ENRICHING KNOWLEDGE SOURCES

FOR NATURAL LANGUAGE UNDERSTANDING

by

Egoitz Laparra

September 2009

Contents

Chapter 1

Motivation

Knowledge mining is emerging as the enabling technology for new forms of information access and multilingual information access, as it combines the last advances in text mining, knowledge acquisition, natural language processing and semantic interpretation. Extracting and modeling this knowlege are become keys issues for those tasks that involves the Natural Language Understanding as Question Answering, information access based on entities, cross-lingual information access, and navigation via cross-document relations. Despite of the difficulty of developing automatic systems to build large and consistent knowledge bases it is neccesary to confront this kind of work to avoid to invest the excesive effort that building these bases manually takes.

1.1 Frame

This project is located whitin the frame of KYOTO project and KNOW2 projecto.

KYOTO(ICT-211423) is an Asian-European project developing a community platform for modeling knowledge and finding facts across languages and cultures. The platform operates as a Wiki system that multilingual and multi-cultural communities can use to agree on the meaning of terms in specific domains. The Wiki is fed with terms that are automatically extracted from documents in different languages. The users can modify these terms and relate them across languages. The system generates complex, language-neutral knowledge structures that remain hidden to the user but that can be used to apply open text mining to text collections. The resulting database of facts will be browseable and searchable. Knowledge is shared across cultures by modeling the knowledge across languages. The system is developed for 7 languages and applied to the domain of the environment, but it can easily be extended to other languages and domains.

KNOW2 will emulate and improve current multilingual information access(MLIA)systems

with research to enable the construction of an integrated environment allowing the cost-effective deployment of vertical information access portals for specific domains.The KNOW project (TIN2006-15049-C03) already enhanced Cross-Lingual Information Retrieval and Question Answering technology with improved concept-based Natural Language Processing technologies. KNOW2 plans to move from general domains to specific domains as a strategy to obtain better performance, and the incorporation of text-mining and collaborative interfaces. In fact the main research objective consists in advancing the state-of-the-art in the integration of text-capture, semantic interpretation, non-standard text treatment (blogs, e-mails, oral transcriptions) and inference and logic reasoning with semantic-based MLIA methods. Given the current state-of-the-art in those areas, we plan to develop intuitive collaborative interfaces which will allow communities of users to improve the systems, including multilingual communities involving Basque, Catalan, English and Spanish. Regarding the expertise and human resources gathered in this project, rather than just piling upc experts from loosely related areas, we have selected on purpose a coordinated groups of researchers from four groups that together form a virtual research laboratory that gathers the necessary critical mass. KNOW2 is formed by an interdisciplinary group including computer-science expertise on natural language processing and industrial applications, and linguistic expertise on the target languages.

The advances of KNOW2 will be demonstrated by quality publications on top-ranking conferences and journals, as well as demonstrators and prototypes on domains such us environment, European parliament and/or geographic texts, including public portals dedicated to popular science (zientzia.net and BasqueResearch, part of AlphaGalileo) owned by Elhuyar, which is a KNOW2 partner. The fact that we apply our state-of-the-art research to real scenarios, and the adoption of the last representation standards and free software licenses will facilitate the technology transfer of the developed technology to industrial environments. The large number of EPOs in this proposal, and their level of commitment, already shows the interest that this proposal raises.

1.2 Contribution

This document presents some novel approaches for building and enriching automatically large, complete and consistent knowldege resources. Specifically, our contribution consist of:

- Enriching and reasoning with ontologies
  - To get complete and consistent annotations of knowledge resources as WordNet
- Enriching knowledge resources

  - Integrating different resources as FrameNet and WordNet

- Aplying these resources in Natural Language Understanding
  - To connect predicate models in different languages as SenSem and FrameNet

1.3 Outline

The rest of this work is structured as follows: The next chapter will introduce some basic concepts as Epistemology or Ontology. It will also introduce the lexical resources and ontologies that are being used in this project. The third chapter will present a summary of the papers published to date and the results obtained. In the fourth chapter will show the conclusions reached in these papers. A bibliography concludes this work.

Chapter 2

Introduction

2.1 Terminology

2.1.1 Epistemology

Epistemology is the philosophical study of the nature of knowledge. It is concerned with the concepts of belief and truth. It analyzes what knowledge can be acquired, how knowledge can be acquired and what it means to acquire knowledge. The eld is immensely complex and brings up numerous puzzles, traps and pitfalls. To overcome them, we will introduce a number of simplifying assumptions. These assumptions will allow us to see reality as a set of true statements for the sake of this thesis.

2.1.2 Knowledge

In general, one distinguishes two types of knowledge: declarative knowledge and procedural knowledge. Declarative knowledge concerns information expressed by statements, such as the information that Paris is in France. Procedural knowledge concerns abilities to perform certain tasks, such as the ability to ride a bicycle. There is some evidence that these two types of knowledge work through different psychological processes in the human mind.

2.1.3 Ontology

The philosophical study concerned with how reality can be structured and described and with existence in general is called Ontology. As in the area of Epistemology, there exist numerous puzzles and pitfalls in the field of Ontology. To overcome these diffculties, we will introduce a number of assumptions. They will allow us to structure the entities of this world into individuals, classes and relations.

2.1.4 Information Retrieval and Information Extraction

While Information Retrieval (IR) and IE are both dealing with some form of text searching, they are quite different in terms of what output or results they produce. IR is the simple classical approach to text searching in the internet. In IR the user enters some words of interest, and then all the documents containing these words are listed. The document list can be ordered accordingly to how many times each search word occurs, how close the different search words are clustered in the document and so on. In this approach, the user has to run many different searches to cover all the possible different search words to describe the fact that she is actually looking for. Also, for every search she might have to read all the articles returned by the search engine, just to see if they really are of interest or not. Information Extraction (IE) seeks to reduce the users workload by adding reasoning to the IR process. With IE the computer will have some knowledge about synonyms and different sentence forms that actually express the same basic facts. That means that the user only has to specify the question that she has, and then the computer will do the tedious work of running several different IR searches, and skimming every single retrieved article to see whether or not it is of interest. The end result from IE can be simple yes/no answers to different questions or it can be specific facts that are extracted from various articles and then used to build databases for quick and easy lookup later.

2.1.5 Natural Language Understanding

In the literature, full parsing and other symbolic approaches are commonly called Natural Language Understanding. Symbolic approaches means using symbols that have a defined meaning both for humans and machines. The other approaches, e.g. statistical, are often called Natural Language Processing. This use of terms tells us that NLU seeks to do something more than just process the text from one format to another. The end goal is to transform the text into something that computers can understand. That means that the computer should be able to answer natural language (e.g. English) questions about the text, and also be able to reason about facts from different texts. The field of NLU is strongly connected to the field of Artificial Intelligence(AI).

2.2 Lexical-Semantic Resources

2.2.1 WordNet

Developed at Princeton University, WordNet [1] is a lexical database for English general domain that currently is one of the most used lexical resources in the area of NLP. WordNet (WN) was born as an attempt to organize lexical information on meanings, unlike conventional dictionaries, where this information is organized by the form of

lexical items.

WN is structured as a semantic network whose nodes, called synsets (synonym sets, or sets of synonyms) constitute the basic unit of meaning. Each consists of a set of lexicalizations representing a meaning and is identified by an offset (byte) and its corresponding POS (Part-of-Speech), which can be (n) for names, (v) for verbs, (a) to adjectives (r) for adverbs:

02152053#n fish#1
01926311#v run#1
02545023#a funny#4
00005567#r automatically#1

The arcs that join these nodes represent the lexical-semantic relations established between synsets. Altogether there are up to 26 different types of relationships, some of the most important are:

Hiperonymy: It is the generic term used to describe a class of specific instances. Y it's a hypernimo of X, if X is a kind of Y.

Example:
tree#n#1 HYPERONYM oak#n#2

Hiponymy: It is the specific term used to designate a member of a class, X is a hyponym of Y if X is a kind of Y. In the case of verbs is called Troponymy.

Example:
oak#n#2 HYPONYM tree#n#1

Antonymy: It is the relationship that binds two senses with opposite meanings.

Example:
active#a#1 ANTONYM_OF inactive#a#2
inactive#a#2 ANTONYM_OF active#a#2

Meronymy: It is the relationship defined as component, substance, or a member of something, X is meronym of Y if X is part of Y.

Example:
car#n#1 HAS_PART window#n#2
milk#n#1 HAS_SUBSTANCE protein#n#1
family#n#1 HAS_MEMBER child#n#2

Holonymy: It is the opposite relationship to the meronymy, Y is holnimo of X if X is a part of Y.

Example:
window#n#2 PART_OF car#n#1
protein#n#1 SUBSTANCE_OF milk#n#1
child#n#2 MEMBER_OF family#n#2

2.2.2 EuroWordNet

The success of WordNet encouraged the creation of new similar projects for other languages. Of these the most prominent has been EuroWordNet [2]. EuroWordNet (EWN) is a multilingual extension of WN, consisting of lexical databases for 8 languages (English, Netherlands, Spanish, Italian, French, German, Czech and Estonian).

After completing the EWN project several wordnets for other languages such as Catalan, Euskera, Portuguese, Greek, Bulgarian, Russian and Swedish began to be developed . Currently, the creation of these wordnets is coordinated by "Global Wordnet Organization".

Interlingual index

The starting point for all these local wordnets was WN1.5 version. They all followed the same linguistic design, kepping the notion of synset and basic semantic relationships. Still, the multilingual nature of this project imposed some modifications.

Each local wordnet of EWN was built separately with the resources available in each language, forming a set of independent modules. The connection between these autonomous systems was made through the Interlingua index (ILI), which was conceived as a superset of the concepts that are common to all languages of EWN. Each index in the ILI is a synset with a syntactic category label, a gloss and the reference to their origin. The synsets of each local wordne are linked to some index of ILI. Thus, EWN can pass from the lexicalization of a concept in a specific language to another lexicalization of the same concept in another language. For example, it is possible to go from the Spanish verb *convertirse* to its counterpart in italian, *diventare*, through its equivalent in the ILI, *to become*:

convertirse (esp) Eq_Near_Synonym *to become*(ILI) Eq_Near_Synonym (it) *diventare*

2.2.3 MCR (Multilingual Central Repository)

The MCR (Multilingual Central Repository) [3] is the result of the merge of different resources (different versions of WordNet, ontologies and knowledge bases) that took place in the MEANING project, from the repository of standard meanings provided by WordNet, following the model proposed by EuroWordNet, thus it includes the interligua relations provided by the set of ILI.

The MCR is a large-scale multilingual resource for a very large number of semantic processes that need a lot of linguistic knowledge. Each reference to a sense of a word will point to concepts that are stored in the MCR.

Its final version is composed by wordnets of five different languages (English, Italian, Spanish, Catalan and Euskera) and contains 1,642,384 unique semantic relationships between concepts. Also enriched with 466,972 semantic features extracted from other sources such as WordNet Domains, Top Concept Ontology or SUMO.

2.2.4 FrameNet

FrameNet [4] is a very rich semantic resource that contains descriptions and corpus annotations of English words following the paradigm of Frame Semantics [5]. In frame semantics, a Frame corresponds to a scenario that involves the interaction of a set of typical participants, playing a particular role in the scenario. FrameNet groups words (lexical units, LUs hereinafter) into coherent semantic classes or frames, and each frame is further characterized by a list of participants (lexical elements, LEs, hereinafter). Different senses for a word are represented in FrameNet by assigning different frames. Currently, FrameNet represents more than 10,000 LUs and 825 frames. More than 6,100 of these LUs also provide linguistically annotated corpus examples. However, only 722 frames have associated a LU. From those, only 9,360 LUs3 where recognized by WN (out of 92%) corresponding to only 708 frames.
LUs of a frame can be nouns, verbs, adjectives and adverbs representing a coherent and closely related set of Word-frame pairs meanings that can be viewed as a small semantic field. For example, the frame EDUCATION TEACHING contains LUs referring to the teaching activity and their participants. It is evoked by LUs like student.n, teacher.n, learn.v, instruct.v, study.v, etc. The frame also defines core semantic roles (or FEs) such as STUDENT, SUBJECT or TEACHER that are semantic participants of the frame and their corresponding LUs.

2.2.5 SenSem

SenSem [6] is a verbal database consisting of a verbal lexical database and a corpus of 700,000 words corresponding to 25.000 sentences and their contexts. The database is built inductively from annotated corpus at syntactic level - syntactic category and

syntactic function - and at semantic level - vebal sense, verbal eventive classes, semantic roles and sentential semantics.
The database SenSem has as main unit the verbal sense with sintactic-semantic information. Specifically, the verbal sense includes the definition of the sense, subcategorization patterns, sentence structures in which it participates (agentive, antiagentive, passive, etc), its asociation to a synset of WordNet, the lexical eventive class, the semantic roles, synonyms in some cases, examples, and the frequency of occurrence of the sense in the corpus. The database includes a total of 998 senses.

2.3 Ontologies

2.3.1 Top Concept Ontology

The TCO [7] was not primarily designed to be used as a repository of lexical semantic information but for clustering, comparing and exchanging concepts across languages in the EWN Project. Nevertheless, most of its semantic features (e.g. Human, Instrument, etc.) have a long tradition in theoretical lexical semantics so they have been usually postulated as semantic components of meanings. The TCO consists of 63 features and it is primarily organized, following [8], in three disjoint types of entities:

- 1stOrderEntity (physical things)
- 2ndOrderEntity (events, states and properties)
- 3rdOrderEntity (unobservable entities)

1st Order entities are further distinguished in terms of four ways of conceptualizing things [9]:

- Form: as an amorphous substance or as an object with a fixed shape (Substance or Object)
- Composition: as a group of self-contained wholes or as a necessary part of a whole (Group or Part)
- Origin: the way in which an entity has come about (Artifact or Natural).
- Function: the typical activity or action is associated to the entity (Comestible, Furniture, Instrument, etc.)

Concepts can be classified in terms of any combination of these four categories. As such, the Top Concepts can be seen more as features than as ontological classes. Nevertheless, most of their subdivisions are disjoint categories: a concept cannot be both Object and Substance, or both Natural and Artifact.

2ndOrderEntitiy lexicalizes nouns and verbs denoting static or dynamic situations. All of the 2nd Order entities are classified using two different classification schemes:

- SituationType
- SituationComponent

SituationType represents a basic classification in terms of the Aktionsart properties of nouns and verbs, as described for instance in Vendler (1967). SituationType can be Static or Dynamic, further subdivided in Property and Relation on the one side, and UnboundedEvent and BoundedEvent on the other. SituationComponent subtypes (e.g. Location, Existence, Cause) emerged empirically when selecting verbal and deverbal Base Concepts (BCs) in EWN. They resemble the cognitive components that play a role in the conceptual structure of events as in Talmy (1985). Each 2ndOrderEntity concept can be classified in terms of a mandatory but unique SituationType and any number of SituationComponent subtypes.

Last, 3rdOrderEntity was not further subdivided.

The TCO has been redesigned twice, first by the EAGLES expert group [10] and then by Vossen (2001). EAGLES expanded the original ontology by adding 74 concepts while the latter made it more flexible, allowing, for instance, to cross-classify features between the three orders of entities.

2.3.2 SUMO

The SUMO [11](Suggested Upper Merged Ontology) is an ontology that was created at Teknowledge Corporation with extensive input from the SUO mailing list, and it has been proposed as a starter document for the IEEE-sanctioned SUO Working Group. The SUMO was created by merging publicly available ontological content into a single, comprehensive, and cohesive structure [11].

2.4 Related Works

2.4.1 YAGO/NAGA

YAGO [12] is an ontology that combines the coverage of Wikipedia with the conceptual hierarchy of WordNet. YAGO builds on entities and relations and currently describes more than 1.7 million entities and 14 million facts. The latter include the Is-A hierarchy as well as non-taxonomic relations between entities (such as hasWonPrize). There are currently 100 different binary relationships in YAGO. The entities and facts about them are mainly extracted from Wikipedia's category system and infoboxes, whereas the class hierarchy is derived from WordNet.

YAGO is based on a clean logical model with a decidable consistency. However, YAGO itself only provides very rudimentary semantics based on merely five basic

axioms, so only limited forms of reasoning are possible. Furthermore, its upper level relies entirely on WordNet, which, as elaborated earlier, has certain limitations when conceived as a formal ontology.

NAGA is a semantic search system that operates on the knowledge graph of YAGO. NAGAs graph-based query language is geared towards expressing queries with additional semantic information. Its scoring model is based on the principles of generative language models, and formalizes several desiderata such as confidence, informativeness and compactness of answers.

Chapter 3

Summary of Papers

3.1 Complete and Consistent Annotation of WordNet using the Top Concept Ontology

Abstract. *This paper presents the complete and consistent ontological annotation of the nominal part of WordNet. The annotation has been carried out using the semantic features defined in the EuroWordNet Top Concept Ontology and made available to the NLP community. Up to now only an initial core set of 1,024 synsets, the so-called Base Concepts, was ontologized in such a way.*
The work has been achieved by following a methodology based on an iterative and incremental expansion of the initial labeling through the hierarchy while setting inheritance blockage points. Since this labeling has been set on the EuroWordNet's Interlingual Index (ILI), it can be also used to populate any other wordnet linked to it through a simple porting process. This feature-annotated WordNet is intended to be useful for a large number of semantic NLP tasks and for testing for the first time componential analysis on real environments. Moreover, the quantitative analysis of the work shows that more than 40% of the nominal part of WordNet is involved in structure errors or inadequacies

The methodology followed for annotating the ILI with the TCO is based on the common assumption that hyponymy corresponds to feature set inclusion (Cruse, 2002) and in the observation that, since wordnets are taken to be crucially structured by hyponymy it is possible to create a rich consistent semantic lexicon inheriting basic features through the hyponymy relations [10]. This methodology confronts two main drawbacks, the hyponymy hierarchy of wordnet is not consistent (Guarino 1998) and there may be multiple inherintace.

In the EuroWordNet project TCO features were asigned to this Basic Concepts. These are genereal concepts that can represent other concepts that are behind them in the hyponymy hierachy, but they do not cover all wordnet. Thus, firs of all, we annotated the gaps of the hierarchy asigning TCO features to the Semantic Files of

WN 1.6:

04 noun.act ⇒ Agentive
05 noun.animal ⇒ Animal
06 noun.artifact ⇒ Artifact
07 noun.attribute ⇒ Property
08 noun.body ⇒ Object; Natural
09 noun.cognition ⇒ Mental

After this, all concepts of wordnet can be annotated by at least one TCO feature by inheriting from those concepts that are already annotated. Now it is possible to find inconsistencies looking for those concepts having features that are defined as dijoints in the TCO. For instance:

Object - Substance
Gas - Liquid - Solid
Artifact - Natural
Animal - Creature - Human - Plant
Dynamic - Static

After a manual check of the incompatibilities obtained it is possible to fix them deleting some of the manually annotated features or blocking some of the hyponymy relations to avoid the inherintace of disjoint features.

The whole process has provided a complete and consistent annotation of the nominal part of WN1.6, which consists of 65,989 nominal synsets, with 116,364 variants or senses. All 227,908 initial incompatibilities were solved by manually adding or removing 13,613 TCO features and establishing 359 blockage points. The final resource has 207,911 synset-feature pairs (an average of 2.66 TCO features per synset), expanded to 427,460 pairs when applying the inheritance of features consistently (an average of 6.48 TCO features per synset). In fact, the synset public_relations_1 has the maximum number of directly assigned features with nine, followed by ballyhoo_1 with eight features. Every TCO feature has been assigned on average 3,300 times. Ranging from Object which is the most widely assigned TCO feature (with 24,905 assignments) to Origin which was only assigned once. The blockage points appear to be distributed along most WordNet levels. However, levels 6, 7 and 8 concentrate most of them (67% of the total, with 86, 87 and 67 blocking points respectively).

Interestingly, every blockage point affects a large number of synsets. Every blockage point subsumes an average of 120.16 synsets. In fact, 28,123 synsets have at least one blockage point in their hypernymy chain (i.e., from itself to WordNet's top). That is, following the TCO ontological incompatibilities, more than 40% of the nominal

part of WordNet is involved in structural errors or inadequacies. However, it seems that most of the them are concentrated in small subparts of the WordNet hierarchy. 18,284 synsets inherit only one blocking point while only 9,839 synsets inherit more than one. On the other side, for instance, 62 synsets inherit 11 blocking points (most of them because of the structural problems of academic_degree_1).

3.2 A New Proposal for Using First-Order Theorem Provers to Reason with OWL DL Ontologies

Abstract. *Existing OWL DL reasoners have been carefully designed to reason with DL ontologies in an efficient way at the expense of lack of expressiveness. In order to overcome this limitation in expressiveness, there have been a few attempts to use first-order logic (FOL) theorem provers, which are known to be less efficient than DL reasoners, to work with DL ontologies. However, these approaches did not still allow full FOL capabilities in queries. In this paper, we introduce a new approach (currently under development) to translate OWL DL ontologies into FOL. The translation has been tested using a simple ontology about animals and some FOL theorem provers. On the basis of these tests, we show that our proposal achieves a good trade-off between expressiveness and simplicity of queries. On one hand, our system is capable of handling any FOL query that cannot be processed by DL reasoners. On the other hand, simple queries are solved in reasonable time by FOL theorem provers in comparison with* ad hoc *DL reasoners.*

Currently, there exist some efficient OWL DL reasoners as FaCT++ or Pellet and some utilities to maintain and edit Description Logic (DL) ontologies, such as Protégé. All this tools allow to use this kind of ontologies for representing knowledge, but the lack of expressivness of OWL DL hinder its use with complex information.

Hoolet is a system developed to improve OWL DL ontologies with the Firs Order Logic expressiveness translating the ontologies into a collection of FOL axioms that can be used to reason with a FOL reasoner as Vampire. But this symste is not able to solve all of FOL queries, for instance those with universally quantifiers:

$$\forall X: \ (\ (Fish \sqsubseteq X \wedge Rodent \sqsubseteq X) \rightarrow Vertebrate \sqsubseteq X \ )$$
$$\exists X: \ (\ (X \not\equiv \bot \wedge \forall Y : (Y \not\equiv X \wedge Y \not\equiv \bot)) \rightarrow \neg(Y \sqsubseteq X) \ )$$

We have proposed a method that allows to get an answer for any FOL queries translating the axioms of the OWL DL ontology to a sound and complete FOL theory. For the moment, we have been able to complete succefully the translation and reasoning process with subclass and disjoint relations.

3.3 Integrating FrameNet and WordNet using a knowledge based WSD algorithm

Abstract. *This paper presents a novel automatic approach to partially integrate FrameNet and WordNet. In that way we expect to extend FrameNet coverage, to enrich WordNet with frame semantic information and possibly to extend FrameNet to languages other than English. The method uses a knowledge-based Word Sense Disambiguation algorithm for matching the FrameNet lexical units to WordNet synsets. Specifically, we exploit a graph-based Word Sense Disambiguation algorithm that uses a large-scale knowledgebase derived from WordNet. We have developed and tested four additional versions of this algorithm showing a substantial improvement over state-of-the-art results.*

Building large and rich enough lexical resources that represent predicate models as FrameNet takes such a great deal of manual effort their coverages are still unsatisfactory. We propose a method to integrate automatically Framenet and WordNet that allows us to:

- Extend the coverage of FN including variants from WN as new LUs
- Enrich WN with new direcet relations between the synsets that belgons to the same frame
- Extend FN to other languages using the Interlingual Index of EuroWordNet

Our method consists of desambiguating the Lexical Units of FrameNet using a knowledge-based algorithm to find with synset corresponds to each LU. The algorithm we have used is called SSI-Disjktra [13] and is a version of the Structural Semantic Interconnections algorithm [14]. The main weakness of this algorithm is that it needs at least one monosemous word in the list of words to interpret. For this reason we have developed four different versions of SSI-Dijkstra algorithm that can give a result even all words in the list are polysemous.

Table 3.1 presents detailed results per Part-of-Speech (POS) of the performance of the different SSI algorithms in terms of Precision (P), Recall (R) and F1 measure (harmonic mean of recall and precision). In bold appear the best results for precision, recall and F1 measures. As baseline, we also include the performance measured on this data set of the most frequent sense according to the WN sense ranking (wn-mfs) that is very competitive in WSD tasks, and it is extremely hard to beat. However, all the different versions of the SSI-Dijkstra algorithm outperform the baseline. Only SSI-Dijkstra obtains lower recall for verbs because of its lower coverage. In fact, SSI-Dijkstra only provide answers for those frames having monosemous LUs, the SSI-Dijkstra variants provide answers for frames having at least two LUs (monosemous or polysemous) while the baseline always provides an answer.

	nouns			verbs			adjectives			all		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
wn-mfs	0.75	0.75	0.75	0.64	0.64	0.64	0.80	0.80	0.80	0.69	0.69	0.69
SSI-dijktra	**0.84**	0.65	0.73	0.70	0.56	0.62	**0.90**	0.82	0.86	**0.78**	0.63	0.69
FSI	0.80	**0.77**	**0.79**	0.66	0.65	0.65	0.89	**0.89**	**0.89**	0.74	**0.73**	0.73
ASI	0,80	**0,77**	**0,79**	0,67	0,65	0,66	0,89	**0,89**	**0,89**	0,75	**0,73**	**0,74**
FSP	0.75	0.73	0.74	**0.71**	**0.69**	**0.70**	0.79	0.79	0.79	0.73	0.72	0.72
ASP	0.72	0.69	0.70	0.68	0.66	0.67	0.75	0.75	0.75	0.70	0.69	0.69

Table 3.1 Results of the different SSI algorithms

As expected, the SSI algorithms present different performances according to the different POS. Also as expected, verbs seem to be more difficult than nouns and adjectives as reflected by both the results of the baseline and the SSI-Dijkstra algorithms. For nouns and adjectives, the best results are achieved by both FSI and ASI variants. The best results for verbs are achieved by FSP, not only on terms of F1 but also on precision.

To our knowledge, on the same dataset, the best results so far are the ones presented by [15]. They presented a novel machine learning approach reporting a Precision of 0.76, a Recall of 0.61 and an F measure of 0.68. In fact, both evaluations are slightly different since they perform 10-fold cross validation on the available data, while we provide results for the whole dataset.

3.4 Evaluation of semiautomatic methods for the connection between FrameNet and SenSem

Abstract. *This paper presents an approach for the automatic connection between predicate model resources in Spanish and English. The objective is to assess the difficulty of the task and to evaluate the performance of different techniques and semi-automated methods. On the one hand we are pursuing a reduction in the effort for the enrichment of these resources, on the other, to increase their coverage and consistency. Thus, we combine manual annotation and two automatic methods in order to establish correspondences between the various semantic units.*

As we said in previous section building large and rich enough resources takes a too expensive manual effort. Connecting different resources is an effective way to get more complete and large knowledge but it is a difficult task because in many times those resources follows different objectives and criterion. This task is even more difficult when the resources are in different languages as SenSem (Spanish) and FrameNet (English)

We have followed the following methodology to connect SenSem and FrameNet:

1. **Connection between FrameNet and WordNet**, this was explained in the

previous chapter.

2. **Connection between SenSem and WordNet** through the synsets asscociated to each sense of SenSem and the synsets associated to the LUs of FrameNet.
3. **Manual validation** of a part (45%) of the correspondences between *frames* of FrameNet and senses of SenSem.
4. **Learning of classifiers** from positive and negative examples of the correspondences frame - sense from SenseMe generated in the previous step.
5. **Use of classifiers to pre-validate correspondences** that had not been validated manually (55%).
6. **Manual validation** of pre-validated correspondences obtained by the classifiers in the previous step, to evaluate the performance of the classifiers.

In total, through this procedure 329 matching candidate pairs between frame and sense have been found through WordNet (It should be borne in mind that only 9325 LUs are recognized by WordNet (a total of 92%) corresponding to only 672 frames), covering a 42% of frames of FrameNet and 44% of the senses of SenSem associated with a synset (SenseMe has several meanings that are not associated with any synset, for them the connection method through WordNet does not apply). Of these 329, 181 (55%) were validated as effective correlations.

Chapter 4

Conclutions and Future Work

As a consecuence of the work explained in the previous chapter we have reached some promising results.

First of all we have proved that is possible to use FOL theorem provers to reason with OWL DL ontologies translating them properly into FOL. This translation allows to overcome the limitations of current FOL theorem provers, which are not *ad hoc* tools for reasoning with ontologies. However, finding a *good* translation is not an easy task. It depends on the way in which FOL theorem provers work and, of course, on the kind of queries to solve. Our translation has taken into account that general purpose FOL theorem provers are resolution-based. However, we have not added redundant information about properties of binary relations (reflexive, transitive property, etc.), which can be inferred from the remaining axioms that result from our translation. Nevertheless, adding this redundant information could help theorem provers to solve some queries in much less time.

Furthermore, it is worth to note that we have empirically proved the complete nature of the resulting theories. More specifically, we automatically check that any ground atom (or its negation) that can be constructed using the predicates and constants from the ontology is a logical consequence of the theory. This fact ensures that, when typing queries, we can solve any of them, by running the query and its negation in parallel.

We have presented a methodology to get the full annotation of the nouns on the EuroWordNet (EWN) Interlingual Index (ILI) with those semantic features constituting the EWN Top Concept Ontology (TCO). This methodology is based on an iterative and incremental expansion of the initial labeling through the hierarchy while setting inheritance blockage points. Since this labeling has been set on the ILI and it is defined as language-independent, it can be also used to populate any other wordnet linked to it through a simple porting process. Moreover, the work shows that more than 40% of the nominal part of WordNet is involved in WordNet's structure

errors or inadequacies. This fact poses significant challenges for relying in WordNet's hierarchy as the unique resource for abstracting semantic classes for NLP.

Further work will focus on the annotation of a corpus oriented to the acquisition of selectional preferences. These selectional preferences will be compared to state-of-the-art synset-generalization semantic preferences. As a result, we expect a qualitative evaluation of the resource. As a side effect, we expect to gain some knowledge for designing an enhanced version of the TCO more suitable for semantically-based NLP.

We have also presented a novel approach to integrate FrameNet and WordNet. The method uses a knowledge based Word Sense Disambiguation (WSD) algorithm called SSI-Dijkstra for assigning the appropriate synset of WordNet to the semantically related Lexical Units of a given frame from FrameNet. This algorithm relies on the use of a large knowledge base derived from WordNet and eXtended WordNet. Since the original SSI-Dijkstra requires set of monosemous or already interpreted words, we have devised, developed and empirically tested four different versions of this algorithm to deal with sets having only polysemous words. The resulting new algorithms obtain improved results over state-of-the-art.

We are currently developping a new version of the SSI-Dijkstra using ASI for nouns and adjectives, and FSP for verbs. We also plan to further extend the empirical evaluation with other available graph based algorithms that have been proved to be petitive in WSD such as UKB [16]. We also plan to disambiguate the Lexical Elements of FrameNet usign the same automatic aproach for a more complete integration with WordNet.

Finally, we have used the method developed to integrate FrameNet and WordNet to connect verbal predicates between SenSem and FrameNet. We also have learned classifiers to determine if a candidate pair sens-frame was really a match. These classifiers offered widely varying results, probably because the learning examples were generated by two judges without prior consensus on the criteria of annotation. After analazying the results some guidelines were created to improve the consistency of the annotation.

The first step for future work will be the creation of more consistent examples to enhance the classifiers learning. After it, we will apply these classifiers to connect the units of SenSeM and FrameNet that were not associated by correlation between synsets. New candidates will be generated, pre-validated by classifiers and manually validated. Once validated, these correspondences will swell the set of examples which must improve the classifiers functioning.

Bibliography

[1] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Five Papers on WordNet," Special Issue of International Journal of Lexicography 3 (1990).

[2] *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* , P. Vossen, ed., (Kluwer Academic Publishers , 1998).

[3] J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen, "The MEANING Multilingual Central Repository," In *Proceedings of GWC*, (Brno, Czech Republic, 2004).

[4] C. Baker, C. Fillmore, and J. Lowe, "The Berkeley FrameNet project," In *COLING/ACL'98*, (Montreal, Canada, 1997).

[5] C. J. Fillmore, "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*, **280,** 20–32 (New York, 1976).

[6] L. Alonso, J. A. Capilla, I. Castelln, A. Fernndez, and G. Vzquez, "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish," Selected papers from RANLP 2005 (2007).

[7] H. Rodríguez, S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna, and A. Roventini, "The top-down strategy for building EuroWordNet: vocabulary coverage, base concepts and top ontology," pp. 45–80 (1998).

[8] *Semantics 1*, J. Lyons, ed., (Cambridge University Press, Cambridge, UK, 1977).

[9] *The Generative Lexicon*, J. Pustejovsky, ed., (MIT Press, Cambridge, MA, 1995).

[10] A. Sanfilippo *et al.*, "Preliminary Recommendations on Lexical Semantic Encoding - Final Report,", 1999.

[11] I. Niles and A. Pease, "Towards a Standard Upper Ontology," In , pp. 2–9 (ACM Press, 2001).

[12] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge," In *16th international World Wide Web conference (WWW 2007)*, (ACM Press, New York, NY, USA, 2007).

[13] M. Cuadros and G. Rigau, "KnowNet: Building a Large Net of Knowledge from the Web," In *22nd International Conference on Computational Linguistics (COLING'08)*, (Manchester, UK, 2008).

[14] R. Navigli and P. Velardi, "Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **27,** 1063–1074 (2005).

[15] S. Tonelli and D. Pighin, "New Features for FrameNet - WordNet Mapping," In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, (Boulder, CO, USA, 2009).

[16] E. Agirre and A. Soroa, "Personalizing PageRank for Word Sense Disambiguation," In *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009)*, (Eurpean Association for Computational Linguistics, Athens, Greece, 2009).