SINAI at Robust WSD Task @ CLEF 2008: When WSD is a good idea for Information Retrieval tasks?

Fernando Martínez-Santiago, José M. Perea-Ortega, Miguel A. García-Cumbreras
SINAI Research Group. Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{dofer,jmperea,magc}@ujaen.es

Abstract

This year we have participated in the first edition of Robust WSD task with the aim of investigating the performance of disambiguation tools applied to Information Retrieval (IR). The main interest of our experimentation is the characterization of queries where WSD is a useful tool. That is, which issues must be fulfilled by a query in order to apply an state-of-art WSD tool? After the interpretation of our experiments, we think that only queries with terms very polysemous and very high IDF value are improved by using WSD.

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; H.3.7 Digital Libraries

General Terms

Disambiguation, Information Retrieval, Experimentation

Keywords

Robust WSD

1 Introduction

Word Sense Disambiguation (WSD) is a traditional task into the discipline of Natural Language Processing (NLP). WSD is the identification process of sense of a word (having a number of different senses) used in a given sentence [1]. Information Retrieval (IR) is a task even older than WSD into the NPL community. IR is defined as the matching of some stated user query against a set of free-text records [2]. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual. User queries can range from multi-sentence full descriptions of an information need to a few words.

Nowadays, the information unit managed by most IR models is the word. A theoretical good idea is the elaboration of IR systems based on concepts better than words or the lemmas of those words. We define a concept as a lexicographic-independent representation of an idea or object. Given a language, it does not care the vocabulary available in order to represent such a concept. Thus, a concept-based IR system translates words into concepts. The advantages of such theoretical system are very interesting:

4 Conclusions and Future Work

State-of-art WSD is not an useful tool for every query, for every term of every query, but we think that some queries could be improved by using WSD. In this paper we investigate queries where WSD gets better results. We find that there are situations where WSD must be used, but these scenarios are very specific. Since some queries are improved by WSD and some queries not at all, if we want to apply WSD in a good way we have to manage two indexes per collection. In addition, the IR system will have to carry out a bit of additional analysis of the user query in order to take a decision about which of both indexes seem more suitable for each user query.

As future work, we think that there are promising ways to improve the obtained results. We want to explore a selective and fragmented evaluation of queries. We think that, given a user query, some words should be disambiguated and others do not. Thus, some words should be evaluated by using a index (the disambiguated one), and some words should be evaluated by using other index (the non-disambiguated one). We think that this line of investigation is promising, but some questions arise: which words should be disambiguated and which queries should not? This question is partially investigated in this text but a more in-depth analysis of results at word level is required. In this way, since we will have to manage simultaneously two indexes, how to calculate the score of each document for a given query? Finally, we think that the combination of this "*fragmented evaluation of queries*" and the application of clustering of senses such as is depicted in [4] will improve this future model proposed.

5 Acknowledgments

This work has been supported by a grant from the Spanish Government, project TIMOM (TIN2006-15265-C06-03), and the RFC/PP2006/Id_514 granted by the University of Jaén.

References

[1] Y. Wilks, B. Slator, and L. Guthrie. *Electric words: dictionaries, computers and meanings*. Cambridge, MA: MIT Press, 1996.

[2] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, London, U.K., 1998.

[3] C. Fellbaum. *WordNet: an electronic lexical database. Language, speech, and communication*. Cambridge, Mass: MIT Press, 1998.

[4] Eneko Agirre and Oier Lopez de Lacalle. Clustering wordnet word senses. In *Recent Advances on Natural Language (RANLP), Borovets, Bulgary*, 2003.

[5] Julio Gonzalo, Felisa Verdejo, and Irina Chugur. Indexing with wordnet synsets can improve text retrieval. pages 38–44, 1998.

[6] Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. Nus-ml:improving word sense disambiguation using topic features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic*, pages 249–252, 2007.

[7] Eneko Agirre and Oier Lopez de Lacalle. Ubc-alm: Combining k-nn with svd for wsd. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic*, pages 342–345, 2007.

[8] M.F. Porter. An algorithm for suffix stripping. In *Program 14*, pages 130–137, 1980.