Multilingual Central Repository (only 13%, *ibidem*).

5 Conclusions

We have presented novel, statistically valid, empirical evidence for the weak version of Zipf's law of meaning distribution on eight languages from four distinct language families (Indo-European, Japonic, Sino-Tibetan and Austronesian).

Zipf's law functions pretty well for mean values in terms of high determination coefficient R-squared, and non-zero slope coefficient γ (stable over the vast range of λ values, but changing while altering frequency lists). The law is, however, inefficient for individual lemmas, because of the lack of model residual normality, despite non-zero correlation coefficient R values.

In the case of Sherlock Holmes stories, this Zipfian catastrophe does not manifest itself only while shifting from bins to individual lemmas, but - surprisingly - also within each whole unabridged corpus containing *hapax legomena*, both for smaller and larger bins.

Slope coefficients that Zipf tended to treat being close to -0.5, in fact, vary largely from language to language, and corpus to corpus, ranging from -0.5 to -0.1.

Acknowledgments

This research was financed by the National Science Centre, Poland, grant number 2018/29/B/HS2/02919, and supported by the Polish Ministry of Education and Science, Project CLARIN-PL, and the NTU Digital Humanities Research Cluster.

References

Eduardo G. Altmann and Martin Gerlach. 2016. Lecture Notes in Morphogenesis. *Creativity and Universality in Language*, pages 7–26.

Konstantin Avrachenkov, Arun Kadavankandy, Liudmila Ostroumova Prokhorenkova, and Andrei Raigorodskii. 2015. PageRank in undirected random graphs. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 151–163. Springer.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Riza Hammam. 2014. The combined Wordnet Bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Francis Bond and Liling Tan. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *Glottometrics*, 22(4):161–174.

Bartosz Broda, Damian Jaworski, and Maciej Piasecki. 2010. Parallel, massive processing in supermatrix - a general tool for distributional semantic analysis of corpus. volume 5, pages 373–379.

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i-Cancho, and Jaume Baixeries. 2019. Polysemy and Brevity versus Frequency in Language. *Computer Speech & Language*, 58:207–237.

Philip Edmonds. 2004. Lexical Disambiguation. In *Elsevier Encyclopedia of Language & Linguistics*, pages 43–62. Elsevier.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Ramon Ferrer-i-Cancho. 2016. The meaning-frequency law in Zipfian optimization models of communication. *Glottometrics*, 35:28–37.

Ramon Ferrer-i-Cancho. 2018. Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3):207–237.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 2525–2529. European Languages Resources Association (ELRA), Istanbul, Turkey.

Antoni Hernández-Fernández, Bernardino Casas, Ramon Ferrer-i-Cancho, and Jaume Baixeries. 2016. Testing the robustness of laws of polysemy and brevity versus frequency. In *International Conference on Statistical Language and Speech Processing*, pages 19–29. Springer.

Bahar Ilgen and Bahar Karaoglan. 2007. Investigation of Zipf's 'law-of-meaning' on Turkish corpora. In *Proceedings of the 22nd International Symposium on Computer and Information Sciences*, pages 1–6. IEEE.

László Lovász and Peter Winkler. 1995. Mixing of random walks and other diffusions on a graph. In *Surveys in combinatorics, 1995*, pages 119–154. Cambridge University Press.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268.

Steven Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Nornadiah Mohd Razali and Yap Bee Wah. 2011. Power comparisons of Shapiro-Wilk,

Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics Vol*, 2(1):21–33.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison Wesley Press.

plWordNet 4.1 – a Linguistically Motivated, Corpus-based Bilingual Resource

Agnieszka Dziob, Maciej Piasecki and Ewa Rudnicka
G4.19 Research Group, Department of Computational Intelligence
Wrocław University of Technology, Wrocław, Poland
{agnieszka.dziob,maciej.piasecki,ewa.rudnicka}@pwr.edu.pl

Abstract

The paper presents the latest release of the Polish WordNet, namely plWordNet 4.1. The most significant developments since 3.0 version include new relations for nouns and verbs, mapping semantic role-relations from the valency lexicon Walenty onto the plWordNet structure and sense-level interlingual mapping. Several statistics are presented in order to illustrate the development and contemporary state of the wordnet.

1 Introduction

plWordNet (Pol. *Słowosieć*) is a very large wordnet of Polish, mapped to Princeton WordNet of English (Miller et al., 1990) and enriched with other links and annotations. Its development started back in 2005, and has been continued since then. In 2016, its complex, mature 3.0 version was presented in (Maziarz et al., 2016). It achieved very large size and coverage of words in Polish corpora. Thus, in our work we focused on increasing the density of the network of wordnet relations, revising the structure wherever necessary and adding new relations in order to improve the description of the lexical system of Polish and to meet the requirements of plWordNet's applications.

The goal of this paper is to present the latest 4.1 release of plWordNet, the result of a linguistically motivated expansion of 3.0 version. We will discuss new synset relations for nouns and verbs, as well as a new relation for lexical units, namely the semantic Collocation relation meant to facilitate the use of plWordNet in Word Sense Disambiguation. A new system of verb classes will be briefly recalled with the focus on its implementation in 4.1. We will also discuss the process of systematic assignment of aspect values, such as perfect, imperfect, and bi-aspectual, to every verbal lexical unit.

Alongside the development of plWordNet, works on its mapping to Princeton WordNet are carried out. The latest version includes the complete mapping of Polish and English noun synsets, the extended mapping of adjective and adverb synsets and the substantial mapping of verb synsets. Moreover, we have started the development of a system of equivalence relations for noun lexical units. Finally, we will present the results of mapping plWordNet lexical units onto the entries of the Polish valency lexicon *Walenty* and their partial manual verification.

2 Linguistic Motivation

Since its origin, plWordNet has been built around the idea of making lexical units[1] (henceforth, LUs) its basic building blocks, using linguistic lexico-semantic relations, and making the wordnet a faithful description of the Polish lexical system, see (Piasecki et al., 2009). This led to a corpus-based wordnet development process (Maziarz et al., 2016), synset composition based on sharing constitutive relations and features, and wordnet model based on the Minimum Commitment Principle, see (Maziarz et al., 2013b).

In plWordNet, the description of lexical meanings is primarily based on lexico-semantic relations directly originating from lexico-semantic relations known from lexicography. As synset relations are abbreviations for the fact of sharing relation between LUs – synset components – synset relation do not differ in their character from relations linking LUs. Glosses and usage examples are treated

[1] A lexical unit is technically defined as a triple: lemma, Part of Speech and sense identifier.

as secondary means of description. Definitions of particular relations directly refer to language data via substitution tests that are then used in the wordnet development.

In plWordNet 3.0 (Maziarz et al., 2016) LUs located in the lower parts of the hypernymy hierarchy were often described by only a few relation links, if not just one. Thus, their meaning descriptions were limited, especially those described by single hyponymy links connecting to the same hypernym. There was no meaning distinction between such LUs. Diversity and density of relation links is crucial for many applications of a wordnet, e.g. comparison of meanings, analysis of selectional preferences (McCarthy and Carroll, 2003, Hajnicz et al., 2016), Word Sense Disambiguation (Agirre and Rigau, 1996, Kędzia et al., 2015), texts semantic indexing (Scott and Matwin, 1998), query expansion in Information Retrieval (Voorhees, 1998, Varelas et al., 2005), or construction of topic descriptors for media monitoring (Johansson et al., 2012).

Taking the above into account, we have proposed an expansion of the plWordNet model by several new relations, described in the next section. In sum, we will have 33 types of synset relations (52 when counting subtypes) and 20 types of LU relations, i.e. not shared among LUs (56 including subtypes).

2.1 Nouns

The system of noun relations in plWordNet 4.1 is based on that of 3.0 release (Maziarz et al., 2011). It has recently been expanded with several new synset relations discovered when analysing instances of the *fuzzynymy* relation. During the many years of plWordNet development, *fuzzynymy* was used as a kind of notebook to record semantic associations that seemed prominent, but irregular from the point of view of the wordnet relation system. Still, not all fuzzynymy relations were renamed into other relations.

Definitional feature is a relation informing about an entity's intrinsic property which defines its membership to a given class of things or people e.g. {*rudzielec 1, marchewka 3, wiewiór a 1*} '≈redhead' →{rudy} 'red', {*upał 1, skwar 1, żar 1, spiekota 1, spieka 1*} '≈heat'→{*upalnie 1, skwarno 1, skwarnie 1*} 'hot', {*abrakadabra 1, metafizyka 3, czarna magia 1*} 'double Dutch' →{*niezrozumiały 1*} 'unclear'. It is a relation between a noun synset and another noun synset or an adjective or adverb synset. This property rarely co-occurs with a given noun in the corpus, but it often appears in its lexical paraphrase.

Area of interest is a noun-noun relation that informs about an object or issue that is lexically constituted as a typical focus for this discipline or area, e.g. {*kardiologia 2*} 'cardiology'→{*układ krwionośny 1, krwiobieg 1, krwioobieg 1*} 'circulatory system'.

Origin is a relation linking a noun with a qualitative adjective derived from a noun denoting the country or culture of origin of the entity denoted by this noun e.g. {*zabaglione 1, zabajone 1, zabaione 1*} 'sabayon'→{*włoski 3, italiański 3, italski 3*} 'Italian', or, when there is no such adjective, with a noun denoting the origin of a given entity. It can be paraphrased as 'something that comes from a country or culture'.

Parameter is a noun-noun relation defined especially for the description of specialist vocabulary and represents a physical, measurable parameter characterising some phenomena, e.g. {*żyzność 1, urodzajność 1, żyzność gleby 1, plenność 1*} 'soil fertility'→{*gleba 1, grunt 3, podłoże 2*} 'ground'. Specialist vocabulary LUs are almost always found in the lower part of the hypernymy hierarchy and are described by few relations besides hyponymy.

Following our earlier positive experience in using a derivationally motivated Role of a hidden predicate relation linking noun LUs since plWordNet 2.0 (Maziarz et al., 2011), we propose to expand this relation to relations between synsets in which the semantic opposition is similar, but the linked synset elements are usually not derivationally associated.

Subject of hidden predicate is a relation between two noun synsets such that the first is a semantic subject of an implicit action intentionally and intrinsically related to an object represented by the second, e.g. {*pulmonolog 1, pneumonolog 2*} 'pulmonologist'→{*układ oddechowy*} 'respiratory system'.

Product|result of hidden predicate, in a similar way, associates two noun synsets such that the first represents a product or result of

an implicit action or process done on an object or substance represented by the second, e.g. {*piwo 1, złoty trunek 1, złocisty trunek 1*} 'beer'→{*brzeczka piwna 1*} 'beer wort'.

Place of hidden predicate links two noun synsets where the first represents a place which is an obligatory, lexically constituted element of an action or process represented by an implicit predicate which is intrinsically related to the entity expressed by the second synset, e.g. {*klinika odwykowa*} 'rehab clinic'→{*nałogowiec 1, uzależniony 1*} 'addict'.

Although synset relations based on the hidden predicate scheme are mostly used for the description of specialist vocabulary, they are quite frequent, i.e. a couple of hundred instances on average, see Sec. 3.

2.2 Collocation

A large wordnet can be successfully used as a knowledge base for Word Sense Disambiguation, but the quality of the resulting system depends a lot on the richness of a network of connections between words from texts via their senses, especially between senses that are likely to co-occur in similar contexts (Leacock et al., 1998). Unfortunately, the coverage of such associations is limited by typical wordnet relations.

Following the above observation, we introduce a *collocation* relation for LUs that links lexical meanings, not words. In contrast to the *definitional feature* relation, which is based on a semantically motivated, paradigmatic feature, *collocation* follows the corpus supported language data and indicates frequent meaning co-occurrences. It can link two LUs of any of part of speech, if they co-occur often enough in corpora. So far we have added 16 979 instances of the collocation relation for all parts of speech to plWordNet (most of them for nouns: 7 838 instances).

Collocation relation was also used for priming selected meanings for words, as a kind of micro-glosses during psychological experiments on collecting emotive evaluations per LUs. For a selected subset of polysemous lemmas, we first drew one LU per lemma as subjects of the experiment. Next, for each selected LU we tried to choose among its possible frequent co-occurrences in such a way that the chosen collocation distinguish the given LU (word meaning) from all the other possible ones for a given lemma. Later, during the experiment, a lemma – representing an intended LU – was presented alongside the collocation to the informants, who were next interviewed about their reactions to several emotive aspects of the LU meaning. *Collocation* as defined and used by us has a pragmatic application: it links meanings, not words, according to their co-occurrence in corpora, while typical statistical analysis of corpora yields only word-form associations. We therefore regard this relation useful for word sense disambiguation. So far, the collocation relation, as it is proposed, has only a utilitarian character. However, we plan further research in this field.

2.3 Adjectives and Adverbs

The description of adjectives in plWordNet 3.0 was based on several synset relations, including *inter-register synonymy*, *hypernymy/hyponymy*, *gradation*, *modifier*, and *value (of the attribute)* (Maziarz et al., 2012). The synset relations were complemented by a set of LU relations, e.g. *predisposition* (with 4 subtypes), *role Adj-V* (7 subtypes), *antonymy* (complementary and gradable), *cross-categorial synonymy* to nouns (2 subtypes), *characteristic*, *markedness*, *role: material* or *state/feature* (derivationally motivated). We found this whole system of relations working well, so except for *definitional feature* proposed in Sec. 2.1 and *collocation* described in Sec. 3, we do not propose any changes to it. Instead, the coverage of several adjective relations was expanded.

Adverbs are treated similarly to adjectives in plWordNet 4.1. Their model in 3.0 version encompassed a set of synset relations almost identical to adjective relations – with the exclusion of *modifier* – and a set of LU relations including: *antonymy* (complementary and gradable), and *cross-categorial synonymy* to adjectives. Similarly to the adjective relations, we keep the adverb relations unchanged.

2.4 Verbs

Verbs in plWordNet 3.0 were organised in a sophisticated system of hierarchical semantic classes that influenced or even determined the verb relation structure. The classifica-

tion encompassed 9 main classes and 4 auxiliary subclasses and was based on the proposal of (Laskowski, 1998), which has never been verified on large language data. This system made plWordNet 3.0 difficult to edit and led to criticism of the excessive proliferation of verb senses (Dziob and Piasecki, 2018b).

Dziob and Piasecki (2018a) proposed a much simpler verb classification for plWordNet consisting of just two main semantic verb classes, namely *static* and *dynamic* verbs. Only this division is reflected in definitions of several selected verb relations. In addition, for dynamic verbs five subclasses were proposed, namely: *distributive*, *accumulative*, *perdurative*, *delimitative* and *action* verbs, but without the obligatory influence on their relations. The decision about the verb class membership of a given LU is done with the help of semantic paraphrases, which simplifies the work of lexicographers and results in a description that is more comprehensible for users. (Dziob and Piasecki, 2018a) proposed a couple of new verb relations and modifications to several relations which we have adopted for plWordNet 4.1 and describe below.

We decided to leave two main subtypes of the *aspectuality* relation: *pure* and *secondary*, which express the basic semantic difference between aspectual pairs in Polish, (Dziob et al., 2017). Yet, since aspect has become a feature assigned to the verb, we have decided against further division of aspectuality and other relations, based only on aspect. Therefore, this system has become simpler. Otherwise, we have introduced a few new types and subtypes of verb relations: for backward relations (*preceding* and *presupposition*) subtypes without subject identity (e.g. *rozwieść się* 'to get divorced'←*małżeństwo* 'marriage') and four new main level relations, based on syntagmatic occurrences and also lexical definitions: *subject*, *object*, *circumstance* and *manner*, see (Dziob and Piasecki, 2018a). An important change is the possibility of linking verbs with adverbs, allowed since 3.1 version.

3 Structure

Since 3.0 version, we have expanded plWordNet both in terms of language material covered and the number of relation links, characterised briefly in this section and shown in Tab. 3. As main goals for the expansion to plWordNet 4.1 we identified: the newest Polish vocabulary (and meaning changes) in relation to the whole lexical system of Polish and specialist terminology (including multi-word expressions) from users' corpora. Concerning the first goal, this is a necessary process for preserving the quality of plWordNet as a comprehensive and up-to-date description of the Polish lexical system. Continuous development of the coverage of a wordnet is an obligatory aspect for the preservation of its quality.

The presence of specialist vocabulary, mostly terminology, in a general dictionary (a large wordnet is often perceived as a large dictionary) is disputable. However, plWordNet is mostly used as a basic language resource in processing, e.g. as the part of the CLARIN language technology infrastructure, and its content should reflect to some extent the vocabulary of texts being processed. As such, the addition of specialist terminology and vocabulary has been a corpus-driven effort.

The development of plWordNet follows the corpus-based wordnet development process proposed in Maziarz et al. (2013a). plWordNet Corpus v10 has been enlarged up to 4.2 billion segments in order to make it a better basis for the acquisition of new lemmas. New colloquial vocabulary was added from sources such as social media, blogs, also the most recent literature. Several much smaller specialist corpora from the CLARIN-PL users were also explored as the sources of language material.

3.1 Changes in Statistics

Since we suspected that adjective and adverb parts were less developed, we compared their content with the plWordNet Corpus. All missing adjectives and adverbs were added and the meanings of many of them were verified that resulted in expanding plWordNet by at least 2300 adjective lemmas (>8,500 adj. LUs) and 2000 adverb lemmas (>3,100 adv. LUs).

As the verb model had been changed and we knew that the coverage for verbs was lower than for other parts of speech, we put special emphasis on a large scale expansion of this sub-database and also on the verification and correction of the existing description of many verbs. More than 11,300 new verb LUs and

Elements	**Verbs**	**Nouns**	**Adv.**	**Adj.**	**All**	↑
plWN 3.0 Lemmas	17 398	126 746	5 719	27 041	**177 003**	–
plWN 3.0 Lexical Units	31 841	167 243	10 416	45 899	**255 733**	–
plWN 3.0 Synsets	21 669	123 985	8 080	39 204	**193 286**	–
plWN 4.1 Lemmas	20 430	134 674	8 042	29 349	**192 495**	8.7%
plWN 4.1 Lexical Units	43 701	178 167	14 088	54 410	**290 366**	13.5%
plWN 4.1 Synsets	32 102	133 747	11 295	47 035	**224 179**	16.0%

Table 1: Basic statistics of plWordNet 4.1 (`http://plwordnet.pwr.edu.pl`)

2,900 new verb lemmas were added. The new LUs were also added to lemmas already present in the 3.0 version to complete the description of their meanings. Manual verification and correction of LUs, synsets and relations was done for most of the already described verbs.

Specialist vocabulary was added in response to requirements of plWordNet applications (esp. in CLARIN) subsuming about 4,000 specialist LUs (mostly nouns, marked by *specialist* register), including many multi-words. The newest vocabulary acquired from plWordNet Corpus v10 was described by more than 1,500 new LUs of all parts of speech. Changes in the noun part are mostly the result of this process.

Tab. 3.1 presents statistics for the proposed noun relations. Because we have started to add new relations to the specialist vocabulary, *area* and (especially) *parameter* are not too frequent relations, but development of this vocabulary is ongoing. We use specialist corpora of CLARIN users and we integrate the vocabulary derived from them by means of these relations. We expect them to be useful especially for describing specialist vocabulary on the lowest levels of the wordnet hierarchy – at least this is the result of our experience up to now.

The second source of new lemmas and lexical units are users' diachronic corpora containing old vocabulary. We include these units in plWordNet only when we can confirm their use in texts, for example in freely available old literature. For this reason and because of the presence of modern vocabulary in our corpora that we write about in Sec. 3., the quantity of *inter-register synonymy* linking synsets with LUs of divergent registers is increasing (in 4.1 version it amounts to 12 223 instances for all parts of speech, of which most for nouns – 7 171). We expect that this process will advance.

3.2 Non-relational Elements and Verification

In 2017 a wordnet editor system called WordnetLoom 2.0 (Naskręt et al., 2018) was enriched with the ability to record comments concerning the correctness of a given LU and synset. Information that has been collected by this system is one of the inputs to the plWordNet verification process started by us. We use also data collected from the diagnostic tools (Piasecki et al., 2016).

The verification of plWordNet is performed on the two levels of LUs and synsets. Both are described by an additional *status* feature whose value is set by a lexicographer after each operation: *verified*, *partially processed*, *new*, *meaning*, *erroneous* and *not processed*, with the last one as a default value. When an editor spots a problem they can describe or comment on it. In this way, statistics concerning the frequency of errors made during the earlier stages of plWordNet development are collected. The most frequent errors are: too small number of meanings for a given lemma, but also too fine-grained granulation of meanings, and wrong stylistic register. The verification and corrective editing that has been performed since the publishing of the 3.0 version is focused on LUs now, as we assume that a *verified* synset must include only *verified* LUs. A fully correct synset must include LUs with proper descriptions, including their relations, and the synset must be described by proper synset relations (compatible with the LUs due to the synset definition assumed in plWordNet). So far 7,976 have been marked by the status *verified* and 5,677 *partially processed*, i.e. verified by a single editor and waiting for the confirmation by the second editor.

The description of LUs in plWordNet is systematically completed by glosses and use ex-

Rel. of hidden pr. (general)	855
Parameter	83
Origin	1324
Area	344
Definitional feature	660

Table 2: Statistics of new relations for nouns.

amples (both added on the level of LUs, not synsets). None of them are necessary from the point of view of a relation-based description of lexical meanings, but they appear helpful for human users and are used in several applications (starting with WSD) as well as in wordnet verification. The number of glosses was increased since 3.0 version by 6,445 and is 170 122, while the number of use examples was increased by 4,763 to 78 001. We assumed that not every LUs must be described by a use example, the priority is given to LUs of polysemous lemmas. However, we aim at achieving a state in which all LUs are characterised by stylistic registers (added to plWordNet at a later stage of its development, as initially it was meant to represent only general language).

Work on glosses and use examples meets the expectations of users who want plWordNet to be more similar to a traditional dictionary in terms of structure, but enriched with relational description. In addition, as already mentioned, the non-relational elements are also useful for natural language engineering.

3.3 Semi-automated Mapping onto Semantic Valence Lexicon

Walenty (Przepiórkowski et al., 2014) is a large lexico-semantic valence dictionary developed independently of plWordNet, but with a lot of cooperation between the two teams. This resulted in its schema referring to plWordNet LUs and semantic selectional preferences often annotated with plWordNet synsets (Hajnicz et al., 2016). Unfortunately, the old, 2.1 version of plWordNet was used for this purpose. Our goal was to automatically map the semantic roles of *Walenty* onto plWordNet in order to increase the density of its relations.

In contrast to FrameNet (Ruppenhofer et al., 2006), automatically linked to Princeton WordNet on the basis of similarity of paraphrases of its units and Princeton WordNet relations (Tonelli and Pighin, 2009), the linking between plWordNet and Walenty was done semi-automatically, with a lot of manual verification. First, we compared 2.1 and 3.0 versions of plWordNet and generated a list of plWordNet synsets whose content differed between the two versions. Next, two rounds of correction were carried out: automatic (based on the comparison of synset content and LU properties) and manual (for synsets which represented too big discrepancies between the two versions). In the latter case, we corrected the discrepancies. The differences between 2.1 and 4.0 synsets were mainly due to the introduction of new LUs or distinguishing new synsets as hyponyms or hypernyms of 2.1 synsets. The final mapping included 2,480 mappings from 2.1 to 4.0 synsets, which allowed us to introduce 17 new relation types to plWordNet. These relations are the equivalents of semantic roles described in the semantic layer of Walenty: *Theme*, *Condition*, *Path*, *Manner*, *Location*, *Purpose*, *Initiator*, *Recipient*, *Attribute*, *Instrument*, *Stimulus*, *Result*, *Measure*, *Time*, *Experiencer*, *Factor*, *Duration*. Both plWordNet and Walenty are the sources that are currently manually verified and corrected with respect to quality and completeness of entry description. The next stage of mapping between the resources was adding the relations on the basis of semantic description in Walenty and plWordNet, but only those with the "checked" status where there were no doubts about their quality and completeness description. In this way, plWordNet was enriched with 3,406 relation instances between plWordNet synsets, showing selectional preferences of units in the semantic layer of Walenty.

4 Alignment to English

A self-contained construction of plWordNet brought about the need of its later alignment to Princeton WordNet. The process started in 2012 and has been continued since then.

I-relation	V		N		Adv		Adj		Total	
I-relation	pl	en	pl	en	pl	en	pl	en	pl	en
I-synonymy	31955	1962	38699	38690	999	999	4338	4339	45991	45990
I-partial syn.	0	2	5821	5698	311	309	1493	1430	7625	7439
I-int.-reg. syn.	205	206	1847	1849	48	48	95	92	2195	2195
I-meronymy	0	1	10785	7944	0	0	0	0	10785	7945
I-hypernymy	79	3447	30736	82315	112	9897	375	44373	31302	140032
I-hyponymy	3433	79	82309	30740	9901	112	44389	374	140032	31305
I-holonymy	0	0	7945	10785	0	0	0	0	7945	10785
I-Type	0	0	7724	623	0	0	0	0	7724	623
I-Instance	0	0	623	7724	0	0	0	0	623	7724
I-allative	40	0	0	0	0	0	0	0	40	0
I-delimitive	157	0	0	0	0	0	0	0	157	0
I-excess	21	0	0	0	0	0	0	0	21	0
I-perdurative	12	0	0	0	0	0	0	0	12	0
I-anticausative	451	0	0	0	0	0	0	0	451	0
I-atenuative	102	0	0	0	0	0	0	0	102	0
I-cumulative	126	0	0	0	0	0	0	0	126	0
I-procesuality	6	0	0	0	0	0	0	0	6	0
I-completive	34	0	0	0	0	0	0	0	34	0
I-inchoative	64	0	0	0	0	0	0	0	64	0
I-distributive	313	0	0	0	0	0	0	0	313	0
I-iterative	37	0	0	0	0	0	0	0	37	0
I-terminative	6	0	0	0	0	0	0	0	6	0
I-ablative	18	0	0	0	0	0	0	0	18	0
I-causative	80	0	0	0	0	0	0	0	80	0
I-c-c-made-of	0	0	2	0	0	0	1067	0	1069	0
I-c-c-resembling	0	0	0	0	1	0	946	0	947	0
I-c-c-related-to	0	0	1	0	97	0	22697	0	22795	0
Total	7139	5697	186493	186368	11469	11365	75401	50608	280502	254038

Table 3: Interlingual relation counts

It took the form of manual mapping that is aligning wordnet nodes (synsets) corresponding in meanings and relation structures via a rich set of interlingual relations, (Rudnicka et al., 2012). It quickly turned out that interlingual synonymy (representing Simple Equivalence, cf. Vossen (2002)) is not enough to link two independently built resources for two quite different languages. English is an analytical Germanic language, while Polish a synthetic Slavic one. Therefore, other Complex Equivalence relations had to be resorted to. In Tab. 3, we present the full list of interlingual relations with their respective counts. The most frequent one is interlingual hyponymy and this tendency occurs across all parts of speech. In the latest 4.1 version of plWordNet, we have expanded the synset mapping between plWordNet and Princeton WordNet.

Moreover, we have also developed the methodology for a more fine-grained sense-level mapping and applied it to a substantial sample of noun lexical units. The methodology is based on a manual verification of the values of equivalence features. These include formal features such as number, countability and gender; semantic-pragmatic features such as sense, lexicalisation of concepts, register, collocations and co-text; and translational features such dictionary listing, dictionary equivalent position, and translation probability. The features are used to define three types of equivalence links: strong, regular and weak.

5 Applications

plWordNet is available on open licence and has been downloaded by more than 1,100 registered users (both individual and institutional). It has also had a quite large number of non-registered users and tens of thousands of users of the on-line browser[2]. On the basis of citations, questionnaires of the registered users, and direct co-operation with users within CLARIN, we can attempt an overview of plWordNet 4.1 applications. First, it was applied in linguistics for an analysis of lexico-semantic fields (Stanulewicz, 2010), analysis of word-forming nests (Lango et al., 2018), derivational processes (Kyjánek, 2018),

[2] http://plwordnet.pwr.edu.pl

identification of semantic classes (Lis, 2012), study on multi-word expressions (support for their extraction, recognition, classification) (Mykowiecka and Marciniak, 2012), and measuring semantic similarity of words (on the basis of their relation structure) (Siemiński, 2012). It found several applications in bilingual lexicography, e.g. in the study on partial equivalences in bilingual dictionaries (Liu, 2018), building a multilingual dictionary of the Yiddish language, as well as development of several bilingual and multilingual dictionaries (Sosnowski and Koseska-Toszewa, 2015). plWordNet was used in applied linguistics, e.g. in studies on the second language learning (Madej and Kiermasz, 2015), clinical research in the lexical system and its disfunction of patients suffering from dementia and Alzheimer disease. It was also utilised in Social Sciences, including an analysis of the language in Polish social media (digital trace, speaker intention, content of blog posts) (Haniewicz et al., 2014, Wawer and Sarzyńska, 2018), analysis of personal self-descriptions (structure and content), and analysis of commercials in media Iwińska-Knop and Krystyańczuk (2016). plWordNet was used to construct new resources, e.g. the system of Polish National Library descriptors was mapped on it, and KPWr Corpus (Broda et al., 2012), Składnica Corpus (Woliński et al., 2011) were annotated by the selected LUs. The most numerous group are applications in Natural Language Engineering, e.g. evaluation of word embedding models on the basis of synonymy tests automatically generated from plWordNet (Piasecki et al., 2018), named entity recognition, text mining and semantic search (Maciołek and Dobrowolski, 2013), text classification and text relation recognition (Brzeski and Boiński, 2014), semantic indexing of text (Karwowski et al., 2018), assignment of descriptive keywords to text documents (as knowledge basis and keyword repository) (Kaleta, 2014), automated structuring of text data (Maciołek and Dobrowolski, 2010), text interpretation in chat bots, text semantic similarity calculation (Siemiński, 2012), antiplagiarism systems (Szmit, 2017), generation of semantically related families/sets of words for Information Retrieval and Internet monitoring and text normalisation, e.g. in the legal domain (Pełech-Pilichowski et al., 2014). As plWordNet is expanded with emotive annotation, cf (Zaśko-Zielińska et al., 2015), it has been applied several times in sentiment analysis and development of sentiment lexicons (Rybiński, 2017). Finally, it was used in *Jasnopis* system for the analysis of text difficulty to extract synonyms and hypernyms of words classified as too difficult for the intended text difficulty level (Dębowski et al., 2015).

6 Further Works

Is it ever possible to complete a wordnet? plWordNet 4.1 size and coverage, as well as its growth since 3.0 version may suggest that it is. However, this is misleading. In the case of a very large wordnet, the focus shifts from mere growth to the improvement of the amount and quality of information expressed for different lexical meanings. We plan to continue the work on increasing the density of relations (especially for LUs described so far by a few, if not single links), continuous maintenance of the wordnet quality by encompassing new lemmas and LUs in a corpus-based way. Instead of incorporating more and more specialist vocabulary, we plan to develop a system of cross-resource mappings envisaged in (Maziarz and Piasecki, 2018) in order to build a system of terminological, ontological and knowledge resources around plWordNet and make it an interface between them and the natural language lexicon. In addition, we also plan to further expand the relation structure towards better support for Word Sense Disambiguation. Moreover, we are going to continue the works on sense-level mappings. While proceeding with manual mapping, we are also going to develop a semi-automatic prompt system.

Acknowledgements

The work co-financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and the project funded by the National Science Centre, Poland under the grant agreement No UMO-2015/18/M/HS2/00100.

References

Eneko Agirre and German Rigau. Word sense disambiguation using conceptual density. In *Proceedings*

of the 16th conference on Computational linguistics-Volume 1, pages 16–22. Association for Computational Linguistics, 1996.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. Kpwr: Towards a free corpus of polish. In *Proceedings of LREC*, volume 12, 2012.

Adam Brzeski and Tomasz Boiński. Towards facts extraction from texts in polish language. 2014.

Łukasz Dębowski, Bartosz Broda, Bartłomiej Nitoń, and Edyta Charzyńska. Jasnopis–a program to compute readability of texts in polish based on psycholinguistic research1. *Natural Language Processing and Cognitive Science _*, page 51, 2015.

Agnieszka Dziob and Maciej Piasecki. Implementation of the verb model in plWordNet 4.0. In *Proceedings of the 9th Global Wordnet Conference*, 2018a. URL `https://pdfs.semanticscholar.org/af21/13bb896f08993f995a68bbfa0ff805e1cbcd.pdf`.

Agnieszka Dziob and Maciej Piasecki. Dynamic verbs in the wordnet of polish. *Cognitive Studies | Études cognitives*, 18, 2018b. URL `https://ispan.waw.pl/journals/index.php/cs-ec/issue/view/98/showToc`.

Agnieszka Dziob, Maciej Piasecki, Marek Maziarz, Justyna Wieczorek, and Marta Dobrowolska-Pigoń. Towards revised system of verb wordnet relations for polish. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*, 2017. URL `ceur-ws.org/Vol-1899/CfWNs_2017_proc6-paper_7.pdf`.

Elzbieta Hajnicz, Anna Andrzejczuk, and Tomasz Bartosiak. Semantic layer of the valence dictionary of polish walenty. In *LREC*, 2016.

Konstanty Haniewicz, Monika Kaczmarek, Magdalena Adamczyk, and Wojciech Rutkowski. A case study of sentiment orientation identification for polish texts. In *2014 European Network Intelligence Conference*, pages 46–51. IEEE, 2014.

Krystyna Iwińska-Knop and Hanna Krystyańczuk. Wykorzystanie big data w badaniu wizerunku marki w świadomości konsumentów. *Ekonomika i Organizacja Przedsiębiorstwa*, (9):28–42, 2016.

Fredrik Johansson, Joel Brynielsson, and Maribel Narganes Quijano. Estimating citizen alertness in crises using social media monitoring and analysis. In *2012 European Intelligence and Security Informatics Conference*, pages 189–196. IEEE, 2012.

Zbigniew Kaleta. Semantic text indexing. *Computer Science*, 15, 2014.

Waldemar Karwowski, Arkadiusz Orłowski, and Marian Rusek. Applications of multilingual thesauri for the texts indexing in the field of agriculture. In *International Multi-Conference on Advanced Computer Systems*, pages 185–195. Springer, 2018.

Paweł Kędzia, Maciej Piasecki, and Marlena Orlińska. Word sense disambiguation based on large scale polish clarin heterogeneous lexical resources. *Cognitive Studies| Études cognitives*, (15), 2015.

Lukáš Kyjánek. Morphological resources of derivational word-formation relations. Technical Report 61 (2018): 49, ÚFAL MFF, Charles Univesity, Prague, 2018.

Mateusz Lango, Magda Sevcikova, and Zdeněk Žabokrtskỳ. Semi-automatic construction of word-formation networks (for polish and spanish). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

Roman Laskowski. Kategorie morfologiczne języka polskiego – charakterystyka funkcjonalna. In Renata Grzegorczykowa, Laskowski Roman, and Henryk Wróbel, editors, *Gramatyka współczesnego języka polskiego. Morfologia*, pages 151–224. Warszawa: Wydawnictwo Naukowe PWN, 1998.

Claudia Leacock, George A Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, 24(1):147–165, 1998.

Magdalena Lis. Polish Multimodal Corpus - a collection of referential gestures. pages 1108–1113. European Language Resources Association, 2012. ISBN 978-2-9517408-7-7.

Lixiang Liu. Partial equivalences in bilingual dictionaries: Classification, causes and compensations. *Lingua*, 214:11–27, 2018.

Przemysław Maciołek and Grzegorz Dobrowolski. Is shallow semantic analysis really that shallow? a study on improving text classification performance. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 455–460. IEEE, 2010.

Przemysław Maciołek and Grzegorz Dobrowolski. Cluo: Web-scale text mining system for open source intelligence purposes. *Computer Science*, 14(1):45–62, 2013.

Monika Madej and Zuzanna Kiermasz. Exploring the attitudes towards word clouds in junior high students with a different multiple intelligence type: A research project. In M. Marczak and M. Hinton, editors, *Contemporary English Language Teaching and Research*, pages 139–157. Newcastle: Cambridge Scholars Publishing, 2015.

Marek Maziarz and Maciej Piasecki. Towards mapping thesauri onto plWordNet. In Francis Bond, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 9th Global Wordnet Conference, Singapore, 8-12 January 2018*. Global WordNet Association, 2018.

Marek Maziarz, Maciej Piasecki, Joanna Rabiega-Wiśniewska, and Stanisław Szpakowicz. Semantic Relations among Nouns in Polish WordNet Grounded in Lexicographic and Semantic Tradition. *Cognitive Studies*, 11:161–181, 2011. http://www.eecs.uottawa.ca/ szpak/pub/Maziarz_et_al_CS2011a.pdf.

Marek Maziarz, Stanisław Szpakowicz, and Maciej Piasecki. Semantic relations among adjectives in polish wordnet 2.0: a new relation set, discussion and evaluation. *Cognitive Studies*, (12), 2012.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. In G. Angelova, K. Bontcheva, and R. Mitkov, editors, *Proc. International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 443–452. INCOMA Ltd. Shoumen, BULGARIA, 2013a.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3): 769–796, 2013b. doi: 10.1007/s10579-012-9209-9.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. plwordnet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL, 2016. URL `http://aclweb.org/anthology/C/C16/`.

Diana McCarthy and John Carroll. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654, 2003.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-Line Lexical Database. *Int. J. of Lexicography*, 3(4):235–244, 1990.

Agnieszka Mykowiecka and Małgorzata Marciniak. Combining Wordnet and Morphosyntactic Information in Terminology Clustering. In *Proc. COLING 2012: Technical Papers COLING 2012, Mumbai, December 2012.*, pages 1951–1962, 2012.

Tomasz Naskręt, Agnieszka Dziob, Maciej Piasecki, Chakaveh Saedi, and António Branco. WordnetLoom – a multilingual wordnet editing system focused on graph-based presentation. In Francis Bond, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the 9th Global Wordnet Conference, Singapore, 8-12 January 2018*. Global Wordnet Association, 2018.

Tomasz Pełech-Pilichowski, Wojciech Cyrul, and Piotr Potiopa. On problems of automatic legal texts processing and information acquiring from normative acts. In *Advances in business ICT*, pages 53–67. Springer, 2014.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2009. URL `http://www.dbc.wroc.pl/dlibra/docmetadata?id=4220&from=publication`.

Maciej Piasecki, Łukasz Burdka, Marek Maziarz, and Michał Kaliński. Diagnostic tools in plwordnet development process. In Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, volume 9561 of *LNCS*, pages 255–273. Springer, 2016. doi: 10.1007/978-3-319-43808-5_20.

Maciej Piasecki, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski, and Paweł Kedzia. Wordnet-based evaluation of large distributional models for polish. In *Proceedings of the 9th Global Wordnet Conference (GWC 2018). Global WordNet Association*, 2018.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. Walenty: Towards a comprehensive valence dictionary of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792. ELRA, 2014.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. A Strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proc. COLING 2012, posters*, pages 1039–1048, 2012.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. Framenet ii: Extended theory and practice. 2006. URL `https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf`.

Krzysztof Rybiński. Sentiment analysis of polish politicians. *e-Politikon. Kwartalnik Naukowy Ośrodka Analiz Politologicznych Uniwersytetu Warszawskiego*, XXIV:162–195, 2017.

Sam Scott and Stan Matwin. Text classification using wordnet hypernyms. *Usage of WordNet in Natural Language Processing Systems*, 1998.

Andrzej Siemiński. Fast algorithm for assessing semantic similarity of texts. *International Jourlnal of Intelligent Information and Database Systems*, 6(5): 495, 2012.

Wojciech Paweł Sosnowski and Violetta Koseska-Toszewa. Multilingualism and dictionaries. *Cognitive Studies/ Études cognitives*, (15), 2015.

Danuta Stanulewicz. Polish terms for 'blue' in the perspective of vantage theory. *Language Sciences*, 32 (2):184 – 195, 2010.

Radosław Szmit. Fast plagiarism detection in large-scale data. In Cham: Springer, editor, *International Conference: Beyond Databases, Architectures and Structures*, pages 329–343, 2017.

Sara Tonelli and Daniele Pighin. New features for framenet: Wordnet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 219–227. Association for Computational Linguistics, 2009.

Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides GM Petrakis, and Evangelos E Milios. Semantic similarity methods in wordnet and their application to information retrieval on the web. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM, 2005.

Ellen M Voorhees. Using wordnet for text retrieval. *WordNet: an electronic lexical database*, pages 285–303, 1998.

Piek Vossen. EuroWordNet General Document Version 3. Technical report, Univ. of Amsterdam, 2002.

Aleksander Wawer and Justyna Sarzyńska. Do we need word sense disambiguation for lcm tagging? In *International Conference on Text, Speech, and Dialogue*, pages 197–204. Springer, 2018.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. A preliminary version of składnica—a treebank of polish. In *Proceedings of the 5th Language & Technology Conference, Poznań*, pages 299–303, 2011.

Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. A Large Wordnet-based Sentiment Lexicon for Polish. In *Proc. RANLP 2015*, page to appear, 2015.

A Comparison of Sense-level Sentiment Scores

Francis Bond,♠ Arkadiusz Janz♢ and Maciej Piasecki♢
♠ Nanyang Technological University, Singapore
♢ Wrocław University of Science and Technology, Poland
bond@ieee.org, {arkadius.janz|maciej.piasecki}@pwr.edu.pl

Abstract

In this paper, we compare a variety of sense-tagged sentiment resources, including SentiWordNet, ML-Senticon, plWordNet emo and the NTU Multilingual Corpus. The goal is to investigate the quality of the resources and see how well the sentiment polarity annotation maps across languages.

1 Introduction

There are several semantic resources with senses annotated by sentiment polarity, e.g. SentiWordNet 3.0 (Baccianella et al., 2010) and even emotions, e.g. WordNet-Affect (Strapparava and Valitutti, 2004; Torii et al., 2011). However, most of them were built on the basis of automated expansion of a small subset of senses described manually. In addition the majority of them were built for a single language, namely English, with ML-SentiCon (Cruz et al., 2014) a notable exception.

This paper present the results of comparing two very different sense-level sentiment resources: a very large semantic lexicon annotated manually for Polish, i.e. plWordNet (Maziarz et al., 2016) expanded with manual emotive annotations (Zaśko-Zielińska et al., 2015); the annotation of two English short stories (*The Adventure of the Speckled Band* and *The Adventure of the Dancing Men* (Conan Doyle, 1892, 1905)) and their Chinese and Japanese translations (Bond et al., 2016a). As the stories have been annotated on the basis of senses, not words – i.e. all words were assigned Princeton WordNet synsets – this opens an unique possibility of cross-lingual comparison of manual sentiment annotation at the level of word senses. These are then compared with SentiWordNet and ML-SentiCon and finally they are all compared to a small gold standard sample Micro-WNOp Corpus (Cerini et al., 2007).

Our technical goal is to analyse the feasibility and technical means of correlation between independently created resources as the first step towards cross-lingual applications. Taking a more fundamental perspective, we want to investigate the level and distribution of correlation between sentiment polarity expression on the sense level between languages. In addition this is also an exercise in utilisation of the interlingual manual mapping between plWordNet and Princeton WordNet that has been built independently.

2 Resources

In this section we describe the resources we used.

2.1 SentiWordNet

SentiWordNet (Esuli and Sebastiani, 2006) annotates a synset with three numerical values in the range $\langle 0, 1 \rangle$ placing the synset in a three dimensional polarity space. The dimensions describe "how objective, positive, and negative the terms contained in the synset are". As the three values must sum to one, there are only two degrees of freedom.

About 10% of the adjectives were manually annotated, each by 3-5 annotators (Baccianella et al., 2010). In SentiWordNet 3.0, the automated annotation process starts with all the synsets which include 7 "paradigmatically positive" and 7 "paradigmatically negative" lemmas.[1] The initial seed is expanded with a random walk algorithm to generate a training set for a committee of classifiers and estimate final polarity scores of synsets. In the end, SentiWordNet 3.0 added automatic sentiment annotation to all of Princeton WordNet 3.0.

[1] good, nice, excellent, positive, fortunate, correct, superior; bad, nasty, poor, negative, unfortunate, wrong, inferior (Turney and Littman, 2003)

2.2 ML-SentiCon

The method proposed in Baccianella et al. (2010) has become the motivation for further work on the development of word-level and sense-level sentiment lexicons. ML-SentiCon (Cruz et al., 2014) expands the idea presented in (Baccianella et al., 2010) by introducing additional sources of information such as WordNet-Affect (Strapparava and Valitutti, 2004) and General Inquirer (Stone et al., 1966) to improve the accuracy and coverage of initial polarity seed. The seed is expanded using the same general approach proposed in Baccianella et al. (2010). However, instead of a single score for each synset, individual scores for each sense are calculated, and then synset scores are calucalted by averaging these.

2.3 plWordNet 4.0 emo

In plWordNet the emotive annotation is assigned not to synsets, but to senses (also known as lexical units: LU), i.e. pairs of lemmas and synsets. These are represented internally as triples of lemma, Part of Speech and sense identifier (number) – every sense belongs to exactly one synset, so a synset represents a sense – a lexical meaning. Senses are fundamental elements of the plWordNet structure, cf (Maziarz et al., 2016).

From the point of view of emotional sentiment polarity, plWordNet senses are divided into *marked* and *neutral*. The first can be also called *polarised*. Polarised senses are assigned the *intensity* of the sentiment polarisation, *basic emotions* and *fundamental human values*. The latter two provide additional characteristics and help annotators to determine the sentiment polarity and its intensity expressed in the 5 grade scale: *strong* or *weak* vs *negative* and *positive*. Each annotator's decision for polarised senses is supported by use examples – a sentence including the given sense and illustrating the postulated sentiment polarity and its strength.

Concerning emotions, due to the compatibility with other wordnet-based annotations, the set of eight basic emotions recognised by Plutchik (Plutchik, 1980) were used (Zaśko-Zielińska et al., 2015). It contains Ekman's six basic emotions (Ekman, 1992): *joy* , *fear*, *surprise*, *sadness*, *disgust*, *anger*, complemented by Plutchik's *trust* and *anticipation*. As a result, negative emotions do not prevail in the set. One sense can be assigned more than one emotion and, as a result, complex emotions can be represented by using the same eight-element set, following the observations of Plutchik (1980).

However, as the comparison we aim for is limited only to sentiment polarity, both emotions and fundamental values will be ignored in comparison.

2.4 NTU Multilingual Corpus

The NTU Multilingual Corpus (Tan and Bond, 2012) has a variety of texts and their translations, many of which are sense annotated.[2]

Two stories from the Sherlock Holmes Canon (*The Adventure of the Speckled Band* and *The Adventure of the Dancing Men*) have been both sense tagged with wordnet senses and annotated for sentiment (Bond et al., 2016a). Princeton Wordnet (Fellbaum, 1998) was used for English, the Chinese Open Wordnet for Chinese (Wang and Bond, 2013) and the Japanese wordnet for Japanese (Bond et al., 2009). These are linked through Princeton WordNet 3.0 (Fellbaum, 1998) with the help of the open multilingual wordnet (Bond and Foster, 2013). In addition, pronouns (Seah and Bond, 2014) and new concepts that were discovered in the corpus during the annotation have been added.

A continuous scale was used for tagging sentiment, with scores from -100 to 100. The tagging tool splits these into seven values by default (-95, -64, -34, 0, 34, 64, 95), and there are keyboard shortcuts to select these values. Three values were chosen for each polarity, in order to be able to show the changes in chunks: *quite good* is less positive than *good* and this is less positive than *very good*. Annotators could select different, more fine-grained values if they desire. The annotators were given several exemplars as guidelines, shown in Table 1. The final column of the table shows examples from the corpus after annotation.

Each of the three texts was annotated by a single native speaker for that language, then the different languages were compared, major differences discussed and, where appropriate, retagged. If they were not sure whether the text segment shows sentiment or not, annotators were instructed to leave it untagged.

In this paper, we only use the sense level annotation, and ignore chunks. Like plWordNet emo, only marked senses are annotated: those senses of

[2]The corpora are searchable here: `http://compling.hss.ntu.edu.sg/ntumc/`. They will be made avaialble for download by the time of the conference.

Score	Example	Chunk Example	Example	Corpus Examples
95	fantastic	very good		perfect, splendidly
64	good	good		soothing, pleasure
34	ok	sort of good	not bad	easy, interesting
0	beige	neutral		puff
-34	poorly	a bit bad		rumour, cripple
-64	bad	bad	not good	hideous, death
-95	awful	very bad		deadly, horror-stricken

Table 1: Exemplars for sentiment scores

words in text that, in context, clearly show positive or negative sentiment were annotated. If a sense is not annotated, then we treat it as an implicit tag of neutral (zero). Operators such as ***very*** and ***not*** were not tagged. Concepts can be multiword expressions, for example *give rise* "produce" or 口を開く *kuchi-wo hiraku* "speak". Each corpus was annotated by a single annotator with linguistic training.

Lang.	Sent.	Words	Concepts	Distinct
English	1,199	23,086	12,972	3,494
Chinese	1,225	24,238	16,285	3,746
Japanese	1,400	27,408	10,095	2,926

Table 2: Size of the Corpus for the three languages

The size of the corpus is shown in Table 2. English is the source language, the translators have separated some long sentences into shorter ones for both Chinese and Japanese. Chinese words are in general decomposed more than English, and the wordnet has fewer multi-word expressions so the corpus has more concepts. Japanese has no equivalent to some common concepts such as ***be*** in *I am happy*, and drops the subject when it is clear from the context and thus has many fewer concepts.

There was some quality control: senses were examined both in context and then out of context. After the initial annotation (done sentence-by-sentence), the annotators were shown the scores organized per word and per sense: where there was a large divergence (greater than one standard deviation), they went back and checked their scores.

Some examples of high and low scoring concepts and their lemmas are given in Table 3. The score for the concept is the average over all the lemmas in all the languages. The concepts are identified with the Interlingual Index (Bond et al., 2016b).[3]

2.5 The Micro-WNOp Corpus

We evaluated the Micro-WNOp Corpus (Cerini et al., 2007) as it is the only **sense-tagged** sentiment lexicon we could find.[4] It was used to evaluate SentiWordNet and build ML-SentiCon, and consists of 1,105 Wordnet synsets chosen from the General Inquirer lexicon (Stone et al., 1966) and annotated by 1–3 annotators.

There are many corpora tagged for sentiment, for example the Stanford Sentiment Treebank (Socher et al., 2013), but few multilingual (Balahur and Turchi, 2014) and no multilingual sentiment corpora for Asian languages. (Prettenhofer and Stein, 2010) contains English, French, German and Japanese product reviews, but they are comparable (reviews of the same product) or machine translated, not translated text, so while useful it is not suitable for studying close correspondences.

3 Comparisons

We are going to compare four languages and two types of resources: a corpus and a lexicon from the perspective of sentiment polarity annotation. In order to make the comparison feasible, we focus on word senses – that can be represented by concepts – and their mappings across languages, as links between the different resources. There are both manually annotated and automatically built (to a very large extent) resources among the compared ones. Finally two types of the sentiment polarity annotations that are represented by the compared resources use similar but slightly different models: the semi-continuous scale, e.g. NTU-MC and the discrete scale, e.g. the five-grades scale of plWordNet emo.

[3] LOD: `http://www.globalwordnet.org/ili/ixxx`.

[4] `http://www-3.unipv.it/wnop/`

Concept	freq	score	English	score	Chinese	score	Japanese	Score
i40833	24	+50	marriage	39	婚事	34	結婚	58
			wedding	34				
i11080	5	+40	rich	33	有钱	34	裕福	66
i72643	4	+33	smile	32	微笑	34	笑み	
i23529	40	−68	die	−80	去世	−60	亡くなる	−63
					死亡	−64	死ぬ	−62
i36562	5	−83	murder	−95	谋杀	−95	殺し	−64
							殺害	−63

Table 3: Examples of high and low scoring concepts from NTU-MC, only total frequencies shown.

3.1 Cross-lingual Comparison inside the Corpus

In this section we take a look at the agreement across the three languages of the NTU-MC. We examined each pair (Chinese-English, Chinese-Japanese and English-Japanese), and measured their correlation using the Pearson product-moment correlation coefficient (ρ), as shown in Table 4. We chose this as it is invariant under separate changes in location and scale. This was calculated over all concepts which appeared in both languages. All three wordnets (Sec. 2.4) use the same conceptual structure, that of Princeton Wordnet. When we compare, it makes no sense to compare senses, as they are language specific.

Instead, we matched concepts, represented by synsets. For each language, we calculated the sentiment score for a synset by averaging over all its senses. When we compare across languages, if a synset appears in the corpus multiple times, we add it to the comparison set as often as the least frequent language. Thus for example, if between Chinese and English, 02433000-a "showing the wearing effects of overwork or care or suffering" appeared three times in Chinese (as 憔悴 *qiáo cuì*) with an average score of -48.5 and twice in English with a score of -64 (as *haggard* and *drawn*), we would count this as *two* occurrences of -48.5 (in Chinese) and -64 (in English). In general, fewer than half of the concepts align directly across any two languages (Bond et al., 2013). Even though we have over 12,000 occurences concepts in English and more in Chinese and Japanese (Table 2) fewer than 7,000 appear in both (Table 4).

Pair	ρ	**# samples**
Chinese-English	.73	6,843
Chinese-Japanese	.77	4,099
English-Japanese	.76	4,163

Table 4: Correlation between the different language pairs

For most concepts, the agreement across languages was high, although rarely identical. There was high agreement for the polarity but not necessarily in intensity/magnitude. For example, for the concept 02433000-a "haggard", the English words *drawn* and *haggard* were given scores of -64, while Chinese 憔悴 *qiáo cuì* was given a weaker score of -34.

An example of different polarity was the English lemma "great" for synset 01386883-a, which received a score of 45.2, whereas the Japanese lemma 大きい for the same synset received a score of 0 (neutral).

In addition, lemmas in the same synset might have another sense that is positive or negative, and this difference causes them to be perceived more or less positively. For example, in English, both *imagine* and *guess* are lemmas under synset 00631737-v, but *imagine* is perceived to be more positive than *guess* because of their other senses. This cross-concept sensitivity can differ from language to language, thus causing further differences. In general, the English annotator was more sensitive to this, which explained much of the difference in the scores. Overall, cross-lingual comparisons of concepts that were lower in agreement were due to both language and annotator differences. The English annotator had generally been more extreme in the rating compared to the Chinese and Japanese annotators.

3.2 Cross-lingual Comparison: Corpus vs Wordnet

NTU-MC and plWN have different sentiment annotation schema. The first one allows for a scale close to continuous: $\langle -100, +100 \rangle$, while the latter uses only 5-degree polarity scale (including *neutral*). In practice, most senses are annotated using the default values, which groups the scores around seven points: three positive and three negative.

NTU-MC annotation was done on the level of word senses represented by PWN synsets. The mapping between plWN and PWN is defined on the level of synsets. Thus, first both annotations in both resources, namely, NTU-MC and plWN had to mapped onto the level of synsets. In the case of NTU-MC we applied the same strategy as above: every synset is assigned a polarity score which is the average across the polarity values assigned to its senses in the corpus (respectively to a given language under examination). This procedure introduces an implicit weighting: more frequent senses have bigger influence on the synset polarity. In addition the polarity values do not need to be constant for a given sense in all its occurrences. So, by averaging them for one synset we additionally balance between small differences resulting from different contexts.

The scale in plWordNet is discrete and semi-continuous in NTU-MC.[5] As any attempt to make the plWordNet scale continuous would be arbitrary (only one dimension and up to three annotations per a sense), we decided to map the NTU-MC scale onto a discrete set of values, namely the five degree scale of plWordNet. First, we generated a histogram of averaged polarity values in which we could observe quasi-Gaussian concentrations of values around ± 34. On the basis of the distribution of values in the histogram we defined thresholds for weak polarity on ± 17. In the case of higher (or lower) polarity of synsets in NTU-MC we could notice that two maxima located around ± 64 and ± 95 were not significantly separated between them, while very distinctively separated from the first one. Thus we decided to treat them as representing one category of strong positive/negative polarity and to set up the threshold for them on ± 54.

In plWordNet in order to obtain synset polarity scores on the basis of sense scores, we cannot simply average them, as the scale consist of only two levels (in each direction) and the average number of senses in a synset is below 2. Thus, the synset polarity is obtained on the basis of simple majority voting[6] from the sense values. In case of a tie, we take the maximum or minimum value, respectively for positive and negative.

In order to identify the corresponding plWordNet and Princeton WordNet synsets, we utilised the manually constructed mapping between both wordnets. It is based on different inter-lingual relations that link synsets and express different levels and forms of meaning correspondence from the very strong correspondence in the case of *I-synonymy* (interlingual) down till, e.g., *I-holonymy* which signals that the target represents a whole that includes the part represented by the source. The mapping procedure organises the inter-lingual relations into a kind of decision lists (one for each Part of Speech) that guide linguists from the strongest relations – also the most informative – to the weakest. The idea was to not leave any synset not mapped, even if only some weak form of correspondence can be expressed. Due to the different types of inter-lingual meaning correspondence, we expected also different levels of correlation between sentiment annotations assigned to the mapped synsets. On the basis of the properties of the inter-lingual relations and the mapping decision lists we divided I-relations into four groups: *synonymic*, *hyponymy*, *hypernymy* and *other*. The first group encompasses *I-synonymy*, *I-partial-synonymy* and *I-interparadigmatic-synonymy* (restricted to *adj–adv* links only).

I-hyponymy is most numerous relation, and expresses that the source synset has more narrow meaning, but mostly it is very close to the meaning represented by the target. The group was extended with *I-inter-register-synonymy* links which share similar properties to *I-hyponymy* links in terms of meaning and polarity.

I-hypernymy is used when the synset of the source wordnet (for which the mapping is built) represents more general meaning than the synset of the target wordnet, so it is a reverse relation to *I-hyponymy*. However, I-hypernymy is further in the mapping decision list than I-hyponymy, so it is used in less clear mapping situations and expresses

[5] I.e. *de facto* discrete on the level of senses and more continuous after averaging

[6] plWN annotation include about 5% of ambiguous senses that can express in some contexts positive or negative polarity. For them both values are taken into account during voting.

significantly weaker correspondence.

The *other* category groups all the rest of inter-lingual relation that are used mostly as a last resort mapping decision, so they signal weak meaning correspondence.

For the comparison we used two different measures. Firstly, in a similar way like for inter-lingual comparison in NTU-MC (see the previous section), we calculated *Pearson's correlation* of the synset scores, setting plWordNet emo's weak to 0.4 and strong to 0.8.

Secondly, we discretised NTU-MC synset (concept) scores to the five grade scale of plWordNet (following the procedure described earlier) and checked the agreement between the resulting values with the sentiment polarity values of plWordNet synsets. Cohen's κ was used to measure the agreement.

The results of both types of comparison are presented in Table 5. The ρ column presents the measured correlation. As sentiment annotations are quite remotely related to each other (done on the level of senses, for two languages, mapped by inter-lingual relations etc.), we decided to measure the agreement in two versions: κ_1 – only the sign of polarity (negative, neutral and positive), and κ_2 – five grade scale. The last column – **#synsets** – tells for how many Polish synsets we managed to establish links to the the synsets annotated in NTU-MC.

Type	ρ	κ_1	κ_2	**#synsets**
Synonymic	**.65**	.60	.53	1,043
Hyponymy	.62	.53	.47	1,271
Hypernymy	.56	.50	.33	147
All links	.63	.55	.48	1,880

Table 5: Correlation and Cohen's Kappa for matched annotations with respect to a type of inter-lingual connection between plWordNet and NTU-MC.

The correlation and agreement are the highest for the synonymic group of inter-lingual relations, as we could expect. The correlation does not drop much for the I-hyponymy group, but the agreement for both non-synonymic relations is significantly lower.

We do not provide results for the *other* category of mapping relations, as we could detect only a small number of links.

Concerning the agreement, it appeared to be good when only the polarity sign is concerned (κ_1), and it is still positive in the case of the full five grade scale (κ_2). The use of the hyponymy and hypernymy categories of links resulted also in a significantly lower, but still positive agreement. All three measures showed continuously decreasing and lower agreement when we apply less and less informative inter-lingual relations.

3.3 Cross-lingual Comparison: Analysis of Discrepancies

Limited agreement between the two manual resources means that there must large number of differences in annotations. In order to understand better the nature of these discrepancies we took a closer look into them into comparisons based on the synonymic inter-lingual relations. Most of the differences in this category result from different levels of the polarity. Only 5.6% of them express significant disagreement, i.e. different sign of polarity. One other co-authors has manually surveyed them to find that there are only 14 cases of two opposite polarity values, and a larger number of cases in which neutral polarity (i.e. the lack of polarity) on one side is mapped on the marked polarity on the other side (67,6%). Concerning the first, the strongest difference type, all such cases are listed in Table 6.

Sense	Pl	MC	Cause
incredible.1 (adj)	-s	+w	err. in plWN
extreme.1 (adj)	-s	+w	more narrow Polish meaning
impassable.1 (adj)	-w	+s	err. in NTU-MC
crazy.2 (adj)	+w	-s	I-part-syn.
grave.1 (adj)	+s	-s	err. in plWN
flare-up.1 (verb)	+s	-w	too strong I-relation
attack.5 (verb)	+w	-s	err. in plWN
blackguard.1 (verb)	-w	+s	err. in NTU-MC
fancy.1 (noun)	-w	+s	err. in NTU-MC
glimmer.2 (noun)	+w	-w	err. in both

Table 6: Survey of the strongest differences between the annotation of plWordNet and NTU-MC, where s = strong, w = weak.

As we could notice in Table 6, there is very little disagreement for nouns, only for adjectives and verbs that are much more difficult for both inter-lingual mapping and emotive annotation. The vast majority of disagreements resulted from the errors in the original annotations, e.g.: *incredible.1* – on

the Polish side the emotive annotation is based on wrong sense interpretation; *extreme.1* – the corresponding Polish sense was interpreted in a more narrow way, with a tendency to negative interpretation of extreme; *impassable.1* – a very likely error in NTU-MC error, it is hard to imagine a positive interpretation of this sense on the basis of the examples from the corpus, etc. The other two discrepancies seem to be caused by the mapping with the help of I-partial-synonymy. It expresses overlapping meaning, so their overlaps do not need to match the assigned sentiment annotations.

For *glimmer.1* it appears as *gleam* in *"See here, mister!" he cried, with a gleam of suspicion in his eyes, "you're not trying to scare me over this, are you?*. The complement *suspicion* is clearly negative but *gleam* is probably neutral, neither resource was perfect, and may have been biased by the context.

We also examined disagreements that involve neutral annotations: that is, in one resource the score is neutral (zero) and in the other is carries sentiment. In almost all cases, the neutral score was wrong. Annotators in NTU-MC were allowed to omit explicit neutral annotation and leave words unannotated in such cases. This resulted in some number of mistakenly skipped words. In a similar way, the vast majority of plWordNet:neutral vs NTU-MC:polarised cases is the combined result of gaps in the plWordNet sentiment annotation and a default rule that all gaps should be treated as neutral cases. The annotation was done for almost 90,000 senses, but this is around half of the wordnet. The default rule works quite well for nouns, where potentially neutral hypernymy branches were intentionally excluded from annotation, but fails definitely for other Parts of Speech.

3.4 Comparison with SentiWordNet and ML-SentiCon

Next, we compared both manually annotated resources, namely, plWordNet and NTU-MC with two resources used in many applications: SentiWordNet (Baccianella et al., 2010) and the newer ML-SentiCon (Cruz et al., 2014), discussed shortly in Sec. 2.1 and 2.2. As it was already mentioned, the sentiment annotation in both these resources were automatically propagated from a small set of manually prepared seeds.

SentiWordNet and ML-SentiCon are annotated on the level of synsets, so we used exactly the same pre-processing of plWordNet and NTU-MC. In the case of plWordNet we used also the same inter-lingual relation to map the Polish synsets onto Princeton WordNet ones. The Pearson's correlation for polarity values is presented in Table 7. Here we are measuring over distinct concepts, with no weighting. For the sentiment lexicons, we give results over the subset in the corpus, and over all synsets.

Pair	ρ	**# samples**
SentiWN – MLSenticon	.51	6,186
	.42	123,845
NTUMC – SentiWN	.42	6,186
NTUMC – MLSenticon	.48	6,186
plWN – SentiWN	.32	22,435
plWN – MLSenticon	.41	22,435
plWN – NTUMC	**.63**	1,880

Table 7: Correlation between the different resources

The results show that none of these four resources agree very well. The automatically created resources related better with each other, but still had a low correlation. Their correlation is significantly smaller than the manually annotated NTU-MC and plWordNet. That is even more significant, when we take into account that the manually annotated resources were created for different languages, are based on different annotation models and we required the help of inter-lingual relations to map them. This whole process had to hamper the observed correlation. Neither automatically built resource closely correlated with the examples seen in context in the corpus and in the plWordNet use examples. However, the newer ML-SentiCon has slightly better agreement.

Examining the examples by hand, many concepts we marked as neutral received a score in these resources (e.g. *be* which is +0.125 in SentiWordNet or *April*, which is -0.125 in ML-SentiCon), while other concepts for which we gave a strong score (e.g *violence* -64) were neutral in these other resources. As our senses were confirmed by manual inspection, we consider our scores to be more accurate.

SentiWordNet and ML-SentiCon were both produced by graph propagation. SentiWordNet from a small number of seeds (around 14) and ML-SentiCon from more. It would be interesting to try to add our new data (suitably normalised) as new

seeds and try to recalculate the scores: a larger pool of seeds should give better results.

3.5 Evaluation with the Micro-WNOp Corpus

The Micro-WNOp Corpus was chosen to evaluate our resources, as it is commonly used and well balanced. First, we calculated the agreement for different annotators in the corpus. In group 1, with three annotators, we calculated annotator one vs the average of two and three, then two vs one and three and three vs one and two (ρ = 0.85, 0.78, 0.83 respectively, mean is 0.82). For group 2 with two annotators we compared them to each other (ρ = 0.94). In each case, we summed positive and negative to get a single score and compared using the Pearson product-moment correlation (ρ). This give us an upper bound for human agreement.

Both plWordNet and NTU-MC have far higher correlations than SentiWN, although with no results for many synsets. This shows the well known effect that hand-built resources are more reliable, but generally sparser.

Pair	ρ	# syn.
Micro-WNOp InterAnnotator	.88	995
Micro-WNOp – plWN	.77	413
Micro-WNOp – NTU-MC	.75	130
Micro-WNOp – SentiWN	.63	1,048
Micro-WNOp – plWN&NTU-MC	**.78**	352

Table 8: Correlation of Micro-WNOp lexicon with other resources

For completeness, we also calculated the correlation between Micro-WNOp and ML-SentiCon $\rho = .96$. However, as Micro-WNOp was used to as training data for ML-SentiCon the evaluation is not meaningful and we do not include it in Table 8.

4 The combined sentiment lexicon

One clear results of this comparison is that comparing the lexicons with each other improves them. Places where there was a difference in polarity or in zero vs non-zero sentiment were almost all errors. Once discovered there are easy to fix, and we have shared the results with the resource creators. Because the scores are different (a continuous score for NTU-MC and a 5 point scale for plWordNet emo) we can combine in two ways: binning NTU-MC or setting values for weak and strong for plWordNet emo (we used 0.4 and 0.8). They can then be combined over all synsets, to give a single resource that should be somewhat more accurate then either alone.

To combine the lexicons we decided to use binning strategy on NTU-MC and Micro-WNOp followed by a simple selection procedure. To represent matched concepts within the same category set we used thresholding function with thresholds being a result of score distribution analysis. In case of NTU-MC the following bins were proposed: $|s| \leq 0.18$ for neutral category, $0.18 < |s| \leq 0.54$ for weak polarity and $|s| > 0.54$ for strong polarity. First we selected a subset of paired synsets annotated both in NTU-MC and plWordNet emo which were compatible in terms of their polarity categories. To reduce the discrepancy between the annotations we also decided to remove all of paired synsets having different polarity categories. In the last step we introduce a group of unmatched synsets with their annotations to extend the coverage of joint lexicon. The final lexicon was evaluated again on Micro-WNOp (Table 7) giving a slight improvement of correlation.

5 Conclusion and Future Work

In this paper we presented a comparison of wordnet-based sense-level sentiment lexicons. We showed that the two manually annotated resources were more accurate than the semi-automatically created resources. We also showed that linking across languages preserved most of the valence (ρ = 0.65 – 0.77 for equivalent synsets). This means that the resources can be used for other languages, linked either directly or through an interlingual index. Finally we showed how they could be improved further by cross-checking and resolving inconsistencies, or by combining them.

In future work, we will: (i) correct the errors in the two resources and recalculate their correlation (as it is sensitive to outliers). (ii) create further sense-annotated sentiment tagged text

- Another Sherlock Holmes story (*The Red-Headed League*)
- Other translations for *The Adventure of the Speckled Band*: we have Bulgarian, Dutch, German, Indonesian, Italian and Polish, and are in the process of annotating them.

and (iii) model the effects of operators on lexemes to allow for compositional changes.

Acknowledgments

This research was partially supported by Fuji Xerox Corporation through joint research on *Multilingual Semantic Analysis*, CLARIN, the EU RISE Project 691152 — RENOIR *Reverse EngiNeering of sOcial Information pRocessing* and the NTU Digital Humanities Research Cluster.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources*, pages 1–8. ACL-IJCNLP 2009, Singapore.

Francis Bond, Tomoko Ohkuma, Luís Morgado da Costa, Yasuhide Miura, Rachel Chen, Takayuki Kuribayashi, and Wenjie Wang. 2016a. A multilingual sentiment corpus for Chinese, English and Japanese. In *6th Emotion and Sentiment Analysis Workshop (at LREC 2016)*. Portorož.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016b. CILI: The collaborative interlingual index. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*, pages 50–57.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013)*, pages 149–158. Sofia. URL `http://www.aclweb.org/anthology/W13-2319`.

Sabrina Cerini, Valentina Compagnoni, Alice Demontis, Maicol Formentelli, and G Gandini. 2007. Micro-wnop: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*, pages 200–210.

Arthur Conan Doyle. 1892. *The Adventures of Sherlock Homes*. George Newnes, London.

Arthur Conan Doyle. 1905. *The Return of Sherlock Homes*. George Newnes, London. Project Gutenberg `www.gutenberg.org/files/108/108-h/108-h.htm`.

Fermín L Cruz, José A Troyano, Beatriz Pontes, and F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of 5th Conference on Language Resources and Evaluation LREC 2006*, pages 417–422.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plwordnet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL. URL `http://aclweb.org/anthology/C/C16/`.

Robert Plutchik. 1980. *Emotion: Theory, research, and experience*, volume Vol. 1. Theories of emotion. Academic, New York.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *48th Annual Meeting of the Association of Computational Linguistics (ACL 10)*, pages 1118–1127. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P10-1114`.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Yoshimitsu Torii, Dipankar Das, Sivaji Bandyopadhyay, and Manabu Okumura. 2011. A Developing Japanese WordNet Affect for Analyzing Emotions. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011), 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 80–86.

Peter D. Turney and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21(4):315–346.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.

Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for Polish. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing – RANLP'2015*, pages 721—730. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria. URL `http://aclweb.org/anthology/R15-1092`.

Portuguese Manners of Speaking

Valeria de Paiva
University of Birmingham
valeria.depaiva@gmail.com

Alexandre Rademaker
IBM Research and FGV/EMAp
alexrad@br.ibm.com

Abstract

Lexical resources need to be as complete as possible. Very little work seems to have been done on adverbs, the smallest part of speech class in Princeton WordNet counting the number of synsets. Amongst adverbs, manner adverbs ending in '-ly' seem the easiest to work with, as their meaning is almost the same as the one of the associated adjective. This phenomenon seems to be parallel in English and Portuguese, where these manner adverbs finish in the suffix '-mente'. We use this correspondence to improve the coverage of adverbs in the lexical resource OpenWordNet-PT, a wordnet for Portuguese.

1 Introduction

Adverbs get a short shrift in Lexical Semantics. They have the smallest number of synsets in Princeton WordNet (PWN) (Fellbaum, 1998), perhaps rightly so, as most of their meaning, at least as far as *manner adverbs* are concerned, can be gleaned from the associated adjective. Manner adverbs tell us about the way something happens or is done. Many English adverbs are formed by adding the suffix -ly to an adjective (e.g. calm + ly = camly). Many Portuguese adverbs are formed in a similar manner by adding the suffix '-mente' to the singular feminine form of the adjective. For example, *sincera* (sincere, feminine) + 'mente' = *sinceramente* (sincerely). (Note the 'a' in *sinceramente*, marking the feminine version of the adjective.) This phenomenon of pairs adjective-adverb, where morphologically only the suffix '-ly' is added to the adjective to form the adverb is very widespread both in English and in Portuguese. Many manner adverbs in Portuguese consist of an adjective plus the suffix '-mente'. These modify a predicate or phrase, specifying how the action unfolds. They can be paraphrased by "one way of Xing" or "a form of X", where X corresponds to the base adjective for the adverb's formation. For example, the adverb *rapidamente/rapidly* lets you qualify that a given action was performed "in a rapid way".

The generic pattern here seems to be that, since most of these adverbs are a kind of "manner", and manner in Portuguese is a feminine word, the adverb tends to be the feminine form of the adjective, plus the suffix '-mente'. This can be used by NLP systems, such as FreeLing(Padro and Stanilovsky, 2012), to avoid keeping duplicate meanings that are almost the same. For example, for *stertorously*, we have the feminine adjective *estertorosa* plus the suffix *-mente*, *estertorosamente*. We take the view that all the reasonably used adjectives and adverbs should be in the lexicon and strive to complete our online, open source Portuguese wordnet(de Paiva et al., 2012). This goal of completing the a wordnet adverbial lexicon, using corpora, we share with (Côrtes et al., 2018). Our work can be seen as taken a subcase of their approach (only "manner" adverbs are considered here), but looking at varied-sized corpora in Portuguese.

Princeton WordNet v3.0 contains 3621 adverb synsets. Out of the total number of adverb synsets, some 2646 synsets have word forms terminating in '-ly', which usually indicates a manner adverb, for example *attributively: in an attributive manner*. Also most of the manner adverbs are, in first approximation, monosemous, i.e. mostly they are in a single synset. Princeton WordNet statistics on polysemy can be found in https://wordnet.princeton.edu/documentation/wnstats7wn. Out of the 3621 adverb synsets, 2554 have no translation in Portuguese in our OpenwordNet-PT lexical resource (de Paiva et al., 2012), at the moment. Out of these 2554 synsets, 1863 have words finishing in '-ly', so we presume that these are manner adverbs. These synsets follow the usual distribution,

mostly (1375) have a single word in English. Then 348 synsets have two words and 109 synsets have three words. Few synsets (31) have 4 or more words.

Total English adverb synsets	3621
EN synsets no-translation	2554
EN synsets with -ly	1863
Single word EN-adverb-ly	1375
Two words EN-adverb-ly	348
Three words EN-adverb-ly	109
Four or more words EN-adverb-ly	31

Table 1: PWN Coverage of manner adverbs

The adverbs with only a word should be the ones that are easy to translate and, in first inspection, seem to behave similarly both in English and in Portuguese. Since the phenomena are so similar in English and Portuguese, an approach based in translations and checking should work well for our goal of completing the lexical resource in Portuguese. Our main goal here is to reduce the number of untranslated synsets as much as we can, as far as manner adverbs finishing in '-ly' (English) and '-mente' (Portuguese) are concerned.

2 PWN, OMW and OWN-PT

Our work is based on using Princeton WordNet (PWN) and the Open Multilingual WordNet (OMW) (Bond and Foster, 2013) to improve the OpenWordNet-PT (de Paiva et al., 2012). PWN is one of the most used lexical resources in Computational Linguistics. It is the most used resource for Word Sense Disambiguation, according to Wikipedia `https://en.wikipedia.org/wiki/WordNet`, but it has many other applications like Natural Language inference, text summarizing, question answering, etc.

The Open Multilingual WordNet (OMW) (Bond and Foster, 2013) provides access to open licensed wordnets in a variety of languages, all linked to the Princeton Wordnet of English (PWN). Their goal is to make it easy to use wordnets in multiple languages. Other wordnets with reasonably large number of adverbial synsets are associated to the OMW, but only the French wordnet WOLF (Wordnet Libre du Français) (Sagot et al., 2009) seems to have had a similar problem to the one we face, of extending the adverbial coverage of synsets from English. For French, as well as for Portuguese, the synsets in previously existing wordnets were not available, because of license issues. However to complete WOLF, Segot et al had the lexicon Lefff and the synonyms database DicoSyn to help them and we have not found similar resources for Portuguese.

OpenWordNet-PT (OWN-PT) is an ongoing project to build up a large wordnet for Portuguese. OWN-PT is still half the size of PWN, and we have done so far only partial evaluations of its coverage. For instance, punctual evaluations of verbs was done in (de Paiva et al., 2016, 2014a) and of (gentilic) adjectives in (Real et al., 2016). Several discussions on nominalizations and their arguments are presented in (Real et al., 2014; Freitas et al., 2014; de Paiva et al., 2014b). As suggested in (Real et al., 2015) the use of a visual interface for OWN-PT helps to discover some interesting subsets of synsets to work on, like the ones ending in "ly" that we investigate here. OWN-PT is the Portuguese wordnet in the OMW project, it has currently 47,932 synsets, 56,928 word forms and 81,374 senses.

3 Kinds of Adverbs

Translation seems to work fairly well for this tightly restricted subset of manner adverb synsets, but there are still some generic issues and problems.

WordNet and other lexical resources have to cope with affixes (prefixes and suffixes) and these seem to be used differently in different languages. For instance Portuguese has no suffix "less" as in *bloodlessly*, *mindlessly*, and *painlessly*. One needs to use a separate preposition *sem/without* to express the same meaning *sem sangue, sem pensar, sem dor*, respectively. Thus corresponding to the suffix "less" in English we need adverbial phrases instead of adverbs for adjetives and adverbs in Portuguese. PWN lists 76 synsets that are manner adverbs ending in "lessly". Of course these can have single word translations in Portuguese, but these are not as direct as we would like them to be. For instance for *bloodlessly* (in the appropriate abstract synset 00418712-r) we can use the similar *pacificamente/pacifically* but the direct translation would be an adverbial locution *sem sangue/without blood*.

English also has several ways of negating an adjective. The preffix "un" appears in several of the manner adverbs, for example in *unofficially*, *unde-*

servedly, and *unmelodiously*. This prefix has no exact corresponding version in Portuguese. Thus while the positive versions e.g. *officially*, *deservedly*, and *melodiously* have exact correspondents in Portuguese (namely *oficialmente*, *merecidamente*, and *melodiosamente*), the negative versions do not exist as single words in Portuguese. Again we need either adverbial expressions (e.g. *não oficialmente* or less direct translations, like *dissonantemente* for *unmelodiously*. For a few words we have the similar prefix "i/in" in *imerecidamente*. Princeton WordNet has 202 synsets for manner adverbs of the form 'un*ly', out of which, 153 have no Portuguese words at all, to begin with.

English has an extremely useful way of creating adjectives from nouns, by adding a "y" or "ly" to the noun. For example *sandy*, *hilly*, and *pearly* mean "like sand", "with many hills" and "of the off-white color of pearls". Some of these are considered informal language and do not appear in PWN, for example *yellowy*, which is in the Collins online dictionary but not in PWN. Some of these adjectives coming from nouns do not seem to have similars in Portuguese, for example *girly*, *womanly*, and *manly*. We can say *femininamente* and *masculinamente* in Portuguese and we can say *womanly* and *manly* in English, but we do not have the adverbs *femininely* and *masculinely* in English, neither do we have *meninamente*, *mulhermente*, or *homenmente* in Portuguese.

For some cases, where the adverb/adjective pair comes from a noun, e.g. *thirstily*, *thirsty*, and *thirst*, in Portuguese we can have the noun and the adjective, but the adverb is much less used. So *sede* (noun) and *sedento* (adjective) are common enough, but *sedentamente* (adverb) is less so. The meaning in this case is obvious and if someone uses the word, they will be understood, but it is not widespread.

Other suffixes that do not translate well are the very Anglo-Saxon ones like "ward" (in the direction of) as in e.g. *heavenward*.

Sometimes an almost exact translation exists, but the words are not from the same root. For example *regally: in a regal manner* corresponds to *regiamente: uma maneira de ser rei*.

We have to be careful with the so called *faux amis*. For instance the work *cruelly* in English has two fairly distinct meanings or synsets: 00232425-r cruelly (excessively; "a cruelly bitter winter") and 00232499-r cruelly (with cruelty; "he treated his students cruelly") The first one is not what *cruelmente* means in Portuguese, it means something like *excessivamente/excessively* only, while the second is exactly the same meaning as the Portuguese word, some kind of evilness. To make it clear, we do say in Portuguese *um inverno cruel*, (a cruel winter), but the adverb is not used in this sense. All of these issues have to be manually checked and we are in the process of doing so.

4 Tools and Evaluation

Evaluation of semi-automatically constructed resources is a thorny subject. There are no canonical ways of evaluating these resources, the usual mechanical turk style evaluation seems both pointless and not objective enough. However, there are plenty of lists of most used words on the web, usually collected by teachers of English as a second language and these may help us guarantee that, at least the most used manner concepts are in place. We used a collection of lists to evaluate this work.

We checked lists of both English and Portuguese adverbs. The first list of English adverbs, comes from the English Club[1], and consists of 130 popular manner adverbs in English, translated automatically and manually verified. We check how many of these 130 adverbs, are already present in OpenWordNet-PT. This is a qualitative measure, as the English adverbs (automatically translated) are sometimes polysemous and translations, even trying to keep roots, whenever possible, can lead to different results. With this list we were hoping to check our coverage for popular manner adverbs, as a list produced for learners of English as a second language tries to cover as many popular meanings as possible. The results were not very inspiring. Out of 130 English verbs translated, 5 were not of the form ending in -mente in Portuguese. Then of the 125 left, 45 were not present in OWN-PT, so we were missing 36% of this set. We had expected to have a number closer to the whole set already covered.

However, many of the adverbs coming out from this list are not very popular in Portuguese. So we decided to check the manner adverbs in the Parallel Universal Dependencies corpora[2], since these exist both in Portuguese and English and

[1] list at https://www.englishclub.com/vocabulary/adverbs-manner.htm

[2] see https://github.com/UniversalDependencies/UD_Portuguese-PUD

they are already aligned. Parallel Universal Dependencies (PUD) treebanks were created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies[3]. There are 1000 sentences in each language, always in the same order. The sentences are taken from the news domain and from Wikipedia. There are usually only a few sentences from each document, selected randomly, not necessarily adjacent, so sentences usually make sense by themselves.

There are 127 manner adverbs ending in -ly in the English Parallel Treebank PUD-EN. Perhaps surprisingly, PWN misses three of these adverbs *proactively, definitively, logistically*. These can arguably be considered newish adverbs, if one looks at the usage distribution provided by Google, e.g. Figure 1.

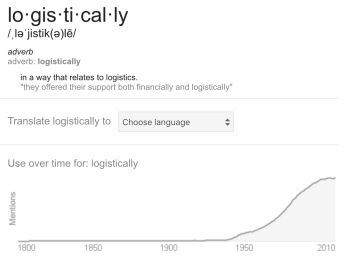


Figure 1: The Universe

We checked whether all 127 English manner adverbs had translations in Portuguese and found that only 12 synsets had not, more in line with what we were expecting. Two of these 12 were not in PWN to begin with, so we had to add word forms to our resource in Portuguese for ten synsets. These synsets correspond to adverbs *reportedly, unusually, thematically, technologically, posthumously, persistently, mindlessly, inimitably, culturally, catastrophically*. Taking advantage of the parallelism of the resources, we then looked at all the "-mente" adverbs in the Parallel Universal Dependencies corpus in Portuguese (PUD-PT). There we found 118 adverbs ending in -mente, of which only 10 were not described in the OpenwordNet-PT. (Of the three adverbs missing in PWN two would need to be added to Portuguese too (*logistically, proactively*, but *definitivamente*, corresponding to *definitively* was already in the base.) The Portuguese manner adverbs added are *tecnologicamente, presumivelmente, postumamente, proativamente, incomumente, escrupulosamente, culturalmente, catastroficamente, arduamente, tematicamente*

Encouraged by these results, which looked more the way we expected, we also looked at the list of the manner adverbs from the largest Universal Dependencies (UD) English corpus, UD-EWT[4]. First we looked at only the hundred most used adverbs in UD-EWT and realized that not many manner ly-adverbs were present. Then we looked at the whole list of ly-adverbs in UD-EWT, which counted 417 unique terms and checked how many we have already in Portuguese, in our resource. Here we come up against one of the problemas of corpus work, in that there are 28 words that are typos, informal language (like "prolly" for 'probably') or mispellings. But 334 were correctly translated, so we needed to add some 60 synset translations.

Finally we turned the tables and looked at Portuguese "mente"-adverbs in our favorite application corpus, the Brazilian Historical Dictionary of Biographies (the acronym in Portuguese is DHBB)(Paiva et al., 2014). Again not so many mente-adverbs in the top 100 most used adverbs, only 20, out of which two we cannot find an easy PWN synset for. (PWN has *parallel*, but not the adverbial form, in Portuguese *paralelamente*. PWN has *provisionally*, which is similar to *interinamente*, but not exactly the same). Then we tried to get all the manner adverbs in this corpus of 323K sentences, but ran out of time to complete the checking. In summary, by looking at all these lists and also by simply checking the glosses for similar adverbs in Portuguese we managed to complete 151 synsets in OWN-PT. While this means that there are still 1727 empty manner adverbs synsets, our error reduction rate was respectable for a manually checked resource. Given the total number of adverb synsets in PWN empty before (2554), completing another 151 synsets gives us 6% more information than we had before.

5 Conclusion

We worked to semi-automatically complete the OpenWordNet-PT, as far as adverbs of manner, finishing in "ly" (in English) or "-mente" (in Portuguese) are concerned. These adverbs are called "deadjectival", as they are usually a "way of X"

[3]check http://universaldependencies.org/conll17/

[4]see http://universaldependencies.org/

where X is an adjective. The phenomenon seems completely parallel in English and Portuguese and hence it seemed to us that it should be an easy task to accomplish.

The task proved somewhat more complicated than envisaged. Out of the 1863 synsets with no Portuguese words we have managed to manually check some 600 and this uncovered some similarities and issues that seem worth pointing out. We also realized that an open dictionary of deadjetival adverbs that we had intended to use as a superset of our synsets (Portal da Língua Portuguesa http://www.portaldalinguaportuguesa.org/) has too many entries that we do not feel confortable listing, as they do not seem very much used in the Portuguese we speak.

On the positive side, the way we found of comparativaly checking the Parallel Universal Dependencies corpus points out a new application of these parallel corpora, to help evaluate coverage of lexical resources. We plan to use it for checking adjectives and verbs in PWN and OWN-PT in the near future. Finally, we need to finish checking all the other thousand adverbial synsets in PWN, which are empty in Portuguese. We also plan to develop our own dictionary of deadjectival pairs that work both in English and Portuguese.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1352–1362.

Priscila Côrtes, Mateus Riva, and Livy Real. 2018. Extending the adverbial coverage of OpenWordnet-PT. In *Proceedings of the PROPOR Conference (PROPOR2018)*, Canela, Brazil. Springer.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Cláudia Freitas, Valeria de Paiva, Alexandre Rademaker, Gerard de Melo, Livy Real, and Anne de Araujo Correia da Silva. 2014. Extending a lexicon of portuguese nominalizations with data from corpora. In *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014*, São Carlos, Brazil. Springer.

Lluis Padro and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Valeria de Paiva, Fabricio Chalub, Livy Real, and Alexandre Rademaker. 2016. Making virtue of necessity: a verb lexicon. In *PROPOR – International Conference on the Computational Processing of Portuguese*, Tomar, Portugal.

Valeria de Paiva, Cláudia Freitas, Livy Real, and Alexandre Rademaker. 2014a. Improving the verb lexicon of openwordnet-pt. In *Proceedings of Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish (ToRPorEsp)*, São Carlos, Brazil. Biblioteca Digital Brasileira de Computação, UFMG, Brazil.

Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. 2012. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In *Proceedings of COLING 2012: Demonstration Papers*, pages 353–360, Mumbai, India. The COLING 2012 Organizing Committee. Published also as Techreport, http://hdl.handle.net/10438/10274.

Valeria de Paiva, Livy Real, Alexandre Rademaker, and Gerard de Melo. 2014b. NomLex-PT: A lexicon of Portuguese Nominalizations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Valeria De Paiva, Dário Oliveira, Suemi Higuchi, Alexandre Rademaker, and Gerard De Melo. 2014. Exploratory information extraction from a historical dictionary. In *IEEE 10th International Conference on e-Science (e-Science)*, volume 2, pages 11–18. IEEE.

Livy Real, Fabricio Chalub, Claudia Freitas, Alexandre Rademaker, and Valeria de Paiva. 2015. Seeing is correcting: curating lexical resources using social interfaces. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 20–29.

Livy Real, Valeria de Paiva, Fabricio Chalub, and Alexandre Rademaker. 2016. Gentle with gentilics. In *Joint Second Workshop on Language and Ontologies (LangOnto2) and Terminology and Knowledge Structures (TermiKS) (co-located with LREC 2016)*, Slovenia.

Livy Real, Alexandre Rademaker, Valeria de Paiva, and Gerard de Melo. 2014. Embedding NomLex-BR nominalizations into OpenWordnet-PT. In *Proceedings of the 7th Global WordNet Conference*, pages 378–382, Tartu, Estonia.

Benoît Sagot, Karën Fort, and Fabienne Venant. 2009. Extending the adverbial coverage of a french wordnet. In *NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources*, page 0.

Completing the Princeton Annotated Gloss Corpus Project

Alexandre Rademaker
IBM Research and FGV/EMAp
alexrad@br.ibm.com

Bruno Cuconato
IBM Research and FGV/EMAp
bcclaro@gmail.com

Henrique Muniz
IBM Research and FGV/EMAp
hnmuniza@gmail.com

Alexandre Tessarollo
FGV/EMAp and Petrobras
alexandretessarollo@gmail.com

Alessandra Cid
FGV/EMAp
alessandracorreacid@gmail.com

Abstract

In the Princeton WordNet Gloss Corpus, the word forms from the definitions ("glosses") in WordNet's synsets are manually linked to the context-appropriate sense in the WordNet. The glosses then become a sense-disambiguated corpus annotated against WordNet version 3.0. The result is also called a semantic concordance, which can be seen as both a lexicon (WordNet extension) and an annotated corpus. In this work we motivate and present the initial steps to complete the annotation of all open-class words in this corpus. Finally, we introduce a freely-available annotation interface built as an Emacs extension, and evaluate a preliminary annotation effort.

1 Introduction

The Princeton WordNet Gloss Corpus is a corpus of the manually annotated synset definitions (glosses) from the Princeton Wordnet (PWN) (Fellbaum, 1998). The corpus is available for download in the PWN website as one of the stand-off packages that supplement the WordNet 3.0 release.[1] Although it has been already recognized as a precious resource, the project of semantically tagging all PWN glosses was not finished. According to the PWN website, the corpus contains 206,711 words (including collocations) yet to be disambiguated. In simple terms, our goal is to complete the disambiguation of all open-class words in this corpus, and here we present our preliminary findings and methodological decisions.

Previous efforts address this same goal in older versions of PWN using automatic or semi-automatic methods (Harabagiu et al., 1999; Moldovan and Novischi, 2004). Here we aim at high-quality human annotation of the glosses, leveraging the lessons learned and directives developed for the project in Princeton but adapting them to our tools and priorities. Data is available at https://github.com/own-pt/glosstag using the same open license used by Princeton for the current version of the data.

The definitional glosses were introduced in PWN primarily to help humans identify the meaning of the synsets, but recently, many word sense disambiguation (WSD) algorithms use the network structure of PWN in combination with the glosses to improve the identification of the most plausible sense for a given word in a corpus (Agirre and Soroa, 2009; Banerjee and Pedersen, 2002; Basile et al., 2007). By semantically disambiguating the words in the glosses, we add pointers from each word to its synset, and this increases the connectivity between the WordNet synsets by approximately an order of magnitude, hopefully improving the performance of these algorithms.

Another reason for such an effort is to ensure the completeness of PWN. By completeness we mean the property of a lexico-semantic resource that all words used in the definitions of the concepts are also themselves explained in this same resource. Hopefully, this completeness could also help us ensure quality in our long-term endeavor, the expansion of PWN to highly technical domains such as those of the geosciences, agriculture, and law. Once more concepts are added or redefined, we will redefine and add glosses that we intend to disambiguate, forcing us to use the newly added senses in a productive cycle of editing, testing, and correcting.

We begin this paper by discussing the original dataset and how we interpreted it converting to a

[1] http://wordnetcode.princeton.edu/glosstag.shtml

more friendly format. Next we describe our annotation interface and some of our implementation decisions. We continue by discussing some of the issues we encountered while sense-tagging the glosses corpus. Finally, we evaluate our ongoing annotation work, and discuss related work and conclude.

2 Sense tagging

Semantically tagging (or sense annotating) a corpus is a task of constructing a semantic concordance – a textual corpus and a lexicon so combined that every content word in the text is linked to its appropriate sense in the lexicon (Miller et al., 1993). Two different strategies for building a semantic concordance are known: the sequential and the targeted approaches.

(Miller et al., 1993) presented one of the first tools developed for supporting the work on building a semantic concordance with PWN, the ConText. The tool was constructed to support sequential tagging. In this approach, the annotator starts with the corpus and proceeds through it word by word. This procedure has the advantage of immediately revealing deficiencies in the lexicon: missing words, missing senses, and indistinguishable definitions. The sequential process was chosen because of their priorities at that time, as they aimed to make substantial improvements in the PWN. Another tool supporting the sequential approach to building semantic concordances was described by (Bond et al., 2015). The tool was introduced after a brief survey on other tools for sense tagging, none of them actively maintained and freely available at that time.

In the targeted approach, the work starts with the lexicon: we focus on a polysemous word, extract all sentences from the corpus in which that word occurs, categorize the instances and write definitions for each missing sense, and create a pointer between each instance of the word and its appropriate sense in the lexicon; we then repeat the process by choosing another word to focus on. The targeted approach has the advantage of concentrating the annotation effort on a single word, producing better definitions. However, the previously listed flaws in the lexicon would not appear so straightforwardly in this targeted strategy. Consequently, this strategy has the potential of being more successful when the lexicon has already reached a more stable stage. The targeted strategy was the one chosen for the Wordnet Gloss Corpus initial phase; it is described in the original annotators' guidelines that we had access to, and we have decided to follow it as close as possible.

The original Wordnet Gloss Corpus project employed an interface called Mantag, implemented in the Perl programming language.[2] Unfortunately, the tool has many dependencies on legacy code that we were not able to solve.

For our continuation of the Wordnet Gloss Corpus annotation project, we decided to implement a serverless application that can be used offline. This decision reflects the prevailing understanding that semantic annotation is a difficult task that is best done individually and in an environment conducive to concentration. In our tool, each annotator can perform their work independently, making annotations on overlapping parts of the corpus or not. The annotators' data can then be consolidated, possibly including discussions aiming at agreement in the cases where annotations diverge. We have also differed in our technology of choice compared to (Bond et al., 2015). Instead of choosing a web framework we have decided to implement our annotation tool in the extensible and free text editor Emacs,[3] taking advantage of the editor's support for multiple platforms and its rich ecosystem. The annotation interface needs no internet access, depends only on Emacs and its libraries, and can be run either from a graphical interface or from a terminal window.

The annotation interface works as follows: given the directory where the data files are stored, it indexes all tokens to be annotated by their lemmas. This index is persisted to disk so that this indexing does not need to be re-run. The user is then prompted for a lemma and (optionally) a PoS tag; if any matching pairs of lemma and PoS tag are found in the index, a new buffer is opened, containing the glosses where the targeted lemma was found. Colours differentiate token's status: pre-annotated tokens are shown in one color, while tokens yet to be annotated are shown in another; tokens annotated in the current iteration are also shown in a different color. Multiword expressions are marked by subscripts in their constituent tokens (whether they are adjacent or not), while sense and PoS annotations are shown as superscripts. The annotation interface offers the user

[2] https://www.perl.org
[3] https://www.gnu.org/software/emacs/

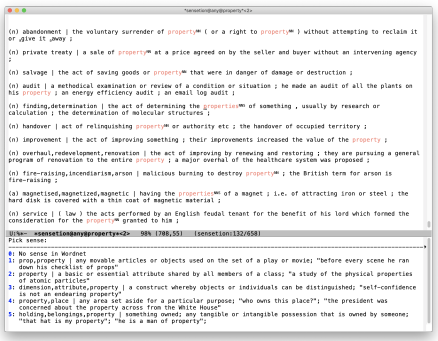


Figure 1: Sense tagging interface

the following capabilities:

1. assign a token zero or more senses;
2. change a token's lemma;
3. mark a token to be ignored (closed-class words or other fragments that are considered meta annotations on the glosses);
4. mark an annotation as having low confidence;
5. create and dismantle multiword expressions.

The first capability is the main functionality of the tool. When selected, the user is asked to confirm the token's PoS tag, and then a dialogue box is shown with all possible senses to that lemma and PoS tag pair, along with their defined terms and glosses. The user can then select or deselect a sense, or explicitly say that there is no sense for that word in WordNet (see Figure 1).

All other commands are there to allow the correct sense annotation of a word. In case its lemma is wrong, there is no way of presenting the user with the correct sense options unless the lemma is corrected; if a token is part of a multiword expression but is not already marked as so, annotators are able to mark it themselves. The dismantling of a multiword expression is necessary for the cases where the token is wrongly assigned as part of a multiword expression, as *rock* and *bass* in Example 2b, where both are marked as part of the multiword expression *rock_bass* (a kind of fish). All commands are available through customizable mnemonic keyboard shortcuts or by a menu.

3 Data Preparation

The Princeton WordNet Gloss Corpus is distributed in two different XML formats: standoff and merged files.[4] We choose to work with the merged files because they are more concise and are precisely described by a document type description (DTD). We have split this data into files containing 100 glosses each, with one annotated gloss per line encoded as an S-expression, a notation for tree-like data.

Every WordNet gloss contains a sense definition. The gloss can be preceded by a domain classification fragment and/or an auxiliary fragment (usually in parenthesis, but not always), and optionally followed by more auxiliary fragments and zero or more examples. In the original XML, all these components are marked up with nesting elements. The tokens are marked up with parts of speech, potential lemma forms, and (optionally) a small set of semantic classes (indicating whether the token is punctuation, abbreviation, acronym, number, year, currency, or some kind of symbol). Collocations are delimited by special markup which can even indicate discontiguous forms. Words and collocations that have been disambiguated are further annotated with WordNet sense keys. To facilitate the implementation of the interface, we have adopted a flat data format where a gloss is a list of tokens, each one of them represented by a property list (see Listing 1). All nesting elements for boundary-marking tokens in the XML files were converted to key-values pairs in the respective tokens. Further details are publicly provided in a README file along with the data itself.

The data conversion is followed by a validation step to ensure that our understanding of the data was right and that no information was lost. Although the XML validation using the DTD takes care of many validation issues, we did find encoding errors and nonexistent sense-keys in the corpus. For the encoding errors, before the conversion, we searched for and replaced invalid characters by UTF-8 legal codes.

Most of the cases of invalid sense keys turned out to be instances of adjective satellites whose sense keys had been wrongly marked with synset type 3 instead of 5, but some cases were tokens marked with

[4] http://wordnetcode.princeton.edu/glosstag.shtml

```
(:ofs "02744323" :pos "n"
 :keys
  (("arterial_road%1:06:00::" . "arterial_road"))
 :gloss "a_major_or_main_route"
 :tokens
  ((:kind :def :action :open)
   (:kind :wf :form "a" :lemma "a" :pos "DT"
    :tag "ignore")
   (:kind :wf :form "major" :pos "JJ" :tag "man"
    :lemma "major%1|major%2|major%3"
    :senses (("major%3:00:06::" . "major")))
   (:kind :wf :form "or" :lemma "or" :pos "CC"
    :tag "ignore")
   (:kind :wf :form "main" :tag "man"
    :lemma "main%1|main%3" :pos "JJ"
    :senses
     (("main%5:00:00:important:00" . "main")))
   (:kind :wf :form "route" :tag "man" :sep ""
    :lemma "route%1|route%2" :pos "NN"
    :senses (("route%1:06:00::" . "route")))
   (:kind :wf :form ";" :pos ":" :tag "ignore"
    :type "punc") (:kind :def :action :close)))
```

Listing 1: Property list encoding of WordNet 3.0 synset `02744323-n`

an undocumented and non-existent sense key `purposefully_ignored%0:00:00::`. The name suggests that this was a virtual sense, created as a way of manually marking tokens as to be ignored in the sense annotation. A case like the annotation of 'ng' in Example 1a[5] seems to support this view. However there is also evidence to believe that the non-existent sense key was created to mark cases where the appropriate sense for a word did not exist in WordNet, as in 'designating' in Example 1b. We also found four cases where a 'purposefully ignored' sense was assigned together with some other sense; these we have revised and corrected manually. These cases include the aforementioned Example 1a (where it had also been tagged as the unit *nanogram*), Example 1c (where *waves* also had been tagged as "(physics) a movement up and down or back and forth"), and Example 1d (where *Mediterranean* was also tagged as "of or relating to or characteristic of or located near the Mediterranean Sea").

(1) a. produced with the back of the tongue touching or near the soft palate (as 'k' in 'cat' and 'g' in 'gun' and 'ng' in 'sing') (01156750-a)

b. designating the player judged to be the most important to the sport; "the most-valuable player award" (01279431-a)

c. atomic events are explained as interactions between particle waves (06107850-n)

d. small dried seedless raisin grown in the Mediterranean region and California (07752966-n)

[5]The identifiers in the end of the examples stand for the synset IDs of PWN 3.0.

4 Challenges

The challenges of this project encompass many aspects: the amount of work, the particularities of the glosses compared to sentences in an ordinary text, and the mismatch between the 'continuous' sense boundaries of words in utterances and the 'discrete' boundaries defined by a lexicon.

The Princeton Wordnet Gloss Corpus contains 117,659 glosses composed by definitions and examples, comprising more than 1,621,129 tokens. So far, 449,355 tokens have been annotated, 118,856 of them automatically. Considering only the taggable tokens, i.e., the open-class words, 206,711 tokens are estimated to remain untagged. From these untagged tokens, we have so far annotated approximately 500 tokens during the development of our tool and training of the annotators. To deal with the amount of work in the next phases of this project, we plan to prioritize the annotation by focusing on domain-specific words most relevant to other projects of our team.

The sense of a word in a text is determined by its context – the more context information we use, the easier is the determination of the right sense of its polysemous words. Compared to other corpora, synset glosses provide relatively little context. Figure 2 summarizes the sizes of glosses (number of characters) by part-of-speech. As we can see, the majority of the glosses has less than 100 characters (76% of them). Moreover, most of the glosses are not complete sentences, e.g. 'secured with bastions or fortifications.' The annotator has to carefully consider the words that are being defined by the gloss and its relations to other synsets, in order to compensate for these obstacles. For some cases, such as 'allomorphs' in 'pertaining to allomorphs' and 'park' in 'The young man

was caught soliciting in the park' the unique viable solution is to allow multiple sense annotation, as described in Section 2.

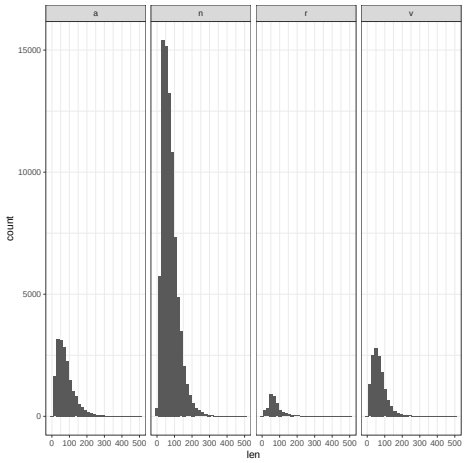


Figure 2: Glosses' sizes (number of characters) by part-of-speech

Regardless of the nature of the sentences, disambiguation of senses is a notoriously hard task that may not be disconnected from the constant revision of the lexical resource being used as a sense inventory. This line of thought is supported by the way previous work on building semantic concordances was conducted. (Miller et al., 1993), carried out sense annotation while expanding and refining PWN, with the annotations continually signalling omissions and inaccuracies in the resource. We have already noted some cases of inconsistencies in PWN such as the case of 'deposit' presented in Figure 3. The dashed red lines point to the two possible senses for the word 'deposit' (bold) in the synset 01576001-v. Note that although the synset 01528069-v seems to be the best option, as it is the direct hyperonym of 01576001-v. The synset 01575675-v is the best matching considering the example in 01576001-v and its hyponym 01988755-v. This situation suggests that 01576001-v should have a different position in the network.

More recently, (Kilgarriff, 1997) has already pointed out 'word senses are only ever relative to a set of interests' and (Rudnicka et al., 2019) emphasizes this point, remarking that dictionaries (or wordnets) and corpora are in two different levels: "Dictionaries and wordnets are metalinguistic

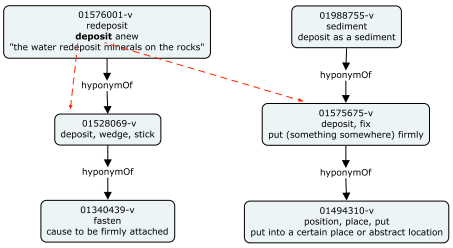


Figure 3: possible senses for an occurrence of the word 'deposit'

generalizations, while corpora are real texts; dictionaries and wordnets include decontextualized isolated items, corpora consist of contextualized continuous text."

Some cases of multiword expressions (MWE) seem to support our belief that sense annotation and PWN maintenance should be joint work. First, we need to define and enforce heuristics to determine when a given word sequence is a multiword expression (being sense annotated as a single entity), and when its component tokens should be annotated individually. The compositionality and conventionality criteria from (Farahmand et al., 2015) may help, however these criteria are not as clear-cut as we would like them to be. Take the case of 'first degree' and the example 'all of the terms in a linear equation are of the *first degree*' in its definition (synset 05861716-n); we can annotate it as 'first degree' (this same sense being defined in the synset where the example is given); but there is no sense for 'second degree', or 'third degree', which are equally valid. This leads us to consider that it should be annotated individually, and that the 'first degree' sense should be removed from PWN.

One can conclude that sense tagging the PWN glosses is a never-ending task, but we believe it is possible (and useful) to achieve definitional completeness in restrictive domains. The question that we face is how to make it feasible and synchronized with the changes in the lexicon (senses, words, and relations). Admittedly, we will need to implement tools for tracking the changes in the dictionary and signal for re-annotation all potentially affected glosses.

Finally, the challenges related to the corpus'

size and to MWEs also interact. To make the annotation process easier we would like to have a certain degree of automation. We have inherited from the original data expressions that have been incorrectly tagged as MWEs, as in the bold words in the sentences 2a and 2b. While it is easy to recognize and fix this kind of error, the other way around is more challenging: identifying an expression that should be added or that is already defined in the lexicon.

(2) a. bearing or producing or containing **calcium** or calcium **carbonate** or calcite (02674398-a)

b. English **rock** star and **bass** guitarist and songwriter who. . . (11167952-n)

5 Evaluation

We have carried out a preliminary annotation effort to test our interface, train our annotators, and refine our guidelines. In this section we report the issues we found and the results we obtained.

We have trained four annotators and instructed them to annotate all glosses in which one of these three words occurred: 'derivation' (9 occurrences in 8 glosses), 'formation' (153 occurrences in 146 glosses), and 'incompatible' (8 occurrences in 7 glosses). The word 'derivation' has eight senses available from the PWN, while 'formation' has seven senses and the adjective 'incompatible' has nine senses. These example words were chosen to balance frequency and polysemy degree.

After the annotation, two sessions of discussion were conducted to consolidate annotation decisions. We must note that because training the annotators was the goal of this experiment, the results presented here are still very preliminary.

Considering all three words, only half of the occurrences presented full agreement among annotators. But partial agreements are reasonably common as we can see in the tables 1 and 2. The 'ctx' column (for context) numbers the occurrences of a given word; when the number is a decimal it means that there is more than one occurrence of that word in the same context. The other columns number the possible sense an annotator could choose, including a label for the absence of a suitable sense in PWN ('N'). Table 1 presents the annotations of the 'derivation' occurrences. In five out of eight contexts, most of the annotators agreed on one sense, e.g. in the last context, all annotators agreed on sense 7 although annotator **T** also assigned sense 2. We can also note that annotators A and B agreed six times even though one of them also annotated an additional sense.

ctx	1	2	3	4	5	6	7	8	N
1		T				A	ABR		
2						ABRT	A		
3				T		BT	AR		
4			BR						AT
5.1				ABRT					
5.2				ABRT					
6		AT				R	BTR		
7	B	BRT	AB	B	B	B	B	B	
8		T					ABRT		

Table 1: Annotations of all 9 occurrences of 'derivation' in 8 glosses. A, B, R and T stands for the annotators' initials.

Particularly interesting is that all of the eight occurrences of 'incompatible' have almost always been annotated with the most generic sense "not compatible" (00508192-a). Nevertheless, annotators reported this to be the hardest among the three words to annotate. Even to reach a consensus on the proper sense afterwards was a hard task. Table 2 also shows that for annotators **A** and **T**, in all of the contexts that they examined, sense 3 and sense 5 are indistinguishable.

ctx	1	2	3	4	5	6	7	8	9	N
1	T	AT	ART		ABRT			B		
2		T	AT		ABRT			B		
3.1			AT		ABRT					
3.2			AT		ABRT					
4			AT		ABR					
5			T		ABRT					
6			AT		ABRT					
7			T		ABR			B		

Table 2: Annotations of all 8 occurrences of 'incompatible' in 7 glosses. A, B, R and T stand for the annotators' initials.

When we consider the word 'formation,' there are 82 occurrences with full agreement (Table 3 lines 1, 4, 7, 9, and 37). Line 1, for example, shows that there was full agreement regarding sense 3,"natural process that causes something to form", in 71 occurrences. This same sense was selected other 29 times with partial agreement. In 21 of these 29 cases, at least one annotator also chose the sense "the act of forming or establishing something". One such case was the gloss for 'electronegativity', which states "(chemistry) the tendency of an atom or radical to attract electrons in the *formation* of an ionic bond" (04944513-

n). However, although only one annotator had assigned 'formation' to the mentioned sense, after discussions among the annotators, others agreed that it could also be assigned to it in the given context.

qt	1	2	3	4	5	6	7	N
71			ABRT					
13			ABR		T			
13		B					ABRT	
6						ABRT		
5		T		ABR				
4						BRT		A
2					ABRT			
2					T	BRT		A
2				ABRT				
2			AB	R	T			
2			B		ABRT			
2		B		A			BRT	
2		BR					ABT	
2		BRT					A	
1						R		ABT
1					BRT		A	
1					T	ABRT		
1					T	AR		B
1					T	BR		A
1			AB		RT			
1			AB	RT				
1			B		ABRT	A		
1			B		T	ABRT		
1			BR		T		A	
1			BT	AR				
1		A	BRT					
1		ABR					BT	
1		ABR			T		B	
1		B					ART	
1		B	R	ABRT				
1		BRT	A					
1		BT					ABR	
1		BT					ABRT	
1		BT		ABR				
1		R	ABR				T	
1		T	B	ABR				
1	ABRT							
1	B					RT		A
1	T		ABR					

Table 3: Annotations of all occurrences of 'formation'. A, B, R and T stand for the annotators' initials. The first column is the number of contexts where the same pattern of annotation appears.

The case of 'formation' is in agreement with results from (Leacock et al., 1993) that say "the degree of difficulty involved in resolving individual senses is a greater performance factor than the degree of polysemy.". It also suggests a two-step sequential approach to annotation: first the annotators agree on each synset's scope and only then do they proceed to the actual annotation process. This two-step approach will be the object of a future investigation.

As for multiword expressions, expressions such as 'military formation', 'geological formation', 'reticular formation', and 'reaction formation' are removed from the above quantitative analysis but we have discussed them. The expression 'military formation' stands out in many glosses. The expression exists as a MWE but a similar expression, 'naval formation' does not, with both appearing in the gloss "the side of military or naval formation". Annotators discussed whether 'naval formation' should be considered a MWE or whether 'military formation' should not be considered one.

An annotator's familiarity with a particular domain also plays a role in the annotation process, affecting both the senses assigned and the decisions regarding which collocations should be considered MWEs. For instance, the expression 'rock formation' is not part of PWN, but it appears many times in the corpus (see Example 3a).

(3) a. a national park in Utah having colorful *rock formations* and desert plants and wildlife (08603525-n)

b. the gradual movement and formation of continents (11434448-n)

Although some of us believe the expression should be added to PWN, it is not in the lexicon yet, and so, three annotators chose the sense '(geology) the geological features of the earth' for the word 'formation' in all occurrences of the expression. This decision was understandable if we consider that the word 'rock', in one of its senses, naturally evokes the domain 'geology'. The same can also be said for the word 'continent' in Example 3b. But one annotator, a geology expert, consistently took the sense 'a particular spatial arrangement' for the word 'formation' in this expression. His decision was based on the strict interpretation of 'geological formation' as a domain-specific concept also described in https://en.wikipedia.org/wiki/Geological_formation and reinforced by the fact that 'geological formation' in PWN has 'physical object' as its hyperonym, not 'formation' (as a process).

Another issue identified in this small experiment was that of an annotator consistency. In the definition of 'male bonding' and 'female bonding,' the word 'formation' appears in a very similar way ('the formation of a close personal relationship between men/women'), but one of the annotators was not consistent in the annotation of its sense in the two glosses. Finally, Tables 1 and 3 show that some annotators have already identified missing senses in PWN (column N).

6 Previous Work

The recognition that PWN contains a substantial amount of knowledge within its glosses was made clear in (Clark et al., 2008a,b). These articles describe the work on some of the standoff files distributed in the PWN website,[6] including the 'logical forms' of the glosses. The authors also mention the use of the 'logical forms' produced years before by another team at the University of Texas.[7] In the Extended Wordnet (Harabagiu et al., 1999), the disambiguation of the glosses was done automatically over the PWN 2.0 glosses. Although their initial plan was to "develop a tool that takes as input the current or future versions of WordNet and automatically generates an extended WordNet that provides several important enhancements intended to remedy the present limitations of WordNet", the project does not seem to be maintained anymore. In (Clark et al., 2008b) the authors reported that the logical forms generated at that stage are not of high quality in general. Further use of the Extended Wordnet was reported in (Castillo et al., 2004).

The most relevant work to our present effort is the original Princeton WordNet Gloss Corpus project, our starting point.[8] Unfortunately, we are not aware of any publications resulting from the project except the README file distributed with the data and the annotators' guidelines. [9]

Here we emphasize the manual process, focus on the creation of a semantic concordance of PWN glosses and PWN itself in the same lines of (Miller et al., 1993). We are also following as close as possible the directives devised by the Princeton team when they started the original Princeton Wordnet Gloss Corpus. Similar to (Moldovan and Novischi, 2004), our primary goal is the development of better word-sense disambiguation methods and algorithms that can take advantage of the annotated glosses for better results on a domain-specific corpus, such as the one described in (Rademaker, 2018).

[6] https://wordnet.princeton.edu/download/standoff-files

[7] http://www.hlt.utdallas.edu/~xwn/

[8] http://wordnetcode.princeton.edu/glosstag.shtml

[9] The authors would like to thank Christiane Fellbaum for sharing the guidelines with us.

7 Conclusion

In this paper we describe our resuming of the Princeton WordNet Glosstag Corpus project.[10] We have assembled a team and created an annotation interface, and have begun our work. As put by (Miller et al., 1993), the semantic annotation of corpora helps improve both the coverage and the precision of the semantic resource being used in the annotation. This work is thus part of our effort in expanding and improving WordNet-like resources in an application-driven and domain-specific way, initially focusing on oil & gas domain applications.

Besides a continuous annotation effort, future work mostly involves improvements in the annotation interface[11] and in annotation methodologies. With respect to the annotation tool, we intend to start supporting the sequential annotation style discussed in Section 2, and to improve its performance. The methodological work involves developing processes for the revision of syntactic annotation (part-of-speech tags, lemmas, and MWE tagging) and for updating the corpus when the underlying WordNet changes. Additionally, we also intend to develop querying and visualization tools to support the annotation and the WordNet's expansion work.

[10] https://github.com/own-pt/glosstag

[11] https://github.com/own-pt/sensetion.el

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational Linguistics and Intelligent Text Processing*, pages 136–145, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pierpaolo Basile, Marco de Gemmis, Anna Lisa Gentile, Pasquale Lops, and Giovanni Semeraro. 2007. Uniba: Jigsaw algorithm for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 398–401, Prague, Czech Republic. Association for Computational Linguistics.

Francis Bond, Luís Morgado da Costa, and Tuan Anh Lê. 2015. Imi — a multilingual semantic annotation

environment. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 7–12, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Mauro Castillo, Francis Real, Jordi Asterias, and German Rigau. 2004. The talp systems for disambiguating wordnet glosses. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 93–96, Barcelona, Spain. Association for Computational Linguistics.

Peter Clark, Christiane Fellbaum, and Jerry Hobbs. 2008a. Using and extending wordnet to support question-answering. In *Proceedings of the 4th Global WordNet Conference (GWC'08)*.

Peter Clark, Christiane Fellbaum, Jerry R Hobbs, Phil Harrison, William R Murray, and John Thompson. 2008b. Augmenting wordnet for deep understanding of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 45–57. Association for Computational Linguistics.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan. 1999. Wordnet 2: a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX99: Standardizing Lexical Resources*, pages 1–8.

Adam Kilgarriff. 1997. "i don't believe in word senses". *Computers and the Humanities*, 31(2):91–113.

Claudia Leacock, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the workshop on Human Language Technology*, pages 260–265. Association for Computational Linguistics.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.

Dan Moldovan and Adrian Novischi. 2004. Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, 18(3):301–317.

Alexandre Rademaker. 2018. Challenges for information extraction in the oil and gas domain. In *Proceedings of the XI Seminar on Ontology Research in Brazil (ONTOBRAS)*, São Paulo, Brazil.

Ewa Rudnicka, Francis Bond, Łukasz Grabowski, Tadeusz Piotrowski, and Maciej Piasecki. 2019. Sense Equivalence in plWordNet to Princeton WordNet Mapping. *International Journal of Lexicography*.

GeoNames Wordnet (gnwn): extracting wordnets from GeoNames

Francis Bond
Nanyang Technological University
bond@ieee.org

Arthur Bond
United World College of Southeast Asia
artcbond@gmail.com

Abstract

This paper introduces a new multilingual lexicon of geographical place names. The names are based on (and linked to) the GeoNames collection. Each location is treated as a new synset, which is linked by `instance_hypernym` to a small set of supertypes. These supertypes are linked to the collaborative interlingual index, based on mappings from GeoDomainWordnet. If a location is already in the interlingual index, then it is also linked to the entry, using mappings from the Geo-Wordnet. Finally, if GeoNames places the location in a larger location, this is linked using the `mero_location` link. Wordnets can be built for any language in GeoNames, we give results for those wordnets in the Open Multilingual Wordnet. We discuss how it is mapped and the characteristics of the extracted wordnets.

1 Introduction

The aim of this paper is to create a large multilingual lexicon of place names, through the use of the vast open source database GeoNames.[1]

Wordnets generally contain open-class words, with only a few proper names. Some names need to be there, as they are derivationally related to open-class words (such as ***Vratislavian*** "a native of Wrocław"). However the general trend is to leave proper names out, and instead link them through other specialist lexicons (Vossen et al., 2016). The goal is for specialists on names to curate names, with wordnets only having to maintain a smaller collection of links.

Another popular approach is to merge completely, into a vast combined resources such as YAGO (Suchanek et al., 2007) or BabelNet (Navigli and Ponzetto, 2012). This can cause problems when a subsidiary resource updates: propagating the corrections into the merged resource is an unsolved problem. For example, the version of GeoNames used in BabelNet 4.0 is from April 2015, a four year difference.[2]

Apart from general merging, there are two main resources made from merging Wordnet with GeoNames. The first, Geo-Wordnet (Buscaldi and Rosso, 2008) links locations in Princeton Wordnet (PWN: Fellbaum, 1998) to GeoNames (we will call this GWN-link). The second, GeoWordnet (Giunchiglia et al., 2010) links the supertypes in GeoNames to PWN synsets (we will call this GWN-super). These are complimentary mappings, but as far as we know, no one has combined them. GeoWordNetDomains (Frontini et al., 2016) further refines the mappings from GeoWordnet and adds some more internal structure. Both GeoWordnet and GeoWordNetDomains link the synsets to English and Italian (the Multiwordnet (Pianta et al., 2002) and Italwordnet (Toral et al., 2010) respectively), but do not consider other languages. All the resources are described in more detail in Section 2.

In this paper, we introduce a method for creating lexicons of placenames for any language in GeoNames: the Geoname Wordnet. Each location is treated as a new synset, which is linked by `instance_hypernym`[3] to a small set of supertypes, linked to the collaborative interlingual index, based on mappings from GeoDomainWordnet. If a location is already in the interlingual index, then it is also linked to the entry, using mappings from the Geo-Wordnet. Finally, we add some additional structure, if GeoNames places the location in a larger location, this is linked using the `mero_location` link. This is described in Section 3. The code to create the lexicons is available at `https://github.com/fcbond/geonames-wordnet`.

We present some statistics of the resulting word-

[1] https://www.GeoNames.org/

[2] `https://babelnet.org/about` accessed on 2019-05-20.

[3] Links are linked to their definition by the Global Wordnet Association Working Group.

net, along with some examples, in Section 4, and finish with some conclusions and ideas for future work in Section 5.

2 Resources

We give descriptions of the major resources we use here. All of GeoNames' information is downloadable and can be found on their website.

2.1 GeoNames

GeoNames is a geographical database, under a Creative Commons license. It boasts over 25 million geographical names, which ultimately are categorised into one of nine categories, and then into one of 645 sub-categories. GeoNames' search engine allows you to search for the location and its accompanying information. Editing these locations are then open to the public, for anyone to correct any mistakes, or perhaps add a new location.

The Wroclaw Panorama for example, is an instance of Geonames' richness of information and features. We show the online result in Figure 2 and a subset of the information available in Figure 1.[4] It immediately comes up with a top 3 items list. The first item was the correct location. It details the location name, type of physical place, which in this case is categorised into the overarching theme of Spots, Buildings or Farms, (see Table 1), as well as the sub-category Monuments (S.MNMT). It is then classified into five potential administrative divisions: Panorama Racławicka is in Wroclaw City, which is in Wroclaw County, which is in Lower Silesia. Lower Silesia is in Poland, but the country is not part of an administrative division, it is simply a country, rendering the location Panorama Racławicka with just three administrative classes.

For the sake of this paper, information regarding coordinates, elevation, timezone, and modification dates of the data points, which GeoNames also offers, have not been used.

The alternative names shown in Figure 1 include a wide variety of languages; how many are featured for each entry has a great deal of variability. Many of the names (almost 40%) are not associated with a language. In the above instance Panorama Racławicka is in Polish, but GeoNames does not indicate that that is the case. In this case, an extra step needs to be done to deduce the language and it is not a trivial task. Names can also be marked with features:

[4]GeoName ids are linked to the GeoNames website.

ID	GN: `11839964`
name	Panorama Racławicka
feature	S.MNMT "Monument"
lat-lon	N 51°06′36″ E 17°02′40″
country	GN: `798544` PL "Poland"
adm1	GN: `3337492` "Lower Silesia"
adm2	GN: `7530801` "Wroclaw County"
adm3	GN: `7531292` "Wrocław"
alt	fr Panorama de Racławice ja パノラマ・ラツワヴィツカ en Racławice Panorama link …/wiki/Racławice_Panorama

Figure 1: GeoNames entry: Panorama Racławicka (links resolved and annotated with labels)

Class	Sub	Description
A	24	country, state, region, …
H	137	stream, lake, …
L	48	parks,area, …
P	18	city, village, …
R	22	road, railroad , …
S	253	spot, building, farm, …
T	98	mountain, hill, rock, …
U	62	undersea …
V	18	forest,heath, …

Table 1: Top Level Feature Classes

PreferredName	an official/preferred name
ShortName	a short name ***California*** for ***State of California***
Colloquial	a colloquial or slang term ***Big Apple*** for ***New York***
Historic	the was used in the past ***Bombay*** for ***Mumbai***

GeoNames also includes non-language data in these fields: external links, mainly to wikipedias and dbpedias; postcodes, airport codes and more. We currently do not use them, but they are a potential source for more translations.

The GeoNames database is built from official public sources, the quality of which may vary. Through a wiki interface, users are invited to manually edit and improve the database by adding or correcting names, move existing features, add new features, etc. Ahlers (2013) showed that there are many inaccuracies, especially in the granularity of coordinates (e.g., due to truncation and low-resolution geocoding in some cases), as well as

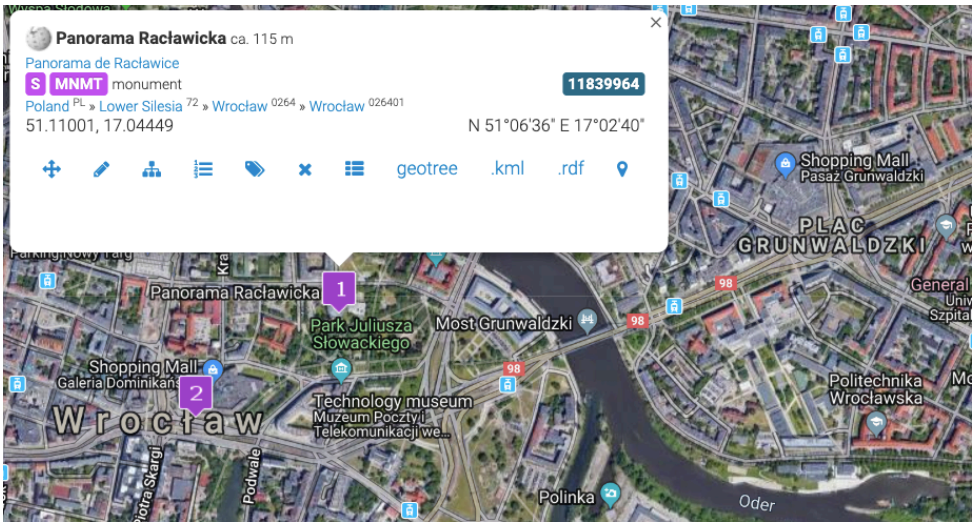


Figure 2: GeoNames entry for the Panorama Racławicka

wrong feature codes, near-identical places, and the placement of places outside their designated countries. However, he also pointed out that there was no other freely available resource with more accuracy.

2.2 Geo-WordNet (GWN-link)

This resource links locations in PWN 3.0 to GeoNames, for example PWN 08997487-n[5] ***Republic of Singapore*** is linked to GeoNames GN: 1880251. There are 1,964 entries so linked.

In the original paper (Buscaldi and Rosso, 2008), locations in PWN 2.0 were linked to Wikipedia to get coordinates. In Geo-WordNet 3.0,[6] the source of geographical data was GeoNames, and the mapping was to PWN 3.0.

This mapping is very useful, but does not extend the vocabulary of PWN, it merely adds more data (links to GeoNames, latitude and longitude).

2.3 GeoWordNet (GWN-super)

GeoWordNet takes a different approach and links the top level categories of GeoNames to wordnet synsets (Giunchiglia et al., 2010). The GeoNames entries are then treated as synsets. This gives an integration of WordNet, GeoNames and the Italian part of MultiWordNet (Pianta et al., 2002).

The GeoWordNet Public Dataset[7] is an impressive collection and contains 3,698,238 entities, 3,698,237 part-of relations between entities, 334 concepts, 182 relations between concepts, 3,698,238 relations between instances and concepts, and 13,562 (English and Italian) alternative entity names.

However, in the data made publicly available, there is no link to the original GeoNames data (so, for example, you cannot look up latitude and longitude). Further, locations that already exist in Multiwordnet are not linked, so for ***Republic of Poland*** a new node is created, and it will appear to be ambiguous: there will be two concepts, one from the wordnets and from GeoNames (although this ambiguity is spurious).

2.4 GeoDomainWordNet

GeoDomainWordNet aims to link the resources more loosely (Frontini et al., 2016). By treating GeoNames as linked open data they make sure that the full up-to-date version of GeoNames will be linked to. They took the GeoNames upper categories and linked them to synsets, either directly, or with a hyponym or meronym relation. For example ***section of lake*** is not a category in wordnet, but can be thought of as a meronym of ***lake***

[5]Wordnet synsets are linked to the Open Multilingual Wordnet.

[6]`http://timm.ujaen.es/recursos/geo-wordnet-3-0/`: note we could not find the data here, but got it from Bogdan Ivanov's NLTK wordnet extensions `https://github.com/bogdan-ivanov/wnext`

[7]Retrieved from here: `http://diversicon-kb.eu/dataset/geowordnet`

PWN 09452395-n. In this way, all categories are connected.

This approach has the same drawback as with GeoWordNet — if a location appears in both GeoNames and PWN (or Italwordnet: Toral et al., 2010), then it it will appear to be ambiguous.

3 GeoNames Wordnet (gnwn)

Our goal is the same as Geo(-)(Domain)Wordnet: to link the data in GeoNames to wordnets. We take advantage of the work they have done already to make what we believe is a better integration in the following ways.

- We make synsets for the feature codes and link them to the Collaborative Interlingual Index (CILI) using the mappings from GeoDomainWordNet, with additional mappings for newly added codes (§ 3.1)
  - the synset names encode the feature code names, so it is easy to retrieve them
- We have a script to build a wordnet for any language in GeoNames in the GWA LMF format (Vossen et al., 2016)
  - synsets are linked as instances of the feature codes
  - synset names encode the GeoName ids, so it is easy to retrieve them
  - synsets that are already in Princeton Wordnet, and thus in the CILI are linked (using the mapping from Geo-Wordnet)
  - GeoName admin codes are linked as location-meronyms — this is completely novel.

The code to and revised mappings used to create the lexicons is available at `https://github.com/fcbond/geonames-wordnet`. Entries that are not already in CILI will not be added — as GeoNames already curates and manages the GeoNames IDs, it is better to not duplicate effort. Instead we will encourage wordnet users to add places to GeoNames.

3.1 The Feature Codes

Figure 3 shows an example of a feature code mapping. There are 645 of these, with 18 newly added.

Figure 4 shows a new entry. In this case there is no corresponding entry in the ILI, so instead the synset is linked to another synset ***power station*** (gnwn-S.PS) which is linked to the ILI (`i57632`).

Synset	gnwn-S.MNMT
Definition	"a commemorative structure or statue"
ILI	`i82178`

Figure 3: Synset for Monument (S.MNMT)

Synset	gnwn-S.PSN
Definition	"nuclear power station"
`hypernym`	gnwn-S.PS

Figure 4: Synset for Nuclear Power Station (S.PSN)

In this way, all supertypes are linked to some existing entry in the wordnets.

3.2 Locations

In this section we give two examples of locations. Each location has a synset (Figures 5 and 6). The synsets each have an `instance_hypernym` and a note giving the GeoNames name (to make it easier to debug the wordnet). The Panorama Racławice synset also has a `mero_location` to the City of Wroclaw.

The Panorama Racławice synset is linked to translations in three languages. We show two of them here: English in Figure 7 and French in Figure 8. A single location can have multiple names (in Wordnets for different languages) or even multiple names in the same language. GeoNames marks preferred names: if a name is so marked, we take it as a vote of higher usage and add one to the frequency count of '1', so that it will be sorted first. Other information about names (short, colloquial, historic) could be encoded as meta information on the variant, this is left for future work.

4 Results

We show the sizes of the wordnets created for all the curated languages in the Open Multilingual Wordnet 2.0 (Bond and Foster, 2013), along with the lemma for the continent of Asia (GN: `6255147`) in Table 2.

Synset	gnwn-11822362
`instance_hypernym`	gnwn-S.MNMT
`mero_location`	gnwn-7531292
note	Panorama Racławice

Figure 5: Synset for Panorama Racławice

Synset	gnwn-798544
`instance_hypernym`	gnwn-A.PCLI
ili	`i83894`
note	Republic of Poland

Figure 6: Synset for Republic of Poland

Lemma	w1
lang	en
writtenForm	Racławice Panorama
partOfSpeech	n
Sense	gnwn-11839964-w1

Figure 7: English Lemma for Racławice Panorama

As can be seen, not all languages are equally well represented. Some languages include transliterations and this thus inflates the number of lemmas. For many languages, the average ambiguity is high.

There are 3,649,522 synsets, 3,129,147 lemmas and 4,587,108 senses, a substantial addition of knowledge.

The last column of Table 2 shows which place names are most common for each of the 40 languages (if there are fewer than 4 or more than one). Some of these are very common names: ***Stormyra*** is well known as the most common place name in Norway, **本町** *hon-machi* "this town" is a common placename in Japanese and *Kampung Baharu* and 新村 *xincun* "new village" are common names in Malay and Chinese. However some results are surprising: Some equivalent of "Washington County" is the most common placename for Estonian, Basque, Italian, Polish and Romanian! This is because many states in the US have a Washington County, and they have all been diligently translated. A more interesting query may have been: what is the most popular placename in a given country, rather than language.

Lemma	w1919
lang	fr
writtenForm	Panorama de Racławice
partOfSpeech	n
Sense	gnwn-11839964-w1919

Figure 8: French Lemma for Racławice Panorama

5 Conclusion and Future Work

We have created a large collection of lexicons of placenames: the Geoname Wordnet. Looking at 40 languages we had over 3.6 million locations with over 4.6 million senses. We can create lexicons for many more languages: all of those in GeoNames. We hope that this is one more step toward a completely open, linguistic knowledge base.

Each location is treated as a new synset, which is linked by `instance_hypernym` to a small set of supertypes based on GeoNames categories. These are linked to the collaborative interlingual index, based on an extended set of mappings from GeoDomainWordnet. If a location is already in the interlingual index, then it is also linked to the entry, using mappings from the Geo-Wordnet. Finally, we added some additional structure, if GeoNames places the location in a larger location, this is linked using the `mero_location` link. The data and code to produce GeoNames Wordnet are released under the MIT licence.

We have some ideas for future work:

- There are translations and definitions for the feature nodes for the languages (bg, nb, nn, no, ru, sv) in GeoNames, and for Italian in GeoDomainWordnet: we should add them.
- Almost half the names have language unknown, we could try to deduce the language perhaps by seeing which country it is in.
- Many of the names are transliterations: e.g. GN: `10630004` has both **庄内町** and it's latin equivalent ***Shōnai-machi*** while GN: `11209749` **小萩** *ohagi* has both hiragana and katakana (**おはぎ** and **オハギ**). The GWA LMF allows us to treat these as variants, but this requires language specific knowledge.
- We need to make sure all locations are merged across languages, we will propose an extension to CILI based on GeoNames IDs.
- We found some errors in the GeoNames database (spaces in fieldnames and so forth). We will fix these online.

Acknowledgements

We would like to thank Nathanael Kusanda, who helped with the coding.

Language	Code	Synsets	Lemmas	Senses	Asia	Most Common
Arabic	ar	232,575	197,679	252,316	آسيا	ال ظهرة
Bulgarian	bg	30,518	29,419	38,059	Азия	Чуката
Catalan	ca	13,857	13,270	14,292	Àsia	Irlanda
Danish	da	3,455	3,444	3,557	Asien	—
German	de	56,548	51,334	58,332	Asien	Neuhof
English	en	599,552	481,369	628,376	Asia	Union Township
Spanish	es	407,846	215,948	439,396	Asia	San Antonio
Estonian	et	4,220	3,914	4,334	Aasia	Washingtoni maakond
Basque	eu	8,860	7,444	9,169	Asia	Washington konderria
Persian	fa	272,151	377,549	492,500	—	Ḩoseynābād
Finnish	fi	48,628	30,641	49,420	Aasia	Isosaari
Irish	ga	3,111	2,893	3,169	an Áise	An Baile Nua
Galician	gl	1,954	2,075	2,125	—	A Rioxa, Guadalaxara
Hebrew	he	14,199	20,985	21,875	אסיה	לבנים יד בית
Hindi	hi	2,166	2,245	2,326	एशिया महाद्वीप	चर्च ऑफ गॉड वर्ल्ड, …
Croatian	hr	2,060	2,098	2,157	—	Sveti Martin, Nova Gora, …
Indonesian	id	322,293	217,548	325,413	Asia	Krajan
Icelandic	is	5,293	4,583	5,590	Asía	Tunga
Italian	it	31,631	32,607	40,492	Asia	Contea di Washington
Japanese	ja	103,881	145,047	184,080	アジア	本町
Lithuanian	lt	33,811	28,831	34,383	Azija	Girelė
Marathi	mr	2,210	2,167	2,271	—	डेटन
Malay	ms	36,993	30,797	37,259	—	Kampung Baharu
Burmese	my	712	728	746	—	—
Dutch	nl	17,316	17,108	17,795	Azië	Bergen
Nynorsk	nn	3,006	2,972	3,056	—	Balearane, London lufthamn, …
Norwegian	no	620,012	423,491	685,065	Asia	Stormyra
Polish	pl	21,456	19,578	21,659	Azja	Hrabstwo Washington
Portuguese	pt	64,161	50,208	65,752	Ásia	Sítio São José
Romanian	ro	7,692	6,703	7,988	—	Comitatul Washington
Sanskrit	sa	658	662	666	—	कुरुक्षेत्रम्, सेंट लूसिया, …
Slovenian	sl	1,640	1,644	1,705	—	Otok
Albanian	sq	3,749	5,613	5,990	—	Novosellë
Thai	th	240,365	168,682	256,304	เอเชีย	หนองบัว
Tswana	tn	7	9	9	—	—
Turkish	tr	40,001	31,978	43,143	Asya	Yeniköy
Venda	ve	11	10	11	—	Kuritiba
Xhosa	xh	32	34	34	—	—
Chinese	zh	740,984	495,549	826,003	亚洲	新村
Zulu	zu	273	291	291	—	—
Total	40	3,649,522	3,129,147	4,587,108	—	—

Table 2: GeoNames Wordnet stastics for various languages
We also show the lemma for Asia, and the most common name in GeoNames for each language

References

Dirk Ahlers. 2013. Assessment of the accuracy of GeoNames gazetteer data. In *Proceedings of the GIR Workshop*, page 74–81.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

Davide Buscaldi and Paolo Rosso. 2008. GeoWordNet: Automatic georeferencing of WordNet. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 1255–1258. Marrakech, Morocco.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Francesca Frontini, Riccardo Del Gratta, and Monica Monachini. 2016. Geodomainwordnet: Linking the geonames ontology to wordnet. In Zygmunt Vetulani, Hans Uszkoreit, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 229–242. Springer International Publishing, Cham.

Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi, and Biswanath Dutta. 2010. Geowordnet: A resource for geo-spatial applications. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *The Semantic Web: Research and Applications*, pages 121–136. Springer Berlin Heidelberg, Berlin, Heidelberg.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago - a core of semantic knowledge. In *16th international World Wide Web conference (WWW 2007)*.

Antonio Toral, Stefania Bracal, Monica Monachini, and Claudia Soria. 2010. Rejuvenating the Italian wordnet: upgrading, standardising, extending. In *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual global wordnet grid. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. 419–426.

New Polysemy Structures in Wordnets Induced by Vertical Polysemy

Ahti Lohk[1], Heili Orav[2], Kadri Vare[2], Francis Bond[3] and **Rasmus Vaik[1]**

[1]Department of Software Science, Tallinn University of Technology, Tallinn, Estonia
[2]Department of Computer Science, University of Tartu, Tartu, Estonia
[3]School of Humanities, Nanyang Technological University, Singapore

<{ahti.lohk, rasmus.vaik}@taltech.ee,
{heili.orav, kadri.vare}@ut.ee, bond@ieee.org>

Abstract

This paper aims to study auto-hyponymy and auto-troponymy relations (or vertical polysemy) in 11 wordnets uploaded into the new Open Multilingual Wordnet (OMW) webpage. We investigate how vertical polysemy forms polysemy structures (or sense clusters) in semantic hierarchies of the wordnets. Our main results and discoveries are new polysemy structures that have not previously been associated with vertical polysemy, along with some inconsistencies of semantic relations analysis in the studied wordnets, which should not be there.

In the case study, we turn attention to polysemy structures in the Estonian Wordnet (version 2.2.0), analyzing them and giving the lexicographers comments. In addition, we describe the detection algorithm of polysemy structures and an overview of the state of polysemy structures in 11 wordnets.

1 Introduction

The advantages of wordnet (Fellbaum, 1998) come from its specific design. On the one hand, it is a machine-readable dictionary, with definitions and examples of concepts, but on the other, a network of concepts in semantic relations (Fellbaum, 1998). This kind of resource makes it easy to figure out how close or far concepts are semantically from each other (semantic distance). Similarly, we can find the sub-concepts, super-concepts or synonyms of a given term. Wordnet offers a lexical-semantic background knowledge base for solving various NLP tasks, in particular for tasks that require semantic analysis.

However, one of the problems that can make wordnet usage difficult is the lexical polysemy in its semantic hierarchies (Freihat, Giunchiglia, & Dutta, 2013; Mihalcea, 2003). Furthermore, the problem is even more acute in the cases of polysemy where the context of two or more lemmas with the same spelling in a semantic network is barely distinguishable. The emergence of such a situation is facilitated by auto-hyponymy and auto-troponymy (Fellbaum, 2002), which fall within the definition of vertical polysemy (Koskela, 2011).

Auto-hyponymy and auto-troponymy in semantic hierarchies of wordnet have already been studied mainly as a criterion for grouping meanings of words (Pociello, Agirre, & Aldezabal, 2011; Pedersen, Agirrezabal, Nimb, Olsen, & Rørmann, 2018), but also for reducing polysemy (i.e. reducing the number of terms for their coarser distinction) (Mihalcea, 2003). In this paper, however, we are going to look at the possible substructures of semantic hierarchies that can only be formed by vertical polysemy.

We discovered that polysemy structures caused by vertical polysemy help us identify both the previously handled basic polysemy structures, such as chain and triangle (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002), but also those that have not previously been associated with vertical polysemy. By studying such polysemy structures, we also were led to cycles and structures containing up to 20 vertical polysemy cases, which we judge are likely to be errors.

The paper is structured as follows: Section 2 gives the theoretical background to understand the main body of the article. Next, Section 3 is dedicates to the overview of polysemy structures from the perspective of previous work. Section 4 describes the algorithm to detect specific polysemy structures from wordnet semantic hierarchies. Section 5 focuses on the case study of Estonian Wordnet. Section 6 gives an overview of the 11 wordnets uploaded to the OMW environment.

Section 7 concludes the paper and presents future work.

2 Theoretical background

This section aims to give some understanding of the theoretical basis of the discussed topic. Here we define the concept of polysemy and provide an overview of different polysemy structures.

2.1 Lexical ambiguity: polysemy and homonymy

We define polysemy to be a specific type of lexical ambiguity where a word or phrase has multiple semantically related meanings (Langemets, 2009). That is to say, they share the same etymology. Every polysemous word or phrase falls into one of three polysemy sub-categories: metonymy, specialization polysemy, or metaphors (Freihat, Giunchiglia, & Dutta, 2013).

In the case of metonymy, the polysemous word *chicken* can be as *a domestic fowl* or *food*.

A specialization polysemy example is the word *programming* where its narrow meaning is *coding* but in a broader meaning, it involves many actions like *inventing and analyzing the algorithm, coding and testing the code*.

In the case of metaphors, the polysemous word *parasite* can be *an animal or a plant* but also *a person.*

Beside the polysemy, another type of lexical ambiguity is homonymy. This concept differs from polysemy in that the meanings of a word or phrase are unrelated. In other words, they do not share the same etymology. For example, the homonymous word *bank* can be a *financial institution* or *edge of a river* (Jia-Fei, 2015).

Sometimes different authors refer to homonymy as contrastive polysemy and polysemy as complimentary polysemy (Weinreich, 1964) and (Freihat, Giunchiglia, & Dutta, 2013), with polysemy being used for both.

As stated, in the case of homonymy, the meanings of a word are unrelated. It implies that homonymous words do not form specific structures in wordnet hierarchies. For that reason, homonymous relationships remain out of our scope for further investigation.

2.2 Polysemy structures

Depending on how polysemy may form substructures in wordnet hierarchies, we divide polysemy structures into three categories (Figures 1-3):

1. Polysemous words in synonym sets have IS-A or MANNER-OF relationship (Figure 1).
2. Multiple inheritance cases in IS-A or MANNER-OF hierarchies. Cases where one synonymous set as child has at least two parents. By the Figure 2, it is important to emphasize that here a sub-term ("milk") has two meanings that come from its parents ("dairy product" and "beverage").
3. Polysemous words in IS-A or MANNER-OF hierarchies are not connected. That is to say, meanings of the words are related but not related in the hierarchical structure (Figure 3).

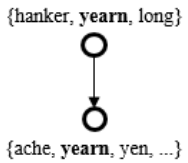


Figure 1: IS-A relationship between polysemous words. The example originates from PrWN 3.1[1]

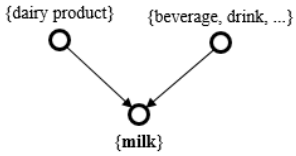


Figure 2: A multiple inheritance case. The example originates from PrWN 3.1

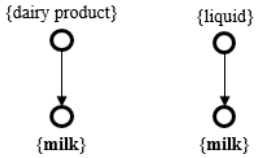


Figure 3: Not directly related polysemous words. The example has translated words and originates from EstWN 2.2.0[2]

Examples of the second (Figure 2) and third (Figure 3) categories are deliberately chosen to point to another important aspect – different wordnet developers can place different meanings of a word differently in the wordnet hierarchy. Furthermore, there are no clear guidelines on how to organize

[1] PrWN (Princeton WordNet)
http://wordnetweb.princeton.edu/perl/webwn

[2] EstWN (Estonian Wordnet)
https://teksaurus.keeleressursid.ee/

polysemous words in the wordnet hierarchy (Verdezoto & Vieu, 2011).

2.3 Vertical polysemy

In this paper, we study only polysemy structures caused by "IS-A" (in the case of noun hierarchy) or MANNER-OF (in the case on verb hierarchies) relationships between polysemous words.

A more appropriate term for describing such a case is **auto-hyponymy** (in noun hierarchies) or **auto-troponymy** (Fellbaum, 2002) and more generally speaking **vertical polysemy** (Koskela, 2011). Auto-hyponymy (also auto-troponymy) is a subset of the hyponymy (troponymy) relation where the superordinate and subordinate synonym sets contain same term (word) as shown in Figure 1. (Koskela, 2011) referring to (Horn, 1984) who says "**auto-hyponymy**, alludes to the fact that a **vertically polysemous** word is effectively its own hyponym." Thus notions of auto-hyponymy (also auto-troponymy) and vertical polysemy are very tightly related and are used here as synonyms. However, as referred by (Koskela, 2011) in the case of vertical polysemy, a polysemous word "with a broader and a narrower sense" can "occupy different levels in a taxonomic hierarchy". That is to say, that there may be not only a parent-child relationship between the polysemic words, but also the relationship of the grandparent-grandchild. However, in our work only parent-child relationship is considered.

3 Previous work: overview of polysemy structures

Previous work in relation to auto-hyponymy (also auto-troponymy) involves finding sense clusters of polysemous words (Peters, Peters, & Vossen, 1998) (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002) and reducing polysemy structures in wordnet semantic hierarchies to transform its term (word) senses from fine-grained to coarse-grained ones (Mihalcea, 2003).

The nearest work for our approach is (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002). The authors' broader goal was to create a bilingual network for Chinese and English, exploring the hierarchies of verbs, since there is approximately twice as much polysemy among the verbs as among the nouns. They aimed to find semantic patterns, hoping that these are helpful in multilingual information retrieval task. In their work, they distinguish five types of patterns calling them specifically the sense clusters of polysemous verbs or polysemic patterns. Very generally, these patterns take into account cases where two or more synonym sets contain the same term in their sets and are tightly related to each other in the semantic structure. The names of these patterns are *sisters*, *twins*, *child*, *chain* and *triangle*. Next, we describe each one of them shortly (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002).

3.1 Sense clusters of the polysemous words

Even though (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002) work focuses only on verb sense clusters we represent examples from both verb and noun hierarchies. That is to say, all of these specific examples are about IS-A/MANNER-OF relations and have been extracted from Princeton WordNet [3] noun and verb hierarchies, first published by (Lohk, 2015), but are still present there.

- **Sisters** are co-hyponym synsets having only one common term (Figure 4). Based on verb analysis of (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002), sisters is the most frequent pattern among the other ones.

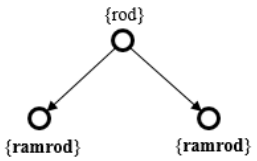


Figure 4: *Polysemic pattern – sisters*

- **Twins** are co-hyponym synsets having at least two common words (Figure 5).

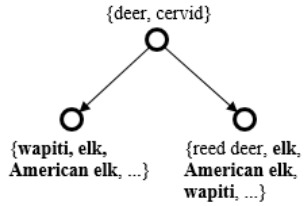


Figure 5: *Polysemic pattern – twins*

- **Child** is the polysemic pattern where the same term exists in a synset and its superordinate (Figure 6).

[3] http://wordnetweb.princeton.edu/perl/webwn

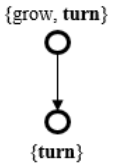


Figure 6: *Polysemic pattern – child*

- **Chain** is a polysemic pattern where the same term appears sequentially in IS-A/MANNER-OF chain three or more times (Figure 7). Based on (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002) analysis, that pattern appeared the least number of times.

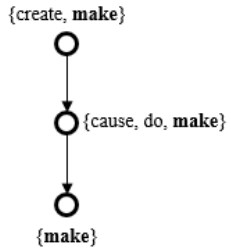


Figure 7: *Polysemic pattern – chain*

- **Triangle** is a polysemic pattern where the same term appears simultaneously in three synsets: in two co-hyponym synsets and their superordinate (Figure 8). Based on (Jen-Yi, Yang, Tseng, & Chu-Ren, 2002) this pattern is the second rarest one.

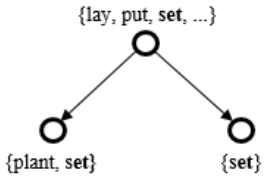


Figure 8: Polysemic pattern – triangle

Auto-troponymy relations characterize the last three patterns. That is to say, that verb term (or word) has both a more general and more specific meaning being simultaneously in troponymy (or MANNER-OF) relation(s). Here we state that these are polysemy structures caused by vertical polysemy.

4 Description of the algorithm

In our study, we find both the structures of polysemy and the statistics that describe these structures in one or another aspect. These statisticians are general enough not to pay attention to them in terms of the algorithm description. Thus, here we describe the only algorithm that finds all polysemy structures caused by vertical polysemy.

4.1 Algorithm

To get better intuition we describe that algorithm roughly through three steps:

Step 1: Separate from wordnet semantical hierarchies (IS-A and MANNER-OF) all pairs of senses (homographic pairs) with their synset id-s sharing the same lemma.

Step 2: For all pairs find equivalent classes. That is to say, find which pairs form connected components.

Step 3: Draw a graph for each connected component.

4.2 An example

To illustrate that we utilize data from Princeton WordNet. In the following example, we have separated sense pairs about "think" in **Step 1**. Even though Princeton WordNet has 13 "think" verb and one noun senses only six of them form pairs.

think-v (eng-630153) | *think-v (eng-631400)*
think-v (eng-630153) | *think-v (eng-741087)*
think-v (eng-630153) | *think-v (eng-741345)*
think-v (eng-691086) | *think-v (eng-691551)*

In **Step 2**, we find all connected pairs or connected components. As a result, two separate classes come up that are used later in separate polysemy structures.

think-v (eng-630153) | *think-v (eng-631400)* - ***1***
think-v (eng-630153) | *think-v (eng-741087)* - ***1***
think-v (eng-630153) | *think-v (eng-741345)* - ***1***
think-v (eng-691086) | *think-v (eng-691551)* - ***2***

In **Step 3**, we draw for every connected component a graph as a picture shown in Figure 9. We call these two graphs polysemy structures caused by vertical polysemy.

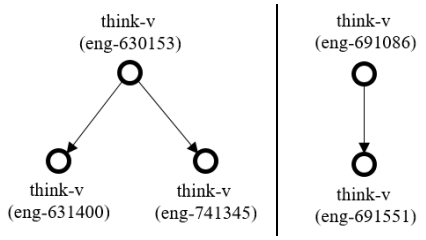


Figure 9: Two polysemy structures. Example from Princeton WordNet.

5 Case study of Estonian Wordnet

In this section, we describe one word and its senses from the Estonian Wordnet as an example of a vertical polysemy structure where it was possible to revise senses and reduce too fine-grained separation of senses.

5.1 Previous developments

In recent years, Estonian Wordnet has been mainly developed as a resource for NLP tasks. While increasing the wordnet's size also the problem of too fine-grained sense distinctions is taken into account. Different methods have been developed to reduce fine-grained senses, for example, the feedback from computer game Alias (Aller, Orav, Vare, & Zupping, 2016) feedback from NLP tasks (Kahusk & Vider, 2002) using the set of test patterns to validate wordnet's hierarchies (Lohk, Norta, Orav, & Võhandu, 2014) (Lohk, 2015) and various results from tasks given to students.

In Estonian, for example, every 10th word carries polysemous meanings. In addition, frequent words have the tendency of being highly polysemous, for example, *aasta* 'year', *asi* 'thing', *jooksma* 'run' etc. Discriminating between word senses is a problem in lexicography and it is considered as one of the hardest tasks. In wordnet, these related polysemous words should be connected via semantic relations.

5.2 An analysis example of a polysemy structure

At this point, we can say that there are 227 cases of polysemy structures in EstWN.

Here we look into one structure as an illustrative example. The word *galerii* 'gallery' in EstWN 2.2.0 has 8 different meaning. Five of them belong to the same hierarchy (Figure 10) and therefore needed attention. As follows, all senses of *gallery* are represented and explained how we tried to modify this hierarchy.

In Estonian 4 senses of the word *gallery* were changed:

- **gallery 3** - *narrow open passage on the top floor of the gallery house (on the upper floors), which is connected with an external staircase and is equipped with a balustrade*. Its synonymous with another synset *arch, arcade*, and it was possible to delete gallery 3.
- **gallery 6** - *a large, autonomous (connection) space in a building, one side of which is designed as arcade or row of windows*. We changed the hypernym from *gallery* to space, room.
- **gallery 7** - *gallery on the sidewalls in the church of Byzantine, Roman and Gothic, which opens by arcade towards the midnight robbery and forms a second high-wall balcony on the arcade*. Here we found to be reasonable delete word *gallery* from synset because others dictionaries do not show this meaning for gallery.
- **gallery 8** - *(reception)room connecting the rooms in the castles*. The hypernym was changed to *room*.

Other senses were left unchanged:

- **gallery 2** - *long, narrow room or covered gear*.

Senses which are not covered by the polysemy structure but are present in Estonian Wordnet:

- **gallery 1** - *building for the art collection, especially for paintings.*
- **gallery 4** - *pillars in the park (shaped as gallery).*
- **gallery 5** - *balcony under the ceiling in the theatre halls.*

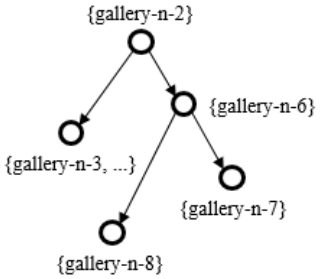


Figure 10: Example of polysemy structure in EstWN

Manual analysis of polysemy structures shows to us that generally the differences in nuance are the cause of auto-hyponomous polysemy structures in wordnets. For example:

- different function (gallery 5 used for receptions and gallery 8 used for art)
- different era (i.e gallery 4)
- different place (gallery 7 as part of theatre)
- different domain (gallery as architectural or landscape gardening or gallery in sports)

Some cases of 'gallery' can be specified, if we use another semantic relation, i.e. the domain-relation. In this case, the 'gallery1' could be associated with art and 'gallery4' with garden design etc.

These test patterns indicate possible inconsistencies, where vertical polysemy causes unjustified fine-grained senses or is otherwise problematic.

6 Polysemy structures in wordnets

In this section, we strive to capture a broader picture of the state of the wordnets in terms of vertical polysemy affected by polysemic structures. For that reason, we highlight the most specific structures caused by vertical polysemy. In addition, we provide tables for wordnets that characterize the structures of polysemy and describe the specificities that arise.
Wordnets we are using here are shown in Table 1.

6.1 Overview of the specific structures

As mentioned before, when we were making preparations to identify polysemy structures from semantic hierarchies, we expected to find structures like the **chain** and the **triangle** or their combinations. However, the results showed something different. These are structures that are not new in nature but have not previously been reflected in the context of vertical polysemy. In this light, we represent these new ones as contribution to the polysemic structures shown in Figures 7 and 8, in particular with the structures shown in Figures 11 and 12.

Next five figures originate from four different wordnets. Every node label contains here only the term common in all nodes of its substructure and synonym set id.

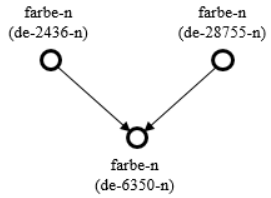


Figure 11: Multiple inheritance case caused by vertical polysemy. Example from Odenet. (*Farbe* in English is 'color')

The most basic structure here is the polysemy structure with multiple inheritance case (Figure 11). Next one (Figure 12) is known as a shortcut. Here, it seems that to multiple inheritance structure one additional link is added. Next three (Figure 13, 14, 15) are shortcut structures with an additional connection that cause the cycle. In Figure 14 purely two shortcut structures are together with an additional link that again causes the cycle.

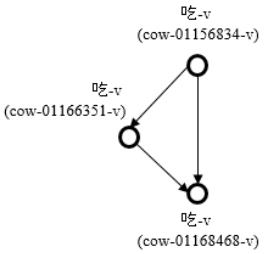


Figure 12: Shortcut structure caused by vertical polysemy. Example from Chinese Open WordNet. (*吃* in English is "eat")

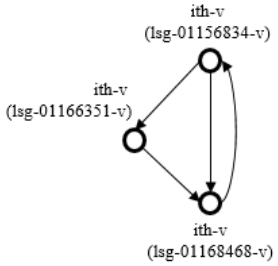


Figure 13: Shortcut structure with cycle. Example from Gaelic Wordnet. (*ith* in English is "eat")

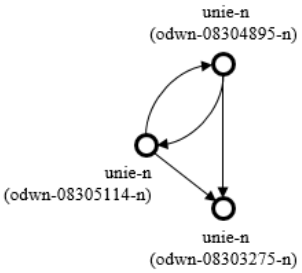


Figure 14: Shortcut structure with cycle. Example from Open Dutch Wordnet. (*unie* is "union")

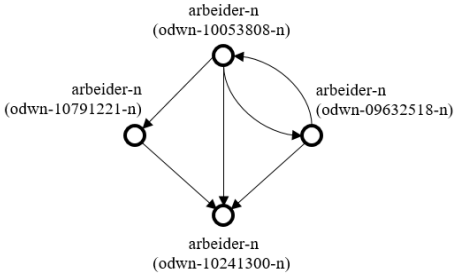


Figure 15: Two shortcut structures with a cycle. Example from Open Dutch Wordnet. (*arbeider* is "worker")

6.2 Statistics describing polysemy structures

The statistical indicators give us a better understanding of the polysemy structures in wordnets.

As mentioned before, cycles are a by-product of our results, which should be the primary goal of developers to eliminate. For this reason, we will not reflect them separately in the following tables. All rows in Tables 1 and 2 are ordered by alphabetically considering names of the wordnets. In particular columns, we represent three of the most extreme values in bold font.

To get a better comparison base, we first give the number of hyponymy relations for each wordnet (Table 1). This will make it clear which wordnets are richer in terms of vertical polysemy. Figure 16 shows the wordnets with the six highest proportion of hyponymy relations associated with vertical polysemy. Table 1 and Figure 16 show that although LSG has a relatively small number of hyponymy relations, this wordnet also has a relatively high number of vertical polysemy relationships. Compared to the three and four columns, we can see which dictionaries have the most varied values in both columns. The more significant difference between these numbers refers to the fact that the vocabulary precedes more pairs of synonyms with more than one word with the same orthography. Here the biggest difference is between NTU-JPN numbers, after that LSG and in third position ODWN.

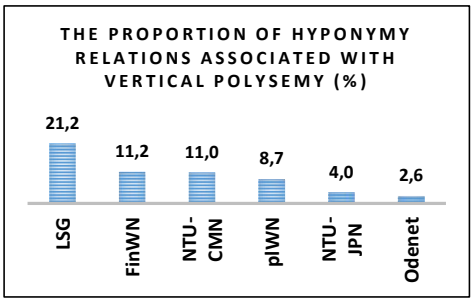


Figure 16: Proportions of hyponymy relations associated with vertical polysemy

Based on Table 1, two wordnet pairs show quite clearly how polysemic relations may vary in different languages. The first pair to consider is NTU-CMN, and the other is NTU-JPN. Both wordnets have exactly the same number of hyponymy relations, but at least in terms of vertical polysemy, NTU-CMN is represented by a much higher number. Their similarity is that they have been developed in parallel. Another pair of similar comparisons is FinWN and plWN-eng. FinWN has been compiled by translating PWN with the help of professional translators, however, FinWN is more diverse in terms of vertical polysemy.

Wordnet	Language	Nr of hyponymy relations	Nr of hyponymy rel.s related to VP[4]	Nr of vertical polysemy relations	Nr of polysemy structures	Unique synsets
Odenet	German	1 594	42	52	49	82
EstWN	Estonian	80 244	254	265	227	453
FinWN	Finnish	91 879	**10 281**	**11 529**	**7 478**	**15 664**
LSG	Irish	19 117	**4 062**	6 424	4 752	**6 094**
NTU-CMN	Chinese	89 376	**9 806**	**13 112**	**9 314**	**15 001**

[4] Vertical polysemy

Wordnet	Language	Nr of hyponymy relations	Nr of hyponymy rel.s related to VP[4]	Nr of vertical polysemy relations	Nr of polysemy structures	Unique synsets
NTU-JPN	Japanese	89 376	3 544	**8 463**	**6 704**	5 676
ODWN	Ducth	**102 789**	1 815	2 510	2 176	3 109
OWN-PT	Portuguese	8 577	4	4	4	8
plWN	Polish	**201 706**	1 743	1 854	1 696	3 252
plWN-eng	English	**97 597**	352	382	357	653
TrWN	Turkish	4 687	9	10	10	18

Table 1: Statistical indicators related to vertical polysemy and polysemy structures

In addition to Figure 16, we can also confirm the proportion of vertical polysemy in the second column of Table 2, which shows how many meanings a word may have among vertical polysemy relations.

Wordnet	Max nr of synsets for a term in vertical polysemy	Nr of **multiple inheritance** cases with vertical polysemy	Nr of **shortcuts** with vertical polysemy	Max nr of relations in a polysemy structure	Nr of nodes in the longest chain if any (nr > 2)
Odenet	4	2	0	3	–
EstWN	9	0	0	4	3
FinWN	**33**	4	**1**	**20**	**4**
LSG	**18**	6	**2**	**17**	**4**
NTU-CMN	**21**	**7**	**1**	**16**	**4**
NTU-JPN	16	3	0	7	**5**
ODWN	7	**31**	**16**	6	3
OWN-PT	2	0	0	1	–
plWN	11	**96**	0	7	3
plWN-eng	7	0	0	6	3
TrWN	2	0	0	1	–

Table 2: Statistical indicators related polysemy structures, multiple inheritance and shortcut structures

7 Conclusion

The study of polysemy in wordnet semantic hierarchies is essential because it is one of the central problems that needs to be considered in the case of distinctive NLP tasks that require semantic analysis. For this reason, we aimed to capture what polysemic structures occur in the wordnets uploaded to OMW.

In more detail, we studied the Estonian Wordnet, where polysemy has been kept under the spotlight for years. That is also the reason why its results were not as extreme as any other wordnets. However, we did find a couple of examples here, yet only one that needed correction. Thus in Estonian Wordnet polysemy structures of auto-hyponyms do not represent major problems of fine-grained senses, only few cases are present. Many of the structures of polysemy are caused by the economy principles of languages, i.e. general meaning is transferred to a more specific meaning or to a domain terminology. Some auto-hyponymy cases can be solved, if we introduced new semantic relations to Estonian Wordnet, for example the domain-relation.

By studying eleven wordnets, we discovered some unexpected polysemic structures. These are structures that by their nature are not new, but are not previously presented as closed chunks caused by vertical polysemy (as substructures in the semantic hierarchy). These polysemic structures are:

- Multiple inheritance
- Shortcut structure
- The longest chain (path)

These structures are unexpectedly frequent in many of the wordnets. The longest chain found

had five vertices, longer than the four that have been previously discussed.

The study of vertical polysemy relations in the 11 different wordnet networks reveals the impact of the individual choices, as it is a choice of the lexicographer, how to organize the polysemic senses in the wordnet hierarchy (e.g. as a configuration of a sister structure or as two children instead). This is the main reason behind the size of polysemic clusters in particular hierarchies.

The code to discover and visualize these structures will be incorporated into the Open Multilingual Wordnet, which can be accessed online or run on your own machine.

Acknowledgements

This work was supported by the Estonian Research Council grant EKTB4 and PSG277.

Reference

Aller, S., Orav, H., Vare, K., & Zupping, S. (2016). Playing Alias - efficiency for wordnet(s). *Global WordNet Conference, Bucharest, 27-30 January 2016* (pp. 16-21). Bucharest: Editura Universităţii "Alexandru Ioan Cuza" din Iaşi.

Fellbaum, D. C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge: MIT Press.

Fellbaum, D. C. (2002). On the Semantics of Troponymy: R. Green, C. A. Bean, & S. H. Myaeng, *The Semantics of Relationships: An Interdisciplinary Perspective* (pp. 23-34). Springer.

Freihat, A. A., Giunchiglia, F., & Dutta, B. (2013). Approaching Regular Polysemy in WordNet. *eKNOW 2013, The Fifth International Conference on* (pp. 63-69). Nice, France: IARIA XPS Press.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference. book: D. Schiffrin, *Meaning, form, and use in context* (pp. 11-42). Washington: Georgetown.

Jen-Yi, L., Yang, C.-H., Tseng, S.-C., & Chu-Ren, H. (2002). The Structure of Polysemy: A study of multi-sense words based on WordNet. *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation* (pp. 320-329). Jeju, Korea: The Korean Society for Language and Information.

Jia-Fei, H. (2015). Previous Researches on Lexical Ambiguity and Polysemy: *Verb Sense Discovery in Mandarin Chinese—A Corpus based Knowledge-Intensive Approach* (pp. 9-21). Springer.

Kahusk, N., & Vider, K. (2002). Estonian WordNet Benefits from Word Sense. *Proceedings of the 1t Global WordNet Conference*, (pp. 26-31). Mysore, India.

Koskela, A. (2011). Metonymy, category broadening and narrowing, and vertical polysemy. rmt: R. Benczes, A. Barcelona, & F. de Mendoza Ibáñez, *Defining Metonymy in Cognitive Linguistics: Towards a consensus view* (pp. 125-146). Amsterdan: John Benjamins Publishing Co.

Langemets, M. (2009). *Systematic Polysemy of Nouns in Estonian and Its Lexicographic Treatment in Estonian Language Resources: In Estonian.* Tallinn: Tallinn University.

Lohk, A. (2015). *A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries.* Tallinn, Estonia: TalTech Press.

Lohk, A., Norta, A., Orav, H., & Võhandu, L. (2014). New Test Patterns to Check the Hierarchical Structure of Wordnets. *Information and Software Technologies : 20th International Conference, ICIST 2014* (pp. 110–120). Druskininkai: Springer.

Mihalcea, R. (2003). Turning WordNet into an Information Retrieval Resource: Systematic Polysemy and Conversion to Hierarchical Codes. *International Journal of Pattern Recognition and Artificial Intelligence, Vol. 17, NO. 05*, 689–704.

Pedersen, B., Agirrezabal, M., Nimb, S., Olsen, S., & Rørmann, I. (2018). Towards a principled approach to senseclustering –a case study ofwordnet and dictionary sensesin Danish. *The 9th Global WordNet Conference* (pp. 183-190). Singapore: Global WordNet Accosiation.

Peters, W., Peters, I., & Vossen, P. (1998). Automatic Sense Clustering in EuroWordNet. *Proceedings of the 1st International Conference on* (pp. 409–416). Granada, Spain: European Language Resources.

Pociello, E., Agirre, E., & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. *Lang Resources & Evaluation*, pp. 121-142.

Snow, R., Prakash, S., Jurafsky, D., & Ng, A. (2007). Learning to Merge Word Senses. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1005-1014). Prague: Association for Computational Linguistics.

Verdezoto, N., & Vieu, L. (2011). Towards Semi-automatic Methods for Improving WordNet. *Proceedings of the Ninth International Conference on Computational Semantics* (pp. 275-284). Oxford, United Kingdom: Association for Computational Linguistics (ACL).

Weinreich, U. (1964). Webster's Third: A Critique of its Semantics American Linguistics, 4, Vol. 30. *International Journal of American Linguistics*, pp. 405-409.

Wordnets Used in This Paper

Open German Wordnet (Odenet). https://ikum.medien-campus.h-da.de/projekt/open-de-wordnet-initiative.

Estonian Wordnet. Orav, Heili; Vare, Kadri; Zupping, Sirli (2018). Estonian Wordnet: Current State and Future Prospects. Proceedings of the 9th Global WordNet Conference Singapore, January 8-12, 2018. Global Wordnet Association: Global Wordnet Association.

FinnWordNet. Lindén K., Carlson. L., (2010). FinnWordNet WordNet påfinska via översättning,LexicoNordica — Nordic Journal of Lexicography, 17 pp 119–140.

Chinese Open Wordnet. Shan Wang and Francis Bond (2013). Building the Chinese Open Wordnet (COW): Starting from Core Synsets. In Proceedings of the 11th Workshop on Asian Language Resources, a Workshop of The 6th International Joint Conference on Natural Language.

Japanese Open Wordnet. Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama and Kyoko Kanzaki (2008). Development of Japanese WordNet. In LREC-2008, Marrakech.

Open Dutch Wordent. Marten Postma and Emiel van Miltenburg and Roxane Segers and Anneleen Schoen and Piek Vossen (2016). Open Dutch WordNet. In Proceedings of the Global WordNet Conference 2016

Brazilian Wordnet. Valeria de Paiva and Alexandre Rademaker (2012). Revisiting a Brazilian wordnet. In Proceedings of Global Wordnet Conference, Matsue.

Polish WordNet. Maciej Piasecki, Stanisław Szpakowicz and Bartosz Broda. (2009). A Wordnet from the Ground Up. Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej, Poland.

enWordnet. Rudnicka, E., Witkowski, W., Kaliński M. (2015). Towards the Extension of Princeton WordNet. Cognitive Studies 15, 335-351.

Turkish Wordnet. R. Ehsani, E. Solak, O. T. Yildiz , Constructing a WordNet for Turkish Using Manual and Automatic Annotation, ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 17, No. 3, Article 24, 2018.

Utilizing Wordnets for Cognate Detection among Indian Languages

Diptesh Kanojia[†,♣,⋆], **Kevin Patel**[†], **Pushpak Bhattacharyya**[†], **Malhar Kulkarni**[†], **Reza Haffari**[⋆]
[†]Indian Institute of Technology Bombay, India
[♣]IITB-Monash Research Academy, India
[⋆]Monash University, Australia
[†]{diptesh, kevin, pb, malhar}@iitb.ac.in
[⋆]reza.haffari@monash.edu

Abstract

Automatic Cognate Detection (ACD) is a challenging task which has been utilized to help NLP applications like Machine Translation, Information Retrieval and Computational Phylogenetics. Unidentified cognate pairs can pose a challenge to these applications and result in a degradation of performance. In this paper, we detect cognate word pairs among ten Indian languages with Hindi and use deep learning methodologies to predict whether a word pair is cognate or not. We identify IndoWordnet as a potential resource to detect cognate word pairs based on orthographic similarity-based methods and train neural network models using the data obtained from it. We identify parallel corpora as another potential resource and perform the same experiments for them.

We also validate the contribution of Wordnets through further experimentation and report improved performance of up to 26%. We discuss the nuances of cognate detection among closely related Indian languages and release the lists of detected cognates as a dataset. We also observe the behaviour of, to an extent, unrelated Indian language pairs and release the lists of detected cognates among them as well.

1 Introduction

Cognates are words that have a common etymological origin (Crystal, 2008). They account for a considerable amount of unique words in many lexical domains, notably technical texts. The orthographic similarity of cognates can be exploited in different tasks involving recognition of translational equivalence between words, such as machine translation and bilingual terminology compilation. For *e.g.,* the German - English cognates, *Blume - bloom* can be identified as cognates with orthographic similarity methods. Detection of cognates helps various NLP applications like IR (Pranav, 2018). Rama et al. (2018) study various cognate detection techniques and provide substantial proof that automatic cognate detection can help infer phylogenetic trees. In many NLP tasks, the orthographic similarity of cognates can compensate for the insufficiency of other kinds of evidence about the translational equivalency of words (Mulloni and Pekar, 2006). The detection of cognates in compiling bilingual dictionaries has proven to be helpful in Machine Translation (MT), and Information Retrieval (IR) tasks (Meng et al., 2001). Orthographic similarity-based methods have relied on the lexical similarity of word pairs and have been used extensively to detect cognates (Ciobanu and Dinu, 2014; Mulloni, 2007; Inkpen et al., 2005). These methods, generally, calculate the similarity score between two words and use the result to build training data for further classification. Cognate detection can also be performed using phonetic features and researchers have previously used consonant class matching (CCM) (Turchin et al., 2010), sound class-based alignment (SCA) (List, 2010) *etc.* to detect cognates in multilingual wordlists. The identification of cognates, here, is based on the comparison of words sound correspondences. Semantic similarity methods have also been deployed to detect cognates among word pairs (Kondrak, 2001). The measure of semantic similarity uses the context around both word pairs and helps in the identification of a cognate word pair by looking of similarity among the collected contexts.

For our work, we can primarily divide words into four main categories *viz.* **True Cognates, False Cognates, False Friends and Non-Cognates**. In Figure 1, we present this clas-

sification with examples from various languages along with their meanings for better understanding. While some false friends are also false cognates, most of them are genuine cognates. *Our primary goal is to be able to identify True Cognates.* Sanskrit (Sa) is known to be the mother of most of the Indian languages. Hindi (Hi), Bengali (Bn), Punjabi (Pa), Marathi (Mr), Gujarati (Gu), Malayalam (Ml), Tamil (Ta) and Telugu (Te) are known to borrow many words from it. Thus, one may observe that *words which belong to the same concept in these languages, if orthographically similar, are True Cognates.* Currently, we include loan words in the dataset used for our work and include them as cognates. Since, eventually we aim to apply our work to Machine Translation and other NLP applications, we believe that this would help establish a better correlation among source-target language pairs. Also, we do not detect false friends and hence restrict the scope of True cognate detection using this hypothesis to Figure 2.

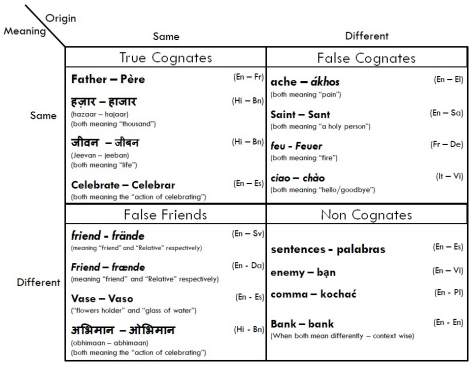


Figure 1: The Cognate Identification Matrix

We utilize the synset information from linked Wordnets to identify words within the same concept and deploy orthographic similarity-based methods to compute similarity scores between them. This helps us identify words with a high similarity score. In case of most of the Indian languages, a sizeable contribution of words/-concepts is loaned from the Sanskrit language. In linked IndoWordnet, each concept is aligned to the other based on an 'id' which can be reliably used as a measure to say that the etymological origin is the same, for both the concepts. Hence, words with the same orthographic similarity can be said to be **'True Cognates'**. Using this methodology, we detect highly similar words and use them as

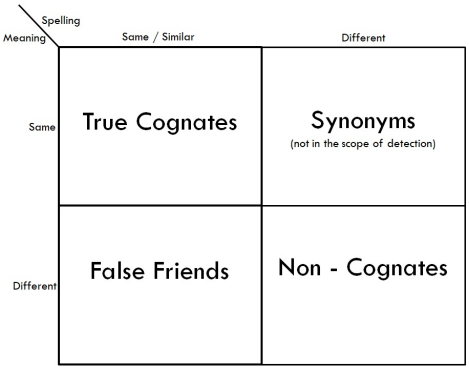


Figure 2: Scope of our work; Detection of True Cognates and False Friends

training data to build models which can predict whether a word pair is cognate or not. The rest of the paper is organized as follows. In Section 2 we describe the related work that has been carried out on cognate detection together with some of its practical applications, while in Section 3 we present our approach and deal in greater detail with our learning algorithms. Once the proposed methodology has been outlined, we step through an evaluation method we devised and report on the results obtained as specified in Section 4. Section 5 concludes our paper with a brief summary and tackling further challenges in the near future.

1.1 Contributions

We make the following contributions in this paper:
1. We perform cognate detection for eleven Indian Languages.
2. We exploit Indian languages behaviour to obtain a list of true cognates (WNdata from WordNet and PCData from Parallel Corpora).
3. We train neural networks to establish a baseline for cognate detection.
4. We validate the importance of Wordnets as a resource to perform cognate detection.
5. We release our dataset (WNdata + PCdata) of cognate pairs publicly for the language pairs Hi - Mr, Hi - Pa, Hi - Gu, Hi - Bn, Hi - Sa, Hi - Ml, Hi - Ta, Hi - Te, Hi - Ne, and Hi - Ur.

2 Related Work

One of the most common techniques to find cognates is based on the manual design of rules describing how orthography of a borrowed word should change, once it has been introduced into

the other language. Koehn and Knight (2000) expand a list of English-German cognate words by applying well-established transformation rules. They also noted that the accuracy of their algorithm increased proportionally with the length of the word since the accidental coexistence of two words with the same spelling with different meanings (we identify them as 'false friends') decreases the accuracy.

Most previous studies on automatic cognate identification do not investigate Indian languages. Most of the Indian languages borrow cognates or "loan words" from Sanskrit. Indian languages like Hindi, Bengali, Sinhala, Oriya and Dravidian languages like Malayalam, Tamil, Telugu, and Kannada borrow many words from Sanskrit. Although recently, Kanojia et al. (2019) perform cognate detection for a few Indian languages, but report results with manual verification of their output. Identification of cognates for improving IR has already been explored for Indian languages (Makin et al., 2007). String similarity-based methods are often used as baseline methods for cognate detection and the most commonly used among them is Edit distance based similarity measure. It is used as the baseline in the early cognate detection papers (Melamed, 1999). Essentially, it computes the number of operations required to transform from source to target cognate.

Research in automatic cognate detection using phonetic aspects involves computation of similarity by decomposing phonetically transcribed words (Kondrak, 2000), acoustic models (Mielke et al., 2012), phonetic encodings (Rama et al., 2015), aligned segments of transcribed phonemes (List, 2012). We study Rama (2016)'s research, which employs a Siamese convolutional neural network to learn the phonetic features jointly with language relatedness for cognate identification, which was achieved through phoneme encodings. Although it performs well on the accuracy, it shows poor results with MRR. Jäger et al. (2017) use SVM for phonetic alignment and perform cognate detection for various language families. Various works on Orthographic cognate detection usually take alignment of substrings within classifiers like SVM (Ciobanu and Dinu, 2014; Ciobanu and Dinu, 2015) or HMM (Bhargava and Kondrak, 2009). We also consider the method of Ciobanu and Dinu (2014), which employs dynamic programming based methods for sequence alignment. Among cognate sets common overlap set measures like set intersection, Jaccard (Järvelin et al., 2007), XDice (Brew et al., 1996) or TF-IDF (Wu et al., 2008) could be used to measure similarities and validate the members of the set.

3 Datasets and Methodology

We investigate language pairs for major Indian languages namely Marathi (Mr), Gujarati (Gu), Bengali (Bn), Punjabi (Pa), Sanskrit (Sa), Malayalam (Ml), Tamil (Ta), Telugu (Te), Nepali (Ne) and Urdu (Ur) with Hindi (Hi). We create two datasets as described below for `<source_lang>-<target_lang>` where the source language is always Hindi. We describe each step in the subsections below.

3.1 Datasets

Dataset 1: WordNet based dataset

We create this dataset (WNData) by extracting synset data from the IndoWordnet database. We maintain all words, in the concept space, in a comma-separated format. We, then, create word lists by combining all possible permutations of word pairs within each synset. For *e.g.,* If synset ID X on the source side (Hindi) contains words S_1W_1 and S_1W_2, and parallelly on the target side (other Indian languages), synset ID X contains T_1W_1 and T_1W_2, we create a word list such as:
S_1W_1, T_1W_1
S_1W_2, T_1W_1
S_1W_1, T_1W_2
S_1W_2, T_1W_2

To avoid redundancy, we remove duplicate word pairs from this list.

Dataset 2: Parallel Corpora based dataset

We use the ILCI parallel corpora for Indian languages (Jha, 2010) and create word pairs list by comparing all words in the source side sentence with all words on the target side sentence. Our hypothesis, here, is that words with high orthographic similarity which occur in the same context window (a sentence) would be cognates with a high probability. Due to the unavailability of ILCI parallel corpora for Sa and Ne, we download these corpora from Wikipedia and align it with the Hindi articles from Hindi Wikipedia. We calculate exact word matches to align articles to each other thus creating comparable corpora and discard unaligned lines from both sides. We, then,

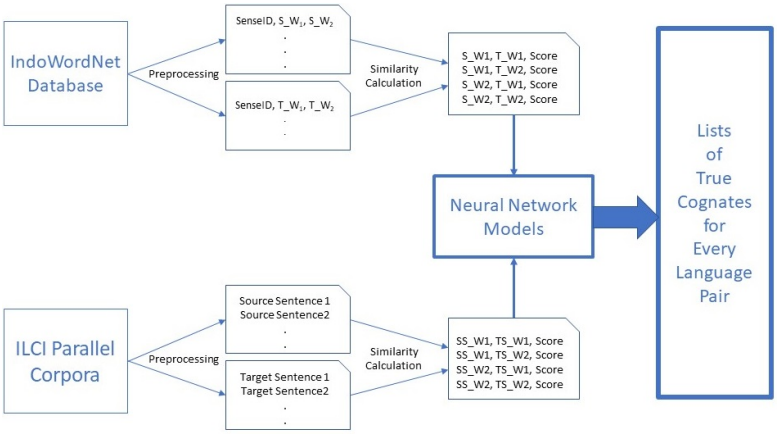


Figure 3: Block Diagram for our experimental setup

create similar word pairs list between Hindi and all the other languages pairs. We removed duplicated word pairs from this list as well and call this data PCData.

3.2 Script Standardization and Text Normalization

The languages mentioned above share a major portion of the most spoken languages in India. Although most of them borrow words from Sanskrit, they belong to different language families. Mr, Gu, Bn, Pa, Ne and Ur belong to the Indo-Aryan family of languages; and Ml, Ta, Te belong to the family of Dravidian languages. They also use different scripts to represent themselves textually. For standardization, we convert all the other written scripts to Devanagari. We perform Unicode transliteration using Indic NLP Library[1] to convert scripts for Bn, Gu, Pa, Ta, Te, Ml, and Ur to Devanagari, for both our datasets. Hi, Mr, Sa, and Ne are already based on the Devanagari script, and hence we only perform text normalization for both our datasets, for these languages. The whole process is outlined in Figure 3.

3.3 Similarity Scores Calculation

We calculate similarity scores for each word on the source side *i.e.,* Hi by matching it with each word on the target side *i.e.,* Sa, Bn, Gu, Pa, Mr, Ml, Ne, Ta, Te, and Ur.

Since we match the words from the same concept space or the same context window, we eliminate the possibility of this word pair carrying different meanings, and hence **a high orthographic similarity score gives us a strong indication of these words falling under the category of True Cognates**. For training neural network models, we then divide the positive and negative labels based on a threshold and follow empirical methods in setting this threshold to 0.5 for both datasets[2]. Using 0.5 as threshold, we obtained the best training performance and hence chose to use this as the threshold for similarity calculation. The various similarity measures used are described in the next subsection.

3.4 Similarity Measures

Normalized Edit Distance Method (NED)

The Normalized Edit Distance approach computes the edit distance (Nerbonne and Heeringa, 1997) for all word pairs in a synset/concept and then provides the output of probable cognate sets with distance and similarity scores. We assign labels for these sets based on the similarity score obtained from the NED method, where the similarity score is (1 - NED score). It is usually defined as a parameterizable metric calculated with a specific set of allowed edit operations, and each operation is assigned a cost (possibly infinite). The score is normalized such that 0 equates to no similarity and 1 is an exact match. NED is equal to the minimum number of operations required to transform

[1] https://anoopkunchukuttan.github.io/indic_nlp_library/

[2] We ran experiments with 0.25, 0.60, and 0.75 as well, and choose 0.5 based on training performance

'word a' to 'word b'. A more general definition associates non-negative weight functions (insertions, deletions, and substitutions) with the operations.

Cosine Similarity (Cos)

The cosine similarity measure (Salton and Buckley, 1988) is another similarity metric that depends on envisioning preferences as points in space. It measures the cosine of the angle between two vectors projected in a multi-dimensional space. In this context, the two vectors are the arrays of character counts of two words. The cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. For example, in information retrieval and text mining, each term is notionally assigned a different dimension and a document is characterised by a vector where the value in each dimension corresponds to the number of times the term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter. This is analogous to the cosine, which is 1 (maximum value) when the segments subtend a zero angle and 0 (uncorrelated) when the segments are perpendicular. In this context, the two vectors are the arrays of character counts of two words.

Jaro-Winkler Similarity (JWS)

Jaro-Winkler distance (Winkler, 1990) is a string metric measuring similar to the normalized edit distance deriving itself from Jaro Distance (Jaro, 1989). It uses a prefix scale P which gives more favourable ratings to strings that match from the beginning, for a set prefix length L. We ensure a normalized score in this case as well. Here, the edit distance between two sequences is calculated using a prefix scale P which gives more favourable ratings to strings that match from the beginning, for a set prefix length L. The lower the JaroWinkler distance for two strings is, the more similar the strings are. The score is normalized such that 1 equates to no similarity and 0 is an exact match.

3.5 Models

3.5.1 Feed Forward Neural Network (FFN)

In this network, we deal with a word as a whole. Words of the source and target languages reside in separate embedding space. The source word passes through the source embedding layer. The target word passes through the target embedding layer. The outputs of both embedding lookups are concatenated. The resulting representation is passed to a fully connected layer with ReLU activations, followed by a softmax layer.

	FFN		RNN	
	D1	D2	D1	D2
Hi-Mr	69.76	85.76	74.76	89.78
Hi-Bn	65.18	81.04	69.18	86.44
Hi-Pa	73.04	78.50	76.04	83.64
Hi-Gu	61.74	79.16	69.84	89.44
Hi-Sa	61.72	85.87	68.92	91.66
Hi-Ml	56.96	74.77	66.96	79.59
Hi-Ta	55.62	61.70	65.62	68.92
Hi-Te	52.78	65.26	62.78	74.83
Hi-Ne	70.20	83.85	80.20	89.63
Hi-Ur	69.99	73.84	76.99	80.12

Table 1: Stratified 5-fold Evaluation using Deep Neural Models on both PCData (D1) and WNData (D2)

3.5.2 Recurrent Neural Network (RNN)

In this network (see Figure 4), we treat a word as a sequence of characters. Characters of the source and the target language reside in separate embedding spaces. The characters of the source word are passed through source embedding layer. The characters of the target word are passed through the target embedding layer. The outputs of both embedding lookups are, then, concatenated. The resulting embedded representation is passed through a recurrent layer. The final hidden state of the recurrent layer is then passed through a fully connected layer with ReLU activation. The resulting output is finally passed through a softmax layer.

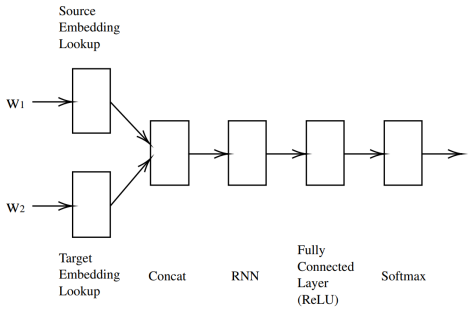


Figure 4: Architecture of a Recurrent Neural Network

	Corp+WN20		Corp+WN40		Corp+WN60		Corp+WN80		Corp+WN100	
	FFN	RNN	FFN	RNN	FFN	RNN	FFN	RNN	FFN	RNN
Hi-Mr	70.12	74.12	73.56	78.37	76.09	81.56	81.34	85.24	86.90	91.87
Hi-Bn	71.06	73.17	73.29	74.98	77.33	76.28	83.99	81.45	82.18	89.58
Hi-Pa	74.16	75.94	76.02	77.39	76.18	79.04	78.04	81.22	80.66	85.64
Hi-Gu	65.26	70.76	71.21	74.83	75.09	79.95	80.14	84.32	81.85	89.81
Hi-Sa	65.93	74.23	69.25	77.51	74.84	79.92	81.03	86.62	88.13	93.86
Hi-Ml	57.75	59.38	56.31	65.67	58.02	71.19	61.01	75.59	69.11	82.54
Hi-Ta	54.63	60.12	56.69	63.38	57.46	66.17	59.36	67.17	60.41	70.62
Hi-Te	53.21	58.18	56.19	63.90	64.15	67.70	65.19	70.65	66.10	74.92
Hi-Ne	70.78	71.23	74.30	78.11	72.19	83.20	79.70	85.01	84.69	90.95
Hi-Ur	69.94	71.25	70.01	72.35	72.03	76.59	71.07	78.27	73.99	80.99

Table 2: Results after combining chunks of WNData with PCData

4 Results

We average the similarity scores obtained using the three methodologies (NED, Cos, and JWS) described above, for each word pair, and then use these as training labels for cognate detection models. We obtain results using the networks described above and report them in Table 1. We calculate average scores for both models and both datasets and show the chart in Figure 5. We observe that RNN outperforms FFN for both the datasets across all language pairs (see Figure 5). We also find that Hi-Sa (see Figure 5) has the best cognate detection accuracy among all language pairs (for both RNN and FFN), which is in line with the fact that they are closely related languages when compared to other Indian language pairs. We observe that average scores for WNData are always higher than average scores for PCData for all language pairs (Figure 5). Also, in line with our observations above, the overall average of RNN scores for both datasets are even higher than average FFN scores (Figure 5).

We perform another set of experiments by combining non-redundant word pairs from both datasets. We add WNData in chunks of 20 per cent to PCData for each language pair and create separate word lists with average similarity scores. We use FFN to train and perform a stratified 5-fold evaluation for each language pair after adding each chunk and show the results in Table 2. After evaluating our results for FFN, we perform the same training and evaluation with RNN. **We observe that adding complete WNData to PCdata improves our performance drastically and given us the best results for almost all cases.**

	WNPairs	CorpPairs	Matches
Hi-Bn	324537	505721	17402
Hi-Pa	260123	465140	16325
Hi-Mr	322013	555719	17698
Hi-Gu	423030	542311	17005
Hi-Sa	669911	248421	10109
Hi-Ml	353104	315234	12392
Hi-Ta	225705	248207	7112
Hi-Te	369872	431869	7599
Hi-Ne	191701	420176	11264
Hi-Ur	99803	420176	6509

Table 3: Total Word Pairs for both datasets and Matches among them

Only in case of Hi-Bn, when using the FFN for training, PCData combined with 80% WNData performs better than 100% Data; possibly due to added sparsity of the additional data. Our hypothesis that adding WNData to PCdata improves the performance holds for all the other cases, including when trained using RNN.

5 Discussion and Analysis

A parallel corpus is a costly resource to obtain in terms of both time and effort. For resource-scarce languages, parallel corpora cannot easily be crawled. We wanted to validate how crucial Wordnets are as a resource and can they act as a substantial dataset in the absence of parallel corpora. In addition to validating the performance of chunks of WNData combined with PCData, we also calculated the exact matches of word pairs from both the datasets and show the results in Table 3. We observed that Hi-Mr had the most

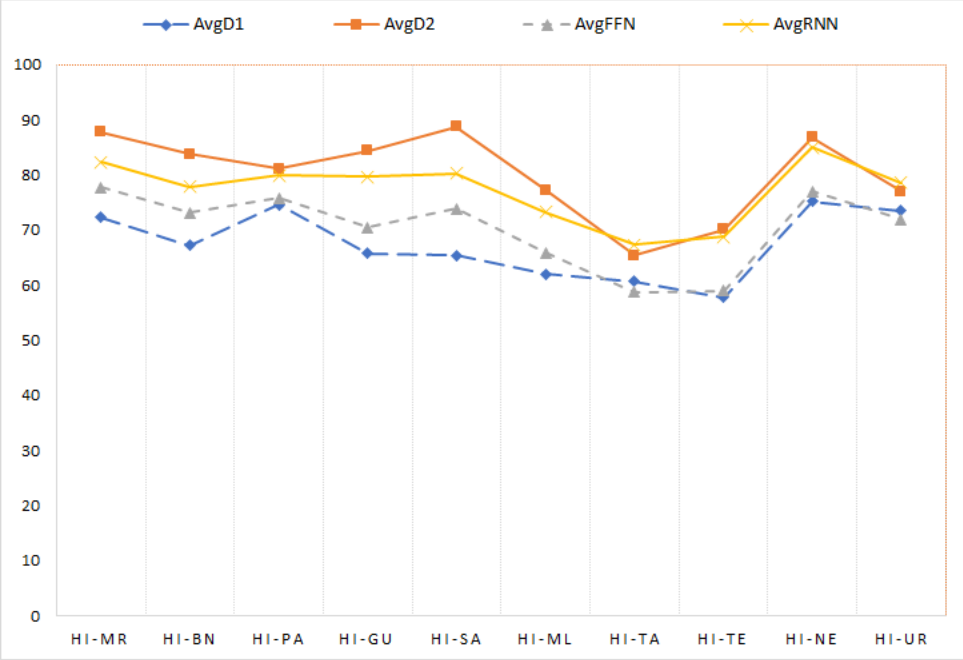


Figure 5: Average Results using Neural Network models on both datasets

Source Word	Target Word	Meaning	Cos	NED	JWS
tadanukool	*tadanusaar*	accordingly	0.500	**0.571**	0.482
yogadaan karna	*yogadaan karane*	to contribute	0.631	**0.636**	0.593
duraatma	*dushtaatama*	evil soul	0.629	**0.700**	0.648

Table 4: Manual analysis of the similarity scores

matched pairs amongst all the languages. PCData is extracted from parallel corpora and is not stemmed for root words, whereas WNData is extracted from IndoWordnet and only contains root words. Despite many words with morphological inflections, we were able to obtain exactly matching words, amongst the datasets. WNData constitutes a fair chunk of root words used in PCData as well, and this validates the fact that models trained on WNData can be used to detect cognate word pairs from any standard parallel corpora as well.

It is a well-established fact that Indian languages are spoken just like they are written and unlike their western counterparts are not spoken and spelled differently. Hence, we choose to perform cognate detection using orthographic similarity methods. This very nature of Indian languages allows us to eliminate the need for using aspects of Phonetic similarity to detect true cognates. Most of the Indian languages borrow words from Sanskrit in either of the two forms - *tatsama* or *tadbhava*. When a word is borrowed in *tatsama* form, it retains its spelling, but in case of *tadbhava* form, the spelling undergoes a minor change to complete change. Before averaging the similarity scores, we tried to observe which of the three (NED, JWS, or Cos) scores would perform better for true cognates known to us in *tadbhava* form with minor spelling changes. We analysed individual word pairs from the data and presented a small sample of our analysis in Table 4. We observe that NED consistently outperforms Cos and JWS for cognate word pairs and confirmed that NED based similarity is the most suited metric for cognate

detection (Rama et al., 2015). We also observe that our methodology can handle word pairs without any changes and with minor spelling changes among cognates, the total of which, constitutes a large portion of the cognates among Indian Language pairs.

6 Conclusion and Future Work

In this paper, we investigate cognate detection for Indian Language pairs (Hi-Bn, Hi-Gu, Hi-Pa, Hi-Mr, Hi-Sa, Hi-Ml, Hi-Ta, Hi-Te, Hi-Ne, and Hi-Ur). A pair of words is said to be Cognates if they are etymologically related; and True Cognates, if they carry the same meaning as well. We know that parallel concepts, bearing the same sense in linked WordNets, are etymologically related. We, then, use the measures of orthographic similarity to find probable Cognates among parallel concepts. We perform the same task for a parallel corpus and then train neural network models on this data to perform automatic cognate detection. We compute a list of True Cognates and release this data along with the data processed previously. We observe that Recurrent Neural Networks are best suited for this task. We observe that Hindi - Sanskrit language pair, being the closest, has the highest percentage of cognates among them. We observe that RNN, which treats the words as a sequence of characters, outperforms FFN for all the language pairs and both the datasets. We validate that Wordnets can play a crucial role in detecting cognates by combining the datasets for improved performance. We observe a minor, but crucial, increase in the performance of our models when chunks of Wordnet data are added to the data generated from the parallel corpora thus confirming that Wordnets are a crucial resource for Cognate Detection task. We also calculate the matches between word pairs from the Wordnet data and the word pairs from the parallel corpora to show that Wordnet data can form a significant part of parallel corpora and thus can be used in the absence of parallel corpora.

In the near future, we would like to use cross-lingual word embeddings, include more Indian languages, and investigate how semantic similarity could also help in cognate detection. We will also investigate the use of Phonetic Similarity based methods for Cognate detection. We shall also study how our cognate detection techniques can help infer phylogenetic trees for Indian languages. We would also like to combine the similarity score by providing them weights based on an empirical evaluation of their outputs and extend our experiments to all the Indian languages.

Acknowledgement

We would like to thank the reviwers for their time and insightful comments which helped us improve the draft. We would also like to thank CFILT lab for its resources which helped us perform our experiments and its members for reading the draft and helping us improve it.

References

Aditya Bhargava and Grzegorz Kondrak. 2009. Multiple word alignment with profile hidden markov models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, pages 43–48. Association for Computational Linguistics.

Chris Brew, David McKelvie, et al. 1996. Word-pair extraction for lexicography. In *Proceedings of the 2nd International Conference on New Methods in Language Processing*, pages 45–55.

Alina Maria Ciobanu and Liviu P Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 99–105.

Alina Maria Ciobanu and Liviu P Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 431–437.

DA Crystal. 2008. Dictionary of linguistics and phonetics 6th edition crystal. *DA Crystal–Oxford: Blackwell Publishing*.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1205–1216.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Anni Järvelin, Antti Järvelin, and Kalervo Järvelin. 2007. s-grams: Defining generalized n-grams for information retrieval. *Information Processing & Management*, 43(4):1005–1019.

Girish Nath Jha. 2010. The tdil program and the indian langauge corpora intitiative (ilci). In *LREC*.

Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholemreza Haffari. 2019. Cognate identification to improve phylogenetic trees for indian languages. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 297–300. ACM.

Philipp Koehn and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *AAAI/IAAI*, pages 711–715.

Grzegorz Kondrak. 2000. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 288–295. Association for Computational Linguistics.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Johann-Mattis List. 2010. Sca: phonetic alignment based on sound classes. In *New Directions in Logic, Language and Computation*, pages 32–51. Springer.

Johann-Mattis List. 2012. Lexstat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125. Association for Computational Linguistics.

Ranbeer Makin, Nikita Pandey, Prasad Pingali, and Vasudeva Varma. 2007. Approximate string matching techniques for effective clir among indian languages. In *International Workshop on Fuzzy Logic and Applications*, pages 430–437. Springer.

I Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Helen M Meng, Wai-Kit Lo, Berlin Chen, and Karen Tang. 2001. Generating phonetic cognates to handle named entities in english-chinese cross-language spoken document retrieval. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.*, pages 311–314. IEEE.

Michelle M Mielke, Rosebud O Roberts, Rodolfo Savica, Ruth Cha, Dina I Drubach, Teresa Christianson, Vernon S Pankratz, Yonas E Geda, Mary M Machulda, Robert J Ivnik, et al. 2012. Assessing the temporal relationship between cognition and gait: slow gait predicts cognitive decline in the mayo clinic study of aging. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 68(8):929–937.

Andrea Mulloni and Viktor Pekar. 2006. Automatic detection of orthographics cues for cognate recognition. In *LREC*, pages 2387–2390.

Andrea Mulloni. 2007. Automatic prediction of cognate orthography using support vector machines. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, pages 25–30. Association for Computational Linguistics.

John Nerbonne and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*.

A Pranav. 2018. Alignment analysis of sequential segmentation of lexicons to improve automatic cognate detection. In *Proceedings of ACL 2018, Student Research Workshop*, pages 134–140.

Taraka Rama, Lars Borin, GK Mikros, and J Macutek. 2015. Comparative evaluation of string similarity measures for automatic language classification.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *arXiv preprint arXiv:1804.05416*.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Peter Turchin, Ilia Peiros, and Murray Gell-Mann. 2010. Analyzing genetic connections between languages by matching consonant classes. *. : .* , (5):117–126.

William E Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage.

Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):13.