Using semantic relatedness and word sense disambiguation for (CL)IR

Eneko Agirre[1], Arantxa Otegi[1], Hugo Zaragoza[2]
[1] IXA NLP Group, University of the 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 on 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 if 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 to further work on this technique, which has attained positive results in the passage retrieval for question answering task at CLEF (ResPubliQA).

.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

Keywords

Robust Retrieval, CLIR, Word Sense Disambiguation, Lexical Relatedness, Document Expansion

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 [3, 4, 5, 7]. WordNet-based approaches need to deal with ambiguity, which proves difficult given the little context available to disambiguate the word in the query effectively. In our experience document expansion works better than topic expansion (see our results of the last edition in [6]). Bearing this in mind, 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 [1] 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. Finally, Section 4 draws the conclusions and mentions future work.

2 Experiments

Our main experimentation strategy consisted on 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. 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[1], 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:

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

- 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 (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.

2.4 Indexing

We indexed the new expanded documents using the MG4J search-engine [2]. 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. Porter stemming was used as per usual.

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 choose 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 neither WSD or semantic relatedness information. We have to mention that we did achieve improvement in the training data, but the difference was not significant[2].

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

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. It is not very clear whether translating the topics from Spanish to English or translating the documents from English to Spain 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 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 on 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 why we have not achieved higher gains using the semantic expansion, as the same strategy obtained remarkable improvements in the passage retrieval task (ResPubliQA).

Acknowledgments

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] E. Agirre, A. Soroa, E. Alfonseca, K. Hall, J. Kravalova, and M. Pasca. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of annual meeting of the North American Chapter of the Association of Computational Linguistics (NAACL)*, Boulder, USA, June 2009.

[2] P. Boldi and S. Vigna. MG4J at TREC 2005. In Ellen M. Voorhees and Lori P. Buckland, editors, *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, number SP 500-266 in Special Publications. NIST, 2005. `http://mg4j.dsi.unimi.it/`.

[3] S. Kim, H. Seo, and H. Rim. Information retrieval using word senses: Root sense tagging approach. In *Proceedings of SIGIR*, 2004.

[4] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of SIGIR*, 2004.

[5] S. Liu, C. Yu, and W. Meng. Word sense disambiguation in queries. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, 2005.

[6] A. Otegi, E. Agirre, and G. Rigau. IXA at CLEF 2008 Robust-WSD task: using Word Sense Disambiguation for (Cross Lingual) Information Retrieval. In *Working Notes of the Cross-Lingual Evaluation Forum*, Aarhus, Denmark, 2008. ISBN 2-912335-43-4, ISSN 1818-8044.

[7] J.R. Pérez-Agüera and H. Zaragoza. UCM-Y!R at CLEF2008 Robust and WSD tasks. In *Working Notes of the Cross-Lingual Evaluation Forum*, Aarhus, Denmark, 2008.