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

Coling 2010: Poster Volume, pages 9–17,
Beijing, August 2010

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.