Analysis of Word Sense Disambiguation-Based Information Retrieval

Jacques Guyot and Gilles Falquet and Saïd Radhouani and Karim Benzineb

Centre Universitaire dInformatique, University of Geneva
Route de Drize 7, 1227 Carouge, Switzerland

Abstract.
Several studies have tried to improve retrieval performances based on automatic Word Sense Disambiguation techniques. So far, most attempts have failed. We try, through this paper, to give a deep analysis of the reasons behind these failures. During our participation at the Robust WSD task at CLEF 2008, we performed experiments on monolingual (English) and bilingual (Spanish to English) collections. Our official results and a deep analysis are described below, along with our conclusions and perspectives.

1 Introduction

Our aim through this paper is not to propose sophisticated strategies to improve retrieval performances using a word sense disambiguation (WSD) algorithm. Rather we mainly want to explore whether WSD (plus the semantic information in WordNet) can be useful in Information Retrieval (IR) and Cross Lingual Information Retrieval (CLIR). Therefore, we carried out a set of experiences in monolingual and bilingual tasks. Then we deeply analyzed the obtained results and formulated some conclusions and perspectives. In the rest of this paper, we first present the steps of the collection processing (Section 2). Then we describe our indexing and searching strategies (Section 3). The obtained results of our experiments are detailed in section 4. Before concluding (Section 6), we discuss the obtained results and provide some perspectives in section 5.

2 Collection processing

The corpus is a news collection, containing 166000 English documents and 160 topics. All topics are available in English and Spanish. Each topic contains three fields: a title (T), a description (D), and a narrative (N). The corpus was disambiguated using two leading WSD systems: the University of the Basque Country (UBC) [1] and the National University of Singapore (NUS) [5], resulting in two different sets. English documents and queries were processed using the English WordNet, while the Spanish topics were annotated using the Spanish WordNet. The disambiguation process consists of annotating documents and queries by

adding sense information to all content words (figure 1). Thus, each occurrence of a word is replaced by an XML element containing the word identifier (TERM ID), an extracted lemma (LEMA), a part-of-speech (POS) tag (noun, verb, adjective, etc.), the original word form (WF), and a list of senses together with their respective scores. The senses are represented by WordNet *synset* codes.

```
<DOC>
    <DOCNO>GH950102-000000</DOCNO>
    <DOCID>GH950102-000000</DOCID>

    <HEADLINE>
        <TERM ID="GH950102-000000-1" LEMA="alien" POS="JJ">
            <WF>Alien</WF>
            <SYNSET SCORE="0.6" CODE="01295935-a"/>
            <SYNSET SCORE="0.4" CODE="00984080-a"/>
        </TERM>

        <TERM ID="GH950102-000000-2" LEMA="treatment" POS="NN">
            <WF>treatment</WF>
            <SYNSET SCORE="0.827904118008605" CODE="00735486-n"/>
            <SYNSET SCORE="0" CODE="03857483-n"/>
            <SYNSET SCORE="0.172095881991395" CODE="00430183-n"/>
            <SYNSET SCORE="0" CODE="05340429-n"/>
        </TERM>
```

Fig. 1. Example of WordNet-based document annotation

3 Indexing and searching strategies

For several reasons, we chose to index the corpus using our IDX-VLI indexer [6]. Indeed, IDX-VLI can gather a wealth of information (positions, etc.), it has built-in operators, and it is remarkably fast. Still, we only used the basic version of that indexer *i.e.*. We did not use any relevance feedback mechanism, context description, or any other sophisticated tool of that sort. We thus avoided interfering with the direct results of the experiment, and we facilitated the result analysis. Documents and queries content were represented using the Okapi BM25 weighting scheme (with default parameters).

3.1 Documents processing

We developed and tested the following document processing strategies that we applied to each <TERM> element within each document annotation:

- NAT: Keep only the word form of each element (*i.e.*, rebuild the original text);
- LEM: Keep only the lemma;
- POS: Keep the lemma and the part-of-speech tag;

– WSD: Keep only the synset that has the best score[1];
– WSDL: Keep the lemma and the best corresponding synset (with the higher score).

During the indexing process, these strategies were applied to all terms including numbers, except for stop-words. Given the poor performance of the POS strategy, we quickly gave up this option.

3.2 Topics processing

The same strategies were applied to the topics, with an extended stop-word list including words such as *report*, *find*, etc. For each topic, we derived three queries:

– T: Includes only the title field;
– TD: Includes the title and the description fields:
– TDN: Includes the title, the description, and the narrative fields.

4 Experimental results

4.1 Lemma-based strategy

In order to come up with a reasonably good baseline, we tested several approaches to build a Boolean pre-filter from a given topic. We didn't want to have a low baseline: when the baseline is low, the probability achieve a better result using WSD becomes high. This happened to Basile et *al.* and Otegi et *al.* when they used WSD for the bilingual robust WSD task at CLEF 2008 [2][8], and for Stokoe et *al.* when they applied their WSD system on a large-scale TREC data collection [20].

Table 1. Baseline results in terms of MAP

Run name	**Run description**	**MAP**
OR	The logical OR of the words (or lemmas)	**0.255**
AND	The logical AND of the words	0.158
NEAR	The logical OR of all pairs (t_i NEAR t_j), where t_i and t_j are two query terms	0.152

The obtained results of the baseline are described in table 1 where columns contain respectively the run's name, the run's description, and the corresponding result in terms of mean average precision (MAP). These tests were conducted on 150 training topics. The best results were given using the OR filtering.

[1] This amounts to considering that the disambiguation algorithm is "perfect". Alternatively, we could have added all synsets that have a score greater than a given threshold.

4.2 WSD-based strategy

In addition to the filtering strategies used for the baseline runs, we performed two more runs based on the hyperonym relationship extracted from WordNet. The results obtained on the training corpus are described in table 2.

Table 2. WSD-based runs results in terms of MAP

Run name	**Run description**	**MAP**
OR	The logical OR of the best synset corresponding to a topic word	**0.224**
AND	The logical AND of the best synset corresponding to a topic word	0.151
NEAR	The logical OR of all pairs (s_i NEAR s_j), where s_i and s_j are the best synsets corresponding to two topic words t_i and t_j	0.125
HYPER	The logical AND of each (s_i OR h_i), where s_i is the best synset corresponding to a topic word t_i, and h_i is the direct hypernym of s_i in WordNet	0.143
ORHYPER	The logical OR of each (s_i OR h_i), where s_i is the best synset corresponding to a topic term t_i, and h_i is the direct hypernym of s_i in WordNet	0.1843

The obtained results on the training corpus showed that the strategy based on the OR-filtering gives the best result. Therefore, we decided to use it for the official runs described in the following section.

4.3 Official results

We carried out several runs in the monolingual and the bilingual task. For the purposes of this paper, however, we present only the most significant ones.

Table 3 contains the official results in terms of MAP for the monolingual task. The first column contains the topic fields used during the corresponding run; the second column contains the results of the lemma-based strategy; and the third column contains the results of the WSD-based strategy (using the NUS disambiguation algorithm). The results clearly demonstrate that the use of WSD techniques does not improve the retrieval performances compared to a lemma-based approach. The best result was obtained using all the topic fields with lemma as indexing unit (0.3917).

The results also demonstrated that the retrieval performances obtained using the NUS disambiguation algorithm are higher than those obtained using the UBC disambiguation algorithm.

The best result obtained using WSD occurred when we combined a WSD-based indexing with a lemma-based indexing (0.3814). However, it is lower than the result obtained using lemma only.

Table 3. Official results in terms of MAP for the monolingual task

Used topic field	Lemma-based strategy	WSD-based strategy (NUS corpus)
T	0.3064	0.2120
TD	0.3664	0.2934
TDN	**0.3917**	0.3269

For the bilingual task, the baseline consisted of translating topics from English to Spanish using Google translator. The obtained results using only the title of the topic gave a MAP of 0.3036. The use of WSD significantly decreases the retrieval performances (0.0846 of MAP using the NUS algorithm).

5 Findings and Discussion

From our experiences, both on the training and the test corpora, we note the following facts:

- Using D and N topics fields increases the MAP in all cases (with and without WSD). This is most probably due to the ranking method that benefits from the additional terms provided by D and N topics fields.
- On the test run with the UBC system, using only synsets (WSD) decreases the MAP: -4.6% using the T field, and -3.1% using TDN. On the training topics, combining lemmas and synsets (WSD + Lemma) slightly improves the MAP (+0.6%). This is the only case where disambiguation brings an improvement.
- Using different disambiguation algorithms for queries and documents noticeably decreases the results. This should not happen if the algorithms were perfect. It demonstrates that disambiguation acts as a kind of "encoding" process on words, and obviously the best results are obtained when the same "encoding", producing the same mistakes, is applied to both queries and documents. Thus, at this stage, the disambiguation algorithms are not interoperable.

We carefully analyzed around 50 queries to better understand what happened with the disambiguation process. For instance, the query whose title is "*El Nio and the weather*" was disambiguated, using the NUS algorithm, as follows:

- "El" was interpreted as the abbreviation "el." of "elevation";
- "Niño" was interpreted as the abbreviation "Ni" of "nickel", probably because the parser failed on the non-ASCII character "ñ";
- "Weather" was correctly interpreted as the "weather" concept.

Although the disambiguation was incorrect, WSD was as good as LEM because the "encoding" was the same in the documents and in the queries. In

addition, WSD was also as good as LEM because there were a few or no documents dealing about nickel that could have produced noise. More generally, when the WSD results were better than the LEM ones, it was not due to semantic processing but to contingencies. For instance, the query title "*Teenage Suicides*" had a better score with WSD because "teenage" was not recognized. Thus, the query became *suicides*, which is narrower than *teenage OR suicide*, and avoided retrieving a large amount of irrelevant documents about teenagers.

The poor performance on Spanish queries is due to: *i*) the above-mentioned lack of interoperability between the different WSD algorithms and *ii*) the low quality of the Spanish WSD itself. This can be illustrated in the following examples:

Topic 41: "Pesticide in baby food" is translated into "Pesticidas en alimentos para bebes", and then converted into the FOOD and DRINK (verb) concepts, because "bebes" is a conjugated form of "beber", which is the Spanish verb for drink.

Topic 43: "El Niño and the weather" is translated into "El Niño y el tiempo", and then converted into the CHILD and TIME concepts, because "Niño" is the Spanish noun for child, and "tiempo" is an ambiguous word meaning both time and weather.

Given those difficulties, outstanding results could not be expected. Looking back on the results, it can be noted that 1793 documents were retrieved out of the 2052 relevant ones (*i.e.*, almost 90% of them). The core issue is to sort out documents so as to reject those whose content does not match users' expectations. A closer look at our results on the training corpus showed that we achieved a solid performance on some topics. This does not mean that our search engine "interpreted" correctly said topics. Rather, it is simply due to the fact that the corpus included only good matches for those topics. Therefore, it was almost impossible to find wrong answers. For instance, on topic 50, which deals with "the Revolt in Chiapas", we retrieved 106 documents out of 107 possible relevant ones, with a MAP of 87%. This is due to the fact that in the corpus, the Chiapas are only known for their revolt (in fact if we Google the word "Chiapas", a good proportion of the results are currently about the Chiapas rebellion). On the other hand, on topic 59, which deals with "Computer Viruses", our search engine retrieved 1 out of 1 possible relevant document, with a MAP of 0.03. This low result is because the 300 documents retrieved before the one we were looking for were indeed about viruses and computers, but did not mention any virus *name* or *damage* as was requested. Therefore term disambiguation does not help search engines to interpret what kinds of documents are expected. A topic, such as the one above, requires the text to be read and correctly interpreted in order to decide whether it is actually a correct match. After a deep analysis, we concluded that the retrieval performance of WSD-based system depends at least on three factors:

1. The quality of the used semantic resource, and in particular its coverage compared with the vocabulary of corpus. This problem can be avoided if we combine WSD-based indexing with keywords-based indexing. So far, the few

works that have been successful are those who proceeded using this method [13][18].

2. The quality (accuracy) of the used disambiguation algorithm: As mentioned by several studies [10][19], the main difficulty to improve retrieval performances is due to the inefficacity of disambiguation algorithm, especially when queries are short (one or two words)[21]. Indeed, it is judicious to think that by using a perfect algorithm (with 100% accuracy), retrieval performances will be at least equal to those obtained by keywords-based approach. We postulate that when a query is large enough (more than two words), the probability that a document containing the query terms in a different context or meaning from the intended definitions one is very low. For instance, it is unlikely that a document containing *mouse*, *cheese*, and *cat* is in fact dealing about a computer mouse. This probably renders WSD useless in many situations. Such a query is similar in nature to the narrative-based tests. On the other hand, the WSD approach could be more applicable when queries include only one or two words (which is the most frequent case in standard searches). So far, the studies regarding this problem have shown that: *i*) ambiguity does not have a strong impact on retrieval performances, especially when queries are quite long (the matching between a query and a document performs already an implicit disambiguation); *ii*) when a disambiguation algorithm is used, it must be very accurate (more than 90%), and *iii*) retrieval performances can be outperformed when indexing is based both on WSD and keywords.
3. The method used to "interpret" the semantic content of documents and queries: in existing approaches, once concepts are extracted, documents and queries are considered as bags of concepts. Therefore, semantic relationships that may exist between the concepts they contain are not exploited. Consequently, documents dealing with a subject close to that of the query could not be found with these approaches. WSD is a very partial semantic analysis that is insufficient to really interpret queries' content. For instance, consider the query "*Computer Viruses*" whose narrative is "Relevant documents should mention the name of the computer virus, and possibly the damage it does." To find relevant documents, a system must recognize phrases that contain virus names ("the XX virus", "the virus named XX", "the virus known as XX", etc.). It should also recognize phrases describing damages ("XX erases the hard disk", "XX causes system crashes", but not "XX propagates through mail messages"). These tasks are very difficult to perform and they are far beyond the scope of WSD. Query expansion (QE) is a possible solution to this problem because they make it possible to extend content representation of the query in order to increase the chance of matching documents [3][14][22]. That said, QE must be controlled in order to carefully choose the concepts to be added to the original query, otherwise the results can be disappointing [3][15]. In [11] and [12], the authors obtained positive results by expanding queries using WSD, but the effect of the use of WSD and QE are not quantified in isolation. In fact, even though the main objective of their study was to evaluate the performance of WSD in IR,

they should have examined the accuracy of their disambiguation method in isolation, so that they could quantify its effect when used in their IR experiments. A more comprehensive study was carried out in [9], in which the authors added additional sense information to both documents and queries using WordNet. Their large-scale experiments on a TREC collection produced promising results, clearly demonstrating the positive effect of WSD on retrieval performances. From our personal experiences, a possible solution to these problems is to use domain knowledge not only for WSD, but also for indexing and searching [17]. We notably showed how the use of semantic relationships could provide a precise representation of documents and query content. Relationships can therefore be used during the information retrieval process in order to allow the system to find a relevant document to a given query, even if it does not share any term with that query [16].

6 Conclusion

Our aim through this paper was to explore whether WSD can be useful in IR and CLIR. Our results confirmed that WSD does not allow for any retrieval performance improvement. It is obvious that these failures are primarily due to the weakness of WSD techniques, but also they depend on many other factors, such as the quality of the semantic resource used by WSD algorithm and the method used to "interpret" the semantic documents and queries content. We think that WSD-based indexing is a promising approach for language-independent indexing and retrieval systems. Although an efficient WSD is essential to create good conceptual indexes, we demonstrated in [7] that ambiguous indexes (with several concepts for some terms) are often sufficient to reach a good multilingual retrieval performance, for the reasons mentioned above. We also revealed that non-trivial queries, like those treated in our study, require adding domain knowledge during indexing and querying process. As shown in our previous work, this can be reached using expressive documents and queries languages, respectively during documents and queries content representation [16].

Acknowledgment

This work was supported by the Swiss National Science Foundation.

References

1. Agirre, E. and Lopez de Lacalle, O.: UBC-ALM: Decombining k-NN with SVD for WSD. Proc. of the 4th International Workshop on Semantic Evaluations (SemEval 2007). Prague, Czech Republic. pp 341-345.
2. Pierpaolo Basile and Annalina Caputo and Giovanni Semeraro: UNIBA-SENSE at CLEF 2008: SEmantic N-levels Search Engine. Working notes of 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008.

3. Mustapha Baziz: Indexation conceptuel le guidée par ontologie pour la recherche d'information. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2005.
4. Mustapha Baziz and Nathalie Aussenac-Gilles and Mohand Boughanem: Désambiguisation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques. Revue des Sciences et Technologies de l'information (RSTI) série ISI 8 (2003) 113-136.
5. Chan, Y. S. and Hwee T. and Zhong, Z.: NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. Proc. of the 4th International Workshop on Semantic Evaluations (SemEval 2007). Prague, Czech Republic. pp 253–256.
6. Jacques Guyot and Gilles Falquet and Karim Benzineb: Construire un moteur d'indexation. Revue Technique et science informatique (TSI), Hermes, Paris, 2006.
7. Jacques Guyot and Saïd Radhouani and Gilles Falquet: Conceptual Indexing for Multilingual Information Retrieval. In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. C. Peters, et al. (Eds.). Lecture Notes in Computer Science, Vol. 4022, Springer.
8. Arantxa Otegi and Eneko Agirre and German Rigau: IXA at CLEF 2008 Robust-WSD Task: using Word Sense Disambiguation for (Cross Lingual) Information Retrieval. Working notes of the 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008.
9. Sang-Bum Kim and Hee-Cheol Seo and Hae-Chang Rim: Information Retrieval Using Word Senses: Root Sense Tagging Approach. Proc. of the 27th annual international ACM SIGIR Conference, pages 258-265, ACM Press, 2004.
10. Robert Krovetz and W. Bruce Croft: Lexical ambiguity and information retrieval. ACM Transactions on Information Systems, 10(2) :115-141, 1992.
11. Shuang Liu and Fang Liu and Clement Yu and Weiyi Meng: An effective approach to document retrieval via utilizing WordNet and recognizing phrases. Proc. of the 27th ACM SIGIR Conference, pages 266-272. ACM Press, 2004.
12. Shuang Liu and Clement Yu and Weiyi Meng: Word Sense dDisambiguation in Queries. Proc. of the 14th ACM CIKM Conference, pages 525-532, 2005.
13. Rada Mihalcea and Dan Moldovan: Semantic indexing using wordnet senses. Proc. of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval, pages 35-45, Morristown, NJ, USA, 2000. Association for Computational Linguistics.
14. Rada Mihalcea and Dan Moldovan: An iterative approach to word sense disambiguation. Proc. of the Thirteenth International Florida Artificial Intelligence Research Society Conference, AAAI Press (2000) 219-223.
15. Yonggang Qiu and Hans-Peter Frei: Concept based query expansion. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, SIGIR, pages 160-169. ACM, 1993.
16. Saïd Radhouani and Gilles Falquet and Jean-Pierre Chevallet: Description Logic to Model a Domain Specific Information Retrieval System. Proc. 19th International Conference on Database and Expert Systems Applications (DEXA'08), Turin, Italy, 1-5 September, 2008. Lecture Notes in Computer Science, Vol. 5181, Springer.
17. Saïd Radhouani and Gilles Falquet: Using External Knowledge to Solve Multi-Dimensional Queries, Proc. 13th Intl Conf. on Concurrent Engineering Research and Applications (CE 2006), Antibes, Sept. 2006. IOS Press.
18. Hinrich Schütze and Jan O. Pedersen: Information Retrieval Based on Word Senses. Fourth Annual Symposium on Document Analysis and Information Retrieval, 1995.

19. Mark Sanderson: Word Sense Disambiguation and Information Retrieval. Proc. of the 17th ACM SIGIR Conference, pages 142-150, 1994.
20. Christopher Stokoe and Michael P. Oakes and John Tait: Word sense disambiguation in information retrieval revisited. Proc. of the 26th ACM SIGIR Conference, pages 159-166. ACM Press, 2003.
21. Ellen Marie Voorhees: Using wordnet to disambiguate word senses for text retrieval. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, SIGIR, pages 171-180. ACM, 1993.
22. Ellen Marie Voorhees: Query expansion using lexical-semantic relations. Proc. of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, Springer-Verlag New York, Inc. (1994) 61-69.