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.

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

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

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)