eman ta zabal zazu

EUSKAL HERRIKO UNIBERTSITATEA
University of the Basque Country

PhD thesis summary

Expansion for information retrieval: contribution of word sense disambiguation and semantic relatedness

Arantxa Otegi Usandizaga

2011

eman ta zabal zazu

EUSKAL HERRIKO UNIBERTSITATEA

University of the Basque Country

Expansion for information retrieval: contribution of word sense disambiguation and semantic relatedness

This summary is a shortened and translated version of the dissertation entitled “Hedapena Informazioaren Berreskurapenean: Hitzen Adiera-Desanbiguazioaren eta Antzekotasun Semantikoaren Ekarpenak”, written by Arantxa Otegi under the supervision of Dr. Eneko Agirre and Dr. Xabier Arregi. It also includes the papers which the candidate has published on the research presented here.

December 2011

Acknowledgments

The Department of Education, Universities and Research of the Basque Government who awarded a pre-doctoral fellowship (BFI06.281) to the author of this PhD dissertation to conduct this research.

Abstract

Information retrieval (IR) aims at searching documents which satisfy the information need of an user. In that way, an IR system informs the user about relevant documents, that is those documents that contain the information they need as formulated in the query. Well-known search engines like Google and Yahoo are prime examples of IR systems.

A perfect IR system should retrieve only, and all, the relevant documents, rejecting the non-relevant ones. However, perfect retrieval systems do not exist. One of the main problems is the so-called vocabulary mismatch problem between query and documents: some documents might be relevant to the query even if the specific terms used differ substantially, or some documents might not be relevant to the query even they have some terms in common. The former is because several words or phrases can be used to express the same idea or item (synonymy). The latter is caused by ambiguity, where one word can have more than one interpretation depending on the context. Owing to these facts, if an IR system relies only on terms occurring in both the query and the document when it comes to deciding whether a document is relevant, it might be difficult to find some of the interesting documents, and also to reject non-relevant documents. It seems fair to think that there will be more chances of successful retrieval if the meaning of the text is also taken into account.

Even though the vocabulary mismatch problem has been widely discussed in the literature from the early days of IR it remains unsolved, and most search engines just ignore it. This PhD dissertation explores whether natural language processing (NLP) can be used to alleviate this problem.

In a nutshell, we expand queries and documents making use of two NLP techniques, word sense disambiguation and semantic relatedness. For each of the mentioned techniques we propose an expansion strategy, in which we obtain synonyms and other related words for the words in the query and documents. We also present, for each case, a method to combine the expansions and original words effectively in an IR system. Furthermore, as the expansion technique we propose is useful for translating queries and documents, we show how a cross lingual information retrieval system could be improved using such an expansion technique.

Our extensive experiments on three datasets show that the expansion methods explored in this dissertation help overcome the mismatch problem, consequently improving the effectiveness of an IR system.

Contents

1 Introduction

The term "information retrieval" was first used by Mooers (1950), who provided the following definition:

> "Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information."

Loosely speaking, the process aimed at searching documents which satisfy the information need of a user has from then on been called information retrieval (IR).

In a more recent definition (Hiemstra 2009), an IR system is "a software programme that stores and manages information on documents, often textual documents but possibly multimedia". It thus informs the users about the documents that contain the information they need. Note that an IR system does not explicitly return information or answer questions, it only retrieves and suggests documents.

The working of IR systems

IR systems perform three main processes as follows (cf. Fig. 1):

(i) Indexing: an index is constructed as a representation of the documents. An index is a data structure which facilitates fast and accurate searches. It takes place offline, and it is not necessary to carry it out more than once, unless the document collection is changed.

(ii) Query formulation: the user formulates a query according to his information need.

(iii) Matching: the query is compared against the document representation (index). The comparison results in a selection of a subset of documents.

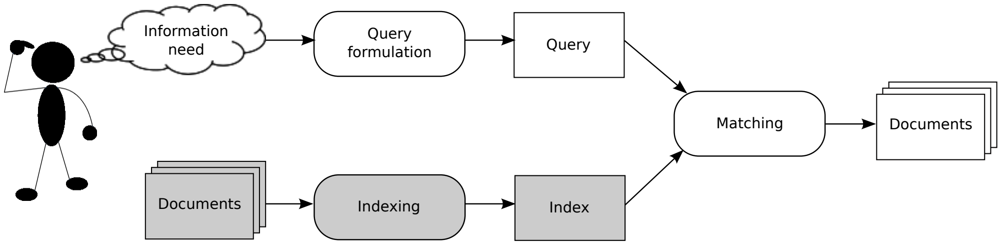


Figure 1 – Schematic diagram of IR systems' processes. The processes which are executed offline are shown in grey.

Some of the documents in the resulting subset will, hopefully, contain an information of value with respect to the information need. These documents are called relevant documents. A perfect system should only retrieve relevant documents, rejecting non-relevant ones. However, perfect retrieval systems do not exist and later we will see some of the shortcomings of current systems. Nowadays, systems usually return a ranked list of documents, in which the documents at the top are those that are most likely to be of interest to the user, or the most likely to be relevant according to the system.

Bush (1945) was the first to propose the use of computers for such retrieval tasks:

> "Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, 'memex' will do. A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Scope of usage

Following the idea of Bush, the first automated systems were developed in the 1950s and 1960s. In those days, most of the IR systems were used to search scientific publications and documents in libraries. They did not use the full content of the documents for searching, but used keywords assigned manually to each document instead.

The field has evolved notably and the current situation is completely different. IR systems are pervasive, due to the widespread use of the Internet and the resulting need of IR systems, the so-called web search engines. The well-known and widely-used web search engines like Google[1] and Yahoo![2] are prime examples of IR systems.

Search methods

The methods to perform IR are also changing as computing power and storage space is increasing. Current systems use almost all the words in the documents when indexing and matching, that is, they take into account the full content of the documents. These systems are known as *full text retrieval* systems. Nevertheless, there are some systems which use a portion of the document, plus some manually assigned keywords, such as the PubMed[3] search engine. This IR system was set up in the 1970s, and it is still in operation. It allows searching of the MEDLINE database, which contains selected publications covering biomedicine and health. For each publication,

[1] http://www.google.es/

[2] http://es.yahoo.com/

[3] http://www.ncbi.nlm.nih.gov/pubmed/

the title, the abstract and the manually assigned keywords (taken from a medical thesaurus) are indexed.

Applications

A search engine is a classical application of IR techniques. There are many types of search engine, although *web search* engines are by far the most used ones. *Vertical search* focuses on a specific topic or media type. *Enterprise search* is used to search across different types of computer files (web pages, email, reports, presentations, spreadsheets...) in an enterprise intranet. *Desktop search* also has to deal with a variety of computer files, but unlike enterprise search the files are located in the personal computer of the user.

Any application that deals with a text document collection or other such unstructured information will need to organise and search that information. Hence, there are some other applications of IR apart from those previously mentioned search engines; for example, *digital libraries* — libraries in which collections are stored in digital formats and are accessed via computers — and *information filtering systems.* As a specific type of the latter we can mention *recommender systems*, systems that recommend information items (film, books, music, events...) that are likely to be of interest to the user.

Tasks

IR techniques are also used for many different tasks, including *document classification*, *question answering*, *multi-document automatic summarisation* and, mainly, *ad hoc retrieval.* Document classification assigns a label or a class to an electronic document, based on its content. Question answering is aimed at automatically answering a question posed in natural language. The goal of multi-document automatic summarisation is to write a summary report, after extracting information from multiple texts written about the same topic. Finally, the task that we are going to address in this thesis is the **ad hoc task**. This task is the most standard IR task, and in it a system **aims to provide documents from within the collection that are relevant to a user query**.

Multimedia documents

Even if more and more IR applications work with non-textual documents (images, videos, audio files or scanned documents), from the earliest days text documents are the most popular objects to search. For this reason, and even if the search system only involves text documents, it is referred to as information retrieval, instead of the more specific *document retrieval* or *text retrieval* system. We will do the same in this dissertation from now on; all our experiments have been done with text documents, and we will use the three terms interchangeably.

Research issues

As we have just seen, IR systems have many applications and, as it extends over new disciplines, many research lines are open: the effectiveness of ranking algorithms, the efficiency of the system (answer time, indexing speed), the capability of incorporating new data into the indexes, scalability (over the amount of data or the number of users), adaptability to new applications, evaluation methodologies, or the vocabulary mismatch problem.

We are going to work on the latter problem in this thesis. To be precise, our research will focus on **the vocabulary mismatch problem when searching text documents in an ad hoc task**.

1.1 The vocabulary mismatch problem

The vocabulary mismatch problem has been widely discussed in the literature from the early days of IR, and yet it remains completely unsolved. Let us discuss the source of this problem.

All languages we use for everyday communication share, at least, these two features:

- richness: more than one word or phrase can be used for express one idea or thing.
- ambiguity: one word can have more than one interpretation depending on the context.

Several researches have confirmed the previous statements. For instance, Furnas et al. (1987) found in their experiments that the probability that two people coincide on the word they spontaneously apply to a given object ranged from 0.07 to 0.18. In respect to ambiguity, for example, 26,896 out of 155,287 words that are in WordNet (17.3%) are polysemous (words that have more than one sense or meaning), and the average numbers of senses of each verb and noun are 2.17 and 1.24, respectively (3.57 and 2.79 without taking into account the words with only one sense)[4].

Owing to these facts, if an IR system tries only to match the character strings in the query with the character strings in the documents during the matching process — that is what basic retrieval models do and what most of the commercial systems used to do until recently — it might be difficult to find some of the interesting documents. The previous characteristic, among others, are the sources of the problems generated during the matching process. These are the linguistic phenomena that an IR system has to face:

[4] These statistics have been taken from http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

– syntactic variants

I'll be at home if it rains / if it rains, I'll be at home

– morphological variants

fish / fishes / fishing / fished / fisher

– morphosyntactic variants

paper roll / roll of paper

– lexical variants: **synonymy**

car / auto / automobile

– semantic variants: **polysemy**

tree: plant / diagram

– cross-language variants: when retrieving documents in another language different from the language used in the query.

Syntactic variants do not cause any problems in current IR systems. This is because most of them use *bag of words* representation, in which a piece of text is characterised as an unordered collection of terms; that is, the order of the words is not taken into account when searching in this model.

Because of morphological and morphosyntactic variants, it would be necessary to search for all the variants of the words used in the query. Instead, most IR systems use a stemmer or a lemmatizer. Using these tools the system can efficiently obtain the stem or the root of a word, and these stems or roots are both used when indexing and searching. For example, after applying a stemmer, the system will use the root fish in both queries and documents where fishes, fishing, fished or fisher would occur.

Synonymy and polysemy are unresolved problems for present systems. These two phenomena influence the retrieval process differently. As a result of synonymy, it might be difficult to retrieve documents in which the same idea of the query is expressed using different words; this could cause retrieving less documents than expected. In contrast, polysemy introduces noise in the retrieved document list, because non-relevant documents with query terms used in a different meaning are retrieved.

Let us illustrate these phenomena with some examples taken from the datasets we have used in our experiments. Firstly, we are going to see some examples of synonymy, as shown in Figs. 2a and 2b. In each of these examples, there is a query (Q) and a document (D) relevant to the given query. The keywords in the query for Fig. 2a are fast, tractor and go, but only one of these (tractor) is in the document. However, there are some other words in the document related to the keywords (such as speed and kilometres per hour) that make the document relevant. Similarly, while the keyword cook of the query in example 2b is not in the document, some other words related to it are used in the document, e.g. recipes and bake. Humans easily understand that these documents have information relevant to the query. In contrast, an IR system performing a simple string match would miss these relevant documents.

These examples show that strict synonymy is not enough to bridge the

Q: How **fast** does a **tractor go**?

D: This Directive shall apply only to **tractors** defined in paragraph 1 which are fitted with pneumatic tyres and which have two axles and a maximum design speed between 6 and 25 kilometres per hour.

(a) Query 96 and document jrc31977L0537/14 from the ResPubliQA dataset.

Q: How do you **cook** an **apple pie**?

D: There are many good recipes for **apple pies** but there are also some important things to remember that are usually not in the recipe. That is you should make sure the bottom of the crust will bake as well and not remain soggy. To do this, coat the inside of the crust with butter before adding the filling and place the baking dish on a dark metal pan so the bottom will get more heat.

(b) Query and document 1005121203620 from the Yahoo! dataset.

Figure 2 – Two examples of the matching problem between the query (Q) and the document (D) due to synonymy.

Title: Computer Mouse RSI
Desc: Find documents that report on computer mouse repetitive strain injuries (RSI).
Narr: Relevant documents report injuries that are caused by the continuous use of a computer mouse. Documents proposing ways to avoid repetitive strain injuries (RSI) when using the computer are also relevant.

(a) Topic 10.2452/064-AH from Robust dataset.

computer mouse rsi repetitive strain injuries

(b) The formulated query using the title and desc fields of the topic.

Figure 3 – A query example from the Robust dataset to illustrate polysemy.

gap between the query and document, as documents contain words which are strongly associated to the query that are not synonyms (speed and kilometres per hour for fast, and recipes and bake for cook). In fact, this dissertation explores the use of lexical relations beyond synonymy. In the rest of the dissertation a very loose definition of synonymy is taken, meaning “related words”. We decided to use the term synonymy to make the text more readable.

We will now turn our attention to polysemy. Fig. 3 shows a query (cf. Fig. 3b) derived from a given information need (cf. Fig. 3a). Some of the terms in this query are polysemous; for example, the words mouse and strain have more than one meaning or sense as shown in Fig. 4. Given this query, a state-of-the-art IR baseline system (which we will use as a baseline for our experiments) retrieves the documents displayed in Fig. 5, among others. When we read these documents, we are going to immediately realise that

mouse-1: any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails.
mouse-2, shiner, black eye: a swollen bruise caused by a blow to the eye.
mouse-3: person who is quiet or timid.
mouse-4, computer mouse: a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad.

(a)

strain-1: (physics) deformation of a physical body under the action of applied forces.
strain-2, stress: difficulty that causes worry or emotional tension; *she endured the stresses and strains of life*.
strain-3, tune, melody, air, melodic line, line, melodic phrase: a succession of notes forming a distinctive sequence; *she was humming an air from Beethoven*.
strain-4, mental strain, nervous strain: (psychology) nervousness resulting from mental stress; *his responsibilities were a constant strain*.
strain-5, breed, stock: a special variety of domesticated animals within a species; *he experimented on a particular breed of white rats*.
strain-6, form, variant, strain, var.: (biology) a group of organisms within a species that differ in trivial ways from similar groups; *a new strain of microorganisms*.
strain-7: injury to a muscle (often caused by overuse), results in swelling and pain.
strain-8, tenor: the general meaning or substance of an utterance; *although I disagreed with him I could follow the tenor of his argument*.
strain-9, striving, nisus, pains: an effortful attempt to attain a goal.
strain-10, straining: an intense or violent exertion.
strain-11, song: the act of singing; *with a shout and a song they marched up to the gates*.

(b)

Figure 4 – WordNet senses of the words mouse and strain (only nouns).

they are not relevant to the given query. These documents are considered to be relevant by the system because there are some query words in them (shown in **this color**), but some of those are used with a different meaning. For instance, the word mouse in the query is used in the computer mouse sense, whereas in the document it refers to a kind of animal. In the case of the word strain, the query refers to an injury in the muscle, while the document in Fig. 5a refers to a variety of an animal, and the document in Fig. 5b refers to a tune or melody of music.

These examples illustrate the problems that arise when the main criterion to classify a document as relevant or non-relevant is whether or not it shares keywords with the query, regardless of the meaning of those words in context.

RESEARCHER ACCUSED OF FAKING DATA; HER STUDY PURPORTED TO USE GENES TO TRANSFER DISEASE RESISTANCE.
(...) Her results were published in the April 25, 1986, issue of the journal Cell in an article co-authored by Nobel laureate David Baltimore. The article "purposed to show that a gene from one **strain** of **mouse** had been transferred to another **strain** of **mouse**, resulting in the latter's production of high levels of antibody molecules it would not normally produce – antibody molecules mimicking the antibody molecules produced by the original **strain**," investigators said in a written statement. (...) after reviewing scientific evidence and performing a **computerized** statistical analysis that showed the false data was not made up of chance errors (...)

(a) Document LA112694-0025 from the Robust dataset.

SOUNDS: LATEST WORK IS BOWEN'S MOST HIGH-PROFILE; COMPOSER AND PERFORMER OF NEW MUSIC SPENT YEARS WORKING ON THE FRINGES.
Listening to the lilting **strains** of Gene Bowen's new album "The Vermilion Sea" (...) the Nordic-looking Bowen has a few guitars, a synthesizer and the all-important **computer** – his main composing tool – and piles of records and CDs. (...) Three years ago, Bowen began his work-in-progress, creating the raw material on synthesizers and **computers**. (...) "My interests came through guitar music and songwriting coupled with interest in folk and ethnic music, where **repetition** is always so important. **Repetition** and texture are almost more important than (...)

(b) Document LA063094-0099 from the Robust dataset.

Figure 5 – Some non-relevant documents retrieved for the given query in the previous example, due to polysemy.

1.2 Lexical semantics for vocabulary mismatch

The problems caused by polysemy and synonymy are the main motivations of this research work. The proposed solutions are drawn from natural language processing (NLP), specifically from the subfield of lexical semantics, which studies among others, word sense disambiguation (WSD) (Agirre and Edmonds, 2006) to deal with polysemy, and semantic relatedness (Budanitsky and Hirst, 2006) to deal with synonymy and other lexical semantic relations.

In order to combine those lexical semantic techniques with IR methods, we will focus our attention on expansion techniques, which consists of adding additional words to the queries or documents. Expansion has been typically used with queries (*query expansion*), but it is also possible to apply it to documents (*document expansion*). In our case the new words will be semantically related to the senses of the words as used in the query or document.

Returning to the example in Fig. 2b, a careful semantic analysis of the query would propose expanding it with words related to cook, such as different manners of cooking (bake, boil or grill among others) and other related words like cooker or recipe. An IR system which properly takes into account the expanded words, would then find that the number of the words in the expanded query that match the document is high and, consequently, the document shown in that example would rank higher.

There are several ways of solving polysemy by means of using WSD or semantic relatedness. The first one is to disambiguate the queries and the documents, that is to tag all the words with its sense using a WSD system. This will allow the IR system to match senses instead of words. Referring again to the previous examples, the query and documents in Figs. 3 and 5 are shown again in Fig. 6, but in the latter one, the words mouse and strain are disambiguated (the sense is specified after the word). It is clear now that the senses in documents 6b and 6c do not match the senses of the query 6a. In contrast, document 6d will be considered relevant, since the senses do match. Note that in these examples we have only disambiguated two words in order to make the examples readable, but a full WSD system will need to disambiguate all the words.

Another option is the expansion to related words, without performing explicit WSD. If the computer manages to understand the semantics behind the query or the document, the expanded representation of the query or document will have more hints on the real meaning, will be semantically richer, and will thus have more words available for matching. The expansion process is likely to introduce some noise, and it is therefore necessary to try to keep a balance between benefit and loss. Returning again to the last example, let us imagine that we expand the query in Fig. 3b with the words electronic device, lesion, wellness, among others. In the same way, imagine that we expand the document in Fig. 5b with the words instrument, singer, vocalist and other words connected with music in general. For this expanded query and document, the number of words that match relative to the number of words would decrease, as the expanded terms in the query are not mentioned in the document and vice versa. As a result, the system would rank the document lower. In contrast, the expansion of the document in Fig. 6d would be very similar to that query, the number of matching words will increase, and the document will be ranked higher.

Although we have presented polysemy and synonymy (including related words) as different phenomena, they are usually closely related. For instance, Fig. 2b was used as an example of synonymy, due to the hypernymy-hyponymy relationship between the words cook and bake. But the word cook is polysemous, because it has more than one meaning as a verb (according to WordNet): the main sense of "prepare a meal" and in the sense of "manipulate, fake, falsify". In fact, bake is related to the "prepare a meal" sense of cook and not to the sense of "manipulate, fake, falsify".

In summary, the examples above try to illustrate that considering the meaning of query and document terms instead of just the strings of characters, the chances for successful retrieval should increase. We thus try to apply NLP techniques, or more specifically, lexical semantic techniques to the ad hoc IR task. In this thesis work, **we are going to use word sense disambiguation and semantic relatedness to better "understand"**

computer **mouse**[mouse-4] **rsi** **repetitive** **strain**[strain-7] **injuries**

(a) Disambiguated query 10.2452/064-AH from the Robust dataset.

RESEARCHER ACCUSED OF FAKING DATA;HER STUDY PURPORTED TO USE GENES TO TRANSFER DISEASE RESISTANCE.
(. . .) Her results were published in the April 25, 1986, issue of the journal Cell in an article co-authored by Nobel laureate David Baltimore. The article ”purposed to show that a gene from one **strain**[strain-5] of **mouse**[mouse-1] had been transferred to another **strain**[strain-5] of **mouse**[mouse-1], resulting in the latter's production of high levels of antibody molecules it would not normally produce – antibody molecules mimicking the antibody molecules produced by the original **strain**[strain-5]," investigators said in a written statement. (. . .) after reviewing scientific evidence and performing a **computerized** statistical analysis that showed the false data was not made up of chance errors (. . .)

(b) Disambiguated document LA112694-0025 from the Robust dataset.

SOUNDS: LATEST WORK IS BOWEN'S MOST HIGH-PROFILE; COMPOSER AND PERFORMER OF NEW MUSIC SPENT YEARS WORKING ON THE FRINGES.
Listening to the lilting **strains**[strain-3] of Gene Bowen's new album ”The Vermilion Sea" (. . .) the Nordic-looking Bowen has a few guitars, a synthesizer and the all-important **computer** – his main composing tool – and piles of records and CDs. (. . .) Three years ago, Bowen began his work-in-progress, creating the raw material on synthesizers and **computers**. (. . .) ”My interests came through guitar music and songwriting coupled with interest in folk and ethnic music, where **repetition** is always so important. **Repetition** and texture are almost more important than (. . .)

(c) Disambiguated document LA063094-0099 from the Robust dataset.

2 FIRMS ADOPT LABELS WARNING **COMPUTER** USERS ABOUT DANGER OF **INJURY**. SAFETY GUIDES PROVIDE USERS WITH TIPS.
Compaq **Computer** Corp. said Tuesday that it will put warning labels on **computer** keyboards this fall, directing people to read a safety guide with tips to avoid hand and wrist **injuries**.(. . .) **Injuries** can range from simple soreness to a tissue swelling that harms nerves in the wrist, a condition known as carpal tunnel syndrome. (. . .) Compaq said Tuesday that there is still no scientifically established link between keyboard design and **injuries**. But it cited growing evidence, chiefly in news accounts, that typing with hands in awkward positions or for a long time can be harmful. (. . .) Microsoft has built a healthy side business in **computer** accessories, such as an ergonomic **mouse**[mouse-4] control. (. . .)

(d) Disambiguated document LA081794-0225 from the Robust dataset.

Figure 6 – Disambiguated query and documents (only the words mouse and strain are disambiguated).

the queries and the documents, using expansion to integrate that information into an IR system. In fact the main goal of this thesis is to use query and document expansion via lexical semantics in order to obtain new words to be inserted into an IR system, with the hope of retrieving more relevant documents at higher ranks. Furthermore, since the expansion technique we are going to propose is useful for translating queries and documents, we will examine whether a cross-language information retrieval system could be improved using the same expansion techniques.

1.3 Prior work in the IXA group

This thesis work has been carried out within the IXA group. The IXA research group of the University of the Basque Country have been working on NLP for more than twenty years. Even though this group mainly focuses on applied research in the Basque language, it also works on research and development of tools in other languages.

Although the group is new in the field of IR, a search service for Basque, called EusBila, has been developed (Leturia et al., 2007). With respect to the field of lexical semantics, various resources and systems have been developed, and several thesis and works have been published, both in the field of WSD (Agirre, 1999; Martinez, 2004; Lopez de Lacalle, 2009) and semantic relatedness (Agirre et al., 2009b). Furthermore, in the field of lexical semantics some resources for Basque have been developed: namely, EuSemcor (semantically tagged Basque corpus) and Euskal WordNet (a Basque WordNet) (Pociello, 2008). Some of the work and tools we have just mentioned have been used in this thesis.

It is important to note that, due to the lack of resources in Basque, we have mainly used English datasets. Nevertheless, the techniques we are going to present in this dissertation are independent of language, and can therefore be applied to any language, in the case of having enough resources for that language in question. Following the group principles, we do not dismiss the future the idea of applying to Basque what we have explored in this thesis, once we have the resources we need.

2 Hypothesis and contributions

The main hypothesis for this dissertation is the following:

> **Does the use of lexical semantics for query and document expansion improve the effectiveness of IR systems in ad hoc tasks?**

In a nutshell, we want to enrich the queries and the documents with semantically related terms in order to alleviate the matching problem. We are going to make use of two lexical semantic techniques, WSD and semantic relatedness. For each of the techniques mentioned, we are going to propose one expansion strategy in which we are going to obtain synonyms and other related words of the words in the queries and the documents. In each case, we are going to present a method for the insertion and exploitation of expansion terms in a state-of-the-art IR system. The IR system will make use of both the words in the original queries and documents and the words obtained from the expansion. On the whole, what we want to achieve is to increase the number of matching words within the query and its relevant documents.

2.1 Research questions and their answers

The central research hypothesis stated above leads to several more specific research questions, whose answers contribute to clarify the main hypothesis. Next, we list those research questions along with answers to them:

– **RQ 1** – ***Does word sense disambiguation and expansion based on synonyms from a lexical knowledge base improve the effectiveness of an IR system?***

In the experiments in Chapter 4 we have carried out query and document expansion making use of topics and documents tagged by a word sense disambiguation system and using WordNet synonyms for expansion[5]. With these resources and without optimising the parameters, we have achieved an improvement in the results on the monolingual (English) task over the baseline system, even if the improvement is not statistically significant .

- 1.1 - *Is this expansion technique suitable for both query and document expansion? Is one more effective than the other?*

We have expanded both the queries and the documents in our experiments. Usually, user queries tend to be short, and, therefore, there is not much content to carry out the disambiguation process. In our experiments we have used the *title* and *description* fields of the topics from the dataset to formulate the queries, and thus the queries are longer than usual. Apart from this, once we have the WSD information, the expansion process is the same for queries and documents. The difference is the method used to insert the expansion terms in the IR system. We have experimented with complex structured queries, in order to determine which was the most effective query to combine original and expansion words, but the best results were obtained by only expanding documents.

[5] Note that in this case we do use strict polysemy, that is we expand to the variants listed as synonyms for the selected sense.

- 1.2 - *What are the different factors affecting the effectiveness of the expansion technique in the IR system?*

  We have experimented with different variants of the expansion technique and some of the findings follow:

  - expansion approach: full (expansion to all synonyms of all senses of each word) vs. best (expansion to the synonyms of the sense with highest WSD score for each word).
    Loosely speaking, the best results are obtained with "full expansion" for query expansion and, in contrast, "best expansion" is the most effective for document expansion.
  - query length, fields of the topic used to formulate the query: *title* vs. *title+description*.
    We have used *title+description* in our main experiments, obtaining small improvements on expansion. We also wanted to explore what happens when only using the *title* field. The conclusions of these experiments are unclear, as we obtained contradictory results: we have not improved the results in the training phase, but, we have obtained a remarkable and statistically significant improvement in the testing phase. Therefore, it is unclear whether or not the expansion is more effective for short queries (2-3 words).
  - the unit used in queries and indices: lemma vs. *synset*[6].
    As new words are introduced in an IR system after the expansion, there is a risk of introducing noise due to incorrect expansion or because some of the words are polysemous. To avoid this problem, we have conducted some experiments using *synsets* for the expansion, instead of words. What we have seen in these experiments is that synsets are not useful.
  - different WSD systems: UBC (Agirre and Lopez de Lacalle, 2007) vs. NUS (Chan et al., 2007).
    We tested the WSD outputs of two systems with no clear conclusion, as we obtained better results with one system in some experiments and vice versa. However, counting all the experiments one by one it seems that NUS performs better. Note that NUS performed slightly better than UBC in the all-words WSD subtask of SemEval-2007 (Pradhan et al., 2007).

  The IR system used in these experiments has several parameters, such as the smoothing parameter, pseudo-relevance feedback pa-

[6] synset: synonym set. The concepts are represented by synsets in WordNet. A synset contains a set of words, each of which has a sense that names those concepts and each of which is therefore synonymous with the other words in the synset.

rameters and the weight of the expanded query. We explored several combinations, but no clear picture emerged.

- 1.3 - *Is this expansion technique suitable for translating queries and documents within cross-language information retrieval?*

As WordNet is available for several languages, and once the synset number of a concept is known it is straightforward to obtain the words that express that concept in other languages. If we take these words from WordNet in a language different from the original one we are translating, in addition to expanding. Thus, the expansion technique presented in Chapter 4 is useful for translating queries and documents. We tested this method in a Spanish-English cross-language information retrieval task, with statistically significant improvements.

– **RQ 2** – ***Does expansion based on semantic relatedness using a lexical knowledge base improve the effectiveness of an IR system?***

In the experiments in Chapters 5 and 6, we have used a WordNet-based graph algorithm to obtain concepts which are semantically related to the each query (or, respectively, each document). The words lexicalising the most closely related concepts were used to expand the query (document). Using the expanded queries and documents, we have shown that its effectiveness on several retrieval task is increased. The results hold for several datasets, with different parameter settings, and also with some other variants, obtaining positive results in general.

- 2.1 - *Is this expansion technique effective for different kinds of retrieval models?*

We have inserted this relatedness-based expansion technique into two IR systems of different types: in a classic probabilistic retrieval model (Chapter 5) and in a language model-based retrieval model (Chapter 6). All in all we have obtained positive results.

- 2.2 - *Is this expansion technique suitable for both query and document expansion? Is one more effective than the other?*

This relatedness-based expansion technique can be used to expand any piece of text and the expansion process is always the same. We have used it for both query and document expansion. We have shown that whether the query or document expansion is the most effective depends on the dataset used, at least when a language model-based retrieval model is used (Chapter 6).

- 2.3 - *How does this expansion technique compare to pseudo-relevance feedback?*

  When a language model-based retrieval model is used, the comparison between the results obtained with, on the one hand, RQE (relatedness-based query expansion) or RDE (relatedness-based document expansion) and, on the other hand, PRF (pseudo-relevance feedback) varies according to the dataset used: RDE and RQE are the most effective ones for Yahoo! and ResPubliQA datasets, but not for Robust dataset. However, analysing the results of each query one by one, we have shown that our expansion models are more effective than PRF for some of the queries. All in all, we can conclude that the relatedness-based expansion models and the PRF model are complementary, in that PRF is better for easy queries and our expansion models are stronger for difficult queries. In fact, GMAP scores on the Robust dataset show that RQE is on a par with the PRF model.

- 2.4 - *What are the different factors affecting the effectiveness of the expansion technique in the IR system?*

  We have performed a detailed analysis of the factors that affect the effectiveness of the relatedness-based expansion techniques, including several parameters (the number of concepts or terms used for the expansion, the weight of the original query and the weight of the expanded index) and their optimisation, length of the documents, the difficulty of the queries and different typologies of the dataset. We have concluded that it is possible to fix the most effective value of these features, although some of them vary depending on the dataset or the retrieval model used. The most important conclusion has to do with parameter optimisation, as this has a big impact on both the baseline and expansion models, where fine-tuning on training data from the same dataset yields the best results for all techniques. The expansion models we have proposed stand out when using sub-optimal parameter settings, which is the case for most real-life IR applications, as in most real cases there is no training dataset available and optimal values from other scenarios do not carry over well.

- 2.5 - *Is this expansion technique suitable for translating queries and documents within cross-language information retrieval?*

  This expansion technique uses graph-based techniques over WordNet which returns synsets. In the English experiments, we take the English variants of the synset, but we could obtain variants in any other language. We tested this approach on a cross-lingual task (Spanish-English) at Robust-WSD 2009.

2.2 Contributions

Our major contribution is that the main research question has a positive answer: **lexical semantics improve the effectiveness of an IR system**. We have tested both WSD and semantic relatedness for query and document expansion, with positive effects. We will next analyse more specific contributions, mentioning the chapter in which it was explored:

- **We have carried out query and document expansion making use of topics and documents enriched with WSD information** (Chapter 4).

  We have added synonyms to each word of the queries and the documents in the expansion process using the English and the Spanish WordNets. We have obtained good results using only this external knowledge for expansion and without optimising the parameters. Moreover, we have shown that this technique is useful for query and document translation within a cross-language information retrieval task.

- **We have carried out query and document expansion making use of semantic relatedness** (Chapters 5 and 6).

  We have proposed a novel expansion technique based on graphs over WordNet, adding concepts (and words) which are related to the text as a whole. This technique does not only obtain expansion terms for the words in the text, but it is able to get concepts which are not explicitly mentioned in the text. We have applied this expansion technique to two IR systems of different types (a classic probabilistic retrieval model and a language model retrieval technique) and, all in all, we have obtained positive results when compared to *query likelihood* and PRF methods. We have also shown that this technique is useful for query and document translation within a cross-language information retrieval task.

- **We have tested the robustness of the semantic relatedness-based expansion techniques over a diverse range of datasets** (Chapters 5 and 6).

  In order to check the performance on datasets of different types, we have used three datasets with different domains, topic typologies and document lengths: (i) Robust, a typical ad hoc dataset on news; (ii) Yahoo!, a dataset that contains questions and answers as posted by real users on diverse topics; and (iii) ResPubliQA, a dataset which was prepared for a passage retrieval task on European Union laws. Our results show that our expansion techniques are robust and achieve a good performance in all three datasets.

- **We have tested the robustness of the semantic relatedness-based expansion techniques over different parameter settings** (Chapters 5 and 6).

  Parameter optimisation has a strong effect on baseline systems of IR and PRF methods, and our expansion techniques are no different. In most real cases there are no training datasets for parameter optimisation, and we have shown that our models are robust in the face of sub-optimal parameters.

- **We have analysed whether there is a link between document length and the effectiveness of relatedness-based expansion techniques** (Chapter 5).

  Using artificially trimmed documents, we have shown that our method is particularly effective for short documents, with few exceptions.

- **We have analysed whether there is a link between the difficulty of the queries and the effectiveness of relatedness-based expansion techniques** (Chapter 6).

  Our analysis shows that PRF is better for easy queries and our model is more effective for hard queries, hinting that there is potential for combination.

- **We have participated in the Robust-WSD task (2008 and 2009 editions) and the ResPubliQA task (2009 and 2010 editions) of CLEF (Cross-Language Evaluation Forum)**[7] (Chapters 4 and 5).

  Our systems ranked highly in these tasks, showing that our expansion techniques are close to the state-of-the-art. In addition, this participation let us compare our systems with other participating systems.

2.3 Future work

We will here summarise some of the research lines that have not yet been fully addressed by this dissertation, as well as some new research lines:

- **Combine the semantic relatedness-based expansion model with PRF**.

  Our analysis indicates that our expansion models based on semantic relatedness and PRF are complementary. We conducted a preliminary

[7] http://clef-campaign.org/

experiment combining both models and got promising results (in Chapter 6), which we would like to explore further.

- **Analyse the concepts we get when using semantic relatedness, and improve the technique**.

  We have used a graph algorithm with default parameter values to obtain the related concepts, since this was the setting obtaining the best results in a word similarity dataset (Agirre et al., 2009b). We only conducted a shallow analysis of the quality of these concepts, but we would like to analyse them further and perhaps improve the semantic relatedness technique.

- **Explore other methods for incorporating the relatedness-based expansion technique in an IR system**.

  We have proposed a straightforward process for the document expansion model: words obtained by the expansion process are added to a second index, giving the option to weight the two indices differently. We would like to explore whether inserting it in a more refined way would yield better results. We foresee two options for this purpose. On the one hand, we could experiment with the BM25F probabilistic retrieval model (Robertson et al., 2004). On the other hand, based on the work in (Mei et al., 2008) and (Huang et al., 2009), we could use the information derived from the expansion to smooth a language model.

- **Analyse the scalability to larger datasets**.

  The largest dataset we have used in our experiments has around a million documents. We believe it is enough for the kind of evaluations we had in mind. However, the most used search engines are web search engines, and as the number of the documents on the web is increasing at a huge rate, it is more and more common to evaluate IR systems with huge datasets. For instance, the biggest dataset used in the last editions of Web Track at the TREC evaluation conference had over a billion web pages[8]. If we want to experiment with such datasets we should firstly analyse the scalability of the graph algorithm we use, as its processing time and space is relatively high.

- **Use external knowledge other than WordNet**.

  The proposed relatedness-based expansion technique makes use of a graph algorithm and a lexical knowledge base. We have chosen WordNet for our experiments since good results were obtained in other experiments by using WordNet and a graph algorithm (Agirre et al., 2009b;

[8] http://plg.uwaterloo.ca/~trecweb/2011.html

Agirre and Soroa, 2009). Even if WordNet is very rich with respect to nouns and verbs, the number of named entities is scarce. Taking into account that such entities might be of a great importance in a retrieval task, it would be better to use an external knowledge base which has many of these entities within it. An alternative to WordNet might be Wikipedia. Wikipedia has lots of articles which are linked between them with anchors. In that way it is feasible to represent Wikipedia with a graph. Moreover, Wikipedia has been used in different semantic relatedness tasks achieving good results (Milne and Witten, 2008; Gabrilovich and Markovitch, 2009).

- **Experiments in other languages**.

  As we have mentioned before, we have experimented mainly in English. We have experimented also in Spanish within a cross-language information retrieval task. Although we have achieved promising results in those cross-language tasks, we believe we should work on the translating techniques further. In addition, we would like to experiment with a Basque dataset. It would be interesting to conduct some domain-specific experiments within a science and technology field and use a specialised ontology called WNTERM as the external knowledge (Pociello et al., 2008).

3 Outline of the dissertation

Below we provide a brief overview of the content of each of the chapters in the dissertation. Note that this summary contains a translation of the introduction and conclusions.

- Chapter 1 – Introduction:

  Firstly, the context and motivation of this research are introduced, as well as the main goal. Afterwards, the research questions we want to answer with this research work are stated. Finally, all the publications related to this thesis work are listed.

- Chapter 2 – State of the art:

  First, due to the importance that the ranking function has in an IR system, the different retrieval models that could be behind a ranking function are outlined. Next, the matching problem and different ways to address it are introduced, followed by several attempts that have been carried out using semantics.

- Chapter 3 – Experimental setup:

  The methodology adopted for this research work is reviewed and some of the basic concepts necessary for understanding the dissertation are presented.

- Chapter 4 – Word sense disambiguation and language model-based IR:

  The experiments we have conducted with the objective of improving the effectiveness of an IR system using WSD are presented. We propose expanding the queries and the documents, adding synonyms using WSD and a lexical knowledge base (WordNet). We have carried out experiments for the English monolingual task and Spanish-English cross-language task; we did these experiments to participate in the *Robust WSD Task @ CLEF 2008* task (Agirre et al., 2009a).

- Chapter 5 – Relatedness and probabilistic IR:

  The experiments we have done with the objective of improving the effectiveness of a probabilistic IR system using semantic relatedness are presented. We propose to expand the documents by adding related words to them using a graph algorithm which is based on WordNet.

- Chapter 6 – Relatedness and language model-based IR:

  The experiments we have done with the objective of improving the effectiveness of a language model-based IR system using semantic relatedness are presented. We propose expanding the queries and the documents using the same technique we have proposed in the previous chapter.

- Chapter 7 – Conclusions and future work:

  The research findings and the contributions of this thesis work are summarised, followed by some possible future areas of research on this subject.

4 Reading guide to the dissertation

The main contributions are published in their respective papers. We will list here these publications, organised according to the dissertation chapters[9]:

[9] Note that the authors are ordered alphabetically in these publications, with exception of the one related to Chapter 6.

- Chapter 4 – Word sense disambiguation and language model-based IR:
    - Agirre E., Otegi A., and Rigau G. **IXA at CLEF 2008 Robust-WSD Task: Using Word Sense Disambiguation for (Cross Lingual) Information Retrieval**. *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, vol. 5706 of Lecture Notes in Computer Science, 118–125. Springer, ISBN 978-3-642-04446-5. 2009.
- Chapter 5 – Relatedness and probabilistic IR:
    - Agirre E., Otegi A., and Zaragoza H. **Using semantic relatedness and word sense disambiguation for (CL)IR**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, vol. 6241 of Lecture Notes in Computer Science, 166–173. Springer, ISBN 978-3-642-15753-0. 2010.
    - Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., Saralegi X., and Zaragoza H. **Elhuyar-IXA: Semantic Relatedness and Cross-Lingual Passage Retrieval**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, vol. 6241 of Lecture Notes in Computer Science, 273–280. Springer, ISBN 978-3-642-15753-0. 2010.
    - Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., and Saralegi X. **Document Expansion for Cross-Lingual Passage Retrieval**. *Proceedings of CLEF 2010 Workshop on Multiple Language Question Answering (MLQA'10)*, ISBN 978-88-904810-0-0. 2010.
    - Agirre E., Arregi X., and Otegi A. **Document expansion based on WordNet for robust IR**. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 9–17, Association for Computational Linguistics, 2010.
- Chapter 6 – Relatedness and language model-based IR:
    - Otegi A., Arregi X., and Agirre E. **Query Expansion for IR using Knowledge-Based Relatedness**. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 1467-1471, ISBN 978-974-466-564-5. 2011.

The following publications, even if not associated with a specific chapter, are also related to the thesis:

- Agirre E., Magnini B., Lopez de Lacalle O., Otegi A., Rigau G., and Vossen P. **SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval**. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 1–6. 2007.

- Agirre E., Di Nunzio G.M., Mandl T., and Otegi A. **CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, Lecture Notes in Computer Science, vol. 6241, 36–49, Springer, ISBN: 978-3-642-15753-0. 2010.

These publications can be found in the appendix of this report. Besides these publications, one further article entitled "Using Knowledge-Based Relatedness for Information Retrieval" has been added to the appendix, which is a working copy and it is related to Chapter 6.

Bibliography

Agirre, E. (1999). *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabi liaz: Dentsitate Kontzeptuala.* PhD thesis, Informatika Fakultatea, UPV-EHU.

Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. (2009a). CLEF 2008: ad hoc track overview. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G., Kurimo, M., Mandl, T., Peñas, A., and Petras, V., editors, *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science*, pages 15–37. Springer.

Agirre, E. and Edmonds, P., editors (2006). *Word Sense Disambiguation: Algorithms and applications*, volume Text, Speech and Language Technology Series, 33. Springer.

Agirre, E. and Lopez de Lacalle, O. (2007). UBC-ALM: combining k-NN with SVD for WSD. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 342–345, Stroudsburg, PA, USA. Association for Computational Linguistics.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41. Association for Computational Linguistics.

Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., and Paşca, M. (2009b). A Study on Similarity and Relatedness Using Distributional and WordN et-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Confe rence of the North American Chapter of the Association for Computational Linguis tics*, NAACL '09, pages 19–27. Association for Computational Linguistics.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computacional Linguistics*, 32:13–47.

Bush, V. (1945). As We May Think. *The Atlantic Monthly*, 176(1):101–108.

Chan, Y. S., Ng, H. T., and Zhong, Z. (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.

Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Gabrilovich, E. and Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34(1):443–498.

Hiemstra, D. (2009). *Information Retrieval: Searching in the 21st Century*, chapter Information Retrieval Models, pages 1–19. John Wiley & Sons, Ltd.

Huang, Y., Sun, L., and Nie, J. (2009). Smoothing document language model with local word graph. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1943–1946. ACM.

Leturia, I., Gurrutxaga, A., Areta, N., Alegria, I., and Ezeiza, A. (2007). EusBila, a search service designed for the agglutinative nature of B asque. In *SIGIR2007- iNEWS'07 workshop.*

Lopez de Lacalle, O. (2009). *Domain-Specific Word Sense Disambiguation.* PhD thesis, Lengoiaia eta Sistema Informatikoak Saila, UPV/EHU.

Martinez, D. (2004). *Supervised Word Sense Disambiguation: Facing Current Challenges*. PhD thesis, Informatika Fakultatea, UPV/EHU.

Mei, Q., Zhang, D., and Zhai, C. (2008). A general optimization framework for smoothing language models on graph structures. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 611–618. ACM.

Milne, D. and Witten, I. H. (2008). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artifical Intellegence (WIKIAI08).*

Mooers, C. N. (1950). Information retrieval viewed as temporal signaling. In *Proceedings of the International Congress of Mathematicians.*

Pociello, E. (2008). *Euskararen ezagutza-base lexikala: Euskal WordNet.* PhD thesis, Euskal Filologia Saila, UPV/EHU.

Pociello, E., Gurrutxaga, A., Agirre, E., Aldezabal, I., and Rigau, G. (2008). WNTERM: Enriching the MCR with a Terminological Dictionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC).*

Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M. (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.

Robertson, S., Zaragoza, H., and Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 42–49, New York, NY, USA. ACM.

Appendix

This appendix includes a copy of the publications related to this dissertation. See Section 4 for the reading guide.

- Agirre E., Otegi A., and Rigau G. **IXA at CLEF 2008 Robust-WSD Task: Using Word Sense Disambiguation for (Cross Lingual) Information Retrieval**. *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, vol. 5706 of Lecture Notes in Computer Science, 118–125. Springer, ISBN 978-3-642-04446-5. 2009.

- Agirre E., Otegi A., and Zaragoza H. **Using semantic relatedness and word sense disambiguation for (CL)IR**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, vol. 6241 of Lecture Notes in Computer Science, 166–173. Springer, ISBN 978-3-642-15753-0. 2010.

- Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., Saralegi X., and Zaragoza H. **Elhuyar-IXA: Semantic Relatedness and Cross-Lingual Passage Retrieval**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, vol. 6241 of Lecture Notes in Computer Science, 273–280. Springer, ISBN 978-3-642-15753-0. 2010.

- Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., and Saralegi X. **Document Expansion for Cross-Lingual Passage Retrieval**. *Proceedings of CLEF 2010 Workshop on Multiple Language Question Answering (MLQA'10)*, ISBN 978-88-904810-0-0. 2010.

- Agirre E., Arregi X., and Otegi A. **Document expansion based on WordNet for robust IR**. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 9–17, Association for Computational Linguistics, 2010.

- Otegi A., Arregi X., and Agirre E. **Query Expansion for IR using Knowledge-Based Relatedness**. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, 1467-1471, ISBN 978-974-466-564-5. 2011.

- Agirre E., Magnini B., Lopez de Lacalle O., Otegi A., Rigau G., and Vossen P. **SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval**. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 1–6. 2007.

- Agirre E., Di Nunzio G.M., Mandl T., and Otegi A. **CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, Lecture Notes in Computer Science, vol. 6241, 36–49, Springer, ISBN: 978-3-642-15753-0. 2010.

- **Using Knowledge-Based Relatedness for Information Retrieval**. *Working copy.*

(*Note that the authors are ordered alphabetically in most of the publications.*)

IXA at CLEF 2008 Robust-WSD Task: using Word Sense Disambiguation for (Cross Lingual) Information Retrieval

Eneko Agirre, Arantxa Otegi, and German Rigau

IXA NLP Group - University of Basque Country. Donostia, Basque Country.
arantza.otegi@ehu.es

Abstract. This paper describes experiments for the CLEF 2008 Robust-WSD task, both for the monolingual (English) and the bilingual (Spanish to English) subtasks. We tried several query and document expansion and translation strategies, with and without the use of the word sense disambiguation results provided by the organizers. All expansions and translations were done using the English and Spanish wordnets as provided by the organizers and no other resource was used. We used Indri as the search engine, which we tuned in the training part. Our main goal was to improve (Cross Lingual) Information Retrieval results using WSD information, and we attained improvements in both mono and bilingual subtasks, with statistically significant differences on the second. Our best systems ranked 4th overall and 3rd overall in the monolingual and bilingual subtasks, respectively.

1 Introduction

Our experiments intended to test whether word sense disambiguation (WSD) information can be beneficial for Cross Lingual Information Retrieval (CLIR). We carried out different expansion and translation strategies of both the topics and documents with and without word sense information. For this purpose, we used thef open source Indri search engine, which is based on the inference network framework and supports structured queries [7].

The remainder of this paper is organized as follows. Section 2 describes the experiments carried out, Section 3 presents the results obtained, Section 4 describes some related work and, finally, Section 5 draws the conclusions and mentions future work.

2 Experiments

In short, our main experimentation strategy consisted on trying several expansion and translation strategies, all of which used the synonyms in the English and Spanish wordnets made available by the organizers as the sole resources (i.e., we did not use any other external resource), with and without word sense information. Our runs have consisted of different combinations of expanded (translated) topics and documents. The steps of our retrieval system are the following. We first expand and translate the documents and topics. In a second step we index the original, expanded and translated document collections. Then we test different query expansion and translation strategies, and finally we search for the queries in the indexes in various combinations. All steps are described sequentially.

2.1 Expansion and translation strategies

WSD data provided to the participants was based on WordNet version 1.6. Each word sense has a WordNet synset assigned with a score. Using those synset codes and the English and Spanish wordnets, we expanded both the documents and the topics. In this way, we generated different topic and document collections using different approaches of expansion and translation, as follows:

- Full expansion of English topics and documents: expansion to all synonyms of all senses.
- Best expansion of English topics and documents: expansion to the synonyms of the sense with highest WSD score for each word, using either UBC or NUS disambiguation data (as provided by organizers).
- Full translation of English documents: translation from English to Spanish of all senses.
- Best translation of English documents: translation from English to Spanish of the sense with highest WSD score for each word, using either UBC or NUS disambiguation data.
- Translation of Spanish topics: translation from Spanish to English of the first sense for each word, taking the English variants from the WordNet.

In the subsequent steps, we used different combinations of these expanded and translated collections.

2.2 Indexing

Once the collections had been pre-processed, they were indexed using Indri. While indexing, the Indri implementation of the Krovetz stemming algorithm was applied to document terms. We created several indexes: one with the original collection words, and one with each collection created after applying different expansion (and translation) strategies, as explained in Section 2.1. No stopword list was used, but only nouns, adjectives, verbs and numbers were indexed.

2.3 Query construction

We constructed queries using the title and description topic fields. Based on the training topics, we excluded some words and phrases from the queries, such as *find, describing, discussing, document, report* for English and *encontrar, describir, documentos, noticias, ejemplos* for Spanish. After excluding those words and taking only nouns, adjectives, verbs and numbers, we constructed several queries for each topic as follows:

1. Original words.
2. Both original words and expansions for the best sense of each word.
3. Both original words and all expansions for each word.
4. Translated words, using translations for the best sense of each word. If a word had no translation, the original word was included in the query.

The first three cases are for the monolingual runs, and the last one for the bilingual run which translated the query. Table 1 shows some examples of each case for the sample topic.

Table 1. Query examples using the title and description fields of a topic. Check Section 2.3 for further explanations.

English topic	*<EN-title>Alternative Medicine</EN-title>* *<EN-desc>Find documents discussing any kind of alternative or natural medical treatment including specific therapies such as acupuncture, homeopathy, chiropractics, or others</EN-desc>*
Spanish topic	*<ES-title>Medicina Alternativa</ES-title>* *<ES-desc>Encontrar documentos que traten sobre algún tipo de tratamiento medico alternativo o naturista, incluyendo terapias concretas como la acupuntura, la homeopatía, la quiropráctica, u otras</ES-desc>*
case 1	`#combine(#1(alternative medicine) kind alternative natural medical treatment including specific therapies acupuncture homeopathy chiropractics others)`
case 2	`#weight(0.6 #combine(#1(alternative medicine) kind alternative natural medical treatment including specific therapies acupuncture homeopathy chiropractics others) 0.4 #combine(#syn(#1(complementary medicine) #1(alternative medicine)) #syn(variety form sort) #syn(option choice) #syn(include) #syn(therapy) #syn(stylostixis) #syn(homoeopathy) #syn(chiropractic)))`
case 3	`#weight(0.6 #combine(#1(alternative medicine) kind alternative natural medical treatment including specific therapies acupuncture homeopathy chiropractics others) 0.4 #combine(#wsyn(1 #1(complementary medicine) 1 #1(alternative medicine)) #wsyn(1 form 1 variety 1 sort) #wsyn(1 option 1 choice) #wsyn(0 nonsynthetic 0 uncontrived 0 misbegot 0 unaffected 0 spurious 0 bastardly 0 lifelike 0 bastard 0 wild 0 rude 0 spontaneous 0 misbegotten 0 unstudied 0 raw) #wsyn(0 aesculapian) #wsyn(0 discussion 0 discourse 0.414874001229255 handling) #wsyn(0 admit 0 #1(let in) 1 include) #wsyn(1 therapy) #wsyn(1 stylostixis) #wsyn(1 homoeopathy) #wsyn(1 chiropractic)))`
case 4	`#combine(#syn(#1(alternative medicine) #1(complementary medicine)) type treatment #syn(medicate medicine) #syn(alternate alternative) #syn(naturistic nudist) include concrete #syn(acupuncture stylostixis) #syn(homeopathy homoeopathy) quiropráctica )`

In the first case, we constructed a simple query combining the original words using the Indri operator `#combine` (see *case 1* in Table 1). Note that multiword expressions (as present in WordNet), such as *alternative medicine*, are added to the query joined with the `#1` operator (ordered window).

For the rest of cases, we have used some other operators available in the structural Indri Query Language. For *case 2*, where we include original words as well as synonyms (obtained after expansion) in the query, we constructed two subqueries, one with original words, and another one with the expanded words. Both subqueries are combined into a single query using the `#weight` operator, where original words are weighted with 0.6, and synonyms with 0.4. We did not fine-tune this weights. We used the synonym operator (`#syn`) to join the expanded words of each sense, as they are meant to be synonyms.

In the case of full expansion (*case 3*), instead of `#syn`, we used `#wsyn` (weighted synonym). This operator allows to give different weights to synonyms, which we took from the score returned by the disambiguation system, that is, each synonym was weighted according to the WSD weight of the corresponding sense of the target word.

For *case 4*, we constructed the query using the first sense of each word of the Spanish topics in order to get their translated English words. In the Spanish topic of the example, as *quiropractica* had not any sense assigned, we could not get its translation and therefore, we included the original Spanish word in the query (see *case 4* in Table 1).

2.4 Retrieval

We carried out several retrieval experiments combining different kinds of indexes with different kinds of queries. We used the training data to perform extensive experimentation, and chose the ones with best MAP results in order to produce the test topic runs. The submitted runs are described in Section 3.

In some of the experiments we applied pseudo-relevance feedback (PRF) with the following default parameters: fbDocs:10, fbTerms:50, fbMu:0 and fbOrigWeight: 0.5. Unfortunately, we did not have time to tune those parameters for the official deadline.

3 Results

Table 2 summarizes the results of our submitted runs. We present them here, as follows:

- monolingual without WSD:
  En2EnNowsd ; original terms in topics; original terms in documents.
  En2EnNowsdPsrel ; same as `En2EnNowsd`, but with PRF.
- monolingual with WSD:
  En2EnNusDocsPsrel ; original terms in topics; both original and expanded terms in documents, using best sense according to NUS word sense disambiguation; PRF.
  En2EnUbcDocsPsrel ; original terms in topics; both original and expanded terms in documents, using best sense according to UBC word sense disambiguation; PRF.
  En2EnFullStructTopNusDocsPsrel ; both original and fully expanded terms in topics; both original and expanded terms in documents, using best sense according to NUS word sense disambiguation; PRF.
- bilingual without WSD:
  Es2EnNowsd ; original terms in topics (in Spanish); translated terms in documents (from English to Spanish).

Es2EnNowsdPsrel ; same as `Es2EnNowsd`, but with PRF.

- bilingual with WSD:
  - **Es2EnNusDocsPsrel** ; original terms in topics (in Spanish); translated terms in documents, using the best sense according to NUS word sense disambiguation; PRF.
  - **Es2EnUbcDocsPsrel** ; original terms in topics (in Spanish); translated terms in documents, using the best sense according to UBC word sense disambiguation; PRF.
  - **Es2En1stTopsNusDocsPsrel** ; translated terms in topics (from Spanish to English) for first sense in Spanish; both original and expanded terms of the best sense according to NUS disambiguation data; PRF.
  - **Es2En1stTopsUbcDocsPsrel** ; translated terms in topics (from Spanish to English) for first sense in Spanish; both original and expanded terms of the best sense according to UBC disambiguation data; PRF.

The results show that the use of WSD data has been effective. With respect to monolingual retrieval, `En2EnUbcDocsPsrel` obtains the best results from our runs, although the difference with respect to `En2WnNowsdPsrel` is not statistically significant[1]. Regarding the bilingual results, `Es2En1stTopsUbcDocsPsrel` is the best, and the difference with respect to `Es2EnNowsdPsrel` is statistically significant. These results confirm the results that we obtained on the training data. Although not shown here, those results showed that the use of WSD led to significantly better results with respect to using all senses (full expansion).

Although it was not our main goal, our systems ranked high in the exercise, making the 7th best in the monolingual no-WSD subtask, 9th in monolingual using WSD, 5th best in the bilingual no-WSD subtask, and 1st in bilingual using WSD. Overall, our best runs ranked 4th overall and 3rd overall in the monolingual and bilingual subtasks, respectively.

Table 2. Results for submitted runs

		runId	map	gmap
monolingual	no WSD	En2EnNowsd	0.3534	0.1488
		En2EnNowsdPsrel	**0.3810**	0.1572
	with WSD	En2EnNusDocsPsrel	0.3862	0.1541
		En2EnUbcDocsPsrel	**0.3899**	0.1552
		En2EnFullStructTopsNusDocsPsrel	0.3890	0.1532
bilingual	no WSD	Es2EnNowsd	0.1835	0.0164
		Es2EnNowsdPsrel	**0.1957**	0.0162
	with WSD	Es2EnNusDocsPsrel	0.2138	0.0205
		Es2EnUbcDocsPsrel	0.2100	0.0212
		Es2En1stTopsNusDocsPsrel	0.2350	0.0176
		Es2En1stTopsUbcDocsPsrel	**0.2356**	0.0172

[1] We used paired Randomization Tests over MAPs with α=0.05

After analyzing the experiments and the results, we have found that the approach of expanding the documents works better than expanding the topics. The extensive experimentation that we performed on the use of structured queries did not yield better results than just expanding the documents.

In our experiments we did not make any effort to deal with hard topics, and we only paid attention to improvements in Mean Average Precision (MAP) metric. In fact, we applied the settings which proved best in training data according to MAP, and we did not pay attention to the Geometric Mean Average Precision (GMAP) values.

4 Related Work

Several teams have managed to successfully use word sense data. Stokoe et al. [6] developed a system that performed sense-based information retrieval which, when used in a large scale IR experiment, demonstrated improved precision over the standard term-based vector space model. They noted that with a word-sense disambiguation accuracy of only 62.1% the experiments showed an absolute increase of 1.73% and a relative increase over TF*IDF of 45.9%. The authors thing that their results support Gonzalo et al. [1] less conservative claim that a breakeven point of 50-60% would be adequate for improved IR performance.

Liu et al. [3] used WordNet to disambiguate word senses of query terms. They employed high-precision disambiguation of query terms for selective query expansion. Whenever the sense of a query term was determined, its synonyms, hyponyms, words from its definition and its compound words were considered for possible additions to the query. Experimental results showed that their approach yielded between 23% and 31% improvements over the best-known results on the TREC 9, 10 and 12 collections for short (title only) queries, without using Web data. In subsequent work [4], they showed that word sense disambiguation together with other components of their retrieval system yielded a result which was 13.7% above than produced by the same system but without disambiguation.

Kim et al. [2] assigned coarse-grained word senses defined in WordNet to query terms and document terms by an unsupervised algorithm which used co-occurrence information constructed automatically. Promising results were obtained when combined with pseudo relevance feedback and state-of-the-art retrieval functions such as BM25.

Finally, Pérez-Agüera and Zaragoza [5] devise a novel way to use word sense disambiguation data. They make explicit some of the term dependence information using a form of structured query, and use a ranking function capable of taking the structure information into account. They combined the use of query expansion techniques and semantic disambiguation to construct the structured queries, yielding queries that are both semantically rich and focused on the query. They report improved results on the same dataset reported here.

Compared to previous work, our own is less sophisticated, but we provide indications that word sense disambiguation on the documents, accompanied by expansion, produces better results than a similar strategy on the queries. All in all, our approach is complementary to other work, and suggests that experimentation on the document side can offer further improvements.

5 Conclusions and future work

We have reported our experiments for the Robust-WSD Track at CLEF. All our runs ended up in good ranking, taking into account that these have been our first experiments in the field of information retrieval. This is remarkable, as we did not use any external resources, except the WSD information and Spanish and English wordnets provided by the organizers. Note also that we did not do any proper parameter tuning (e.g. in the relevance feedback step) on the training part.

Our main goal was to get better (CL)IR results using WSD and we achieved it, obtaining remarkable gains in bilingual IR, and smaller gains in monolingual IR. We discovered that using WSD information for document expansion is a good strategy, in contrast to most previous IR work, which has focused on WSD of topics.

For the future, we plan to improve the bilingual results, mainly incorporating external resources like bilingual dictionaries. Our main goal will be to pursue more sophisticated methods for expansion and indexing of documents using WSD information, beyond the simple combinations tried in this paper.

Acknowledgements

This work has been supported by KNOW (TIN2006-15049-C03-01) and KYOTO (ICT-2007-211423). Arantxa Otegi's work is funded by a PhD grant from the Basque Government.

References

1. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing With WordNet Synsets Can Improve Text Retrieval. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (1998)
2. Kim, S., Seo, H., Rim, H.: Information retrieval using word senses: Root sense tagging approach. Proceedings of SIGIR 2004
3. Liu, S., Liu, F., Yu, C., Meng, W.: An effective approach to document retrieval via utilizing WordNet and recognizing phrases. Proceedings SIGIR 2004
4. Liu, S., Yu, C., Meng, W.: Word Sense Disambiguation in Queries. Proceedings of ACM Conference on Information and Knowledge Management, CIKM 2005
5. Pérez-Agüera, J.R., Zaragoza, H.: UCM-Y!R at CLEF 2008 Robust and WSD tasks. In this volume (2009)
6. Stokoe, C., Oakes, M.P., Tait, J.: Word Sense Disambiguation in Information Retrieval Revisited. Proceedings of SIGIR 2003
7. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. Proceedings of the International Conference on Intelligence Analysis (2005)

Using Semantic Relatedness and Word Sense Disambiguation for (CL)IR

Eneko Agirre[1], Arantxa Otegi[1], and Hugo Zaragoza[2]

[1] IXA NLP Group - University of Basque Country. Donostia, Basque Country.
{e.agirre,arantza.otegi}@ehu.es
[2] Yahoo! Researech, Barcelona, Spain.
hugoz@yahoo-inc.com

Abstract. In this paper we report the experiments for the CLEF 2009 Robust-WSD task, both for the monolingual (English) and the bilingual (Spanish to English) subtasks. Our main experimentation strategy consisted of expanding and translating the documents, based on the related concepts of the documents. For that purpose we applied a state-of-the art semantic relatedness method based on WordNet. The relatedness measure was used with and without WSD information. Even though we obtained positive results in our training and development datasets, we did not manage to improve over the baseline in the monolingual case. The improvement over the baseline in the bilingual case is marginal. We plan further work on this technique, which has attained positive results in the passage retrieval for question answering task at CLEF (ResPubliQA).

1 Introduction

Our goal is to test whether Word Sense Disambiguation (WSD) information can be beneficial for Cross Lingual Information Retrieval (CLIR) or monolingual Information Retrieval (IR). WordNet has been previously used to expand the terms in the query with some success [5] [6] [7] [9]. WordNet-based approaches need to deal with ambiguity, which proves difficult given the little context available to disambiguate the words in the query effectively. In our experience document expansion works better than topic expansion (see our results for the previous edition of CLEF in [8]). Bearing this in mind, in this edition we have mainly focused on documents, using a more elaborate expansion strategy. We have applied a state-of-the-art semantic relatedness method based on WordNet [3] in order to select the best terms to expand the documents. The relatedness method can optionally use the WSD information provided by the organizers.

The remainder of this paper is organized as follows. Section 2 describes the experiments carried out. Section 3 presents the results obtained and Section 4 analyzes the results. Finally, Section 5 draws conclusions and mentions future work.

2 Experiments

Our main experimentation strategy consisted of expanding the documents, based on the related concepts of the documents. The steps of our retrieval system are the following. We first expand/translate the topics. In a second step we extract the related concepts of the documents, and expand the documents with the words linked to these concepts in WordNet. Then we index these new expanded documents, and finally, we search for the queries in the indexes in various combinations. All steps are described sequentially.

2.1 Expansion and Translation Strategies of the Topics

WSD data provided to the participants was based on WordNet version 1.6. In the topics each word sense has a WordNet synset assigned with a score. Using those synset codes and the English and Spanish wordnets, we expanded the topics. In this way, we generated different topic collections using different approaches of expansion and translation, as follows:

- Full expansion of English topics: expansion to all synonyms of all senses.
- Best expansion of English topics: expansion to the synonyms of the sense with highest WSD score for each word, using either UBC or NUS disambiguation data (as provided by organizers).
- Translation of Spanish topics: translation from Spanish to English of the first sense for each word, taking the English variants from WordNet.

In both cases we used the Spanish and English wordnet versions provided by the organizers.

2.2 Query Construction

We constructed queries using the title and description topic fields. Based on the training topics, we excluded some words and phrases from the queries, such as *find, describing, discussing, document, report* for English and *encontrar, describir, documentos, noticias, ejemplos* for Spanish.

After excluding those words and taking only nouns, adjectives, verbs and numbers, we constructed several queries for each topic using the different expansions of the topics (see Section 2.1) as follows:

- Original words.
- Both original words and expansions for the best sense of each word.
- Both original words and all expansions for each word.
- Translated words, using translations for the best sense of each word. If a word had no translation, the original word was included in the query.

The first three cases are for the monolingual runs, and the last one for the bilingual run which translated the query.

2.3 Expansion and Translation Strategies of the Documents

Our document expansion strategy was based on semantic relatedness. For that purpose we used UKB[3], a collection of programs for performing graph-based Word Sense Disambiguation and lexical similarity/relatedness using a pre-existing knowledge base, in this case WordNet 1.6.

Given a document, UKB returns a vector of scores for each concept in WordNet. The higher the score, the more related is the concept to the given document. In our experiments we used different approaches to represent each document:

- using all the synsets of each word of the document.
- using only the synset with highest WSD score for each word, as given by the UBC disambiguation data [2] (provided by the organizers).

In both cases, UKB was initialized using the WSD weights: each synset was weighted with the score returned by the disambiguation system, that is, each concept was weighted according to the WSD weight of the corresponding sense of the target word.

Once UKB outputs the list of related concepts, we took the highest-scoring 100 or 500 concepts and expanded them to all variants (words in the concept) as given by WordNet. For the bilingual run, we took the Spanish variants. In both cases we used the Spanish and English wordnet versions provided by the organizers.

The variants for those expanded concepts were included in two new fields of the document representation; 100 concepts in the first field and 400 concepts in the second field. This way, we were able to use the original words only, or also the most related 100 concepts, or the original words and the most related 500 concepts. We will get back to this in Section 2.4 and Section 2.5.

Figure 2 shows a document expansion for the document in Figure 1. The second column in Figure 2 is the vector of related concepts (synsets values) returned by UKB for the mentioned document. The vector in the example is sorted by the score for each concept (first column). So the concepts that are shown on it are the most related concepts for that document. The words in the third column are the variants for each concept taken from WordNet. We also added these words to another index. The terms in bold in the example are the words that appear in the document. And the terms in italic are the new terms that we obtain by means of the expansion.

2.4 Indexing

We indexed the new expanded documents using the MG4J search-engine [4]. MG4J makes it possible to combine several indices over the same document collection. We created one index for each field: one for the original words, one for the expansion of the top 100 concepts, and another one for the expansion of the following 400 concepts. The Porter stemmer was used with default settings.

[3] The algorithm is publicly available at http://ixa2.si.ehu.es/ukb/

```
HUNTINGTON BANK ROBBERY NETS $780
A man walked into a bank Friday, warned a teller that he had a gun and
made off with $780, police said.
Huntington Beach Police Sgt. Larry Miller said the teller at the World
Savings and Loan Assn., 6902 Warner Ave., did not see a weapon during
the robbery, which occurred at 4:35 p.m.
The robber escaped out the west door of the building. Police have no
suspects in the case.
```

Fig. 1. Document example

0.0071192807	06093563 $-n \Longrightarrow$	*constabulary*, *law*, **police**, *police force*
0.007016694	02347413 $-n \Longrightarrow$	**building**, *edifice*
0.00701617062	07635368 $-n \Longrightarrow$	**teller**, *vote counter*
0.00700878272	06646591 $-n \Longrightarrow$	**huntington**
0.0070066648	00499726 $-n \Longrightarrow$	**robbery**
0.006932565	00235191 $-v \Longrightarrow$	*come about, go on, hap, happen*, **occur**, *pass, pass off, take place*
0.006929787	03601056 $-n \Longrightarrow$	*arm*, **weapon**, *weapon system*
0.006903118	01299603 $-v \Longrightarrow$	**walk**
0.006898292	02588950 $-n \Longrightarrow$	**door**
0.006894822	02778084 $-n \Longrightarrow$	**gun**
0.006892254	09651550 $-n \Longrightarrow$	**loan**
0.0068790509	06739108 $-n \Longrightarrow$	**beach**
0.0068660484	10937709 $-n \Longrightarrow$	**p.m.**, *pm, post meridiem*
0.006831742	10883362 $-n \Longrightarrow$	*fri*, **friday**
0.0068182234	07422992 $-n \Longrightarrow$	*mugger*, **robber**
0.00676897472	07410610 $-n \Longrightarrow$	**miller**
0.0058595173	00126393 $-n \Longrightarrow$	*economy*, **saving**
0.0055009496	00465486 $-v \Longrightarrow$	**suspect**
0.0053402969	00589833 $-v \Longrightarrow$	**warn**
0.005200375	07391044 $-n \Longrightarrow$	*adult male*, **man**
...	...	...

Fig. 2. Example for an expansion

2.5 Retrieval

We carried out several retrieval experiments combining different kind of queries with different kind of indices. We used the training data to perform extensive experimentation, and chose the ones with best MAP results in order to produce the test topic runs.

The different kind of queries that we had prepared are those explained in Section 2.2. Our experiments showed that original words were getting good results, so in the test runs we used only the queries with original words.

MG4J allows multi-index queries, where one can specify which of the indices one wants to search in, and assign different weights to each index. We conducted different experiments, by using the original words alone (the index made of original words) and also by using one or both indices with the expansion of concepts, giving different weight to the original words and the expanded concepts. The best weights were then used in the test set, as explained in the following Section.

We used the BM25 ranking function with the following parameters: 1.0 for *k1* and 0.6 for *b*. We did not tune these parameters.

The submitted runs are described in Section 3.

3 Results

Table 1 summarizes the results of our submitted runs. The IR process is the same for all the runs and the main differences between them is the expansion strategy. The characteristics of each run are as follows:

- monolingual without WSD:
  - **EnEnNowsd**: original terms in topics; original terms in documents.
- monolingual with WSD:
  - **EnEnAllSenses100Docs**: original terms in topics; both original and expanded terms of 100 concepts, using all senses for initializing the semantic graph. The weight of the index that included the expanded terms: 0.25.
  - **EnEnBestSense100Docs**: original terms in topics; both original and expanded terms of 100 concepts, using best sense for initializing the semantic graph. The weight of the index that included the expanded terms: 0.25.
  - **EnEnBestSense500Docs**: original terms in topics; both original and expanded terms of 500 concepts, using best sense for initializing the semantic graph. The weight of the index that included the expanded terms: 0.25.
- bilingual without WSD:
  - **EsEnNowsd**: translated terms in topics (from Spanish to English); original terms in documents (in English).
- bilingual with WSD:
  - **EsEn1stTopsAllSenses100Docs**: translated terms in topics (from Spanish to English); both original and expanded terms of 100 concepts, using all senses for initializing the semantic graph. The weight of the index that included the expanded terms: 0.15.
  - **EsEn1stTopsBestSense500Docs**: translated terms in topics (from Spanish to English); both original and expanded terms of 100 concepts, using best sense for initializing the semantic graph. The weight of the index that included the expanded terms: 0.15.
  - **EsEnAllSenses100Docs**: original terms in topics (in Spanish); both original terms (in English) and translated terms (in Spanish) in documents, using all senses for initializing the semantic graph. The weight of the index that included the expanded terms: 1.00.

- **EsEnBestSense500Docs**: original terms in topics (in Spanish); both original terms (in English) and translated terms (in Spanish) in documents, using best sense for initializing the semantic graph. The weight of the index that included the expanded terms: 1.60.

The weight of the index which was created using the original terms of the documents was 1.00 for all the runs.

Table 1. Results for submitted runs

		runId	map	gmap
monolingual	no WSD	EnEnNowsd	**0.3826**	0.1707
	with WSD	EnEnAllSenses100Docs	0.3654	0.1573
		EnEnBestSense100Docs	0.3668	0.1589
		EnEnBestSense500Docs	**0.3805**	0.1657
bilingual	no WSD	EsEnNowsd	**0.1805**	0.0190
	with WSD	EsEn1stTopsAllSenses100Docs	0.1827	0.0193
		EsEn1stTopsBestSense500Docs	**0.1838**	0.0198
		EsEnAllSenses100Docs	0.1402	0.0086
		EsEnBestSense500Docs	0.1772	0.0132

Regarding monolingual results, we can see that using the best sense for representing the document when initializing the semantic graph achieves slightly higher results with respect to using all senses. Besides, we obtained better results when we expanded the documents using 500 concepts than using only 100 (compare the results of the runs `EnEnBestSense100Docs` and `EnEnBestSense500Docs`). However, we did not achieve any improvement over the baseline with either WSD or semantic relatedness information. We have to mention that we did achieve improvement in the training data, but the difference was not significant[4].

With respect to the bilingual results, `EsEn1stTopsBestSense500Docs` obtains the best result, although the difference with respect to the baseline run is not statistically significant. This is different to the results obtained using the training data, where the improvements using the semantic expansion were remarkable (4.91% of improvement over MAP). It is not very clear whether translating the topics from Spanish to English or translating the documents from English to Spanish is better, since we got better results in the first case in the testing phase (see runs called `...1stTops...` in the Table 1), but not in the training phase.

In our experiments we did not make any effort to deal with hard topics, and we only paid attention to improvements in Mean Average Precision (MAP) metric. In fact, we applied the settings which proved best in training data according to MAP. Another option could have been to optimize the parameters and settings according to Geometric Mean Average Precision (GMAP) values.

[4] We used paired Randomization Tests over MAPs with α=0.05

4 Analysis

In this section we focus on comparison, on the one hand, between different approaches of using WSD data for IR, and on the other hand, between different collections used to test the document expansion strategies for IR.

The expansion strategy we used in the previous edition of the task consisted of expanding documents with synonyms based on WSD data and it provided consistent improvements over the baseline, both in monolingual and bilingual tasks [8]. With the document expansion strategy presented in this paper we achieve gains over the baseline in monolingual task using training data and in bilingual task both in training and testing phases.

With respect to using different datasets, we found that using semantic relatedness to expand documents can be effective for the passage retrieval task (ResPubliQA) [1]. The strategy used in it differs from the one explained here, as the expansion is done using the variants of the synsets, rather than the synsets themselves. After the competition, we applied this expansion strategy to the dataset of the Robust task and the monolingual results raised up to 0.3875.

5 Conclusions and Future Work

We have described our experiments and the results obtained in both monolingual and bilingual tasks at Robust-WSD Track at CLEF 2009. Our main experimentation strategy consisted of expanding the documents based on a semantic relatedness algorithm.

The objective of carrying out different expansion strategies was to study if WSD information and semantic relatedness could be used in an effective way in (CL)IR. After analyzing the results, we have found that those expansion strategies were not very helpful, especially in the monolingual task.

For the future, we want to analyze expansion using variants of the related concepts, as it attained remarkable improvements in the passage retrieval task (ResPubliQA) [1].

Acknowledgments

This work has been supported by KNOW2 (TIN2009-14715-C04-01) and KYOTO (ICT-2007-211423). Arantxa Otegi's work is funded by a PhD grant from the Basque Government. Part of this work was done while Arantxa Otegi was visiting Yahoo! Research Barcelona.

References

1. Agirre, E., Ansa, O., Arregi, X., Lopez de Lacalle, M., Otegi, A., Saralegi, X., Zaragoza, H.: Elhuyar-IXA: Semantic Relatedness and Cross-Lingual Passage Retrieval. In this volume (2010)

2. Agirre, E., Lopez de Lacalle, O.: UBC-ALM: Combining k-NN with SVD for WSD. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic (2007) 341345
3. Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M.: A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. Proceedings of annual meeting of the North American Chapter of the Association of Computational Linguistics (NAACL), Boulder, USA (2009)
4. Boldi, P., Vigna, S.: MG4J at TREC 2005. The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, NIST Special Publications, SP 500-266 (2005). `http://mg4j.dsi.unimi.it/`
5. Kim, S., Seo, H., Rim H.: Information Retrieval using word senses: Root sense tagging approach. Proceedings of SIGIR 2004
6. Liu, S., Liu, F., Yu, C., Meng, W.: An effective approach to document retrieval via utilizing WordNet and recognizing phrases. Proceedings of SIGIR 2004
7. Liu, S., Yu, C., Meng, W.: Word Sense Disambiguation in Queries. Proceedings of ACM Conference on Information and Knowledge Management (CIKM) (2005)
8. Otegi, A., Agirre, E., Rigau, G.: IXA at CLEF 2008 Robust-WSD Task: using Word Sense Disambiguation for (Cross Lingual) Information Retrieval. Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science Vol. 5706 (2009)
9. Pérez-Agüera, J.R., Zaragoza, H.: Query Clauses and Term Independence. Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science Vol. 5706 (2009)

Elhuyar-IXA: Semantic Relatedness and Cross-Lingual Passage Retrieval

Eneko Agirre[1], Olatz Ansa[1], Xabier Arregi[1], Maddalen Lopez de Lacalle[2],
Arantxa Otegi[1], Xabier Saralegi[2], Hugo Zaragoza[3]

[1] IXA NLP Group, University of the Basque Country. Donostia, Basque Country
{e.agirre,olatz.ansa,xabier.arregi,arantza.otegi}@ehu.es
[2] R&D, Elhuyar Foundation. Usurbil, Basque Country
{maddalen,xabiers}@elhuyar.com
[3] Yahoo! Research. Barcelona, Spain
hugoz@yahoo-inc.com

Abstract. This article describes the participation of the joint Elhuyar-IXA group in the ResPubliQA exercise at QA&CLEF. In particular, we participated in the English–English monolingual task and in the Basque–English cross-lingual one. Our focus has been threefold: (1) to check to what extent information retrieval (IR) can achieve good results in passage retrieval without question analysis and answer validation, (2) to check Machine Readable Dictionary (MRD) techniques for the Basque to English retrieval when faced with the lack of parallel corpora for Basque in this domain, and (3) to check the contribution of semantic relatedness based on WordNet to expand the passages to related words. Our results show that IR provides good results in the monolingual task, that our crosslingual system performs lower than the monolingual runs, and that semantic relatedness improves the results in both tasks (by 6 and 2 points, respectively).

Keywords: Cross-lingual passage retrieval, semantic relatedness, MRD, word concurrences.

1 Introduction

The joint team was formed by two different groups, on the one hand the Elhuyar Foundation, and on the other hand the IXA NLP group. This collaboration allowed us to tackle the English–English monolingual task and the Basque–English cross-lingual one in the ResPubliQA track.

With respect to the Basque-English task, we met the challenge of retrieving English passages for Basque questions. We tackled this problem by translating the lexical units of the questions into English. The main setback is that no parallel corpus was available for this pair of languages, given that there is no Basque version of the JRC-Acquis collection. So we have explored an approach which does not use parallel corpora when translating queries, which could also be interesting for other less resourced languages. In our opinion, bearing in mind the idiosyncrasy of the

European Union, it is worthwhile dealing with the search of passages that answer questions formulated in unofficial languages.

Question answering systems typically rely on a passage retrieval system. Given that passages are shorter than documents, vocabulary mismatch problems are more important than in full document retrieval. Most of the previous work on expansion techniques has focused on pseudo-relevance feedback and other query expansion techniques. In particular, WordNet has been used previously to expand the terms in the query with little success [2, 3, 4]. The main problem is ambiguity, and the limited context available to disambiguate the word in the query effectively. As an alternative, we felt that passages would provide sufficient context to disambiguate and expand the terms in the passage. In fact, we do not do explicit word sense disambiguation, but rather apply a state-of-the-art semantic relatedness method [5] in order to select the best terms to expand the documents.

2 System Overview

2.1 Question Pre-processing

We analysed the Basque questions by re-using the linguistic processors of the *Ihardetsi* question-answering system [1]. This module uses two general linguistic processors: the lemmatizer/tagger named *Morfeus* [6], and the Named Entity Recognition and Classification (NERC) processor called *Eihera* [7]. The use of the lemmatizer/tagger is particularly suited to Basque, as it is an agglutinative language. It returns only one lemma and one part of speech for each lexical unit, which includes single word terms and multiword terms (MWTs) (those included in the Machine Readable Dictionary (MRD) introduced in the next subsection). The NERC processor, *Eihera*, captures entities such as *person*, *organization* and *location*. The numerical and temporal expressions are captured by the lemmatizer/tagger. The questions thus analyzed are passed to the translation module.

English queries were tokenized without further analysis.

2.2 Translation of the Query Terms (Basque-English Runs)

Once the questions had been linguistically processed, we translated them into English. Due to the scarcity of parallel corpora for a small language or even for big languages in certain domains, we have explored a MRD-based method. These approaches have inherent problems, such as the presence of ambiguous translations and also out-of-vocabulary (OOV) words. To tackle these problems, some techniques have been proposed such as structured query-based techniques [8, 9] and concurrences-based techniques [10, 11]. These approaches have been compared for Basque by obtaining best MAP (Mean Average Precision) results with structured queries [12]. However, structured queries were not supported in the retrieval algorithm used (see Section 2.3), so we adopted a concurrences-based translation selection strategy.

The translation process designed comprises two steps and takes the keywords (Name Entities, MWTs and single words tagged as noun, adjective or verb) of the question as source words.

In the first step the translation candidates of each source word are obtained. The translation candidates for the lemmas of the source words are taken from a bilingual eu-en MRD composed from the Basque-English *Morris* dictionary[1], and the *Euskalterm* terminology bank[2] which includes 38,184 MWTs. After that, OOV words and ambiguous translations are dealt with. The number of OOV words quantified out of a total of 421 keywords for the 77 questions of the development set was 42 (10%). Nevertheless, it must be said that many of these OOV words were wrongly tagged lemmas and entities. We deal with OOV words by searching for their cognates in the target collection. The cognate detection is done in two phases. Firstly, we apply several transliteration rules to the source word. Then we calculate the Longest Common Subsequence Ratio (LCSR) among words with a similar length (+-10%) from the target collection (see Figure 1). The ones which reach a previously established threshold (0.9) are selected as translation candidates. The MWTs that are not found in the dictionary are translated word by word, as we realized that most of the MWTs could be translated correctly in that way, exactly 91% of the total MWTs identified by hand in the 77 development questions.

err- ---> r- *erradioterapeutiko=radioterapeutiko*

k ---> c *radioterapeutiko=radioterapeutico*

LCSR(radioterapeutico, radioterapeutic) = 0.9375

Fig. 1. Example of cognate detection.

In the second step, we select the best translation of each source keyword according to an algorithm based on target collection concurrences. This algorithm sets out to obtain the translation candidate combination that maximizes their global association degree. We take the algorithm proposed by Monz and Dorr [11]

Initially, all the translation candidates are equally likely. Assuming that t is a translation candidate of the set of all candidates $tr(s_i)$ for a query term s_i given by the MRD, then:

Initialization step:

$$w_T^0(t \mid s_i) = \frac{1}{|tr(s_i)|} \tag{1}$$

In the iteration step, each translation candidate is iteratively updated using the weights of the rest of the candidates and the weight of the link connecting them.

[1] English/Basque dictionary including 67,000 entries and 120,000 senses.

[2] Terminological dictionary including 100,000 terms in Basque with equivalences in Spanish, French, English and Latin.

Iteration step:

$$w_T^n(t \mid s_i) = w_T^{n-1}(t \mid s_i) + \sum_{t' \in inlink(t)} w_L(t, t') \cdot w_T^{n-1}(t' \mid s_i) \qquad (2)$$

where $inlink(t)$ is the set of translation candidates that are linked to *t*, and $w_L(t, t')$ is the association degree between *t* and *t'* on the target passages measured by Log-likelihood ratio. These concurrences were calculated by taking the target passages as window.

After re-computing each term weight they are normalized.

Normalization step:

$$w_T^n(t \mid s_i) = \frac{w_T^n(t \mid s_i)}{\sum_{m=1}^{|tr(s_i)|} w_T^n(t_{i,m} \mid s_i)} \qquad (3)$$

The iteration stops when the variations of the term weights become smaller than a predefined threshold.

We have modified the iteration step by adding a factor $w_F(t, t')$ to increase the association degree $w_L(t, t')$ between translation candidates *t* and *t'* whose corresponding source words $so(t), so(t')$ are close to each other (distance *dis* in words is low) in the source query *Q*, or even belong to the same Multi-Word Unit ($smw(so(t), so(t')) = 1$). As the global association degree between translation candidates is estimated from the association degree of pairs of candidates, we score positively these two characteristics when the association degree for a pair of candidates is calculated. Thus, the modified association degree $w'_L(t, t')$ between *t* and *t'* will be calculated in this way:

$$w'_L(t, t') = w_L(t, t') \cdot w_F(t, t') \qquad (4)$$

$$w_F(t, t') = \frac{\max_{s_i, s_j \in Q} dis(s_i, s_j)}{dis(so(t), so(t'))} \cdot 2^{smw(so(t), so(t'))} \qquad (5)$$

$$smw(s, s') = \begin{cases} 1 & \{s, s'\} \subseteq Z \quad \text{where } Z \in MWU \\ 0 & \end{cases} \qquad (6)$$

2.3 Passage Retrieval

The purpose of the passage retrieval module is to retrieve passages from the document collection which are likely to contain an answer. The main feature of this module is that the passages are expanded based on their related concepts, as explained in the following sections.

2.3.1 Document Preprocessing and Application of Semantic Relatedness

Given that the system needs to return paragraphs, we first split the document collection into paragraphs. Then we lemmatized and part-of-speech (POS) tagged those passages using the OpenNLP open source software[3].

After preprocessing the documents, we expanded the passages based on semantic relatedness. To this end, we used UKB[4], a collection of programs for performing graph-based Word Sense Disambiguation and lexical similarity/relatedness using a pre-existing knowledge base [5], in this case WordNet 3.0.

Given a passage (represented using the lemmas of all nouns, verbs, adjectives and adverbs), UKB returns a vector of scores for concepts in WordNet. Each of these concepts has a score, and the higher the score, the more related the concept is to the given passage. Given the list of related concepts, we took the highest-scoring 100 concepts and expanded them to all variants (words that lexicalize the concepts) in WordNet. An example of a document expansion is shown in Figure 2.

We applied the expansion strategy only to passages which had more than 10 words (half of the passages), for two reasons: the first one is that most of these passages were found not to contain relevant information for the task (e.g. “Article 2”, “Having regard to the proposal from the Commission” or “HAS ADOPTED THIS REGULATION”), and the second is that we thus saved some computation time.

2.3.2 Indexing

We indexed the new expanded documents using the MG4J search-engine [13]. MG4J makes it possible to combine several indices over the same document collection. We created one index for the original words and another one with the variants for the most related 100 concepts. This way, we were able to use the original words only, or alternatively, to also include the expanded words during the retrieval. Porter stemmer was used.

2.3.3 Retrieval

We used the BM25 ranking function with the following parameters: 1.0 for *k1* and 0.6 for *b*. We did not tune these parameters. MG4J allows multi-index queries, where one can specify which of the indices one wants to search in, and assign different weights to each index. We conducted different experiments, by using only the index made of original words and also by using the index with the expansion of concepts, giving different weights to the original words and the expanded concepts. The weight of the index which was created using the original words from the passages was 1.00 for all the runs. 1.00 was also the weight of the index that included the expanded words for the monolingual run, but it was 1.78 for the bilingual run. These weights were fixed following a training phase with the English development questions provided by the organization, and after the Basque questions had been translated by hand (as no development Basque data was released).The submitted runs are described in the next section.

[3] http://opennlp.sourceforge.net/

[4] The algorithm is publicly available at http://ixa2.si.ehu.es/ukb/

3 Description of Runs

We participated in the English-English monolingual task and the Basque-English cross-lingual task. We did not analyze the English queries for the monolingual run, and we just removed the stopwords. For the bilingual runs, we first analyzed the questions (see Section 2.1), then we translated the question terms from Basque to English (see Section 2.2), and, finally, we retrieved the relevant passages for the translated query terms (see Section 2.3).

As we were interested in the performance of passage retrieval on its own, we did not carry out any answer validation, and we just chose the first passage returned by the passage retrieval module as the response. We did not leave any question unanswered.

For both tasks, the only difference between the submitted two runs is the use (or not) of the expansion in the passage retrieval module. That is, in the first run ("run 1" in Table 1), during the retrieval we only used the original words that were in the passage. In the second run ("run 2" in Table 1), apart from the original words, we also used the expanded words.

4 Results

Table 1 summarizes the results of our submitted runs, explained in Section 3.

Table 1. Results for submitted runs.

submitted runs		#answered correctly	#answered incorrectly	c@1
English - English	run 1	211	289	0.42
	run 2	240	260	**0.48**
Basque - English	run 1	78	422	0.16
	run 2	91	409	**0.18**

The results show that the use of the expanded words (run 2) was effective for both tasks, improving the final result by 6 % in the monolingual task.

Figure 2 shows an example of a document expansion which was effective for answering the English question number 32: "*Into which plant may genes be introduced and not raise any doubts about* <u>*unfavourable consequences*</u> *for people's health?*"

In the second part of the example we can see some words that we obtained after applying the expansion process explained in Section 2.3.1 to the original passage showed in the example too. As we can see, there are some new words among the expanded words that are not in the original passage, such as *unfavourable* or *consequence*. Those two words were in the question we mentioned before (number

32). That could be why we answered that question correctly when using the expanded words (in run 2), but not when using the original words only.

original passage: *Whereas the Commission, having examined each of the objections raised in the light of Directive 90/220/EEC, the information submitted in the dossier and the opinion of the Scientific Committee on Plants, has reached the conclusion that there is no reason to believe that there will be any adverse effects on human health or the environment from the introduction into maize of the gene coding for phosphinotricine-acetyl-transferase and the truncated gene coding for beta-lactamase;*

some expanded words: *cistron factor gene coding cryptography secret_writing ... acetyl acetyl_group acetyl_radical ethanoyl_group ethanoyl_radical beta_lactamase penicillinase ... ec eec eu europe european_community european_economic_community european_union ... directive directing directional guiding citizens_committee committee environment environs surround surroundings corn ... maize zea_mays health wellness health adverse contrary homo human human_being man adverse inauspicious untoward gamboge ... unfavorable* ***unfavourable*** *... set_up expostulation objection remonstrance remonstration dissent protest believe light lightly belief feeling impression notion opinion ... reason reason_out argue jurisprudence law* ***consequence*** *effect event issue outcome result upshot ...*

Fig. 2. Example of a document expansion (doc_id: *jrc31998D0293-en.xml,* p_id*: 17)*.

As expected, the best results were obtained in the monolingual task. With the intention of finding reasons to explain the significant performance drop in the bilingual run, we analyzed manually 100 query translations obtained in the query translation process of the 500 test queries, and detected several types of errors arising from both the question analysis process and from the query translation process. In the question analysis process, some lemmas were not correctly identified by the lemmatizer/tagger, and in other cases some entities were not returned by the lemmatizer/tagger causing us to lose important information for the subsequent translation and retrieval processes. In the query translation process, leaving aside the incorrect translation selections, the words appearing in the source questions were not exactly the ones that figured in many queries that had been correctly translated. In most cases this happened because the English source query word was not a translation candidate in the MRD. If we assume that the answers contain words that appear in the questions and therefore in the passage that we must return, this will negatively affect the final retrieval process.

5 Conclusions

The joint Elhuyar-Ixa team has presented a system which works on passage retrieval alone, without any question analysis and answer validation steps. Our English-English results show that good results can be achieved by means of this simple strategy. We experimented with applying semantic relatedness in order to expand passages prior to indexing, and the results are highly positive, especially for English-English. The performance drop in the Basque-English bilingual runs is significant, and is caused by the accumulation of errors in the analysis and translation of the query mentioned.

Acknowledgments

This work has been supported by KNOW (TIN2006-15049-C03-01), imFUTOURnet (IE08-233) and KYOTO (ICT-2007-211423). Arantxa Otegi's work is funded by a PhD grant from the Basque Government. Part of this work was done while Arantxa Otegi was visiting Yahoo! Research Barcelona.

References

1. Ansa, O., Arregi, X., Otegi, A., Soraluze. A.: Ihardetsi: A Basque Question Answering System at QA@CLEF 2008. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, pp. 369-376. ISSN 0302-9743 ISBN 978-3-642-04446. (2009)
2. Kim, S., Seo, H., Rim, H.: Information retrieval using word senses: Root sense tagging approach. In: Proceedings of SIGIR. (2004)
3. Liu, S., Yu, C., Meng, W.: Word Sense Disambiguation in Queries. In: Proceedings of the 14th ACM Conference on Information and Knowledge Management, CIKM. (2005)
4. Pérez-Agüera, J.R., Zaragoza, H.: Query Clauses and Term Independence. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, pp. 369-376. ISSN 0302-9743 ISBN 978-3-642-04446. (2009)
5. Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M.: A study on similarity and relatedness using distributional and WordNet-based approaches. In: Proceedings of the annual meeting of the North American Chapter of the Association of Computational Linguistics (NAACL), Boulder, USA (2009)
6. Ezeiza, N., Aduriz, I., Alegria, I., Arriola, J.M., Urizar, R.: Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In: COLING-ACL, pp.380–384. (1998)
7. Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., Urizar. R.: Development of a Named Entity Recognizer for an Agglutinative Language. In: IJCNLP, (2004)
8. Darwish, K., Oard., D.W.: Probabilistic structured Query Methods. In: Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 338–344. (2003)
9. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55-63. (1998)
10. Ballesteros, L., Bruce Croft, W.: Resolving Ambiguity for Cross-language Retrieval. In: Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71. (1998)
11. Monz, C., Dorr, B.J.: Iterative translation disambiguation for cross-language Information Retrieval. In: Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 520-527. (2005)
12. Saralegi, X., López de Lacalle, M.: Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries v. Target Co-occurrence Based Selection. In: 6th TIR workshop. (2009)
13. Boldi, P., Vigna, S.: MG4J at TREC 2005. In: Voorhees, E.M., Buckland, L.P. (eds.) The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, number SP 500-266 in Special Publications. NIST. http://mg4j.dsi.unimi.it/. (2005)

Document Expansion for Cross-Lingual Passage Retrieval

Eneko Agirre[1], Olatz Ansa[1], Xabier Arregi[1], Maddalen Lopez de Lacalle[2],
Arantxa Otegi[1], Xabier Saralegi[2]

[1] IXA NLP Group, University of the Basque Country. Donostia, Basque Country
{e.agirre,olatz.ansa,xabier.arregi,arantza.otegi}@ehu.es

[2] R&D, Elhuyar Foundation. Usurbil, Basque Country
{m.lopezdelacalle,x.saralegi}@elhuyar.com

Abstract. This article describes the participation of the joint Elhuyar-IXA group in the ResPubliQA exercise at QA&CLEF 2010. In particular, we participated in the English–English monolingual task and in the Basque–English cross-lingual one. Our focus was threefold: (1) to check to what extent information retrieval (IR) can achieve good results in passage retrieval without question analysis and answer validation, (2) to check dictionary techniques for Basque to English retrieval when faced with the lack of parallel corpora for Basque in this domain, and (3) to check the contribution of semantic relatedness based on WordNet to expand the passages to related words. Our results show that IR provides good results in the monolingual task, that our performance drop in the cross-lingual system was much greater than in previous CLIR experiments, and that expansion improves the results in the monolingual task.

Keywords: Cross-lingual passage retrieval, semantic relatedness, word co-occurrences.

1 Introduction

Like last year, the team consisted of two different groups: the Elhuyar Foundation, and the IXA NLP group. Last year we participated in the CLEF 2009 ResPubliQA task by submitting two English-English monolingual runs and two Basque-English cross-lingual runs. It should be mentioned that we were the only team who participated in a cross-lingual task.

Following the positive experience of last year's participation it seemed interesting to continue sharing our experience and knowledge on QA-oriented (CL)IR. Like last year, we participated in the English-English monolingual task and Basque-English cross-lingual task.

With respect to the Basque-English task, we met the challenge of retrieving English passages for Basque questions. We tackled this problem by translating the lexical units of the questions into English. The main setback is that no parallel corpus was available for this pair of languages, given that there is no Basque version of the JRC-Acquis and the Europarl collections. So we explored an approach which does not use parallel corpora when translating queries, which could also be interesting for other less resourced languages. In our opinion, bearing in mind the idiosyncrasy of the European Union, it is worthwhile tackling the search for passages that answer questions formulated in non-official languages.

Question answering systems typically rely on a passage retrieval system. Given that passages are shorter than documents, vocabulary mismatch problems are more significant than in full document retrieval. Most of the previous work on expansion techniques has focused on pseudo-relevance feedback and other query expansion techniques. In particular, WordNet has been used previously to expand the terms in the query with little success [1, 2, 3]. The main problem is ambiguity, and the limited context available to disambiguate the word in the query effectively. As an alternative, we felt intuitively that passages would provide sufficient context to disambiguate and expand the terms in the passage. In fact, we did not do explicit word sense disambiguation, but rather applied a state-of-the-art semantic relatedness method [4] in order to select the best terms to expand the documents.

2 System Overview

2.1 Question pre-processing

We analysed the Basque questions by re-using the linguistic processors included in the *Ihardetsi* question-answering system [5]. This system uses two general linguistic processors: the lemmatizer/tagger named *Morfeus* [6], and the Named Entity Recognition and Classification (NERC) processor called *Eihera* [7]. The use of the lemmatizer/tagger is particularly suited to Basque, as it is an agglutinative language. It provides the corresponding lemma and part of speech of each lexical unit, which also includes both single words and multiword units (MWU). The numerical and temporal expressions are also captured by the lemmatizer/tagger. The NERC processor, Eihera, captures entities such as persons, organizations and locations. The questions thus analyzed are passed to the translation module once the function words are removed. In the case of English, queries were just tokenized without further analysis.

2.2 Translation of the query terms (Basque-English runs)

Once the questions had been linguistically processed, they were translated into English using a dictionary-based method. According to the literature, parallel corpora-based translation methods provide the best translation quality, but these are scarce for small languages like Basque or even for major languages in certain domains. So, a dictionary-based translation approach was chosen. To tackle translation ambiguity produced by the dictionary translation, some techniques have been proposed in the literature, such as structured query-based techniques [8, 9] and co-occurrences-based techniques [10, 11, 12]. According to previous pieces of work [13], structured queries offer better MAP than co-occurrences-based methods on Basque-English CLIR experiments only when dealing with long queries [13]. However, the questions to evaluate in ResPubliQA are short, and structured queries were not supported in the retrieval algorithm used (see Section 2.4), so we adopted a co-occurrences-based translation selection strategy. The dictionary-based translation process designed comprises two main steps, taking the keywords (named entities, MWU and single words tagged as noun, adjective or verb) of the question as source words:

1. Obtaining translation candidates: In the first step the translation candidates of each source word are obtained from a bilingual eu-en dictionary comprising the Basque-English Morris dictionary[1], and the Euskalterm terminology bank[2] which includes 38,184 MWUs. After that, Out-Of-Vocabulary words are solved by searching for their cognates in the target collection. The cognate detection is done in two phases. First, several transliteration rules are applied to the source word. Then, the Longest Common Subsequence Ratio is calculated with respect to all the words from the target collection. Those that reach a previously established threshold (0.9) are selected as translation candidates.

2. Solving ambiguous candidates: The selection of the best translation for each source keyword is performed by an algorithm based on the maximum association degree, explained on detail in [14]. The association degree is computed by calculating co-occurrences of word pairs in the target collection. The algorithm obtains the set of translation candidates that maximizes the association degree between each other in the target collection. This maximization problem is solved by an Expectation Maximization-type greedy algorithm made up of initialization, iteration and normalization steps:

Initially, all the translation candidates provided by the dictionary are equally likely.

In the **iteration** step, the weight of each translation candidate is iteratively updated according to the association degree it has regarding the rest of the source word translation candidates. This association degree is pondered using the weights obtained on the previous iteration. The association degree between two translation candidates is measured by the Log-likelihood ratio using the target collection as a corpus. A factor is included in order to increase the association degree between translation candidates whose source words are near each other in the source query, and whose source words belong to the same MWU.

Finally, after re-computing each term weight, all of them are normalized. The algorithm stops when the difference between the term weights corresponding to previous and current iteration become lower than a predefined threshold.

[1] English/Basque dictionary including 67,000 entries and 120,000 senses.

[2] Terminological dictionary including 100,000 terms in Basque with equivalences in Spanish, French, English and Latin.

2.3 Document Pre-processing and Expansion

Given that the aim of the task was to retrieve a paragraph that contains an answer for each question, we first split the document collection into paragraphs.

One of the main features of our system is that the passages are expanded based on their related concepts according to the background information in WordNet [15]. We selected those concepts that are most closely related to the passage as a whole. For this purpose, we used a technique based on random walks over the graph representation of WordNet 3.0 concepts and relations [4], whose implementation is publicly available[3].

Given a passage and the graph-based representation of WordNet, we obtained a ranked list of WordNet concepts as follows:

1. We first pre-processed the passage to obtain the lemmas and parts of speech of the open category words using the OpenNLP open source software[4]. It should be noted that the lemmatizer/tagger Morfeus used for Basque questions works only with the Basque language.
2. We then assigned a uniform probability distribution to the terms found in the passage. The rest of the nodes were initialized to zero.
3. We computed personalized PageRank [16] over the graph, using the previous distribution as the reset distribution, and producing a probability distribution over WordNet concepts. The higher the probability for a concept, the more related it is to the given passage.

In order to select the expansion terms, we chose the 100 highest scoring concepts, and got all the words that lexicalize the given concept. An example of a document expansion is shown in Fig. 1.

We applied the expansion strategy only to passages which had more than 10 words, for two reasons: the first one was that most of the shorter passages were found not to contain relevant information for the task (e.g. “Article 2” or “Having regard to the proposal from the Commission”), and the second was that we thus saved some computation time.

The same expansion strategy has been used in some of our previous work with promising results [17].

2.4 Including Expansions in a Retrieval System

Once we had the list of words for document expansion, we created one index for the words in the original documents and another index with the expansion terms. We used the MG4J search engine [18] as it enables several indices over the same document collection to be combined. This way, we were able to use the original words only, or to include the expansion words during retrieval as well.

We used the BM25 ranking function, which has two free parameters (b and k_1) [19]. In the implementation of BM25 of the MG4J search engine, the two indices are

[3] http://ixa2.si.ehu.es/ukb/

[4] http://opennlp.sourceforge.net/

combined linearly, where the relative weight of the expanded index can be specified setting up the free λ parameter. Further information about the scoring function and the combination of the index we used can be found in [17].

3 Experimental Setup

We participated in the English-English monolingual task and the Basque-English cross-lingual task. For the monolingual run, we did not analyze the English questions, we carried out the passage retrieval only after expanding the documents, as explained in Sections 2.3 and 2.4. For the bilingual runs, we first analyzed the questions (see Section 2.1), then we translated the question terms from Basque to English (see Section 2.2), and, finally, we retrieved the relevant passages for the translated query terms (see Sections 2.3 and 2.4). For both languages, stop words were removed from the queries and a stemming pre-process based on the Porter algorithm was applied to the query and document words.

As we were interested in the performance of passage retrieval on its own, we did not carry out any answer validation, and we just chose the first passage returned by the passage retrieval module as the response. We did not leave any question unanswered.

For both tasks, the only difference between the two runs submitted is the use (or not) of the expansion in the passage retrieval phase. In other words, in the first run (referenced as “run 1” in the tables throughout this paper), apart from the original words that were in the passages, we also used the expanded words during the retrieval. In the second run (referenced as “run 2” in the tables throughout this paper), we only used the original words that were in the passages.

The BM25 parameters and the λ parameter (see Section 2.4) for both languages were fixed after a training phase with the question set from the previous edition of ResPubliQA [20]. Table 1 lists the parameter values used for each run.

Table 1**.** Free parameters described in Section 2.4. λ is not used in run 2.

Submitted runs		b	k_1	λ
English - English	run 1	0.17	0.30	0.22
	run 2	0.09	0.53	-
Basque - English	run 1	0.35	0.34	0.57
	run 2	0.71	0.23	-

4 Results

This section describes the results obtained in our ResPubliQA 2010 participation and discusses the performance of our document expansion approach and the translation of query terms approach.

Table 2 shows the official results of the four runs we submitted. The Mean Reciprocal Rank (MRR) measure is also shown in the table. We use * to indicate statistical significance at 99% confidence level, based on the Paired Randomization Test [21].

Table 2. Results for submitted runs

Submitted runs		#answered correctly	#answered incorrectly	c@1	MRR
English - English	run 1	130	70	**0.65**	**0.6067***
	run 2	123	77	0.62	0.5658
Basque - English	run 1	66	134	0.33	0.2742
	run 2	72	128	**0.36**	**0.2958**

Table 3 lists, for each language pair, the number of questions answered correctly in run 1 alone (i.e. using expansions), in run 2 alone (i.e. not using expansions) and in both runs, respectively.

Table 3. Comparison between the two runs per language pair

Language pairs	#answered correctly only in run 1	#answered correctly only in run 2	#answered correctly in both runs
English - English	9	2	121
Basque - English	5	11	61

4.1 Analysis of the Document Expansion Approach

Regarding monolingual results ("English-English" row in Table 2), we can see that the number of correct answers is higher in run 1 than in run 2. Since the only difference between the two runs was that run 1 used expanded words of the passages, the results indicate that the use of document expansion is beneficial. It should be noted that the improvement in MRR in run 1 compared with run 2 is statistically significant. To be precise, the correct answer set in run 1 was 130, and 123 in run 2, where the intersection of both sets was 121 (see Table 3).

The results of cross-lingual runs ("Basque-English" row in Table 2) show that the use of the expanded words did not improve the results, but the differences between both runs are not statistically significant. To our surprise, 72 questions were correctly answered without expansion, 6 more than when it was used. However, the answers to 5 questions were only found by the run enriched with expansions (see Table 3). As we obtained improvements using expansions in the training phase and also at ResPubliQA 2009 [14], further analysis of our cross-lingual approach is needed in order to determine why the use of expanded words is favourable only for some settings.

Fig. 1 shows an example of a document expansion which was effective for answering the English question number 32 of the training set: "*Into which plant may genes be introduced and not raise any doubts about unfavourable consequences for people's health?*"

In the second part of the example we can see some words that we obtained after applying the expansion process explained in Section 2.3 to the original passage also shown in the example. As we can see, there are some new words among the expanded words that are not in the original passage, such as *unfavourable* or *consequence*. Those two words were in the question referred to above (number 32). That could be why our system answered that question correctly when using the expanded words, but not when using the original words alone.

original passage: *Whereas the Commission, having examined each of the objections raised in the light of Directive 90/220/EEC, the information submitted in the dossier and the opinion of the Scientific Committee on Plants, has reached the conclusion that there is no reason to believe that there will be any adverse effects on human health or the environment from the introduction into maize of the gene coding for phosphinotricine-acetyl-transferase and the truncated gene coding for beta-lactamase;*

some expanded words: *cistron factor gene coding cryptography secret_writing ... acetyl acetyl_group acetyl_radical ethanoyl_group ethanoyl_radical beta_lactamase penicillinase ... ec eec eu europe european_community european_economic_community european_union ... directive directing directional guiding citizens_committee committee environment environs surround surroundings corn ... maize zea_mays health wellness health adverse contrary homo human human_being man adverse inauspicious untoward gamboge ... unfavorable* ***unfavourable*** *... set_up expostulation objection remonstrance remonstration dissent protest believe light lightly belief feeling impression notion opinion ... reason reason_out argue jurisprudence law* ***consequence*** *effect event issue outcome result upshot ...*

Fig. 1. Example of a document expansion (doc_id: *jrc31998D0293-en.xml,* p_id*: 17)*.

4.2 Analysis of the Query Terms Translation Approach

Compared with the monolingual run, the cross-lingual task yielded worse results. 50% of the monolingual performance was achieved for run 1, and 58% for run 2 (see table 3). This drop in performance for the cross-lingual task is worse than the one reported in a similar CLIR experiment [22] with the same cross-lingual method, where 74% of monolingual results were achieved. In that work, the drop in performance in our system was produced mainly because of the lack of recall of the dictionary. The source word appeared on the dictionary, but translations for the corresponding sense did not. This case falls between ambiguity and Out-Of-Vocabulary word. In the experiment carried out in this paper, in addition to the dictionary recall problem, many Out-Of-Vocabulary words corresponding to acronyms were detected. This adversely affects the retrieval performance since the cognate-based method does not solve them. Irrespective of the translation method, the accumulation of errors (i.e. question analysis, automatic lemmatization and entities detection) is another factor which explains the deterioration in the system performance in the cross-lingual task.

Despite this difference between the monolingual and cross-lingual task, some questions were answered correctly only in the cross-lingual runs (see Table 4).

Table 4. Number of questions answered correctly in the monolingual run alone, in the cross-lingual run alone, and in both runs

	Number of questions answered correctly		
	Only in the Monolingual Run	Only in the Cross-lingual Run	In both runs
run 1	75	11	55
run 2	64	13	59

We compared the translations of the test questions provided by our system with the source English questions. Our system translations helped to retrieve the correct passage in those cases because of the following isolated reasons:

a) Some relevant Out-Of-Vocabulary words are translated by cognate detection as they appear spelled in the correct passage (e.g. "*Zimmerman*" was translated to "*Zimmermann*" instead of "*Zimmerman*" as in the source English question).

b) Some words are translated as they appear in the correct passage, but different from spelling in the source English question (e.g. in question number 42, the Basque keyword "*zuzendari*" was translated by our system into "*manager*" which appears in the correct passage, instead of "*director*" as in the source English question).

c) The wrong translation of a word helps to retrieve the appropriate passage because it appears accidentally in the passage.

d) The translations provided by our system give a better distribution of weights by allowing the chance retrieval of the appropriate passage.

5 Conclusions

This paper describes the participation of the joint Elhuyar-IXA team at ResPubliQA

2010. For that purpose we used a system which works with passage retrieval alone, without any question analysis and answer validation steps.

Our English-English results show that good results can be achieved by means of this simple strategy. After expanding the passages based on semantic relatedness and tuning the retrieval system parameters, we obtained improvements for the English-English task. The drop in performance in the Basque-English bilingual runs is significant, and is caused by the accumulation of errors in the analysis and translation of the query. The use of expanded words was not effective for the cross-lingual task. A possible reason is the following: the co-occurrence-based translation selection algorithm uses as the target collection the one without expanded words to calculate the association degree between translation candidates, and consequently, the final translations are adapted to the original collection. Then, when expanded words are added to the passages, instead of helping the retrieval, they could add noise.

Acknowledgments

This work has been supported by KNOW2 (TIN2009-14715-C04-01), imFUTOURnet (IE08-233) and KYOTO (ICT-2007-211423). Arantxa Otegi's work is funded by a PhD grant from the Basque Government.

References

1. Kim, S., Seo, H., Rim, H.: Information retrieval using word senses: root sense tagging approach. In: Proceedings of SIGIR. (2004)
2. Liu, S., Yu, C., Meng, W.: Word Sense Disambiguation in Queries. In: Proceedings of the 14th ACM Conference on Information and Knowledge Management, CIKM. (2005)
3. Pérez-Agüera, J.R., Zaragoza, H.: Query Clauses and Term Independence. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, pp. 369-376. ISSN 0302-9743 ISBN 978-3-642-04446. (2009)
4. Agirre, E., Soroa, A., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M.: A study on similarity and relatedness using distributional and WordNet-based approaches. In: Proceedings of the annual meeting of the North American Chapter of the Association of Computational Linguistics (NAACL), Boulder, USA (2009)
5. Ansa, O., Arregi, X., Otegi, A., Soraluze. A.: Ihardetsi: A Basque Question Answering System at QA@CLEF 2008. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, pp. 369-376. ISSN 0302-9743 ISBN 978-3-642-04446. (2009)
6. Ezeiza, N., Aduriz, I., Alegria, I., Arriola, J.M., Urizar, R.: Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In: COLING-ACL, pp.380–384. (1998)
7. Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., Urizar. R.: Development of a Named Entity Recognizer for an Agglutinative Language. In: IJCNLP, (2004)
8. Darwish, K., Oard, D.W.: Probabilistic Structured Query Methods. In: Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in

Information Retrieval, pp. 338–344. (2003)

9. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55-63. (1998)
10. Ballesteros, L., Bruce Croft, W.: Resolving Ambiguity for Cross-language Retrieval. In: Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 64–71. (1998)
11. Gao, J., Nie, J.Y., Xun, E., Zhang, J., Zhou, M., Huang, C.: Improving Query Translation for Cross-language Information Retrieval Using Statistical Models. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 96-104. (2001)
12. Monz, C., Dorr, B.J.: Iterative translation disambiguation for cross-language information retrieval. In: Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 520-527. (2005)
13. Saralegi, X., López de Lacalle, M.: Comparing Different Approaches to Treat Translation Ambiguity in CLIR: Structured Queries v. Target Co-occurrence-Based Selection. In: 6th TIR workshop. (2009)
14. Agirre, E., Ansa, O., Arregi, X., Lopez de Lacalle, M., Otegi, A., Saralegi, X., Zaragoza. H.: Elhuyar-IXA: semantic relatedness and cross-lingual passage retrieval. Working Notes of the Cross-Lingual Evaluation Forum, Corfu, Greece. (2009)
15. Fellbaum, C.: WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press, Cambridge, Mass. (1998)
16. Haveliwala, T. H.: Topic-sensitive PageRank. In: Proceedings of WWW'02, pages 517-526. (2002)
17. Agirre, E., Arregi, X., Otegi, A.: Document Expansion Based on WordNet for Robust IR. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING). To appear (2010)
18. Boldi, P., Vigna, S.: MG4J at TREC 2005. In: Voorhees, E.M., Buckland, L.P. (eds.) The Fourteenth Text Retrieval Conference (TREC 2005) Proceedings, number SP 500-266 in Special Publications. NIST. http://mg4j.dsi.unimi.it/. (2005)
19. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4):333-389. (2009)
20. Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, A., Forăscu, C., Alegria, I., Giampiccolo, D., Moreau, N., Osenova, P.: Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. Working Notes for the CLEF 2009 Workshop. (2009)
21. Smucker, M. D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of CIKM 2007, Lisbon, Portugal. (2007)
22. Saralegi, X., Lopez de Lacalle, M.: Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In the 7th International Conference on Language Resources and Evaluations (LREC). Malta. (2010)

Document Expansion Based on WordNet for Robust IR

Eneko Agirre
IXA NLP Group
Univ. of the Basque Country
e.agirre@ehu.es

Xabier Arregi
IXA NLP Group
Univ. of the Basque Country
xabier.arregi@ehu.es

Arantxa Otegi
IXA NLP Group
Univ. of the Basque Country
arantza.otegi@ehu.es

Abstract

The use of semantic information to improve IR is a long-standing goal. This paper presents a novel Document Expansion method based on a WordNet-based system to find related concepts and words. Expansion words are indexed separately, and when combined with the regular index, they improve the results in three datasets over a state-of-the-art IR engine. Considering that many IR systems are not robust in the sense that they need careful fine-tuning and optimization of their parameters, we explored some parameter settings. The results show that our method is specially effective for realistic, non-optimal settings, adding robustness to the IR engine. We also explored the effect of document length, and show that our method is specially successful with shorter documents.

1 Introduction

Since the earliest days of IR, researchers noted the potential pitfalls of keyword retrieval, such as synonymy, polysemy, hyponymy or anaphora. Although in principle these linguistic phenomena should be taken into account in order to obtain high retrieval relevance, the lack of algorithmic models prohibited any systematic study of the effect of this phenomena in retrieval. Instead, researchers resorted to distributional semantic models to try to improve retrieval relevance, and overcome the brittleness of keyword matches. Most research concentrated on Query Expansion (QE) methods, which typically analyze term co-occurrence statistics in the corpus and in the highest scored documents for the original query in order to select terms for expanding the query terms (Manning et al., 2009). Document expansion (DE) is a natural alternative to QE, but surprisingly it was not investigated until very recently. Several researchers have used distributional methods from similar documents in the collection in order to expand the documents with related terms that do not actually occur in the document (Liu and Croft, 2004; Kurland and Lee, 2004; Tao et al., 2006; Mei et al., 2008; Huang et al., 2009). The work presented here is complementary, in that we also explore DE, but use WordNet instead of distributional methods.

Lexical semantic resources such as WordNet (Fellbaum, 1998) might provide a principled and explicit remedy for the brittleness of keyword matches. WordNet has been used with success in psycholinguistic datasets of word similarity and relatedness, where it often surpasses distributional methods based on keyword matches (Agirre et al., 2009b). WordNet has been applied to IR before. Some authors extended the query with related terms (Voorhees, 1994; Liu et al., 2005), while others have explicitly represented and indexed word senses after performing word sense disambiguation (WSD) (Gonzalo et al., 1998; Stokoe et al., 2003; Kim et al., 2004). More recently, a CLEF task was organized (Agirre et al., 2008; Agirre et al., 2009a) where queries and documents were semantically disambiguated, and participants reported mixed results.

This paper proposes to use WordNet for document expansion, proposing a new method: given a full document, a random walk algorithm over the WordNet graph ranks concepts closely related to the words in the document. This is in contrast to previous WordNet-based work which focused on WSD to replace or supplement words with their senses. Our method discovers important concepts, even if they are not explicitly mentioned in the document. For instance, given a document mentioning *virus*, *software* and *DSL*, our method suggests related concepts and associated words such us *digital subscriber line*, *phone company* and *computer*. Those expansion words are indexed separately, and when combined with the regular index, we show that they improve the results in three datasets over a state-of-the-art IR engine (Boldi and Vigna, 2005). The three datasets used in this study are ResPubliQA (Peñas et al., 2009), Yahoo! Answers (Surdeanu et al., 2008) and CLEF-Robust (Agirre et al., 2009a).

Considering that many IR systems are not robust in the sense that they need careful fine-tuning and optimization of their parameters, we decided to study the robustness of our method, exploring some alternative settings, including default parameters, parameters optimized in development data, and parameters optimized in other datasets. The study reveals that the additional semantic expansion terms provide robustness in most cases.

We also hypothesized that semantic document expansion could be most profitable when documents are shorter, and our algorithm would be most effective for collections of short documents. We artificially trimmed documents in the Robust dataset. The results, together with the analysis of document lengths of the three datasets, show that document expansion is specially effective for very short documents, but other factors could also play a role.

The paper is structured as follows. We first introduce the document expansion technique. Section 3 introduces the method to include the expansions in a retrieval system. Section 4 presents the experimental setup. Section 5 shows our main results. Sections 6 and 7 analyze the robustness and relation to document length. Section 8 compares to related work. Finally, the conclusions and future work are mentioned.

2 Document Expansion Using WordNet

Our key insight is to expand the document with related words according to the background information in WordNet (Fellbaum, 1998), which provides generic information about general vocabulary terms. WordNet groups nouns, verbs, adjectives and adverbs into sets of synonyms (synsets), each expressing a distinct concept. Synsets are interlinked with conceptual-semantic and lexical relations, including hypernymy, meronymy, causality, etc.

In contrast with previous work, we select those concepts that are most closely related to the document as a whole. For that, we use a technique based on random walks over the graph representation of WordNet concepts and relations.

We represent WordNet as a graph as follows: graph nodes represent WordNet concepts (synsets) and dictionary words; relations among synsets are represented by undirected edges; and dictionary words are linked to the synsets associated to them by directed edges. We used version 3.0, with all relations provided, including the gloss relations. This was the setting obtaining the best results in a word similarity dataset as reported by Agirre et al. (2009b).

Given a document and the graph-based representation of WordNet, we obtain a ranked list of WordNet concepts as follows:

1. We first pre-process the document to obtain the lemmas and parts of speech of the open category words.
2. We then assign a uniform probability distribution to the terms found in the document. The rest of nodes are initialized to zero.
3. We compute personalized PageRank (Haveliwala, 2002) over the graph, using the previous distribution as the reset distribution, and producing a probability distribution over WordNet concepts The higher the probability for a concept, the more related it is to the given document.

Basically, personalized PageRank is computed by modifying the random jump distribution vector in the traditional PageRank equation. In our case, we concentrate all probability mass in the concepts corresponding to the words in the document.

Let G be a graph with N vertices $v_1, \ldots, v_N$ and d_i be the outdegree of node i; let M be a $N \times N$ transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if a link from i to j exists, and zero otherwise. Then, the calculation of the *PageRank vector* $\mathbf{Pr}$ over G is equivalent to resolving Equation (1).

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)\mathbf{v} \qquad (1)$$

In the equation, $\mathbf{v}$ is a $N \times 1$ vector and c is the so called *damping factor*, a scalar value between 0 and 1. The first term of the sum on the equation models the voting scheme described in the beginning of the section. The second term represents, loosely speaking, the probability of a surfer randomly jumping to any node, e.g. without following any paths on the graph. The damping factor, usually set in the $[0.85..0.95]$ range, models the way in which these two terms are combined at each step.

The second term on Eq. (1) can also be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible, and thus guarantees that PageRank calculation converges to a unique stationary distribution.

In the traditional PageRank formulation the vector $\mathbf{v}$ is a stochastic normalized vector whose element values are all $\frac{1}{N}$, thus assigning equal probabilities to all nodes in the graph in case of random jumps. In the case of personalized PageRank as used here, $\mathbf{v}$ is initialized with uniform probabilities for the terms in the document, and 0 for the rest of terms.

PageRank is actually calculated by applying an iterative algorithm which computes Eq. (1) successively until a fixed number of iterations are executed. In our case, we used a publicly available implementation[1], with default values for the damping value (0.85) and the number of iterations (30). In order to select the expansion terms, we chose the 100 highest scoring concepts, and get all the words that lexicalize the given concept.

Figure 1 exemplifies the expansion. Given the short document from Yahoo! Answers (cf. Section 4) shown in the top, our algorithm produces the set of related concepts and words shown in the bottom. Note that the expansion produces synonyms, but also other words related to concepts that are not mentioned in the document.

3 Including Expansions in a Retrieval System

Once we have the list of words for document expansion, we create one index for the words in the original documents and another index with the expansion terms. This way, we are able to use the original words only, or to also include the expansion words during the retrieval.

The retrieval system was implemented using MG4J (Boldi and Vigna, 2005), as it provides state-of-the-art results and allows to combine several indices over the same document collection. We conducted different runs, by using only the index made of original words (baseline) and also by using the index with the expansion terms of the related concepts.

BM25 was the scoring function of choice. It is one of the most relevant and robust scoring functions available (Robertson and Zaragoza, 2009).

$$w_{Dt}^{BM25} := \qquad (2)$$
$$\frac{tf_{D_t}}{k_1\left((1-b)+b\frac{dl_D}{avdl_D}\right)+tf_{Dt}} idf_t$$

where tf_{D_t} is the term frequency of term t in document D, dl_D is the document length, idf_t is the inverted document frequency (or more specifically the RSJ weight, (Robertson and Zaragoza, 2009)), and k_1 and b are free parameters.

The two indices were combined linearly, as follows (Robertson and Zaragoza, 2009):

$$score(d,e,q) := \qquad (3)$$
$$\sum_{t \in q \cap d} w_{D_t}^{BM25} + \lambda \sum_{t \in q \cap e} w_{E_t}^{BM25}$$

where *D* and *E* are the original and expanded indices, *d, e* and *q* are the original document, the expansion of the document and the query respectively, t is a term, and λ is a free parameter for the relative weight of the expanded index.

[1] http://ixa2.si.ehu.es/ukb/

You should only need to turn off virus and anti-spy not uninstall. And that's done within each of the softwares themselves. Then turn them back on later after installing any DSL softwares.

06566077-n → *computer software, package*, **software**, *software package, software program, software system*
03196990-n → *digital subscriber line*, **dsl**
01569566-v → *instal*, **install**, *put in, set up*
04402057-n → line, phone line, suscriber line, telephone circuit, telephone line
08186221-n → phone company, phone service, telco, telephone company, telephone service
03082979-n → computer, computing device, computing machine, data processor, electronic computer

Figure 1: Example of a document expansion, with original document on top, and some of the relevant WordNet concepts identified by our algorithm, together with the words that lexicalize them. Words in the original document are shown in bold, synonyms in italics, and other related words underlined.

4 Experimental Setup

We chose three data collections. The first is based on a traditional news collection. DE could be specially interesting for datasets with short documents, which lead our choice of the other datasets: the second was chosen because it contains shorter documents, and the third is a passage retrieval task which works on even shorter paragraphs. Table 1 shows some statistics about the datasets.

One of the collections is the English dataset of the **Robust** task at CLEF 2009 (Agirre et al., 2009a). The documents are news collections from LA Times 94 and Glasgow Herald 95. The topics are statements representing information needs, consisting of three parts: a brief title statement; a one-sentence description; a more complex narrative describing the relevance assessment criteria. We use only the title and the description parts of the topics in our experiments.

The **Yahoo! Answers** corpus is a subset of a dump of the Yahoo! Answers web site[2] (Surdeanu et al., 2008), where people post questions and answers, all of which are public to any web user willing to browse them. The dataset is a small subset of the questions, selected for their linguistic properties (for example they all start with "how {to||do||did||does||can||would||could||should}"). Additionally, questions and answers of obvious low quality were removed. The document set was created with the best answer of each question (only one for each question).

	docs	length	q. train	q. test
Robust	166,754	532	150	160
Yahoo!	89610	104	1000	88610
ResPubliQA	1,379,011	20	100	500

Table 1: Number of documents, average document length, number of queries for train and test in each collection.

The other collection is the English dataset of **ResPubliQA** exercise at the Multilingual Question Answering Track at CLEF 2009 (Peñas et al., 2009). The exercise is aimed at retrieving paragraphs that contain answers to a set of 500 natural language questions. The document collection is a subset of the JRC-Acquis Multilingual Parallel Corpus, and consists of 21,426 documents for English which are aligned to a similar number of documents in other languages[3]. For evaluation, we used the gold standard released by the organizers, which contains a single correct passage for each query. As the retrieval unit is the passage, we split the document collection into paragraphs. We applied the expansion strategy only to passages which had more than 10 words (half of the passages), for two reasons: the first one was that most of these passages were found not to contain relevant information for the task (e.g. "Article 2" or "Having regard to the proposal from the Commission"), and the second was that we thus saved some computation time.

In order to evaluate the quality of our expansion in practical retrieval settings, the next Section re-

[2] Yahoo! Webscope dataset "ydata-yanswers-manner-questions-v1_0" http://webscope.sandbox.yahoo.com/

[3] Note that Table 1 shows the number of paragraphs, which conform the units we indexed.

		base.	expa.	Δ
Robust	MAP	.3781	**.3835*****	1.43%
Yahoo!	MRR	.2900	**.2950*****	1.72%
	P@1	.2142	**.2183*****	1.91%
ResPubliQA	MRR	.3931	**.4077*****	3.72%
	P@1	.2860	**.3000****	4.90%

Table 2: Results using default parameters.

		base.	expa.	Δ
Robust	MAP	.3740	**.3823****	2.20%
Yahoo!	MRR	.3070	**.3100*****	0.98%
	P@1	.2293	**.2317***	1.05%
ResPubliQA	MRR	.4970	.4942	-0.56%
	P@1	.3980	.3940	-1.01%

Table 3: Results using optimized parameters.

port results with respect to several parameter settings. Parameter optimization is often neglected in retrieval with linguistic features, but we think it is crucial since it can have a large effect on relevance performance and therefore invalidate claims of improvements over the baseline. In each setting we assign different values to the free parameters in the previous section, k_1, b and λ.

5 Results

The main evaluation measure for Robust is mean Average Precision (MAP), as customary. In two of the datasets (Yahoo! and ResPubliQA) there is a single correct answer per query, and therefore we use Mean Reciprocal Rank (MRR) and Mean Precision at rank 1 (P@1) for evaluation. Note that in this setting MAP is identical to MRR. Statistical significance was computed using Paired Randomization Test (Smucker et al., 2007). In the tables throughout the paper, we use * to indicate statistical significance at 90% confidence level, ** for 95% and *** for 99%. Unless noted otherwise, base. refers to MG4J with the standard index, and expa. refers to MG4J using both indices. Best results per row are in bold when significant. Δ reports relative improvement respect to the baseline.

5.1 Default Parameter Setting

The values for k_1 and b are the default values as provided in the w^{BM25} implementation of MG4J, 1.2 and 0.5 respectively. We could not think of a straightforward value for λ. A value of 1 would mean that we are assigning equal importance to original and expanded terms, which seemed an overestimation, so we used 0.1. Table 2 shows the results when using the default setting of parameters. The use of expansion is beneficial in all datasets, with relative improvements ranging from 1.43% to 4.90%.

Setting	System	k_1	b	λ
Default	base.	1.20	0.50	-
	expa.	1.20	0.50	0.100
Robust	base.	1.80	0.64	-
	expa.	1.66	0.55	0.075
Yahoo!	basel.	0.99	0.82	-
	expa.	0.84	0.87	0.146
ResPubliQA	base.	0.09	0.56	-
	expa.	0.13	0.65	0.090

Table 4: Parameters as in the default setting or as optimized in each dataset. The λ parameter is not used in the baseline systems.

5.2 Optimized Parameter Setting

We next optimized all three parameters using the train part of each collection. The optimization of the parameters followed a greedy method called "promising directions" (Robertson and Zaragoza, 2009). The comparison between the baseline and expansion systems in Table 3 shows that expansion helps in Yahoo! and Robust, with statistical significance. The differences in ResPubliQA are not significant, and indicate that expansion terms were not helpful in this setting.

Note that the optimization of the parameters yields interesting effects in the baseline for each of the datasets. If we compare the results of the baseline with default settings (Table 2) and with optimized setting (Table 3), the baseline improves MRR dramatically in ResPubliQA (26% relative improvement), significantly in Yahoo! (5.8%) and decreases MAP in Robust (-0.01%). This disparity of effects could be explained by the fact that the default values are often approximated using TREC-style news collections, which is exactly the genre of the Robust documents, while Yahoo uses shorter documents, and ResPubliQA has the shortest documents.

Table 4 summarizes the values of the parameters in both default and optimized settings. For k_1, the optimization yields very different values. In Robust the value is similar to the default value, but

		base.	expa.	Δ	λ
Rob	MAP	.3781	**.3881*****	2.64%	0.18
Y!	MRR	.2900	**.2980*****	2.76%	0.27
	P@1	.2142	**.2212*****	3.27%	
ResP.	MRR	.3931	**.4221*****	7.39%	0.61
	P@1	.2860	**.3180****	11.19%	

Table 5: Results obtained using the λ optimized setting, including actual values of λ.

in ResPubliQA the optimization pushes it down below the typical values cited in the literature (Robertson and Zaragoza, 2009), which might explain the boost in performance for the baseline in the case of ResPubliQA. When all three parameters are optimized together, the values λ in the table range from 0.075 to 0.146. The values of the optimized λ can be seem as an indication of the usefulness of the expanded terms, so we explored this farther.

5.3 Exploring λ

As an additional analysis experiment, we wanted to know the effect of varying λ keeping k_1 and b constant at their default values. Table 5 shows the best values in each dataset, which that the weight of the expanded terms and the relative improvement are highly correlated.

5.4 Exploring Number of Expansion Concepts

One of the free parameters of our system is the number of concepts to be included in the document expansion. We have performed a limited study with the default parameter setting on the Robust setting, using 100, 500 and 750 concepts, but the variations were not statistically significant. Note that with 100 concepts we were actually expanding with 268 words, with 500 concepts we add 1247 words and with 750 concepts we add 1831 words.

6 Robustness

The results in the previous section indicate that optimization is very important, but unfortunately real applications usually lack training data. In this Section we wanted to study whether the parameters can be carried over from one dataset to the other, and if not, whether the extra terms found by DE would make the system more robust to those sub-optimal parameters.

train			base.	expa.	Δ
Rob.	def.	MAP	.3781	**.3835*****	1.43%
	Rob.	MAP	.3740	**.3823****	2.20%
	Y!	MAP	.3786	.3759	-0.72%
	Res.	MAP	.3146	**.3346*****	6.35%
Y!	def.	MRR	.2900	**.2950*****	1.72%
	Rob.	MRR	.2920	.2920	0.0%
	Y!	MRR	.3070	**.3100****	0.98%
	Res.	MRR	.2600	**.2750*****	5.77%
ResP.	def.	MRR	.3931	**.4077*****	3.72%
	Rob.	MRR	.3066	**.3655*****	19.22%
	Y!	MRR	.3010	**.3459*****	14.93%
	Res.	MRR	.4970	.4942	-0.56%

Table 6: Results optimizing parameters with training from other datasets. We also include default and optimization on the same dataset for comparison. Only MRR and MAP results are given.

Table 6 includes a range of parameter settings, including defaults, and optimized parameters coming from the same and different datasets. The values of the parameters are those in Table 4. The results show that when the parameters are optimized in other datasets, DE provides improvement with statistical significance in all cases, except for the Robust dataset when using parameters optimized from Yahoo! and vice-versa.

Overall, the table shows that our DE method either improves the results significantly or does not affect performance, and that it provides robustness across different parameter settings, even with sub-optimal values.

7 Exploring Document Length

The results in Table 6 show that the performance improvements are best in the collection with shortest documents (ResPubliQA). But the results for Robust and Yahoo! do not show any relation to document length. We thus decided to do an additional experiment artificially shrinking the document in Robust to a certain percentage of its original length. We create new pseudo-collection with the shrinkage factors of 2.5%, 10%, 20% and 50%, keeping the first N% words in the document and discarding the rest. In all cases we used the same parameters, as optimized for Robust.

Table 7 shows the results (MAP), with some clear indication that the best improvements are obtained for the shortest documents.

	length	base.	expa.	Δ
2.5%	13	.0794	.0851	7.18%
10%	53	.1757	.1833	4.33%
20%	107	.2292	.2329	1.61%
50%	266	.3063	.3098	1.14%
100%	531	.3740	.3823	2.22%

Table 7: Results (MAP) on Robust when artificially shrinking documents to a percentage of their length. In addition to the shrinking rate we show the average lengths of documents.

8 Related Work

Given the brittleness of keyword matches, most research has concentrated on Query Expansion (QE) methods. These methods analyze the user query terms and select automatically new related query terms. Most QE methods use statistical (or distributional) techniques to select terms for expansion. They do this by analyzing term co-occurrence statistics in the corpus and in the highest scored documents of the original query (Manning et al., 2009). These methods seemed to improve slightly retrieval relevance on average, but at the cost of greatly decreasing the relevance of difficult queries. But more recent studies seem to overcome some of these problems (Collins-Thompson, 2009).

An alternative to QE is to perform the expansion in the document. Document Expansion (DE) was first proposed in the speech retrieval community (Singhal and Pereira, 1999), where the task is to retrieve speech transcriptions which are quite noisy. Singhal and Pereira propose to enhance the representation of a noisy document by adding to the document vector a linearly weighted mixture of related documents. In order to determine related documents, the original document is used as a query into the collection, and the ten most relevant documents are selected.

Two related papers (Liu and Croft, 2004; Kurland and Lee, 2004) followed a similar approach on the TREC ad-hoc document retrieval task. They use document clustering to determine similar documents, and document expansion is carried out with respect to these. Both papers report significant improvements over non-expanded baselines. Instead of clustering, more recent work (Tao et al., 2006; Mei et al., 2008; Huang et al., 2009) use language models and graph representations of the similarity between documents in the collection to smooth language models with some success. The work presented here is complementary, in that we also explore DE, but use WordNet instead of distributional methods. They use a tighter integration of their expansion model (compared to our simple two-index model), which coupled with our expansion method could help improve results further. We plan to explore this in the future.

An alternative to statistical expansion methods is to use lexical semantic knowledge bases such as WordNet. Most of the work has focused on query expansion and the use of synonyms from WordNet after performing word sense disambiguation (WSD) with some success (Voorhees, 1994; Liu et al., 2005). The short context available in the query when performing WSD is an important problems of these techniques. In contrast, we use full document context, and related words beyond synonyms. Another strand of WordNet based work has explicitly represented and indexed word senses after performing WSD (Gonzalo et al., 1998; Stokoe et al., 2003; Kim et al., 2004). The word senses conform a different space for document representation, but contrary to us, these works incorporate concepts for all words in the documents, and are not able to incorporate concepts that are not explicitly mentioned in the document. More recently, a CLEF task was organized (Agirre et al., 2009a) where terms were semantically disambiguated to see the improvement that this would have on retrieval; the conclusions were mixed, with some participants slightly improving results with information from WordNet. To the best of our knowledge our paper is the first on the topic of document expansion using lexical-semantic resources.

We would like to also compare our performance to those of other systems as tested on the same datasets. The systems which performed best in the Robust evaluation campaign (Agirre et al., 2009a) report 0.4509 MAP, but note that they deployed a complex system combining probabilistic and monolingual translation-based models. In ResPubliQA (Peñas et al., 2009), the official evaluation included manual assessment, and we cannot therefore reproduce those results. Fortunately, the organizers released all runs, but only the first ranked document for each query was included, so we could only compute P@1. The P@1 of best run was 0.40. Finally (Surdeanu et al., 2008) report MRR figure around 0.68, but they evaluate only in the questions where the correct answer is retrieved by answer retrieval in the top 50 answers, and is thus not comparable to our setting.

Regarding the WordNet expansion technique we use here, it is implemented on top of publicly available software[4], which has been successfully used in word similarity (Agirre et al., 2009b) and word sense disambiguation (Agirre and Soroa, 2009). In the first work, a single word was input to the random walk algorithm, obtaining the probability distribution over all WordNet synsets. The similarity of two words was computed as the similarity of the distribution of each word, obtaining the best results for WordNet-based systems on the word similarity dataset, and comparable to the results of a distributional similarity method which used a crawl of the entire web. Agirre et al. (2009) used the context of occurrence of a target word to start the random walk, and obtained very good results for WordNet WSD methods.

9 Conclusions and Future Work

This paper presents a novel Document Expansion method based on a WordNet-based system to find related concepts and words. The documents in three datasets were thus expanded with related words, which were fed into a separate index. When combined with the regular index we report improvements over MG4J using w^{BM25} for those three datasets across several parameter settings, including default values, optimized parameters and parameters optimized in other datasets. In most of the cases the improvements are statistically significant, indicating that the information in the document expansion is useful. Similar to other expansion methods, parameter optimization has a stronger effect than our expansion strategy. The problem with parameter optimization is that in most real cases there is no tuning dataset available. Our analysis shows that our expansion method is more effective for sub-optimal parameter settings, which is the case for most real-live IR applications. A comparison across the three datasets and using artificially trimmed documents indicates that our method is particularly effective for short documents.

As document expansion is done at indexing time, it avoids any overhead at query time. It also has the advantage of leveraging full document context, in contrast to query expansion methods, which use the scarce information present in the much shorter queries. Compared to WSD-based methods, our method has the advantage of not having to disambiguate all words in the document. Besides, our algorithm picks the most relevant concepts, and thus is able to expand to concepts which are not explicitly mentioned in the document. The successful use of background information such as the one in WordNet could help close the gap between semantic web technologies and IR, and opens the possibility to include other resources like Wikipedia or domain ontologies like those in the Unified Medical Language System.

Our method to integrate expanded terms using an additional index is simple and straightforward, and there is still ample room for improvement. A tighter integration of the document expansion technique in the retrieval model should yield better results, and the smoothed language models of (Mei et al., 2008; Huang et al., 2009) seem a natural choice. We would also like to compare with other existing query and document expansion techniques and study whether our technique is complementary to query expansion approaches.

Acknowledgments

This work has been supported by KNOW2 (TIN2009-14715-C04-01) and KYOTO (ICT-2007-211423) projects. Arantxa Otegi's work is funded by a PhD grant from the Basque Government. Part of this work was done while Arantxa Otegi was visiting Yahoo! Research Barcelona.

References

Agirre, E. and A. Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proc. of*

[4] http://ixa2.si.ehu.es/ukb

EACL 2009, Athens, Greece.

Agirre, E., G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. 2008. CLEF 2008: Ad-Hoc Track Overview. In *Working Notes of the Cross-Lingual Evaluation Forum*.

Agirre, E., G. M. Di Nunzio, T. Mandl, and A. Otegi. 2009a. CLEF 2009 Ad Hoc Track Overview: Robust - WSD Task. In *Working Notes of the Cross-Lingual Evaluation Forum*.

Agirre, E., A. Soroa, E. Alfonseca, K. Hall, J. Kravalova, and M. Pasca. 2009b. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proc. of NAACL*, Boulder, USA.

Boldi, P. and S. Vigna. 2005. MG4J at TREC 2005. In *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, number SP 500-266 in Special Publications. NIST.

Collins-Thompson, Kevyn. 2009. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of CIKM '09*, pages 837–846.

Fellbaum, C., editor. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, Mass.

Gonzalo, J., F. Verdejo, I. Chugur, and J. Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings ACL/COLING Workshop on Usage of WordNet for Natural Language Processing*.

Haveliwala, T. H. 2002. Topic-sensitive PageRank. In *Proceedings of WWW '02*, pages 517–526.

Huang, Yunping, Le Sun, and Jian-Yun Nie. 2009. Smoothing document language model with local word graph. In *Proceedings of CIKM '09*, pages 1943–1946.

Kim, S. B., H. C. Seo, and H. C. Rim. 2004. Information retrieval using word senses: root sense tagging approach. In *Proceedings of SIGIR '04*, pages 258–265.

Kurland, O. and L. Lee. 2004. Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of SIGIR '04*, pages 194–201.

Liu, X. and W. B. Croft. 2004. Cluster-based retrieval using language models. In *Proceedings of SIGIR '04*, pages 186–193.

Liu, S., C. Yu, and W. Meng. 2005. Word sense disambiguation in queries. In *Proceedings of CIKM '05*, pages 525–532.

Manning, C. D., P. Raghavan, and H. Schütze. 2009. *An introduction to information retrieval*. Cambridge University Press, UK.

Mei, Qiaozhu, Duo Zhang, and ChengXiang Zhai. 2008. A general optimization framework for smoothing language models on graph structures. In *Proceedings of SIGIR '08*, pages 611–618.

Peñas, A., P. Forner, R. Sutcliffe, A. Rodrigo, C. Forăscu, I. Alegria, D. Giampiccolo, N. Moreau, and P. Osenova. 2009. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In *Working Notes of the Cross-Lingual Evaluation Forum*.

Robertson, S. and H. Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Singhal, A. and F. Pereira. 1999. Document expansion for speech retrieval. In *Proceedings of SIGIR '99*, pages 34–41, New York, NY, USA. ACM.

Smucker, M. D., J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of CIKM 2007*, Lisboa, Portugal.

Stokoe, C., M. P. Oakes, and J. Tait. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of SIGIR '03*, page 166.

Surdeanu, M., M. Ciaramita, and H. Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of ACL 2008*.

Tao, T., X. Wang, Q. Mei, and C. Zhai. 2006. Language model information retrieval with document expansion. In *Proceedings of HLT/NAACL*, pages 407–414, June.

Voorhees, E. M. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR '94*, page 69.

Query Expansion for IR using Knowledge-Based Relatedness

Arantxa Otegi
IXA NLP Group
Univ. of the Basque Country
arantza.otegi@ehu.es

Xabier Arregi
IXA NLP Group
Univ. of the Basque Country
xabier.arregi@ehu.es

Eneko Agirre
IXA NLP Group
Univ. of the Basque Country
e.agirre@ehu.es

Abstract

The limitations of keyword-only approaches to information retrieval were recognized since the early days, specially in cases where different but closely-related words are used in the query and the relevant document. Query expansion techniques like pseudo-relevance feedback rely on the target document set in order to bridge the gap between those words, but they might suffer from topic drift. This paper explores the use of knowledge-based semantic relatedness in order to bridge the gap between query and documents. We performed query expansion, with positive effects over some language modeling baselines.

1 Introduction

The potential pitfalls of keyword retrieval have been noted since the earliest days of Information Retrieval (IR). Keyword retrieval proves ineffective when different but closely-related words are used in the query and the relevant document. The use of different words creates a lexical gap between the query and the document.

In order to bridge the gap, IR has resorted to distributional semantic models. Most research concentrated on Query Expansion (QE) methods, which typically analyze term co-occurrence statistics in the corpus and/or in the highest scored documents in order to select terms for expanding the query terms (Manning et al., 2009). The work presented here is complementary, in that we explore QE, but we use an approach based on semantic relatedness instead of distributional methods.

In a closely related work, (Agirre et al., 2010) proposed a WordNet-based document expansion method using random walks: given a document, a random walk algorithm over the WordNet graph, inspired in (Agirre et al., 2009b), ranks concepts closely related to the words in the document. Note that the method can return concepts which are not explicitly mentioned in the document. The highest ranking concepts were then selected to expand the document.

In this work, we explore an alternative method to exploit relatedness, query expansion, so we thus run the relatedness algorithm over the queries and we expand the queries. We adopt a language modeling framework to implement the query likelihood and pseudo-relevance feedback baselines, as well as our relatedness-based query expansion method.

In order to test the performance of our method we selected several datasets with different domains, topic typologies and document lengths. Given the relevance among the community using WordNet-related methods, we selected the Robust-WSD dataset from CLEF (Agirre et al., 2009a), which is a typical ad-hoc dataset on news. As we think that our method is specially relevant for short queries and/or short documents, we also selected the Yahoo! Answers dataset, which contains questions and answers as phrased by real users on diverse topics (Surdeanu et al., 2008), and ResPubliQA, a paragraph retrieval task on European Union laws organized at CLEF (Peñas et al., 2009).

The results show that our method provide improvements in all three datasets, when compared to the query likelihood baseline, and that they compare favorably to pseudo-relevance feedback in two datasets.

The paper is structured as follows. We first briefly introduce related work. We then mention the random walk model for query expansion. The design of the experiments is presented in Section 4. Section 5 shows our results, and, finally, Section 6 presents the conclusions.

2 Related Work

Query expansion methods analyze user query terms and incorporate related terms automatically (Voorhees, 1994). They are usually divided into local and global methods.

Local methods adjust a query relative to the documents that initially appear to match the query (Manning et al., 2009). Pseudo-relevance Feedback (PRF) is one of the most widely used expansion methods (Rocchio, 1971; Xu and Croft, 1996). This method assumes top-ranked documents to be relevant (and sometimes, also that low-ranked documents are irrelevant), and selects additional query terms from the top-ranked documents.

Global methods are techniques for expanding query terms without checking the results returned by the query. These methods analyze term co-occurrence statistics in the entire corpus or use external knowledge sources to select terms for expansion (Manning et al., 2009). For example, techniques using Word Sense Disambiguation (WSD) techniques and synonyms from WordNet have been used for query expansion with some success (Voorhees, 1994; Liu et al., 2005).

The query expansion method proposed in this paper is a global expansion technique based on WordNet, but in contrast to the previous work based on WordNet, it does not perform WSD and adds related words beyond synonyms.

(Agirre et al., 2010) is the work which is closest to ours. They use the same WordNet-based relatedness method in order to expand documents, following the BM25 probabilistic method for IR, obtaining some improvements, specially when parameters had not been optimized. In contrast to their work, we investigate methods to apply relatedness to query expansion, and we compare the results with pseudo-relevance feedback. Besides, we found that a language modeling (Ponte and Croft, 1998) approach to IR combined with inference networks (Turtle and Croft, 1991) offered more flexibility for query expansion.

Our work stems from the use of random walks over the WordNet graph to compute the relatedness between pairs of words (Hughes and Ramage, 2007). In this work a single word was input to the random walk algorithm, obtaining the probability distribution over all WordNet synsets. The similarity of two words was computed as the similarity of the distributions of each word. In later work, (Agirre et al., 2009b) tested different configurations of the graph, and obtained the best results for a WordNet-based system, comparable to the results of a distributional similarity method which used a crawl of the entire web. The same authors later released their UKB software, which is the one we use here.

3 Relatedness-based Query Expansion (RQE)

The key insight of our model is to expand the query with related words according to the background information in WordNet (Fellbaum, 1998), which provides generic information about general vocabulary terms.

In contrast with previous work using WordNet, we select those concepts that are most closely related to the query as a whole. To this end, we follow the approach in (Agirre et al., 2010), which, based on random walks over the graph representation of WordNet concepts and relations, obtains concepts related to the documents. We use the same settings and implementation for the graph algorithm, which is publicly available[1]. Details are omitted here due to lack of space, please refer to (Agirre et al., 2010).

In order to select the expansion terms, we choose the top N highest scoring concepts, and get all the words that lexicalize the given concept. We explored several values of N, and tune it in order to get the optimum value, as discussed in Section 4. For instance, given a query like "*What is the lowest speed in miles per hour which can be shown on a speedometer?*", our method suggests related terms like *vehicle*, *distance* and *mph*.

Our retrieval model runs queries which contain the original terms of the query and the expansion terms. Documents are ranked by their probability of generating the whole expanded query (Q_{RQE}), which is given by:

$$P_{RQE}(Q_{RQE} \mid \Theta_D) = P(Q \mid \Theta_D)^{w} P(Q' \mid \Theta_D)^{1-w} \tag{1}$$

where w is the weight given to the original query and Q' is the expansion of query Q.

The query likelihood probability is estimated following the multinomial distribution:

$$P(Q \mid \Theta_D) = \prod_{i=1}^{|Q|} P(q_i \mid \Theta_D)^{\frac{1}{|Q|}} \tag{2}$$

[1] http://ixa2.si.ehu.es/ukb/

where q_i is a query term of query Q and $|Q|$ is the length of Q. And following the Dirichlet smoothing (Zhai and Lafferty, 2001) we have

$$P(q_i \mid \Theta_D) = \frac{tf_{q_iD} + \mu \frac{tf_{q_iC}}{|C|}}{|D| + \mu} \qquad (3)$$

where tf_{q_iD} and tf_{q_iC} are the frequency of the query term q_i in the document D and the entire collection, respectively, and μ is the smoothing free parameter.

The probability of generating the expansion terms is defined as

$$P(Q' \mid \Theta_D) = \prod_{q'_i}^{|Q'|} P(q'_i \mid \Theta_D)^{\frac{w_i}{W}} \qquad (4)$$

where q'_i is a expansion term, $W = \sum_{i=1}^{|Q'|} w_i$ and w_i is the weight we give to a expansion term, which we can see as the relatedness between the original query Q and the expansion term, and is computed as

$$w_i = P(q' \mid Q) = \sum_{j=1}^{N} P(q' \mid c_j)P(c_j \mid Q) \qquad (5)$$

where c is a concept returned by the expansion algorithm, N is the number of concepts we chose for the expansion, $P(q' \mid c_j)$ is estimated using the sense probabilities estimated from Semcor (i.e. how often the query term q' occurs with sense c_j), and $P(c_j \mid Q)$ is the similarity weight that the mentioned expansion algorithm assigned to c_j concept.

4 Experiments

In order to test the performance of our method we selected several datasets with different domains, topic typologies and document lengths. Table 1 shows some statistics for each.

The first is the English dataset of the **Robust-WSD** task at CLEF 2009 (Agirre et al., 2009a), a typical ad-hoc dataset on news. This dataset has been widely used among the community interested on WSD and WordNet-related methods. The documents in the Robust-WSD comprise news collections from LA Times 94 and Glasgow Herald 95.

The **Yahoo! Answers** corpus is a subset of a dump of the Yahoo! Answers web site, where people post questions and answers, all of which are public to any web user willing to browse them (Surdeanu et al., 2008). The document set was created with the best answer of each question (only one for each question). We use the dataset as released by its authors[2].

	docs	length	q. train	q. test	length
Robust	166,754	532	150	160	8.6
Yahoo!	89,610	104	1,000	30,000	11.7
ResPubliQA	1,379,011	20	100	500	12.2

Table 1: Number of documents, average document length, number of queries for train and test in each collection, and average query length.

	QL	PRF				RQE		
	μ	μ	d	t	w	μ	N	w
Rob	1000	1000	10	50	0.3	2000	100	0.5
Yah	200	200	2	20	0.8	200	50	0.7
Res	100	100	10	30	0.8	100	125	0.7

Table 2: Optimal values in each dataset for free parameters.

The other collection is the English dataset of **ResPubliQA** exercise at the Multilingual Question Answering Track at CLEF 2009 (Peñas et al., 2009). The exercise is aimed at retrieving paragraphs that contain answers to a set of 500 natural language questions.

Our experiments were performed using the Indri search engine (Strohman et al., 2005), which is a part of the open-source Lemur toolkit[3].

To determine whether the query expansion model we developed is useful to improve retrieval performance, we set up a number of experiments in which we compared our expansion model with other retrieval approaches. We used two baseline retrieval approaches for comparison purposes. One of the baselines is the default query likelihood (**QL**) language modeling method implemented in the Indri search engine. The other one is pseudo-relevance feedback (**PRF**) using a modified version of Lavrenko's relevance model (Lavrenko and Croft, 2001), where the final query is a weighted combination of the original and expanded queries, analogous to Eq. 1. As in our own model presented in the previous section, we chose the Dirichlet smoothing method for the baselines. We consider **QL** and **PRF** to be strong, reasonable baselines.

All the methods have several free parameters. The PRF model has three: number of documents (d) and terms (t), and w (cf. Eq. 1). The RQE

[2] Check the features of the dataset at Yahoo! Webscope dataset: http://webscope.sandbox.yahoo.com/ ("ydata-yanswers-manner-questions-v1_0")

[3] http://www.lemurproject.org

		QL	PRF	Δ QL	RQE	Δ QL	Δ PRF
Rob	MAP	33.22	**36.69**	10.44% ***	33.67	1.36%	-8.22% ***
	GMAP	13.21	**14.38**	8.90% ***	14.34	8.59% **	-0.29%
	P@5	42.50	**43.63**	2.65%	42.25	-0.59%	-3.15%
	P@10	35.31	**37.38**	5.84% ***	35.81	1.42%	-4.18% *
Yah	MRR	26.36	26.40	0.15%	**27.22**	3.26% ***	3.11% ***
	P@5	6.67	6.63	-0.56% **	**6.88**	3.21% ***	3.79% ***
	P@10	3.95	3.96	0.25%	**4.10**	3.91% ***	3.65% ***
Res	MRR	48.77	46.33	-5.00% ***	**49.78**	2.07%	7.44% ***
	P@5	12.44	12.00	-3.54% *	**12.68**	1.93%	5.67% ***
	P@10	**6.80**	6.78	-0.29%	6.78	-0.29%	0.00%

Table 3: Results of all methods. Δ columns show relative improvement with respect to QL or PRF.

model has two parameters: w (cf. Eq.. 1) and N the number of concepts for the expansion (Eq. 5). In addition, all methods use Dirichlet smoothing, which has a smoothing parameter μ. We used the train part of each dataset to tune all these parameters via a simple grid-search. The μ parameter was tested on the [100,1200] range for ResPubliQA and Yahoo! and [100,2000] for Robust, with increments of 100. The w parameter ranged over [0,1] with 0.1 increments. The d parameter ranged over [2,50] and the t and N in the range [1,200] (we tested 10 different values in the respective ranges). The parameter settings that maximized mean average precision for each model and each collection are shown in Table 2.

5 Results

Our main results are shown in Table 3. The main evaluation measure for Robust is Mean Average Precision (MAP), as customary. In two of the datasets (Yahoo! and ResPubliQA), there is a single correct answer per topic, and therefore we use Mean Reciprocal Rank (MRR). We also report Mean Precision at ranks 5 and 10 (P@5 and P@10). GMAP is also included (we will introduce and mention it afterwards). Statistical significance was computed using Paired Randomization Test (Smucker et al., 2007). In the tables throughout the paper, we use * to indicate statistical significance at 90% confidence level, ** for 95% and *** for 99%.

QL and PRF. The first two columns in Table 3 shows the results for QL and PRF and the performance difference between them. The results for PRF are mixed. It is very effective in the Robust dataset, with dramatic improvements, specially in MAP. All differences are statistical significant, except for P@5. In Yahoo! the improvement is small in MRR and P@10, without statistical significance, but P@5 is lower. In ResPubliQA the results are bad, with statistical significant degradation in MRR.

RQE. Continuing rightwards with Table 3, the following columns show the results for RQE, together with its difference with respect to QL and PRF. Note that figures in bold mean the best performance for each metric. It can be seen that, although RQE is not effective for Robust, it is the best method for Yahoo! and ResPubliQA. Moreover, the improvements over QL, and also over PRF, for Yahoo! are all statistical significant.

PRF is known to perform well for some topics and datasets but not for others. Table 3 includes results for the geometrical mean, GMAP (Robertson, 2006), in the Robust dataset, as it is not relevant in the other datasets. GMAP tries to promote systems which are able to perform well for all topics, in contrast to systems that perform better in some but worse in others. The figures show that RQE approximate the performance of PRF, showing that it perform better for difficult topics.

Combining PRF and RQE. In a preliminary experiment, we added the expansion terms produced both by RQE and PRF, obtaining a **MAP of 37.67** in the Robust collection, the best result. We would like to explore the potential for combination further in the future.

6 Conclusions

Motivated by the recent success of knowledge-based methods in word similarity and relatedness tasks (Agirre et al., 2009b), we explored a generic method to improve IR results using WordNet-based query expansion, and compared it to baseline query likelihood and pseudo-relevance feedback methods.

Our results on a diverse range of ad-hoc datasets with different domains, topic typologies and document lengths show that our method improves over a query likelihood baseline in all three datasets, while Pseudo Relevance Feedback is beneficial in only two datasets. Our method compares favorably to PRF in two datasets, and, in a preliminary experiment, the combination of PRF and our method yielded the best results in the third dataset.

In the future, we would like to analyze the differences between PRF and our method, and explore further combinations. We would also like to use our method on domains where large lexical resources are available, such as UMLS (Humphreys et al., 1998) and linked data repositories

Acknowledgments

This work has been supported by KNOW2 (TIN2009-14715-C04-01). Arantxa Otegi's work is funded by a PhD grant from the Basque Government. Part of this work was done while Arantxa Otegi was visiting ILPS group of the University of Amsterdam.

References

E. Agirre, G. M. Di Nunzio, T. Mandl, and A. Otegi. 2009a. CLEF 2009 Ad Hoc Track Overview: Robust - WSD Task. In *Working Notes of the Cross-Lingual Evaluation Forum*.

E. Agirre, A. Soroa, E. Alfonseca, K. Hall, J. Kravalova, and M. Pasca. 2009b. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proc. of NAACL*, Boulder, USA.

E. Agirre, X. Arregi, and A. Otegi. 2010. Document expansion based on WordNet for robust IR. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 9–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, Mass.

T. Hughes and D. Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of EMNLP-CoNLL-2007*, pages 581–589.

L. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 1(5):1–11.

V. Lavrenko and W. B. Croft. 2001. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

S. Liu, C. Yu, and W. Meng. 2005. Word sense disambiguation in queries. In *Proceedings of CIKM '05*, pages 525–532.

C. D. Manning, P. Raghavan, and H. Schütze. 2009. *An introduction to information retrieval*. Cambridge University Press, UK.

A. Peñas, P. Forner, R. Sutcliffe, A. Rodrigo, C. Forăscu, I. Alegria, D. Giampiccolo, N. Moreau, and P. Osenova. 2009. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In *Working Notes of the Cross-Lingual Evaluation Forum*.

J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

S. Robertson. 2006. On GMAP: and other transformations. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 78–83, New York, NY, USA. ACM.

J. J. Rocchio. 1971. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall.

M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of CIKM 2007*, Lisboa, Portugal.

T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. 2005. Indri: a language-model based search engine for complex queries. Technical report, in Proceedings of the International Conference on Intelligent Analysis.

M. Surdeanu, M. Ciaramita, and H. Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of ACL 2008*.

H. Turtle and W. B. Croft. 1991. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9:187–222, July.

E. M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR '94*, page 69.

J. Xu and W. B. Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA. ACM.

C. Zhai and J. Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 334–342, New York, NY, USA. ACM.

SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval

Eneko Agirre
IXA NLP group
University of the Basque Country
Donostia, Basque Counntry
e.agirre@ehu.es

Bernardo Magnini
ITC-IRST
Trento, Italy
magnini@itc.it

Oier Lopez de Lacalle
IXA NLP group
University of the Basque Country
Donostia, Basque Country
jibloleo@ehu.es

Arantxa Otegi
IXA NLP group
University of the Basque Country
Donostia, Basque Country
jibotusa@ehu.es

German Rigau
IXA NLP group
University of the Basque Country
Donostia, Basque Country
german.rigau@ehu.es

Piek Vossen
Irion Technologies
Delftechpark 26
2628XH Delft, Netherlands
Piek.Vossen@irion.nl

Abstract

This paper presents a first attempt of an application-driven evaluation exercise of WSD. We used a CLIR testbed from the Cross Lingual Evaluation Forum. The expansion, indexing and retrieval strategies where fixed by the organizers. The participants had to return both the topics and documents tagged with WordNet 1.6 word senses. The organization provided training data in the form of a pre-processed Semcor which could be readily used by participants. The task had two participants, and the organizer also provide an in-house WSD system for comparison.

1 Introduction

Since the start of Senseval, the evaluation of Word Sense Disambiguation (WSD) as a separate task is a mature field, with both lexical-sample and all-words tasks. In the first case the participants need to tag the occurrences of a few words, for which hand-tagged data has already been provided. In the all-words task all the occurrences of open-class words occurring in two or three documents (a few thousand words) need to be disambiguated.

The community has long mentioned the necessity of evaluating WSD in an application, in order to check which WSD strategy is best, and more important, to try to show that WSD can make a difference in applications. The use of WSD in Machine Translation has been the subject of some recent papers, but less attention has been paid to Information Retrieval (IR).

With this proposal we want to make a first try to define a task where WSD is evaluated with respect to an Information Retrieval and Cross-Lingual Information Retrieval (CLIR) exercise. From the WSD perspective, this task will evaluate all-words WSD systems indirectly on a real task. From the CLIR perspective, this task will evaluate which WSD systems and strategies work best.

We are conscious that the number of possible configurations for such an exercise is very large (including sense inventory choice, using word sense induction instead of disambiguation, query expansion, WSD strategies, IR strategies, etc.), so this first edition focuses on the following:

- The IR/CLIR system is fixed.
- The expansion / translation strategy is fixed.
- The participants can choose the best WSD strategy.

- The IR system is used as the upperbound for the CLIR systems.

We think that it is important to start doing this kind of application-driven evaluations, which might shed light to the intricacies in the interaction between WSD and IR strategies. We see this as the first of a series of exercises, and one outcome of this task should be that both WSD and CLIR communities discuss together future evaluation possibilities.

This task has been organized in collaboration with the Cross-Language Evaluation Forum (CLEF[1]). The results will be analyzed in the CLEF-2007 workshop, and a special track will be proposed for CLEF-2008, where CLIR systems will have the opportunity to use the annotated data produced as a result of the Semeval-2007 task. The task has a webpage with all the details at `http://ixa2.si.ehu.es/semeval-clir`.

This paper is organized as follows. Section 2 describes the task with all the details regarding datasets, expansion/translation, the IR/CLIR system used, and steps for participation. Section 3 presents the evaluation performed and the results obtained by the participants. Finally, Section 4 draws the conclusions and mention the future work.

2 Description of the task

This is an application-driven task, where the application is a fixed CLIR system. Participants disambiguate text by assigning WordNet 1.6 synsets and the system will do the expansion to other languages, index the expanded documents and run the retrieval for all the languages in batch. The retrieval results are taken as the measure for fitness of the disambiguation. The modules and rules for the expansion and the retrieval will be exactly the same for all participants.

We proposed two specific subtasks:

1. Participants disambiguate the corpus, the corpus is expanded to synonyms/translations and we measure the effects on IR/CLIR. Topics[2] are not processed.
2. Participants disambiguate the topics per language, we expand the queries to synonyms/translations and we measure the effects on IR/CLIR. Documents are not processed

The corpora and topics were obtained from the ad-hoc CLEF tasks. The supported languages in the topics are English and Spanish, but in order to limit the scope of the exercise we decided to only use English documents. The participants only had to disambiguate the English topics and documents. Note that most WSD systems only run on English text.

Due to these limitations, we had the following evaluation settings:

IR with WSD of topics , where the participants disambiguate the documents, the disambiguated documents are expanded to synonyms, and the original topics are used for querying. All documents and topics are in English.

IR with WSD of documents , where the participants disambiguate the topics, the disambiguated topics are expanded and used for querying the original documents. All documents and topics are in English.

CLIR with WSD of documents , where the participants disambiguate the documents, the disambiguated documents are translated, and the original topics in Spanish are used for querying. The documents are in English and the topics are in Spanish.

We decided to focus on CLIR for evaluation, given the difficulty of improving IR. The IR results are given as illustration, and as an upperbound of the CLIR task. This use of IR results as a reference for CLIR systems is customary in the CLIR community (Harman, 2005).

2.1 Datasets

The English CLEF data from years 2000-2005 comprises corpora from ’Los Angeles Times’ (year 1994) and ’Glasgow Herald’ (year 1995) amounting to 169,477 documents (579 MB of raw text, 4.8GB in the XML format provided to participants, see Section 2.3) and 300 topics in English and Spanish (the topics are human translations of each other). The relevance judgments were taken from CLEF. This

[1] `http://www.clef-campaign.org`

[2] In IR topics are the short texts which are used by the systems to produce the queries. They usually provide extensive information about the text to be searched, which can be used both by the search engine and the human evaluators.

might have the disadvantage of having been produced by pooling the results of CLEF participants, and might bias the results towards systems not using WSD, specially for monolingual English retrieval. We are considering the realization of a post-hoc analysis of the participants results in order to analyze the effect on the lack of pooling.

Due to the size of the document collection, we decided that the limited time available in the competition was too short to disambiguate the whole collection. We thus chose to take a sixth part of the corpus at random, comprising 29,375 documents (874MB in the XML format distributed to participants). Not all topics had relevant documents in this 17% sample, and therefore only 201 topics were effectively used for evaluation. All in all, we reused 21,797 relevance judgements that contained one of the documents in the 17% sample, from which 923 are positive[3]. For the future we would like to use the whole collection.

2.2 Expansion and translation

For expansion and translation we used the publicly available Multilingual Central Repository (MCR) from the MEANING project (Atserias et al., 2004). The MCR follows the EuroWordNet design, and currently includes English, Spanish, Italian, Basque and Catalan wordnets tightly connected through the Interlingual Index (based on WordNet 1.6, but linked to all other WordNet versions).

We only expanded (translated) the senses returned by the WSD systems. That is, given a word like ‘car’, it will be expanded to ‘automobile’ or ‘railcar’ (and translated to ’auto’ or ‘vagón’ respectively) depending on the sense in WN 1.6. If the systems returns more than one sense, we choose the sense with maximum weight. In case of ties, we expand (translate) all. The participants could thus implicitly affect the expansion results, for instance, when no sense could be selected for a target noun, the participants could either return nothing (or NOSENSE, which would be equivalent), or all senses with 0 score. In the first case no expansion would be performed, in the second all senses would be expanded, which is equivalent to full expansion. This fact will be mentioned again in Section 3.5.

Note that in all cases we never delete any of the words in the original text.

In addition to the expansion strategy used with the participants, we tested other expansion strategies as baselines:

noexp no expansion, original text

fullexp expansion (translation in the case of English to Spanish expansion) to all synonyms of all senses

wsd50 expansion to the best 50% senses as returned by the WSD system. This expansion was tried over the in-house WSD system of the organizer only.

2.3 IR/CLIR system

The retrieval engine is an adaptation of the TwentyOne search system (Hiemstra and Kraaij, 1998) that was developed during the 90’s by the TNO research institute at Delft (The Netherlands) getting good results on IR and CLIR exercises in TREC (Harman, 2005). It is now further developed by Irion technologies as a cross-lingual retrieval system (Vossen et al.,). For indexing, the TwentyOne system takes Noun Phrases as an input. Noun Phases (NPs) are detected using a chunker and a word form with POS lexicon. Phrases outside the NPs are not indexed, as well as non-content words (determiners, prepositions, etc.) within the phrase.

The Irion TwentyOne system uses a two-stage retrieval process where relevant documents are first extracted using a vector space matching and secondly phrases are matched with specific queries. Likewise, the system is optimized for high-precision phrase retrieval with short queries (1 up 5 words with a phrasal structure as well). The system can be stripped down to a basic vector space retrieval system with an tf.idf metrics that returns documents for topics up to a length of 30 words. The stripped-down version was used for this task to make the retrieval results compatible with the TREC/CLEF system.

The Irion system was also used for preprocessing. The CLEF corpus and topics were converted to the TwentyOne XML format, normalized, and named-entities and phrasal structured detected. Each of the target tokens was identified by an unique identifier.

2.4 Participation

The participants were provided with the following:

[3] The overall figures are 125,556 relevance judgements for the 300 topics, from which 5700 are positive

1. the document collection in Irion XML format
2. the topics in Irion XML format

In addition, the organizers also provided some of the widely used WSD features in a word-to-word fashion[4] (Agirre et al., 2006) in order to make participation easier. These features were available for both topics and documents as well as for all the words with frequency above 10 in SemCor 1.6 (which can be taken as the training data for supervised WSD systems). The Semcor data is publicly available [5]. For the rest of the data, participants had to sign and end user agreement.

The participants had to return the input files enriched with WordNet 1.6 sense tags in the required XML format:

1. for all the documents in the collection
2. for all the topics

Scripts to produce the desired output from word-to-word files and the input files were provided by organizers, as well as DTD's and software to check that the results were conformant to the respective DTD's.

3 Evaluation and results

For each of the settings presented in Section 2 we present the results of the participants, as well as those of an in-house system presented by the organizers. Please refer to the system description papers for a more complete description. We also provide some baselines and alternative expansion (translation) strategies. All systems are evaluated according to their Mean Average Precision [6] (MAP) as computed by the `trec_eval` software on the pre-existing CLEF relevance-assessments.

3.1 Participants

The two systems that registered sent the results on time.

PUTOP They extend on McCarthy's predominant sense method to create an unsupervised method of word sense disambiguation that uses automatically derived topics using Latent Dirichlet Allocation. Using topic-specific synset similarity measures, they create predictions for each word in each document using only word frequency information. The disambiguation process took aprox. 12 hours on a cluster of 48 machines (dual Xeons with 4GB of RAM). Note that contrary to the specifications, this team returned WordNet 2.1 senses, so we had to map automatically to 1.6 senses (Daude et al., 2000).

UNIBA This team uses a a knowledge-based WSD system that attempts to disambiguate all words in a text by exploiting WordNet relations. The main assumption is that a specific strategy for each Part-Of-Speech (POS) is better than a single strategy. Nouns are disambiguated basically using hypernymy links. Verbs are disambiguated according to the nouns surrounding them, and adjectives and adverbs use glosses.

ORGANIZERS In addition to the regular participants, and out of the competition, the organizers run a regular supervised WSD system trained on Semcor. The system is based on a single k-NN classifier using the features described in (Agirre et al., 2006) and made available at the task website (cf. Section 2.4).

In addition to those we also present some common IR/CLIR baselines, baseline WSD systems, and an alternative expansion:

noexp a non-expansion IR/CLIR baseline of the documents or topics.

fullexp a full-expansion IR/CLIR baseline of the documents or topics.

wsdrand a WSD baseline system which chooses a sense at random. The usual expansion is applied.

1st a WSD baseline system which returns the sense numbered as 1 in WordNet. The usual expansion is applied.

wsd50 the organizer's WSD system, where the 50% senses of the word ranking according to the WSD system are expanded. That is, instead of expanding the single best sense, it expands the best 50% senses.

3.2 IR Results

This section present the results obtained by the participants and baselines in the two IR settings. The

[4] Each target word gets a file with all the occurrences, and each occurrence gets the occurrence identifier, the sense tag (if in training), and the list of features that apply to the occurrence.

[5] `http://ixa2.si.ehu.es/semeval-clir/`

[6] `http://en.wikipedia.org/wiki/Information_retrieval`

	IRtops	IRdocs	CLIR
no expansion	0.3599	0.3599	0.1446
full expansion	0.1610	0.1410	0.2676
UNIBA	0.3030	0.1521	0.1373
PUTOP	0.3036	0.1482	0.1734
wsdrand	0.2673	0.1482	0.2617
1st sense	0.2862	0.1172	0.2637
ORGANIZERS	0.2886	0.1587	0.2664
wsd50	0.2651	0.1479	0.2640

Table 1: Retrieval results given as MAP. IRtops stands for English IR with topic expansion. IRdocs stands for English IR with document expansion. CLIR stands for CLIR results for translated documents.

second and third columns of Table 1 present the results when disambiguating the topics and the documents respectively. Non of the expansion techniques improves over the baseline (no expansion).

Note that due to the limitation of the search engine, long queries were truncated at 50 words, which might explain the very low results of the full expansion.

3.3 CLIR results

The last column of Table 1 shows the CLIR results when expanding (translating) the disambiguated documents. None of the WSD systems attains the performance of full expansion, which would be the baseline CLIR system, but the WSD of the organizer gets close.

3.4 WSD results

In addition to the IR and CLIR results we also provide the WSD performance of the participants on the Senseval 2 and 3 all-words task. The documents from those tasks were included alongside the CLEF documents, in the same formats, so they are treated as any other document. In order to evaluate, we had to map automatically all WSD results to the respective WordNet version (using the mappings in (Daude et al., 2000) which are publicly available).

The results are presented in Table 2, where we can see that the best results are attained by the organizers WSD system.

3.5 Discussion

First of all, we would like to mention that the WSD and expansion strategy, which is very simplistic, degrades the IR performance. This was rather expected, as the IR experiments had an illustration goal, and are used for comparison with the CLIR experiments. In monolingual IR, expanding the topics is much less harmful than expanding the documents. Unfortunately the limitation to 50 words in the queries might have limited the expansion of the topics, which make the results rather unreliable. We plan to fix this for future evaluations.

Senseval-2 all words			
	precision	recall	coverage
ORGANIZERS	0.584	0.577	93.61%
UNIBA	0.498	0.375	75.39%
PUTOP	0.388	0.240	61.92%
Senseval-3 all words			
	precision	recall	coverage
ORGANIZERS	0.591	0.566	95.76%
UNIBA	0.484	0.338	69.98%
PUTOP	0.334	0.186	55.68%

Table 2: English WSD results in the Senseval-2 and Senseval-3 all-words datasets.

Regarding CLIR results, even if none of the WSD systems were able to beat the full-expansion baseline, the organizers system was very close, which is quite encouraging due to the very simplistic expansion, indexing and retrieval strategies used.

In order to better interpret the results, Table 3 shows the amount of words after the expansion in each case. This data is very important in order to understand the behavior of each of the systems. Note that UNIBA returns 3 synsets at most, and therefore the wsd50 strategy (select the 50% senses with best score) leaves a single synset, which is the same as taking the single best system (wsdbest). Regarding PUTOP, this system returned a single synset, and therefore the wsd50 figures are the same as the wsdbest figures.

Comparing the amount of words for the two participant systems, we see that UNIBA has the least words, closely followed by PUTOP. The organizers WSD system gets far more expanded words. The explanation is that when the synsets returned by a WSD system all have 0 weights, the wsdbest expansion strategy expands them all. This was not explicit in the rules for participation, and might have affected the results.

A cross analysis of the result tables and the number of words is interesting. For instance, in the IR exercise, when we expand documents, the results in

		English	Spanish
No WSD	noexp	9,900,818	9,900,818
	fullexp	93,551,450	58,491,767
UNIBA	wsdbest	19,436,374	17,226,104
	wsd50	19,436,374	17,226,104
PUTOP	wsdbest	20,101,627	16,591,485
	wsd50	20,101,627	16,591,485
Baseline	1st	24,842,800	20,261,081
WSD	wsdrand	24,904,717	19,137,981
ORG.	wsdbest	26,403,913	21,086,649
	wsd50	36,128,121	27,528,723

Table 3: Number of words in the document collection after expansion for the WSD system and all baselines. wsdbest stands for the expansion strategy used with participants.

the third column of Table 1 show that the ranking for the non-informed baselines is the following: best for no expansion, second for random WSD, and third for full expansion. These results can be explained because of the amount of expansion: the more expansion the worst results. When more informed WSD is performed, documents with more expansion can get better results, and in fact the WSD system of the organizers is the second best result from all system and baselines, and has more words than the rest (with exception of wsd50 and full expansion). Still, the no expansion baseline is far from the WSD results.

Regarding the CLIR result, the situation is inverted, with the best results for the most productive expansions (full expansion, random WSD and no expansion, in this order). For the more informed WSD methods, the best results are again for the organizers WSD system, which is very close to the full expansion baseline. Even if wsd50 has more expanded words wsdbest is more effective. Note the very high results attained by random. These high results can be explained by the fact that many senses get the same translation, and thus for many words with few translation, the random translation might be valid. Still the wsdbest, 1st sense and wsd50 results get better results.

4 Conclusions and future work

This paper presents the results of a preliminary attempt of an application-driven evaluation exercise of WSD in CLIR. The expansion, indexing and retrieval strategies proved too simplistic, and none of the two participant systems and the organizers system were able to beat the full-expansion baseline. Due to efficiency reasons, the IRION system had some of its features turned off. Still the results are encouraging, as the organizers system was able to get very close to the full expansion strategy with much less expansion (translation).

For the future, a special track of CLEF-2008 will leave the avenue open for more sophisticated CLIR techniques. We plan to extend the WSD annotation to all words in the CLEF English document collection, and we also plan to contact the best performing systems of the SemEval all-words tasks to have better quality annotations.

Acknowledgements

We wish to thank CLEF for allowing us to use their data, and the CLEF coordinator, Carol Peters, for her help and collaboration. This work has been partially funded by the Spanish education ministry (project KNOW)

References

E. Agirre, O. Lopez de Lacalle, and D. Martinez. 2006. Exploring feature set combinations for WSD. In *Proc. of the SEPLN*.

J. Atserias, L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, and P. Vossen. 2004. The MEANING Multilingual Central Repository. In *Proceedings of the 2.nd Global WordNet Conference, GWC 2004*, pages 23–30. Masaryk University, Brno, Czech Republic.

J. Daude, L. Padro, and G. Rigau. 2000. Mapping WordNets Using Structural Information. In *Proc. of ACL*, Hong Kong.

D. Harman. 2005. Beyond English. In E. M. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, pages 153–181. MIT press.

D. Hiemstra and W. Kraaij. 1998. Twenty-One in ad-hoc and CLIR. In E.M. Voorhees and D. K. Harman, editors, *Proc. of TREC-7*, pages 500–540. NIST Special Publication.

P. Vossen, G. Rigau, I. Alegria, E. Agirre, D. Farwell, and M. Fuentes. Meaningful results for Information Retrieval in the MEANING project. In *Proc. of the 3rd Global Wordnet Conference*.

CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task

Eneko Agirre[1], Giorgio Maria Di Nunzio[2], Thomas Mandl[3], and
Arantxa Otegi[1]

[1] Computer Science Department, University of the Basque Country, Spain
{e.agirre,arantza.otegi}@ehu.es
[2] Department of Information Engineering, University of Padua, Italy
{dinunzio}@dei.unipd.it
[3] Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. The Robust-WSD at CLEF 2009 aims at exploring the contribution of Word Sense Disambiguation to monolingual and multilingual Information Retrieval. The organizers of the task provide documents and topics which have been automatically tagged with Word Senses from WordNet using several state-of-the-art Word Sense Disambiguation systems. The Robust-WSD exercise follows the same design as in 2008. It uses two languages often used in previous CLEF campaigns (English, Spanish). Documents were in English, and topics in both English and Spanish. The document collections are based on the widely used LA94 and GH95 news collections. All instructions and datasets required to replicate the experiment are available from the organizers website (http://ixa2.si.ehu.es/clirwsd/). The results show that some top-scoring systems improve their IR and CLIR results with the use of WSD tags, but the best scoring runs do not use WSD.

1 Introduction

The Robust-WSD task at CLEF 2009 aims at exploring the contribution of Word Sense Disambiguation to monolingual and multilingual Information Retrieval. The organizers of the task provide documents and topics which have been automatically tagged with Word Senses from WordNet using several state-of-the-art Word Sense Disambiguation systems. The task follows the same design as in 2008.

The robust task ran for the fourth time at CLEF 2009. It is an Ad-Hoc retrieval task based on data of previous CLEF campaigns. The robust task emphasizes the difficult topics by a non-linear integration of the results of individual topics into one result for a system, using the geometric mean of the average precision for all topics (GMAP) as an additional evaluation measure [13,14]. Given the difficulty of the task, training data including topics and relevance assessments was provided for the participants to tune their systems to the collection.

For the second year, the robust task also incorporated word sense disambiguation information provided by the organizers to the participants. The task follows the 2007 joint SemEval-CLEF task [2] and the 2008 Robust-WSD exercise [3], and has the aim of exploring the contribution of word sense disambiguation to monolingual and cross-language information retrieval. The goal of the task is to test whether WSD can be used beneficially for retrieval systems, and thus participants were required to submit at least one baseline run without WSD and one run using the WSD annotations. Participants could also submit four further baseline runs without WSD and four runs using WSD.

The experiment involved both monolingual (topics and documents in English) and bilingual experiments (topics in Spanish and documents in English). In addition to the original documents and topics, the organizers of the task provided both documents and topics which had been automatically tagged with word senses from WordNet version 1.6 using two state-of-the-art word sense disambiguation systems, UBC [1] and NUS [7]. These systems provided weighted word sense tags for each of the nouns, verbs, adjectives and adverbs that they could disambiguate.

In addition, the participants could use publicly available data from the English and Spanish wordnets in order to test different expansion strategies. Note that given the tight alignment of the Spanish and English wordnets, the wordnets could also be used to translate directly from one sense to another, and perform expansion to terms in another language.

The datasets used in this task can be used in the future to run further experiments. Check `http://ixa2.si.ehu.es/clirwsd` for information of how to access the datasets. Topics and relevance judgements are freely available. The document collection can be obtained from ELDA purchasing the CLEF Test Suite for the CLEF 2000-2003 Campaigns – Evaluation Package. As an alternative, the website offers the unordered set of words in each document, that is, the full set of documents where the positional information has been eliminated to avoid replications of the originals. Lucene indexes for the later are also available from the website.

In this paper, we first present the task setup, the evaluation methodology and the participation in the different tasks (Section 2). We then describe the main features of each task and show the results (Sections 3 - 5). The final section provides a brief summing up. For information on the various approaches and resources used by the groups participating in this task and the issues they focused on, we refer the reader to the rest of the papers in the Robust-WSD part of the Ad Hoc section of these Proceedings.

2 Task Setup

The Ad Hoc task in CLEF adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments in the late 1960s [8]. The **tasks** offered are studied in order to effectively measure textual document retrieval under specific conditions. The **test collections** are made up of **documents**, **topics** and **relevance assessments**. The topics consist of a set of statements simulating information needs from which the systems derive the queries to search the document collections. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the recall and precision measures.

2.1 Test Collections

The Documents. The robust task used existing CLEF news collections but with word sense disambiguation (WSD) information added. The word sense disambiguation data was automatically added by systems from two leading research laboratories, UBC [1] and NUS [7]. Both systems returned word senses from the English WordNet, version 1.6.

The document collections were offered both with and without WSD, and included the following[1]:

- LA Times 94 (with word sense disambiguated data); ca 113,000 documents, 425 MB without WSD, 1,448 MB (UBC) or 2,151 MB (NUS) with WSD;
- Glasgow Herald 95 (with word sense disambiguated data); ca 56,500 documents, 154 MB without WSD, 626 MB (UBC) or 904 MB (NUS) with WSD.

The Topics. Topics are structured statements representing information needs. Each topic typically consists of three parts: a brief title statement; a one-sentence description; a more complex narrative the relevance assessment criteria. Topics are prepared in xml format and identified by means of a Digital Object Identifier (DOI)[2] of the experiment [12] which allows us to reference and cite them.

The WSD robust task used existing CLEF topics in English and Spanish as follows:

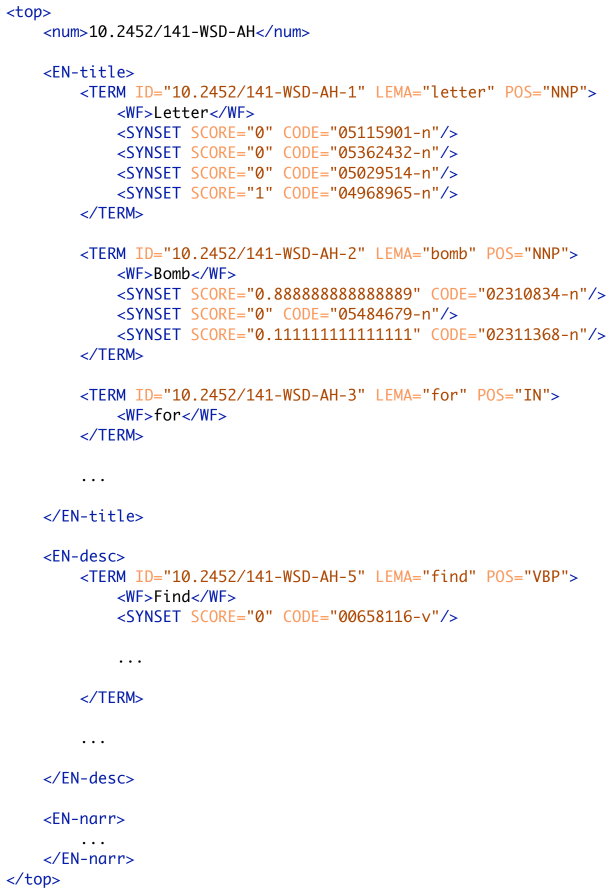
- CLEF 2001; Topics 10.2452/41-AH – 10.2452/90-AH; LA Times 94
- CLEF 2002; Topics 10.2452/91-AH – 10.2452/140-AH; LA Times 94
- CLEF 2003; Topics 10.2452/141-AH – 10.2452/200-AH; LA Times 94, Glasgow Herald 95
- CLEF 2004; Topics 10.2452/201-AH – 10.2452/250-AH; Glasgow Herald 95
- CLEF 2005; Topics 10.2452/251-AH – 10.2452/300-AH; LA Times 94, Glasgow Herald 95
- CLEF 2006; Topics 10.2452/301-AH – 10.2452/350-AH; LA Times 94, Glasgow Herald 95

Topics from years 2001, 2002 and 2004 were used as training topics (relevance assessments were offered to participants), and topics from years 2003, 2005 and 2006 were used for the test.

All topics were offered both with and without WSD. Topics in English were disambiguated by both UBC [1] and NUS [7] systems, yielding word senses from

[1] A sample document and dtd are available at `http://ixa2.si.ehu.es/clirwsd/`

[2] `http://www.doi.org/`



```
<top>
    <num>10.2452/141-WSD-AH</num>

    <EN-title>
        <TERM ID="10.2452/141-WSD-AH-1" LEMA="letter" POS="NNP">
            <WF>Letter</WF>
            <SYNSET SCORE="0" CODE="05115901-n"/>
            <SYNSET SCORE="0" CODE="05362432-n"/>
            <SYNSET SCORE="0" CODE="05029514-n"/>
            <SYNSET SCORE="1" CODE="04968965-n"/>
        </TERM>

        <TERM ID="10.2452/141-WSD-AH-2" LEMA="bomb" POS="NNP">
            <WF>Bomb</WF>
            <SYNSET SCORE="0.888888888888889" CODE="02310834-n"/>
            <SYNSET SCORE="0" CODE="05484679-n"/>
            <SYNSET SCORE="0.111111111111111" CODE="02311368-n"/>
        </TERM>

        <TERM ID="10.2452/141-WSD-AH-3" LEMA="for" POS="IN">
            <WF>for</WF>
        </TERM>

        ...

    </EN-title>

    <EN-desc>
        <TERM ID="10.2452/141-WSD-AH-5" LEMA="find" POS="VBP">
            <WF>Find</WF>
            <SYNSET SCORE="0" CODE="00658116-v"/>

            ...

        </TERM>

        ...

    </EN-desc>

    <EN-narr>
        ...
    </EN-narr>
</top>
```

Fig. 1. Example of Robust WSD topic: topic `10.2452/141-WSD-AH`.

WordNet version 1.6. A large-scale disambiguation system for Spanish was not available, so we used the first-sense heuristic, yielding senses from the Spanish wordnet, which is tightly aligned to the English WordNet version 1.6 (i.e., they share synset numbers or sense codes). An excerpt from a topic is shown in Figure 1, where each term in the topic is followed by its senses with their respective scores as assigned buy the automatic WSD system[3].

Relevance Assessment. The number of documents in large test collections such as CLEF makes it impractical to judge every document for relevance. Instead approximate recall values are calculated using pooling techniques. The robust WSD task used existing relevance assessments from previous years. The

[3] Full sample and dtd are available at `http://ixa2.si.ehu.es/clirwsd/`

relevance assessments regarding the training topics were provided to participants before competition time.

The total number of assessments was 66,441 documents of which 4,327 were relevant. The distribution of the pool according to each year was the following:

- CLEF 2003: 23,674 documents, 1,006 relevant;
- CLEF 2005: 19,790 document, 2,063 relevant;
- CLEF 2006: 21,247 document, 1,258 relevant;

Seven topics had no relevant documents at all: 10.2452/149-AH, 10.2452/161-AH, 10.2452/166-AH, 10.2452/186-AH, 10.2452/191-AH, 10.2452/195-AH, 10.2-452/321-AH. Each topic had an average of about 28 relevant documents and a standard deviation of 34, a minimum of 1 relevant document and a maximum of 229 relevant documents per topic.

2.2 Result Calculation

Evaluation campaigns such as TREC and CLEF are based on the belief that the effectiveness of *Information Retrieval Systems (IRSs)* can be objectively evaluated by an analysis of a representative set of sample search results. For this, effectiveness measures are calculated based on the results submitted by the participants and the relevance assessments. Popular measures usually adopted for exercises of this type are Recall and Precision. Details on how they are calculated for CLEF are given in [6].

The robust task emphasizes the difficult topics by a non-linear integration of the results of individual topics into one result for a system, using the geometric mean of the average precision for all topics (GMAP) as an additional evaluation measure [13,14].

The individual results for all official Ad Hoc experiments in CLEF 2009 are given in the one of the Appendices of the CLEF 2009 Working Notes [9].

2.3 Participants and Experiments

As shown in Table 1, 10 groups submitted 89 runs for the Robust tasks:

- 8 groups submitted monolingual non-WSD runs (25 runs out of 89);
- 5 groups also submitted bilingual non-WSD runs (13 runs out of 89).

All groups submitted WSD runs (51 out of 89 runs):

- 10 groups submitted monolingual WSD runs (33 out of 89 runs)
- 5 groups submitted bilingual WSD runs (18 out of 89 runs)

Table 2 provides a breakdown of the number of participants and submitted runs by task. Note that jaen submitted a monolingual non-WSD run as if it was a WSD run, and that alicante missed to send their non-WSD run on time. The figures below are the official figures.

Table 1. CLEF 2009 Ad Hoc Robust participants

participant	task	No. experiments
alicante	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	3
darmstadt	AH-ROBUST-MONO-EN-TEST-CLEF2009	5
darmstadt	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	5
geneva	AH-ROBUST-MONO-EN-TEST-CLEF2009	5
geneva	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	1
geneva	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	2
ixa	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	1
ixa	AH-ROBUST-MONO-EN-TEST-CLEF2009	1
ixa	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	4
ixa	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	3
jaen	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	2
know-center	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	3
know-center	AH-ROBUST-MONO-EN-TEST-CLEF2009	3
know-center	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	3
know-center	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	3
reina	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	5
reina	AH-ROBUST-MONO-EN-TEST-CLEF2009	5
reina	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	5
reina	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	5
ufrgs	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	1
ufrgs	AH-ROBUST-MONO-EN-TEST-CLEF2009	1
ufrgs	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	1
uniba	AH-ROBUST-BILI-X2EN-TEST-CLEF2009	3
uniba	AH-ROBUST-MONO-EN-TEST-CLEF2009	3
uniba	AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009	5
uniba	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	5
valencia	AH-ROBUST-MONO-EN-TEST-CLEF2009	2
valencia	AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009	4

Table 2. Number of runs per track.

Track	**# Part.**	**# Runs**
Robust Mono English Test	8	25
Robust Mono English Test WSD	10	33
Robust Biling. English Test	5	13
Robust Biling. English Test WSD	5	18

3 Results

Table 3 shows the best results for the monolingual runs, and Table 4 shows the best results for the bilingual runs. In the following pages, Figures 2 and 3 compare the performances of the best systems in terms of average precision of the top participants of the Robust Monolingual and Monolingual WSD, and Figures 4 and 5 compare the performances of the best participants of the Robust Bilingual and Bilingual WSD.

Table 3. Best entries for the robust monolingual task.

Track	Rank	Participant	Experiment DOI	MAP	GMAP
English	**1st**	darmstadt	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.DARMSTADT.DA_4	45.09%	20.42%
	2nd	reina	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.REINA.ROB2	44.52%	21.18%
	3rd	uniba	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.UNIBA.UNIBAKRF	42.50%	17.93%
	4th	geneva	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.GENEVA.ISIENNATTDN	41.71%	17.88%
	5th	know-center	10.2415/AH-ROBUST-MONO-EN-TEST-CLEF2009.KNOW-CENTER.ASSO	41.70%	18.64%
English WSD	**1st**	darmstadt	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.DARMSTADT.DA_WSD_4	45.00%	20.49%
	2nd	uniba	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.UNIBA.UNIBAKEYSYNRF	43.46%	19.60%
	3rd	know-center	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.KNOW-CENTER.ASSOWSD	42.22%	19.47%
	4th	reina	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.REINA.ROBWSD2	41.23%	18.38%
	5th	geneva	10.2415/AH-ROBUST-WSD-MONO-EN-TEST-CLEF2009.GENEVA.ISINUSLWTDN	38.11%	16.26%

Table 4. Best entries for the robust bilingual task.

Track	Rank	Participant	Experiment DOI	MAP	GMAP
Es-En	**1st**	reina	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.REINA.BILI2	38.42%	15.11%
	2nd	uniba	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.UNIBA.UNIBACROSSKEYRF	38.09%	13.11%
	3rd	know-center	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.KNOW-CENTER.BILIASSO	28.98%	06.79%
	4th	ufrgs	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.UFRGS.BILINGUAL	27.65%	07.37%
	5th	ixa	10.2415/AH-ROBUST-BILI-X2EN-TEST-CLEF2009.IXA.ESENNOWSD	18.05%	01.90%
Es-En WSD	**1st**	uniba	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.UNIBA.UNIBACROSSKEYSYNRF	37.53%	13.82%
	2nd	geneva	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.GENEVA.ISINUSWSDTD	36.63%	16.02%
	3rd	reina	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.REINA.BILIWSD2	30.32%	09.38%
	4th	know-center	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.KNOW-CENTER.BILIASSOWSD	29.64%	07.05%
	5th	ixa	10.2415/AH-ROBUST-WSD-BILI-X2EN-TEST-CLEF2009.IXA.ESEN1STTOPSBESTSENSE500DOCS	18.38%	01.98%

The comparison of the bilingual runs with respect to the monolingual results yield the following:

- ES → EN: 85.2% of best monolingual English IR system (MAP);
- ES → EN WSD: 83.3% of best monolingual English IR system (MAP);

3.1 Statistical Testing

When the goal is to validate how well results can be expected to hold beyond a particular set of queries, statistical testing can help to determine what differences between runs appear to be real as opposed to differences that are due to sampling issues. We aim to identify whether the results of the runs of a task are significantly different from the results of other tasks. In particular, we want to test whether there is any difference between applying WSD techniques or not. Significantly different in this context means that the difference between the performance scores for the runs in question appears greater than what might be expected by pure chance. As with all statistical testing, conclusions will be qualified by an error probability, which was chosen to be 0.05 in the following.

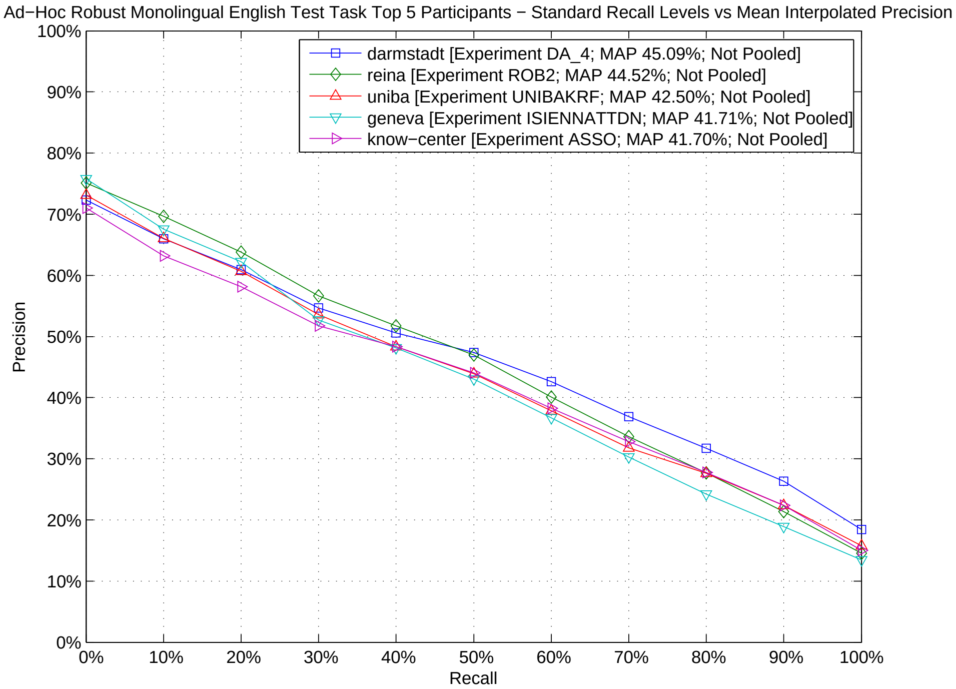


Fig. 2. Mean average precision of the top 5 participants of the Robust Monolingual English Task.

We have designed our analysis to follow closely the methodology used by similar analyses carried out for Text REtrieval Conference (TREC) [23].

We used the MATLAB Statistics Toolbox, which provides the necessary functionality plus some additional functions and utilities.

Two tests for goodness of fit to a normal distribution were chosen using the MATLAB statistical toolbox: the Lilliefors test and the Jarque-Bera test. In the case of the CLEF tasks under analysis, both tests indicate that the assumption of normality is not violated for most of the data samples (in this case the runs for each participant).

The two tests were:

– Robust Monolingual vs Robust WSD Monolingual;
– Robust Bilingual vs Robust WSD Bilingual.

In both cases, the t-test confirmed that the mean of the two distributions are different and, in particular, the mean of the monolingual distribution is greater than the mean of the robust monolingual WSD, and the same happens for the bilingual. This suggests some loss of performances due to the effect of the word sense disambiguation in both monolingual and bilingual tasks. However, there

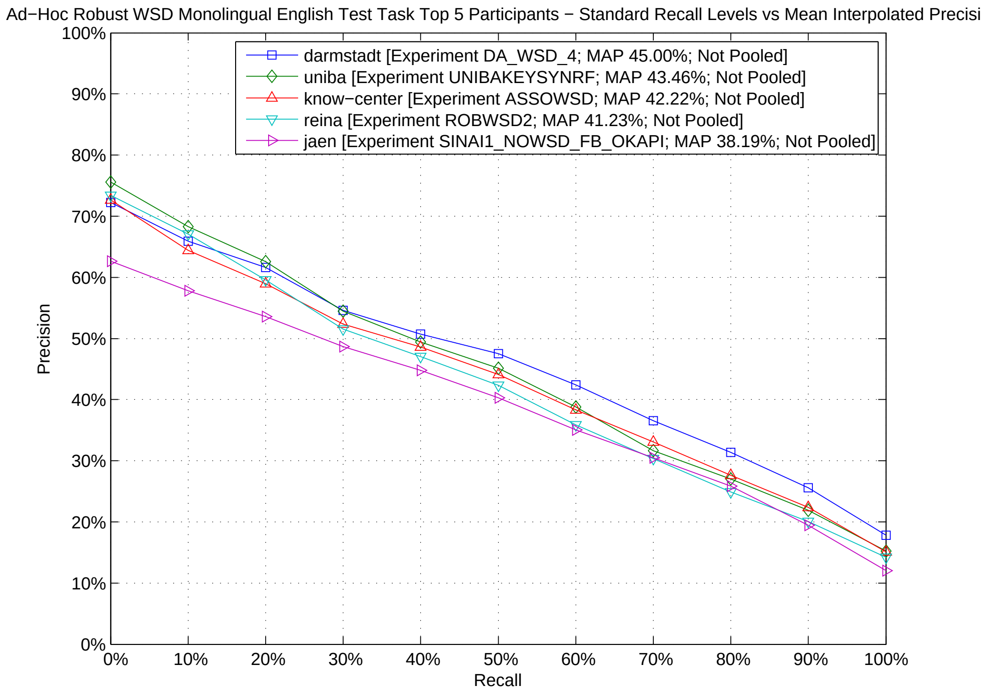


Fig. 3. Mean average precision of the top 5 participants of the Robust WSD Monolingual English Task.

are a few topics where the WSD techniques significantly improve the effectiveness of the retrieval; these are the cases worth studying from a WSD point of view.

3.2 Analysis

In this section we focus on the comparison between WSD and non-WSD runs. Overall, the best MAP and GMAP results in the monolingual system were for two distinct runs which did not use WSD information. Several participants were able to obtain their best MAP and GMAP scores using WSD information. In the bilingual experiments, the best results in MAP was for non-WSD runs, but two participants were able to profit from the WSD annotations. As it is difficult to summarize the behavior of all participants below, we will only mention the performance of the best teams, as given in Tables 3 and 4. The interested reader is directed to the working notes of each participant for additional details.

In the monolingual experiments, cf. Table 3, the best results overall in MAP was for darmstadt. Their WSD runs scored very similar to the non-WSD runs, with a slight decrease of MAP (0.09 percentage points) and a slight increase of GMAP (0.07 percentage points) [15]. The second best MAP score and best GMAP was attained by reina [16] without WSD, with their WSD systems show-

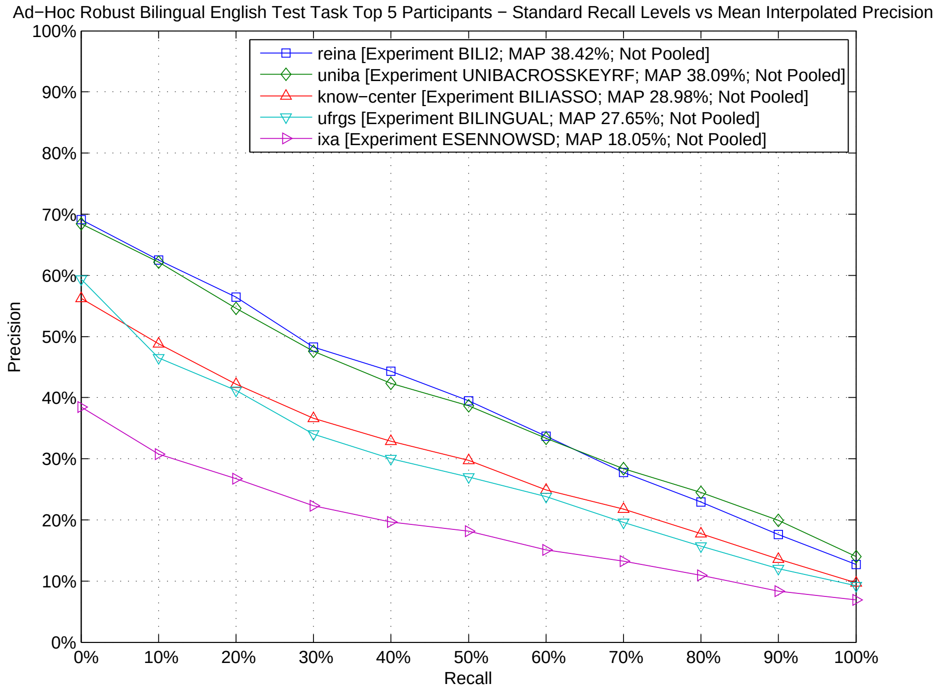


Fig. 4. Mean average precision of the top 5 participants of the Robust Bilingual English Task.

ing a considerable performance drop. The third best MAP and second GMAP where obtained by uniba [4] using WSD. This team showed a 0.94 increase in MAP and 1.67 increase in GMAP with respect to their best non-WSD run. Another team showing high MAP and GMAP values was know-center [11], which attained 0.52 improvements in MAP and 0.83 increase in GMAP with the use of WSD. Finally, geneva [10] also attained good results, but their WSD system also had a considerable drop in both MAP and GMAP. All in all, regarding the use of WSD in the monolingual task, two teams exhibited modest gains, two teams had quite large performance drops, and the teams reporting best results had very similar results.

In the bilingual experiments, cf. Table 4, the best results overall in MAP were for reina with a system which did not use WSD annotations [16]. The best GMAP was for geneva using WSD [10]. Unfortunately, they did not submit any non-WSD run. Uniba [4] got the second best MAP, with better MAP for the non-WSD run and better GMAP for the WSD run. The differences were small in both cases (0.56 in MAP, 0.71 in GMAP). Those three teams had the highest results, well over 35% MAP, and the rest got more modest performances. know-center [11] reported better results using WSD information (0.66 MAP, 0.26

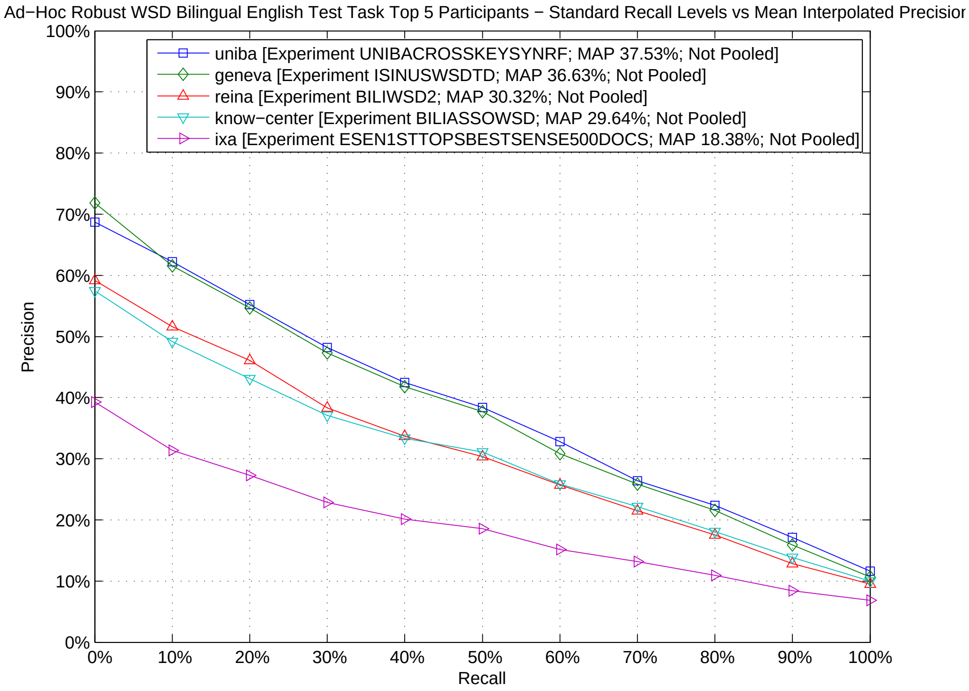


Fig. 5. Mean average precision of the top 5 participants of the Robust WSD Bilingual English Task.

GMAP). Ufrgs [5] only submitted the WSD result. Finally ixa got low results, with small improvements using WSD information (0.33 MAP, 0.08 GMAP).

All in all, the exercise showed that some teams did improve results using WSD (close to 1 MAP point and more than 1 GMAP point in monolingual, and below 1 MAP/GMAP point in bilingual), but the best results for both monolingual and bilingual tasks were for systems which did not use WSD.

4 Conclusions

This new edition of the robust WSD exercise has measured to what extent IR systems could profit from automatic word sense disambiguation information. The conclusions on the monolingual subtask are similar to the conclusions of 2008. The evidence for using WSD in monolingual IR is mixed, with some top scoring groups reporting small improvements in MAP and GMAP, but with the best overall scores for systems not using WSD.

Regarding the cross-lingual task, the situation is very similar, but the improvements reported by using WSD are smaller.

Instructions and datasets to replicate the results (including Lucene indexes) are available from `http://ixa.si.ehu.es/clirwsd`.

5 Acknowledgements

The robust task was partially funded by the Ministry of Education (project KNOW TIN2006-15049) and the European Commission (project KYOTO ICT-2007-211423). We want to thank Oier Lopez de Lacalle, who runs the UBC WSD system, and Yee Seng Chan, Hwee Tou Ng and Zhi Zhong, who run the NUS WSD system. Their generous contribution was invaluable to run this exercise.

References

1. Agirre, E., Lopez de Lacalle, O.: UBC-ALM: Combining k-NN with SVD for WSD. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval 2007), Prague, Czech Republic (2007) 341–345
2. Agirre, E., Magnini, B., Lopez de Lacalle, O., Otegi, A., Rigau, G., Vossen, P.: SemEval-2007 Task01: Evaluating WSD on Cross-Language Information Retrieval. In Proceedings of CLEF 2007 Workshop, Budapest, Hungary (2007).
3. Agirre, E., Di Nunzio, G.M., Ferro, N., Peters, C., Mandl, T.: CLEF 2008: Ad Hoc Track Overview. In Borri, F., Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2009 Workshop, http://www.clef-campaign.org/
4. Basile, P., Caputo, A., Semeraro, G.: UNIBA-SENSE at CLEF 2009: Robust WSD task. In this volume.
5. Borges, T.B., Moreira, V.P.: UFRGS@CLEF2009: Retrieval by Numbers In this volume.
6. Braschler, M., Peters, C.: CLEF 2003 Methodology and Metrics. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany (2004) 7–20
7. Chan, Y. S., Ng, H. T., Zhong, Z.: 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 2007), Prague, Czech Republic (2007) 253–256
8. Cleverdon, C.: The Cranfield Tests on Index Language Devices. In Sparck Jones, K., Willett, P., eds.: Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997) 47–59
9. Di Nunzio, G.M., Ferro, N.: Appendix C: Results of the Robust Task. In Borri, F., Nardi, A., Peters, C., eds.: Working Notes for the CLEF 2009 Workshop, http://www.clef-campaign.org/ (2008)
10. Guyot, J., Falquet, G., Radhouani, S.: UniGe at CLEF 2009 Robust WSD Task. In this volume.
11. Kern, R., Juffinger, A., Granitzer, M.: Application of Axiomatic Approaches to Crosslanguage Retrieval. In this volume.
12. Paskin, N., ed.: The DOI Handbook – Edition 4.4.1. International DOI Foundation (IDF). http://dx.doi.org/10.1000/186 (2006)
13. Robertson, S.: On GMAP: and Other Transformations. In Yu, P.S., Tsotras, V., Fox, E.A., Liu, C.B., eds.: Proc. 15th International Conference on Information and Knowledge Management (CIKM 2006), ACM Press, New York, USA (2006) 78–83
14. Voorhees, E.M.: The TREC Robust Retrieval Track. SIGIR Forum **39** (2005) 11–20

15. Wolf, E., Bernhard, D., Gurevych, I.: Combining Probabilistic and Translation-Based Models for Information Retrieval based on Word Sense Annotations Information Retrieval. In this volume.
16. Zazo, A., Figuerola, C.G., Alonso Berrocal, J.L., Gomez, R.: REINA at CLEF 2009 Robust-WSD Task: Partial Use of WSD Information for Retrieval. In this volume.

Using Knowledge-Based Relatedness for Information Retrieval

Abstract

Traditional information retrieval (IR) systems use keywords to index and retrieve documents. The limitations of keywords were recognized since the early days, specially when different but closely-related words are used in the query and the relevant document. Query expansion techniques like pseudo-relevance feedback (PRF) and document clustering techniques rely on the target document set in order to bridge the gap between those words. This paper explores the use of WordNet-based semantic relatedness in order to bridge the gap between query and documents. We performed both query expansion and document expansion, with positive effects over a language modeling baseline on three datasets (Robust, Yahoo!, ResPubliQA), and over PRF on two of those datasets (Yahoo! and ResPubliQA). Our analysis shows that our models and PRF are complementary, in that PRF is better for easy queries and our models are stronger for difficult queries. We also show that our models are more robust in face of sub-optimal parameters. Finally, in preliminary work, we present a combined system wich outperforms all individual techniques, showing promise to further improve results in the future. Our methods can be easily applied to other relevant knowledge sources like medical ontologies or linked-data repositories.

1 Introduction

The potential pitfalls of keyword retrieval have been noted since the earliest days of information retrieval (IR). Keyword retrieval proves ineffective when different but closely-related words are used in the query and the relevant document. The use of different words creates a lexical gap between the query and the document.

To exemplify this problem, Figure 1 shows some examples taken from the datasets used in this paper. In each example, there is a query (Q) and its relevant document (D), which answers the question using other related words.

Q: How *fast* does a *tractor go*?

D: This Directive shall apply only to *tractors* defined in paragraph 1 which are fitted with pneumatic tyres and which have two axles and a maximum design *speed* between 6 and 25 *kilometres per hour*.

(a)

Q: How do you *cook* an *apple pie*?

D: There are many good *recipes* for *apple pies* but there are also some important things to remember that are usually not in the recipe. That is you should make sure the bottom of the crust will *bake* as well and not remain soggy. To do this, coat the inside of the crust with butter before adding the filling and place the baking dish on a dark metal pan so the bottom will get more heat.

(b)

Figure 1: Examples of lexical gap from ResPubliQA and Yahoo! datasets

For example, the question in Fig. 1a contains *fast*, *tractor* and *go*. Only one of these words appears in the document (*tractor*), but other words related to the query are also present (*speed* and *kilometres per hour*). Something similar happens on Fig. 1b example. Instead of the keyword *cook*, related words like *recipes* or *bake* are used in the document.

In order to bridge the gap, IR has resorted to distributional semantic models. Most research concentrated on Query Expansion (QE) methods, which typically analyze term co-occurrence statistics in the corpus and/or in the highest scored documents in order to select terms for expanding the query terms (Manning et al., 2009). Pseudo-relevance feedback (PRF) is one of the most notorious techniques in this area. Document expansion (DE) is a natural alternative to QE. Several researchers have used distributional methods from similar documents in the collection in order to expand the documents with related terms that do not actually occur in the document (Liu and Croft, 2004; Kurland and Lee, 2004; Tao et al., 2006; Mei et al., 2008; Huang et al., 2009). The work presented here is complementary, in that we explore QE and DE, but use WordNet instead of distributional methods. In the future, the complementarity of the approaches could be profited to further improve performance by combining them.

WordNet has been used with great success in psycholinguistic datasets of word similarity and relatedness, where it often surpasses distributional methods based on keyword matches (Agirre et al., 2009b). It has also been applied to IR before. Some authors extended the query with synonyms from WordNet (Voorhees, 1994; Liu et al., 2005), while others have explicitly represented and indexed word senses after performing word sense disambiguation (WSD) (Gonzalo et al., 1998; Stokoe et al., 2003; Kim et al., 2004). More recently, a CLEF task was organized[1] where queries and documents where semantically disambiguated. Some high-scoring participants reported significant improvements when using WordNet information.

This paper proposes to use WordNet for query and document expansion. Given a full document, a random walk algorithm over the WordNet graph, inspired in (Agirre et al., 2009b), ranks concepts closely related to the words in the document. This is in contrast to previous WordNet-based work which focused on WSD to replace or supplement words with their senses. Our method discovers important concepts, even if they are not explicitly mentioned in the query or document. Our work follows closely (Agirre et al., 2010), which used the same WordNet-based relatedness algorithm for document expansion, but we investigate methods to apply relatedness to query expansion, and we perform a comparison with regard to pseudo-relevance feedback.

In this work we adopt a language modeling framework to implement the query likelihood and pseudo-relevance feedback baselines, as well as our relatedness-based query expansion and document expansion methods. In order to test the performance of our method we selected several datasets with different domains, topic typologies and document lengths. Given the relevance among the community using WordNet-related methods, we selected the Robust-WSD dataset from CLEF (Agirre et al., 2009a), which is a typical ad-hoc dataset on news. As we think that our method is specially relevant for short queries and/or short documents, we also selected the Yahoo! Answers dataset, which contains questions and answers as phrased by real users on diverse topics (Surdeanu et al., 2008), and ResPubliQA, a paragraph retrieval task on European Union laws organized at CLEF (Peñas et al., 2009).

The results show that our methods provide improvements in all three datasets, when compared to the query likelihood baseline, and that they compare favorably to PRF in two datasets. The analysis suggests that our models and PRF are complementary, in that PRF improves results for easy queries and our models are stronger for difficult queries. We also show that our models are more robust in face of sub-optimal parameters. Finally, in preliminary work, we present a combined system which outperforms all individual techniques, showing promise to further improve results in the future.

The paper is structured as follows. We first introduce the random walk model and the relatedness-based models for query and document expansion. Section 3 presents the experimental setup. Section 4 shows our main results, and analyzes diverse factors. Section 5 reviews related work. Finally, the conclusions and future work are mentioned.

2 Relatedness-based Expansion Models

In this section we describe the relatedness-based method to expand queries and documents, followed by the expansion models we propose for information retrieval.

2.1 Obtaining Expansion Terms

The key insight of our model is to expand the query or the document with related words according to the background information in WordNet (Fellbaum, 1998), which provides generic information about general vocabulary terms. WordNet groups nouns, verbs, adjectives and adverbs into sets of synonyms (synsets), each expressing a distinct concept. Synsets are interlinked with conceptual-semantic and lexical relations, including hypernymy, meronymy, causality, etc.

In contrast with previous work using WordNet, we select those concepts that are most closely related to the text as a whole. As we will see in the following sections, this text could be a query or a document. For that, we use a technique based on random walks over the graph representation of WordNet concepts and relations (Hughes and Ramage, 2007).

We represent WordNet as a graph as follows: graph nodes represent WordNet concepts (synsets) and dictio-

[1] http://ixa2.si.ehu.es/clirwsd/

nary words; relations among synsets are represented by undirected edges; and dictionary words are linked to the synsets associated to them by directed edges. We used version 3.0, with all relations provided, including the gloss relations. This was the setting obtaining the best results in a word similarity dataset as reported by Agirre et al. (2009b).

Given a text and the graph-based representation of WordNet, we obtain a ranked list of WordNet concepts as follows: (1) We first pre-process the text to obtain the lemmas and parts of speech of the open category words. (2) We then assign a uniform probability distribution to the terms found in the text. The rest of nodes are initialized to zero. (3) We compute personalized PageRank (Haveliwala, 2002) over the graph, using the previous distribution as the reset distribution, and producing a probability distribution over WordNet concepts. The higher the probability for a concept, the more related it is to the given text.

Basically, personalized PageRank is computed by modifying the random jump distribution vector in the traditional PageRank equation. In our case, we concentrate all probability mass in the concepts corresponding to the words in the text.

Let G be a graph with N vertices $v_1, \ldots, v_N$ and d_i be the outdegree of node i; let M be a $N \times N$ transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if a link from i to j exists, and zero otherwise. Then, the calculation of the *PageRank vector* $\mathbf{Pr}$ over G is equivalent to resolving Equation (1).

$$\mathbf{Pr} = cM\mathbf{Pr} + (1-c)\mathbf{v} \tag{1}$$

In the equation, $\mathbf{v}$ is a $N \times 1$ vector and c is the so called *damping factor*, a scalar value between 0 and 1. The first term of the sum on the equation models the voting scheme described in the beginning of the section. The second term represents, loosely speaking, the probability of a surfer randomly jumping to any node, e.g. without following any paths on the graph. The damping factor, usually set in the $[0.85..0.95]$ range, models the way in which these two terms are combined at each step.

The second term on Eq. (1) can also be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible, and thus guarantees that PageRank calculation converges to a unique stationary distribution.

In the traditional PageRank formulation the vector $\mathbf{v}$ is a stochastic normalized vector whose element values are all $\frac{1}{N}$, thus assigning equal probabilities to all nodes in the graph in case of random jumps. In the case of personalized PageRank as used here, $\mathbf{v}$ is initialized with uniform probabilities for the terms in the document, and 0 for the rest of terms.

PageRank is actually calculated by applying an iterative algorithm which computes Eq. (1) successively until a fixed number of iterations are executed. In our case, we used a publicly available implementation[2], with the default values provided by the software, i.e. a damping value of 0.85, and 30 iterations.

In order to select the expansion terms, we chose the top N highest scoring concepts, and get all the words that lexicalize the given concept. When expanding the documents (see Section 2.2) we follow the work in (Agirre et al., 2010), and fix N to 100. When expanding the queries (cf. Section 2.3) we explored several values of N, and tune it in order to get the optimum value, as discussed in Section 3.

For instance, given a query like "*What is the lowest speed in miles per hour which can be shown on a speedometer?*", our method suggests related terms like the following: *vehicle*, *distance* and *mph*.

2.2 Relatedness-based Document Expansion (RDE)

The relatedness-based document expansion approach requires that the document collection has been preprocessed to obtain a list of most related terms for each document, following the method explained in Section 2.1. These related terms are indexed separately. Documents are ranked by their probability of generating the query (Ponte and Croft, 1998), where this probability is estimated as a weighted combination of query likelihoods from the different document representations:

$$P_{RDE}(Q \mid \Theta_{RDE}) = P(Q \mid \Theta_D)^w P(Q \mid \Theta_E)^{1-w} \tag{2}$$

where Θ_D and Θ_E are the language models estimated from the original document representation and the expanded document representation, respectively, and w is the weight given to the original document language model set in the [0..1] range. Query likelihood is estimated following the multinomial distribution (we show the document model, but the expansion model is analogous):

$$P(Q \mid \Theta_D) = \prod_{i=1}^{|Q|} P(q_i \mid \Theta_D)^{\frac{1}{|Q|}} \tag{3}$$

where q_i is a query term of query Q and $|Q|$ is the length of Q. And following the Dirichlet smoothing (Zhai and Lafferty, 2001) we have

$$P(q_i \mid \Theta_D) = \frac{tf_{q_i D} + \mu \frac{tf_{q_i C}}{|C|}}{|D| + \mu} \tag{4}$$

[2] http://ixa2.si.ehu.es/ukb/

where tf_{q_iD} and tf_{q_iC} are the frequency of the query term q_i in the document D and the entire collection, respectively, and μ is the smoothing free parameter.

2.3 Relatedness-based Query Expansion (RQE)

In this approach, we expand each query with the terms obtained following the expansion technique described in Section 2.1. Thus, we retrieve documents based on the expanded query, which contains the original terms of the query and the expansion terms. Documents are ranked by their probability of generating the whole expanded query ($Q_{\scriptscriptstyle RQE}$), which is given by:

$$P_{\scriptscriptstyle RQE}(Q_{\scriptscriptstyle RQE} \mid \Theta_D) = P(Q \mid \Theta_D)^w P(Q' \mid \Theta_D)^{1-w} \quad (5)$$

where w is the weight given to the original query and Q' is the expansion of query Q. The query likelihood probability $P(Q \mid \Theta_D)$ is again calculated following a multinomial distribution and Dirichlet smoothing, as specified in Equation 3 and Equation 4. The probability of generating the expansion terms is defined as

$$P(Q' \mid \Theta_D) = \prod_{q'_i}^{|Q'|} P(q'_i \mid \Theta_D)^{\frac{w_i}{W}} \quad (6)$$

where q'_i is a expansion term, $W = \sum_{i=1}^{|Q'|} w_i$ and w_i is the weight we give to a expansion term, which we can see as the relatedness between the original query Q and the expansion term, and is computed as

$$w_i = P(q' \mid Q) = \sum_{j=1}^{N} P(q' \mid c_j) P(c_j \mid Q) \quad (7)$$

where c is a concept returned by the expansion algorithm (see Section 2.1), N is the number of concepts we chose for the expansion, $P(q' \mid c_j)$ is estimated using the sense probabilities estimated from Semcor (i.e. how often the query term q' occurs with sense c_j), and $P(c_j \mid Q)$ is the similarity weight that the mentioned expansion algorithm assigned to c_j concept.

3 Experiments

In order to test the performance of our method we selected several datasets with different domains, topic typologies and document lengths.

The first is the English dataset of the **Robust-WSD** task at CLEF 2009 (Agirre et al., 2009a), a typical adhoc dataset on news. This dataset has been widely used among the community interested on WSD and WordNet-related methods for the following reasons. Note that we need to reuse existing relevance judgments (customary on standard datasets), which were pooled among participants of the task, and thus systems that are based on different expansion strategies (e.g. WSD or WordNet) might return relevant documents which were not available in the pool that was manually judged at competition time. For this reason, the organizers of the Robust-WSD dataset used relevance judgments obtained pooling both monolingual and multilingual runs. The organizers of the exercise hoped that the inclusion of multilingual runs, with a larger variability due to translation strategies, would include relevance judgments for query-document pairs where different wording had been used (Agirre et al., 2009a).

	docs	length	q. train	q. test	length
Robust	166,754	532	150	160	8.6
Yahoo!	89,610	104	1,000	30,000	11.7
ResPubliQA	1,379,011	20	100	500	12.2

Table 1: Number of documents, average document length, number of queries for train and test in each collection, and average query length.

	QL	PRF				RDE		RQE		
	μ	μ	d	t	w	μ	w	μ	N	w
Rob	1000	1000	10	50	0.3	1200	0.8	2000	100	0.5
Yah	200	200	2	20	0.8	200	0.8	200	50	0.7
Res	100	100	10	30	0.8	100	0.7	100	125	0.7

Table 2: Optimal values in each dataset for free parameters.

The documents in the Robust-WSD comprise news collections from LA Times 94 and Glasgow Herald 95. The topics are statements representing information needs, consisting of three parts: a brief title statement; a one-sentence description; a more complex narrative describing the relevance assessment criteria. Following the rules of the Robust-WSD task, we use the title and the description parts of the topics in our experiments.

As we think that our method is specially relevant for short queries and/or short documents, we also evaluated our methods on the Yahoo! Answers dataset, which contains questions and answers as phrased by real users on diverse topics (Surdeanu et al., 2008), and ResPubliQA, a paragraph retrieval task on European Union laws organized at CLEF (Peñas et al., 2009).

The **Yahoo! Answers** corpus is a subset of a dump of the Yahoo! Answers web site, where people post questions and answers, all of which are public to any web user willing to browse them[3] (Surdeanu et al., 2008). Before releasing the dataset, the Yahoo team filtered the dataset as follows: (1) It comprised a subset of the questions, selected for their linguistic properties (for example they all start with "how {to — do — did — does — can — would

[3] Yahoo! Webscope dataset "ydata-yanswers-manner-questions-v1_0" http://webscope.sandbox.yahoo.com/

— could — should}”). (2) Questions and answers of obvious low quality were removed. (3) The document set was created with the best answer of each question (only one for each question). We use the dataset as released by its authors.

The other collection is the English dataset of **ResPubliQA** exercise at the Multilingual Question Answering Track at CLEF 2009 (Peñas et al., 2009). The exercise is aimed at retrieving paragraphs that contain answers to a set of 500 natural language questions. The document collection is a subset of the JRC-Acquis Multilingual Parallel Corpus, and consists of 21,426 documents for English which are aligned to a similar number of documents in other languages[4]. For evaluation, we used the gold standard released by the organizers, which contains a single correct passage for each query.

Table 1 shows some statistics for the three datasets.

Our experiments were performed using the Indri search engine (Strohman et al., 2005), which is a part of the open-source Lemur toolkit[5].

To determine whether the two expansion models we developed are useful to improve retrieval performance, we set up a number of experiments in which we compared our expansion models with other retrieval approaches. We used two baseline retrieval approaches for comparison purposes. One of the baselines is the default query likelihood (**QL**) language modeling method implemented in the Indri search engine. The other one is pseudo-relevance feedback (**PRF**) using a modified version of Lavrenko’s relevance model (Lavrenko and Croft, 2001), where the final query is a weighted combination of the original and expanded queries, analogous to Eq. 5. As in our own model presented in the previous sections, we chose the Dirichlet smoothing method for the baselines. We consider **QL** and **PRF** to be strong, reasonable baselines.

All the methods have several free parameters. The PRF model has three parameters: number of documents (d) and terms (t), and w (cf. Eq. 5). The RDE model also has w (cf. Eq. 2). The RQE model has two parameters: w (cf. Eq.. 5) and N the number of concepts for the expansion (Eq. 7). In addition, all methods use Dirichlet smoothing, which has a smoothing parameter μ. We used the train part of each dataset to tune all these parameters via a simple grid-search. The μ parameter was tested on the [100,1200] range for ResPubliQA and Yahoo! and [100,2000] for Robust, with increments of 100. The w parameter ranged over [0,1] with 0.1 increments. The d parameter ranged over [2,50] and the t and N in the range [1,200] (we tested 10 diff. values in the respective ranges). The parameter settings that maximized mean average precision for each model and each collection are shown in Table 2.

	PRF			RDE			RQE		
	Rob	Yah	Res	Rob	Yah	Res	Rob	Yah	Res
Rob		90.3	81.2		100.0	101.0		95.7	92.4
Yah	93.7		101.7	100.0		101.0	100.5		99.6
Res	92.7	99.1		99.3	99.3		100.9	101.3	
ave	93.1%			100.1%			98.4%		

Table 4: Effectiveness ratios for *inter-collections generalization* (based on MAP or MRR). The first column specifies the training dataset, the rows the test dataset. Empty slots correspond to the reference (100.0%). The average row shows the macro-average of all differences above it.

4 Results

In this section we present the results for the baseline query likelihood model (QL), the pseudo relevance feedback model (PRF) and our relatedness-based expansion models: query expansion (RQE) and document expansion (RDE).

Our main results are shown in Table 3. The main evaluation measure for Robust is Mean Average Precision (MAP), as customary. In two of the datasets (Yahoo! and ResPubliQA), there is a single correct answer per topic, and therefore we use Mean Reciprocal Rank (MRR). Note that in this setting MAP is identical to MRR. We also report Mean Precision at ranks 5 and 10 (P@5 and P@10). GMAP is also included, we will introduce and mention it in Section 4.1. Statistical significance was computed using Paired Randomization Test (Smucker et al., 2007). In the tables throughout the paper, we use * to indicate statistical significance at 90% confidence level, ** for 95% and *** for 99%.

QL and PRF The first two columns in Table 3 shows the results for QL and PRF and the performance difference between them. The highest results are obtained for the ResPubliQA dataset, followed by Robust and Yahoo!. The results for PRF are mixed. It is very effective in the Robust dataset, with dramatic improvements, specially in MAP. All differences are statistical significant, except for P@5. In Yahoo! the improvement is small in MRR and P@10, without statistical significance, but P@5 is lower. In ResPubliQA the results are bad, with statistical significant degradation in MRR.

RDE and RQE Continuing leftwards with Table 3, the following columns show the results for RDE and RQE, together with their difference with respect to QL. RDE improves QL in nearly all datasets and measures, except ResPubliQA with P@5. The improvement is strongest

[4] Note that Table 1 shows the number of paragraphs, which conform the units we indexed.

[5] http://www.lemurproject.org

		QL	PRF	Δ QL		RDE	Δ QL		RQE	Δ QL		RDE	Δ PRF		RQE	Δ PRF	
Robust	MAP	33.22	**36.69**	10.44%	***	**33.87**	1.95%	**	**33.67**	1.36%		33.87	-7.69%	***	33.67	-8.22%	***
	GMAP	13.21	**14.38**	8.90%	***	**13.51**	2.26%		**14.34**	8.59%	**	13.51	-6.10%	**	14.34	-0.29%	
	P@5	42.50	**43.63**	2.65%		**43.00**	1.18%		42.25	-0.59%		43.00	-1.43%		42.25	-3.15%	
	P@10	35.31	**37.38**	5.84%	***	**35.56**	0.71%		**35.81**	1.42%		35.56	-4.85%	***	35.81	-4.18%	*
Yahoo!	MRR	26.36	**26.40**	0.15%		**27.52**	4.42%	***	**27.22**	3.26%	***	**27.52**	4.26%	***	**27.22**	3.11%	***
	P@5	6.67	6.63	-0.56%	**	**6.91**	3.64%	***	**6.88**	3.21	***	**6.91**	4.22%	***	**6.88**	3.79%	***
	P@10	3.95	**3.96**	0.25%		**4.12**	4.29%	***	**4.10**	3.91%	***	**4.12**	4.03%	***	**4.10**	3.65%	***
ResPubliQA	MRR	48.77	46.33	-5.00%	***	**49.26**	1.02%		**49.78**	2.07%		**49.26**	6.33%	***	**49.78**	7.44%	***
	P@5	12.44	12.00	-3.54%	*	12.36	-0.64%		**12.68**	1.93%		**12.36**	3.00%		**12.68**	5.67%	***
	P@10	6.80	6.78	-0.29%		**6.94**	2.06%		6.78	-0.29%		**6.94**	2.36%		6.78	0.00%	

Table 3: Results of all methods. Δ columns show relative improvement with respect to QL or PRF. Bold means better than QL (middle columns) and PRF (rightmost columns).

in Yahoo!. In Robust the increase in MAP is small but significant.

4.1 Comparison with Respect to PRF

The rightmost columns in Table 3 repeat the results of RDE and RQE, together with the comparison with respect to PRF. Note that figures in bold mean better performance than PRF. We can see that the best results vary across datasets, with PRF yielding the best results for Robust, RDE for Yahoo! and RQE for ResPubliQA. Both RDE and RQE improve over PRF in Yahoo! and ResPubliQA, with mostly statistically significant differences.

PRF is known to perform well for some topics and datasets but not for others. Table 3 includes results for the geometrical mean, GMAP (Robertson, 2006), in the Robust dataset, as it is not relevant in the other datasets. GMAP tries to promote systems which are able to perform well for all topics, in contrast to systems that perform better in some but worse in others. The figures show that RDE and RQE approximate the performance of PRF, showing that they perform better for difficult topics. We will analyze this in more detail below.

4.2 Hard vs. Easy Questions

In order to study the behavior of our expansion models with respect to easy and hard topics, we performed pairwise comparisons between methods, and plot the performance of each topic according to the MAP (or MRR) obtained by two methods (Fig. 2). For instance, Fig. 2a plots Robust topics according to the performance of PRF (vertical axis) and QL (horizontal axis). The best fitting lineal trend line shows that PRF improves over QL irrespective of the performance of QL for the topic. In the other two collections (Figs. 2f and 2k), it seems that PRF drops performance for easier questions (e.g. those with high MAP for QL).

In the case of RDE (Figs. 2b, 2g, 2l), we find that there is some performance gain for difficult topics, at the cost of performance losses for easier topics. A similar behavior is observed for RQE (Figs. 2c, 2h, 2m).

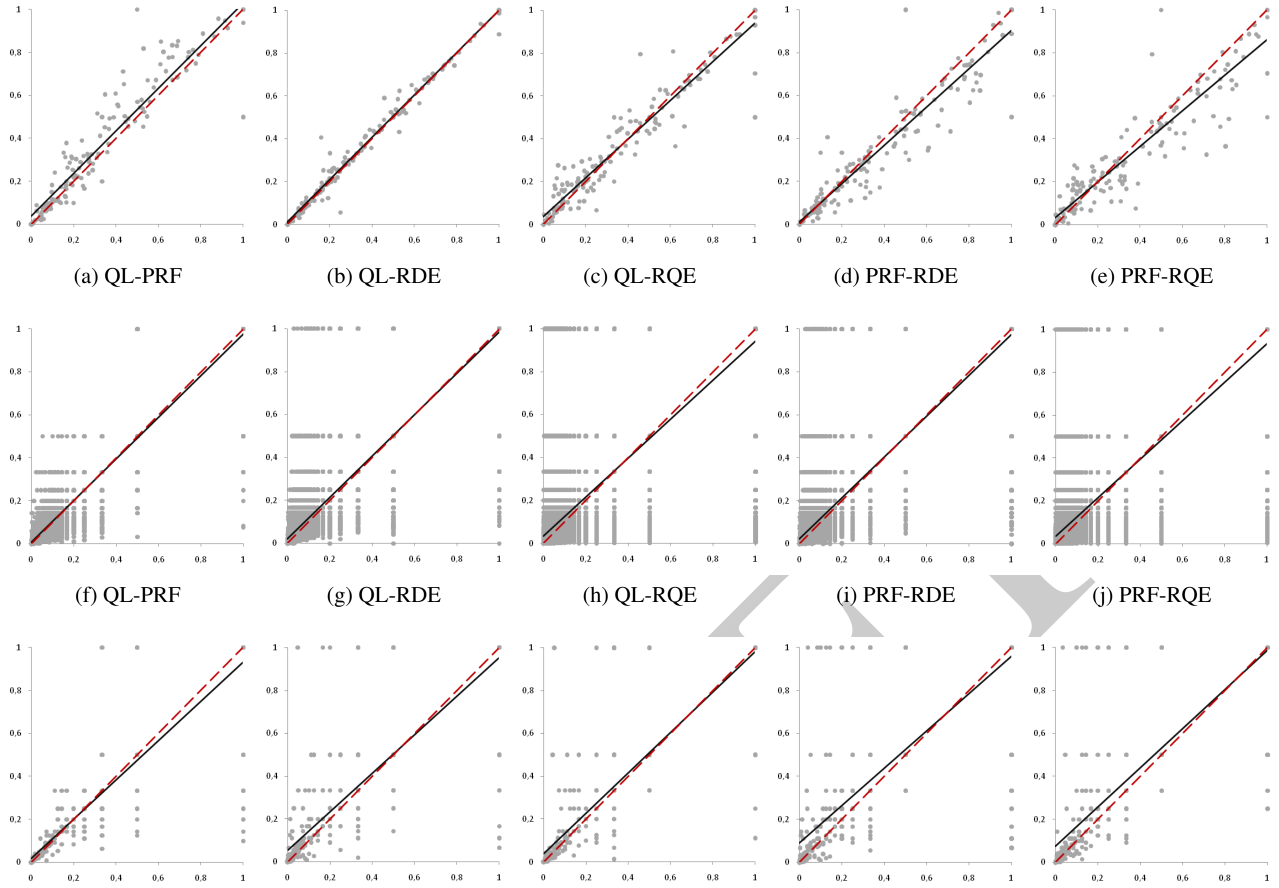
In fact, the performance gains for PRF seem complementary to the gains for RDE and RQE. The rightmost columns plot RDE vs. PRF (Figs. 2d, 2i, 2n) and RQE vs. PRF (Figs. 2e, 2j, 2o), where in both cases the trend line shows benefits for using PRF for high MAP topics and RDE and RQE for difficult topics (i.e. those with low MAP).

4.3 Parameter Optimization

In Table 2 we showed the optimum parameters for each technique and dataset, according to cross-validation results. In most practical situations, there is no training data to adjust the parameters, and optimal values from other scenarios are used. This analysis was named *inter-collections generalization* in (Metzler, 2006). Metzler proposed to measure generalization properties of a model by computing effectiveness ratio, which is the ratio of the observed effectiveness of a (trained) model to the optimal effectiveness. Thus, an effectiveness ratio of 100% represents a model that generalizes optimally. We take a simpler approach, and apply the idea directly to the MAP/MRR values, obtaining a MAP/MRR ratio for each combination of training/testing datasets, and macro-averaging across all possible combinations. The ratios in Table 4 show that RDE is the least sensible to optimization (it actually improves the results), with RQE losing some performance and PRF with the biggest loss.

Note that in order to keep the analysis simpler, we kept μ at the optimal values. The smoothing parameter μ has a direct relation with document length, and can be thus adjusted according to past experiences easily.

One important parameter when expanding queries is the number of terms to be expanded. Figure 3 shows the behaviour of PRF and RQE with respect of the number of query terms, when keeping the other parameters fixed. Fig. 3a shows that PRF can suffer some up and downs until it reaches 50 terms, where it plateaus. The up and downs vary across the three datasets. They are relatively small, except for Robust, where there is a steady increase up to 50 terms, and then a small degradation. Note that the best results are for 50 terms, but a different number of terms (30, 125) would cause a reduction of around 2



(a) QL-PRF (b) QL-RDE (c) QL-RQE (d) PRF-RDE (e) PRF-RQE

(f) QL-PRF (g) QL-RDE (h) QL-RQE (i) PRF-RDE (j) PRF-RQE

(k) QL-PRF (l) QL-RDE (m) QL-RQE (n) PRF-RDE (o) PRF-RQE

Figure 2: MAP and MRR plots on Robust, Yahoo! and ResPubliQA (rows) of PRF, RDE and RQE compared to QL (three leftmost columns, QL in x axis) and RDE and RQE compared to PRF (two rightmost columns, PRF in x axis). Best fitting lineal trend lines (solid lines) are also shown.

absolute points.

The case for RQE (cf. Fig. 3b) is more regular across datasets. RQE grows steadily according to the number of terms until it plateaus in all three datasets, using around 50 expansion concepts.

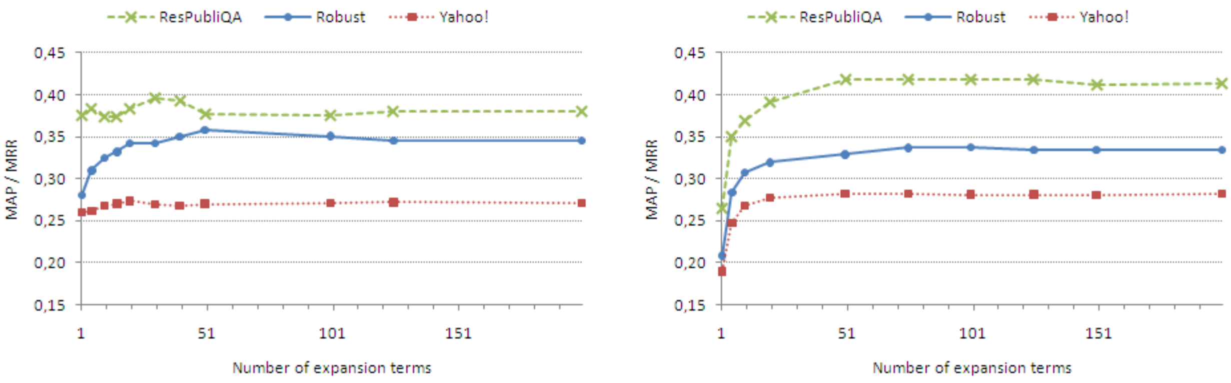
4.4 Variations in Each Dataset

The presence of non-relevant documents at top-ranks affects PRF negatively, in a phenomenon known as topic drift (Mitra et al., 1998). The performance differences experimented by PRF could be explained partially by topic drift. While the Robust dataset tends to contain many documents (news) related to each topic (thus being a good target collection for PRF), Yahoo! Answers contains completely unrelated documents describing answers to questions of all sorts, thus being amenable to topic drift. In fact, the optimization in the training dataset for Yahoo! chooses to expand terms on the two top documents only, compared to 10 documents in Robust (cf. Table 2). It has also been reported that PRF shows greater advantages for shorter queries (Xu and Croft, 2000), and the Robust dataset contains shorter queries compared to ResPubliQA and Yahoo!.

Regarding the behaviour of RQE and RDE, it is still too early to know which factors affect their performance. From the previous sections we have learned that they tend to perform worse in easy queries, but other than that they seem to be robust and stable in a number of settings.

4.5 Preliminary Experiments on Combinations

The analysis in the previous sections shows that the proposed methods are complementary to PRF. In order to test whether the combination of methods would be productive, we performed a preliminary experiment combining PRF and RQE in the Robust collection. We added



(a) PRF

(b) RQE

Figure 3: Results on three datasets for different number of expansion terms

the question expansion terms produced by RQE to the expansion terms produced by PRF, yielding a **MAP of 37.67** and a **GMAP of 15.43**, outperforming all individual methods. Given the fact that we did not test sophisticated combinations nor perform parameter optimization (we just applied the values in Table 2), we would expect results to improve further in the future.

4.6 Computational Cost

Improved performance comes at a computational cost. If the processing of the Robust test set (160) questions takes 22 seconds for the QL baseline on a server with two Intel QuadCore Xeon X5460 processors at 3160MHz with 32 GB of memory; PRF takes 7 minutes and 20 seconds; RDE one minute; and RQE 22 minutes and 45 seconds. The larger cost for PRF and RQE at query time comes from the added complexity of examining additional terms in the expanded query. Given that RQE is using more terms than PRF, the cost is higher.

In addition, running the random walk on one query or document takes 6 seconds. In the case of RDE, the process can be easily parallelised. In the case of RQE, query time computations could be speed up using less iterations, or if we had precomputed the random walks for each word in advance. In the later case, at query time one would just need to do a linear combination of the probability vectors of the words in the query. For the future, we would like to check whether there is any performance loss involved.

5 Related Work

Query expansion (QE) methods analyze user query terms and incorporate related terms automatically (Voorhees, 1994), and are usually divided into local and global methods. Local methods adjust a query relative to the documents that initially appear to match the query (Manning et al., 2009). Pseudo-relevance feedback (PRF) is one of most widely used expansion methods (Rocchio, 1971; Xu and Croft, 1996). This method assumes top-ranked documents to be relevant (and sometimes, also that low-ranked documents are irrelevant), and selects additional query terms from the top-ranked documents. Since Rocchio presented an algorithm for relevance feedback (Rocchio, 1971), lots of variations have been developed. The TREC 2008 Relevance Feedback Track results confirmed that relevance feedback consistently improves different kinds of retrieval models, but the amount of relevance information needed to improve results and the use or not of non-relevant information varied among systems (Buckley and Sanderson, 2008).

Global methods are techniques for expanding query terms without checking the results returned by the query. These methods analyze term co-occurrence statistics in the entire corpus or use external knowledge sources to select terms for expansion (Manning et al., 2009). For example, synonyms from WordNet after performing word sense disambiguation (WSD) have been used for query expansion with some success (Voorhees, 1994; Liu et al., 2005).

The query expansion method proposed in this paper is a global expansion technique based on WordNet, but in contrast to the previous work based on WordNet it does not perform WSD and adds related words beyond synonyms.

An alternative to QE is to perform the expansion in the document. Document Expansion (DE) was first proposed in the speech retrieval community (Singhal and Pereira, 1999), where the task is to retrieve speech transcriptions which are quite noisy. Singhal and Pereira proposed to enhance the representation of a noisy document by adding to the document vector a linearly weighted mixture of related documents. In order to determine related documents, the original document is used as a query into the collection, and the ten most relevant documents are selected. Two related papers (Liu and Croft, 2004; Kurland and Lee, 2004) followed a similar approach on the TREC ad-hoc document retrieval task. They use document clustering to determine similar documents, and document expansion is carried out with respect to these. Both papers report significant improvements over non-expanded baselines. Instead of clustering, more recent work (Tao et al., 2006; Mei et al., 2008; Huang et al., 2009) use language models and graph representations of the similarity between documents in the collection to smooth language models with some success.

The document expansion method presented here is complementary to those methods, in that we also explore DE, but use WordNet instead of distributional methods. The comparison with respect to other DE techniques and the exploration of potential combinations will be the focus of future research.

Another strand of WordNet-based IR work has explicitly represented and indexed word senses after performing WSD (Gonzalo et al., 1998; Stokoe et al., 2003; Kim et al., 2004). The word senses conform a different space for document representation, but contrary to us, these works incorporate concepts for all words in the documents, and are not able to incorporate concepts that are not explicitly mentioned in the document. More recently, a CLEF task was organized (Agirre et al., 2009a) where terms were semantically disambiguated to see the improvement that this would have on retrieval; the conclusions were mixed, with some participants slightly improving results with information from WordNet.

(Agirre et al., 2010) is the work which is closest to ours. They use the same WordNet-based relatedness method in order to expand documents, following the BM25 probabilistic method for IR, obtaining some improvements, specially when parameters had not been optimized. In contrast to their work, we adopt an approach combining inference network (Turtle and Croft, 1991) and language modeling (Ponte and Croft, 1998). In addition to document expansion, we also test question expansion and perform a more elaborate analysis, including the comparison to PRF.

Our work stems from the use of random walks over the WordNet graph to compute the relatedness between pairs of words (Hughes and Ramage, 2007). In this work a single word was input to the random walk algorithm, obtaining the probability distribution over all WordNet synsets. The similarity of two words was computed as the similarity of the distributions of each word. In later work, (Agirre et al., 2009b) tested different configurations of the graph, and obtained the best results for a WordNet-based system, comparable to the results of a distributional similarity method which used a crawl of the entire web. The same authors later released their UKB software, which is the one we use here.

6 Conclusions

In this paper we explore a generic method to improve IR results using structured knowledge, both doing query expansion and document expansion. Our work has been motivated by the success of knowledge-based methods in word similarity and relatedness tasks (Agirre et al., 2009b). Note that distributional similarity is closely related to query expansion and clustering techniques for IR. In the first case, techniques such as pseudo-relevant feedback (PRF) expand the query with terms which are deemed to be related to the query according to the retrieved documents (Xu and Croft, 1996). In the second case, documents are clustered, and terms from related documents are used to re-estimate counts and to expand the documents with new terms (Singhal and Pereira, 1999).

Our expansion method is based on random walks over a graph-representation of a knowledge base. The random walks return sets of concepts which are related to the input query (or document), even if those concepts are not explicitly mentioned in the texts. The query (or document) is then expanded using the terms lexicalizing the related concepts. In this work we focused on WordNet, but any other knowledge structure could be used.

We adopted a language modeling framework to implement the query likelihood and pseudo-relevance feedback baselines, as well as our relatedness-based query expansion (RQE) and document expansion (RDE) methods, where the expansion terms for documents are indexed separately. We wanted to check the performance on a diverse range of ad-hoc datasets with different domains, topic typologies and document lengths: Robust-WSD dataset from CLEF (ad-hoc dataset on news which got the attention of the WSD community), Yahoo! Answers (questions and answers as phrased by real users on diverse topics) and ResPubliQA (a paragraph retrieval task on European Union laws).

Our two methods provide improvements in all three datasets, when compared to the query likelihood baseline. PRF is beneficial in two datasets, but degrades performance in ResPubliQA. RDE and RQE compare favorably to PRF in two datasets, but perform worse in Robust. Our analysis shows that our models and PRF are complementary, in that PRF is better for easy queries and our models are stronger for difficult queries. In fact, GMAP scores show that RQE is comparable to PRF in Robust, and a first tentative combination of PRF and RQE improved results further. In addition we show that our models are more robust in face of sub-optimal parameters.

In the future, we would like to evaluate separately the concepts obtained from the random walks, in order to study which are the words that have good expansions that contribute to improved performance. We also plan to exploit the ability of RQE and RDE to perform well on difficult queries, perhaps combining them with PRF techniques. We would also like to explore the relation with clustering and document clustering techniques. Given the very positive results obtained with WordNet, we would like to explore other knowledge bases and resources. It holds special promise in the case of domains for which rich lexical resources are available (e.g. medicine, with UMLS (Humphreys et al., 1998)), but also opens new avenues to integrate the growing number of structured knowledge being made available following linked data initiatives.

References

E. Agirre, G. M. Di Nunzio, T. Mandl, and A. Otegi. 2009a. CLEF 2009 Ad Hoc Track Overview: Robust - WSD Task. In *Working Notes of the Cross-Lingual Evaluation Forum*.

E. Agirre, A. Soroa, E. Alfonseca, K. Hall, J. Kravalova, and M. Pasca. 2009b. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proc. of NAACL*, Boulder, USA.

E. Agirre, X. Arregi, and A. Otegi. 2010. Document expansion based on WordNet for robust IR. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 9–17, Stroudsburg, PA, USA. Association for Computational Linguistics.

C. Buckley and M. Sanderson. 2008. Relevance feedback track overview: Trec 2008. In Ellen M. Voorhees and Lori P. Buckland, editors, *TREC*, volume Special Publication 500-277. National Institute of Standards and Technology (NIST).

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database and Some of its Applications*. MIT Press, Cambridge, Mass.

J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings ACL/COLING Workshop on Usage of WordNet for Natural Language Processing*.

T. H. Haveliwala. 2002. Topic-sensitive PageRank. In *Proceedings of WWW '02*, pages 517–526.

Y. Huang, L. Sun, and J. Nie. 2009. Smoothing document language model with local word graph. In *Proceedings of CIKM '09*, pages 1943–1946.

T. Hughes and D. Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of EMNLP-CoNLL-2007*, pages 581–589.

L. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 1(5):1–11.

S. B. Kim, H. C. Seo, and H. C. Rim. 2004. Information retrieval using word senses: root sense tagging approach. In *Proceedings of SIGIR '04*, pages 258–265.

O. Kurland and L. Lee. 2004. Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of SIGIR '04*, pages 194–201.

V. Lavrenko and W. B. Croft. 2001. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

X. Liu and W. B. Croft. 2004. Cluster-based retrieval using language models. In *Proceedings of SIGIR '04*, pages 186–193.

S. Liu, C. Yu, and W. Meng. 2005. Word sense disambiguation in queries. In *Proceedings of CIKM '05*, pages 525–532.

C. D. Manning, P. Raghavan, and H. Schütze. 2009. *An introduction to information retrieval*. Cambridge University Press, UK.

Q. Mei, D. Zhang, and C. Zhai. 2008. A general optimization framework for smoothing language models on graph structures. In *Proceedings of SIGIR '08*, pages 611–618.

Donald Metzler. 2006. Estimation, sensitivity, and generalization in parameterized retrieval models. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 812–813, New York, NY, USA. ACM.

M. Mitra, A. Singhal, and C. Buckley. 1998. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 206–214, New York, NY, USA. ACM.

A. Peñas, P. Forner, R. Sutcliffe, A. Rodrigo, C. Forăscu, I. Alegria, D. Giampiccolo, N. Moreau, and P. Osenova. 2009. Overview of ResPubliQA 2009: Question Answering Evaluation over European Legislation. In *Working Notes of the Cross-Lingual Evaluation Forum*.

J. M. Ponte and W. B. Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

S. Robertson. 2006. On GMAP: and other transformations. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 78–83, New York, NY, USA. ACM.

J. J. Rocchio. 1971. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall.

A. Singhal and F. Pereira. 1999. Document expansion for speech retrieval. In *Proceedings of SIGIR '99*, pages 34–41, New York, NY, USA. ACM.

M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of CIKM 2007*, Lisboa, Portugal.

C. Stokoe, M. P. Oakes, and J. Tait. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of SIGIR '03*, page 166.

T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. 2005. Indri: a language-model based search engine for complex queries. Technical report, in Proceedings of the International Conference on Intelligent Analysis.

M. Surdeanu, M. Ciaramita, and H. Zaragoza. 2008. Learning to Rank Answers on Large Online QA Collections. In *Proceedings of ACL 2008*.

T. Tao, X. Wang, Q. Mei, and C. Zhai. 2006. Language Model Information Retrieval with Document Expansion. In *Proceedings of HLT/NAACL*, pages 407–414, June.

H. Turtle and W. B. Croft. 1991. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst.*, 9:187–222, July.

E. M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of SIGIR '94*, page 69.

J. Xu and W. B. Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA. ACM.

Jinxi Xu and W. Bruce Croft. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18:79–112, January.

C. Zhai and J. Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 334–342, New York, NY, USA. ACM.