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

Arantxa Otegi, Eneko Agirre, German Rigau
IXA NLP Group - University of the Basque Country
Donostia, Basque Country
aotegui004@ikasle.ehu.es

Abstract

This paper describes the participation of the IXA NLP group at the CLEF 2008 Robust-WSD Task. This is our first time at CLEF, and we participated at both 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, although the improvement was only significant for the bilingual subtask. As a secondary goal, our best systems ranked 4th overall and 3rd overall in the monolingual and bilingual subtasks, respectively.

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

General Terms

Measurement, Performance, Experimentation

Keywords

Robust Retrieval, CLIR, Word Sense Disambiguation, Query Expansion, Structured Queries

1 Introduction

The objective of the robust task in previous editions was to give preference to systems which achieve good stable performance over all queries. This year it also included an additional goal: to test whether word sense disambiguation (WSD) can be used beneficially by retrieval systems. All our experiments have been focused on the second goal and we experimented on how WSD data can be exploited in order to improve retrieval. In this sense, 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 the open source Indri search engine, which is based on the inference network framework and supports structured queries [1].

The remainder of this paper is organized as follows. Section 2 describes the experiments carried out, Section 3 presents the results obtained and, finally, Section 4 draws the conclusions and 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 combination 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 tested different query expansion and translation strategies, and finally we search for the queries in the indexes in various combinations. We will see each in turn.

2.1 Expansion and translation strategies

WSD data provided to the participants was based on WordNet version 1.6. Each word sense have 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 all synonyms of only the highest scored sense for each word using the two different expansion collections using UBC and 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 only the highest scored sense for each word using the two different translation collections using UBC and NUS disambiguation data.
- Translation of Spanish topics: translation from Spanish to English of the first sense for each word.

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 phrased 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. Only original words.
2. Both original words and all expansions for each word.
3. Both original words and only expansions for the best sense of each word.
4. Only translated words, using all translations for each word. If a word had not a translation, that original word was included in the query.
5. Only translated words, using translations for the best sense of each word. If a word had not a translation, that original word was included in the query.

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

In the first case, we constructed a simple query combining only the original words using the Indri operator `#combine`. In the other cases, we have used some of the operators available in the structured query language. For example, when we wanted to use original words as well as synonyms (obtained after expansion) in the same query, we constructed two subqueries (one with original words, and another one with the expanded words). Then we integrated both subqueries in the same query using the `#weight` operator and giving a weight of 0.6 to the original word's subquery and 0.4 to the other subquery. We used also the `#syn` operator to join the expanded words of each sense, as they are meant to be synonyms. In the case of full expansion, instead of `#syn`, we used `#wsyn` (weighted synonym). This operator allows to give different weights to synonyms. So, as in the previous case, we joined synonyms, and also, we weighted each synonym-set with the score that the disambiguation system had assigned to each sense. Finally, multiword expressions, such as `prime minister` are added to the query joined with the `#1` operator (ordered window).

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 created the test topic runs. The submitted runs are described in Section 3.

In some of the experiments we applied pseudo-relevance feedback (PRF) with these 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 1 summarizes the results of our submitted runs, 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 topics; both original and expanded terms in documents, using best sense according to NUS word sense disambiguation; PRF.

  En2EnUbcDocsPsrel original terms topics; both original and expanded terms in documents, using best sense according to UBC word sense disambiguation; PRF.

  En2EnFullStructTopNusDocsPsrel original and fully expanded terms in topics; both both original and expanded terms in documents, using best sense according to NUS word sense disambiguation; PRF.

		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

Table 1: Results for submitted runs

- 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 no significant difference is found with respect to `En2WnNowsdPsrel`[1]. Regarding the bilingual results, `Es2En1stTopsUbcDocsPsrel` is the best, and it is significantly better than `Es2EnNowsdPsrel`. These results confirm the results we got in the training data. Although not shown here, the results of using WSD are significantly better in the training data 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 systems ranked 4th overall and 3rd overall in the monolingual and bilingual subtasks, respectively.

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. That is why the Geometric Mean Average Precision (GMAP) values are lower.

[1] Paired Randomization Tests over MAPs with α=0.05 have been used along this work

4 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, and given 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 the WSD information for documents was a good strategy, in contrast to most of previous IR work, which has focused on WSD of topics.

For the future we plan to improve the bilingual results, mainly incorporating external resources. We tried straightforward methods to exploit WSD information for the expansion and indexing of the documents. We would like to pursue more sophisticated methods.

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] H. Turtle Strohman, D. Metzler and W.B. Croft. Indri: A language model-based search engine for complex queries. *Proceedings of the International Conference on Intelligence Analysis*, 2005.