NLEL at CLEF 2009 Robust WSD Task

Davide Buscaldi and Paolo Rosso

Natural Language Engineering Lab,

ELiRF Research Group, DSIC

Universidad Politécnica de Valencia, Spain

{dbuscaldi, prosso}@dsic.upv.es

Abstract

This report describes our approach to the Robust - Word Sense Disambiguation task. We applied the same index expansion technique used in 2008 for the Question Answering WSD task, with the addition of pseudo (blind) relevance feedback. In our approach, a WordNet expanded index is generated from the disambiguated document collection. This index contains synonyms, hypernyms and holonyms of the disambiguated words contained in documents. Query words are searched for in both the expanded WordNet index and the default index. The results show that the use of the extended index did not prove useful, obtaining $14 - 16\%$ less in MAP with respect to the base system.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation, Text Analysis

Keywords

Information Retrieval, Word Sense Disambiguation

1   Introduction

In 2008 we participated in the QA-WSD task using an index expansion method based on WordNet hypernyms, synonyms and holonyms, which exploited the disambiguated collection [1]. The results did not show any relevant difference between the use of disambiguation or not, although we observed that passages returned using the disambiguated collection and our method tended to be shorter with respect to the base system. We took the opportunity presented by the Robust WSD Task at CLEF 2009 to test the same method in this task. A novelty for this participation was the introduction of a naïve Pseudo Relevance Feedback[3, 4] method, consisting in the expansion of the query with the top 5 terms (according to their tf.idf weights) resulting from the unexpanded query.

In the following section, we describe the RobustWorSE (Robust Wordnet Search Engine) system. In section 3 we describe the characteristics of our submissions and discuss the obtained results.

2   The RobustWorSE System

The core of the system is a standard Lucene[1] search engine (version 2.4.1). During the indexing phase, we create two indices: the first one (*text*) contains all the terms of the sentence; the second one (expanded index, or *wn* index) contains all the synonyms of the disambiguated words (we consider the sense with the highest score to be the "right" sense). In the case of nouns and verbs, it contains also their hypernyms. For nouns, the holonyms (if available) are also added to the index, in a similar way to the GeoWorSE system that participated in the 2008 GeoCLEF track [2]. For instance, let us consider the following sentence from document `GH951115-000080`:

> Splitting the left from the Labour Party would weaken the battle for progressive policies inside the Labour Party.

The underlined words are those that have been disambiguated in the collection. For these words we can found their synonyms and related concepts in WordNet, as listed in Table 1.

Table 1: Expansion of the index terms of the example sentence. NA : not available (the relationship is not defined for the Part-Of-Speech of the related word).

lemma	ass. sense	synonyms	hypernyms	holonyms
split	4	separate part	move	NA
left	1	–	position place	–
Labour Party	2	labor party	political party party	–
weaken	1	–	change alter	NA
battle	1	conflict fight engagement	military action action	war warfare
progressive	2	reformist	NA	NA
policy	2	–	argumentation logical argument line of reasoning line	–

Therefore, the *wn* index will contain the following terms: *separate, part, move, position, place, labor party, political party, party, change, alter, conflict, fight, engagement, war, warfare, military action, action, reformist, argumentation, logical argument, line of reasoning, line.*

During the search phase, in the default configuration, the *text* is searched for question terms. The top 5 resulting documents are analysed to extract up to 5 keywords that are used to expand the query. The keywords are

selected according to their *tf.idf* weight. Inverse document frequency is calculated over the entire document collection.

In the WSD configuration, search is carried out in a similar way, with the difference that every noun and adjective is also searched for in the *wn* index.

In Table 2 we show the expansion terms obtained for the topic `147-AH` : "*Oil accidents and birds*", using the two different configurations. From the example it is possible to notice that weights of the terms from the WordNet query resulted higher than those obtained with the base query.

---

[1]http://lucene.apache.org

Table 2: Terms extracted for pseudo relevance feedback, topic 147-AH. Original query: "Oil accidents birds".

mode	term	tf.idf weight
No-WSD	gero	52.07
	pigeon	31.68
	fli	29.21
	spill	28.66
	wildlife	24.24
WSD	spill	200.60
	pipeline	174.10
	river	64.05
	arco	63.93
	fish	61.82

3 Experiments

We submitted four runs with the WSD system, two using the NUS labeled collection and two with the UBC labeled collection. For each collection, we submitted one run using only the topic title and another one using both the title and the description. As baseline, we submitted two non-WSD runs, one in the configuration "title only" and one in the configuration "title and description".

In Table 3 we show the results obtained by the two non-WSD runs and the four WSD runs.

Table 3: Results obtained by RobustWorSE at the CLEF 2009 Robust WSD track. TD: Title and Description. TO: Title Only. NUS: NUS labelled collection. UBC: UBC labelled collection.

run ID	WSD	type	avg. MAP	avg. R-Prec
NLEL0901	n	TD	40.26%	38.72%
NLEL0906	n	TO	33.42%	32.98%
NLEL0902	n	TD NUS	27.14%	26.57%
NLEL0904	n	TD UBC	26.05%	25.59%
NLEL0903	n	TO NUS	17.48%	17.63%
NLEL0905	n	TO UBC	17.53%	18.67%

The results show that the use of the disambiguated collection did worsen the results obtained by the base system. There are differences of $\sim 16\%$ in MAP between the normal and WSD runs in the title only configuration, and up to 14.21% between in TD configuration. There is little difference ($\sim 1\%$ in TD configuration) between the use of the NUS disambiguated collection and the UBC disambiguated collection.

4 Conclusions

The index expansion method proved to be particularly ineffective, reducing the MAP of the base system up to $\sim 16\%$. We still have to investigate the specific reasons of such a negative behaviour, and the role of the pseudo relevance feedback in the obtained results.

Acknowledgements

References

[1] Davide Buscaldi and Paolo Rosso. Some experiments in question answering with a disambiguated document collection. In *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, pages 442–447. Springer, 2009.

[2] Davide Buscaldi and Paolo Rosso. Using geowordnet for geographical information retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*, volume 5706 of *Lecture Notes in Computer Science*, pages 863–866. Springer, 2009.

[3] S. E. Robertson. On term selection for query expansion. *J. Doc.*, 46(4):359–364, 1990.

[4] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM.