Thinking Beyond the Nouns – Computing Semantic Relatedness Across Parts of Speech

Iryna Gurevych
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg
http://www.eml-research.de/~gurevych

Semantic relatedness is any kind of lexical-functional association between two words. For example, given the words “surfen”, “Internet” and “Apfel”, human judges are likely to classify “surfen - Internet” as being more closely related than “Internet - Apfel” or “surfen - Apfel”. Information about semantic relatedness of words is often required in various natural language processing (NLP) applications ranging from spelling correction (Budanitsky & Hirst, 2005) and sense tagging (McCarthy et al., 2004) to information retrieval (Gurevych 2005a).

In previous work, GermaNet was successfully used in order to approximate human judgments of semantic relatedness and evaluated on a dataset with German nouns (Gurevych, 2005b; Gurevych & Niederlich, 2005). However, from the point of view of robust NLP applications, this work has been subject to a number of limitations. Firstly, experimental work based on a limited number of word pairs (65) does not necessarily translate into good performance in functioning NLP systems when large amounts of unrestricted discourse have to be processed. Secondly, NLP applications have been confined to using information about semantic relatedness of nouns only. In this case, the discourse is represented as a set of noun concepts, which does not capture important information from the text contained in verbs or adjectives.

We hypothesize that measuring the relevancy of a document given a query in an information retrieval application on the basis of semantic relatedness will be improved, if the underlying representations include other kinds of content words. This requires that metrics of semantic relatedness operating on GermaNet are extended such that they allow to measure semantic relatedness of words belonging to different parts of speech.

The present contribution is aimed to enhance the understanding of approaches to compute semantic relatedness in a number of ways. We introduce a new dataset containing 350 word pairs of different parts of speech created in a corpus-based manner. Human subjects annotated the word pairs for their semantic relatedness. In the next step, we applied a number of existing metrics (Gurevych, 2005a; Resnik, 1995; Jiang & Conrath, 1997; Lin, 1998) and a new metric based on path lengths to compute semantic relatedness for the new dataset.

The paper will analyse the performance of individual metrics with respect to the new task of computing semantic relatedness across parts of speech. We will discuss the coverage of GermaNet for 350 corpus-based word pairs and implications for pre-processing the data. Further structural analysis of GermaNet will determine what kind of knowledge relevant to compute semantic relatedness is currently captured in GermaNet and how it could be extended to cover more types of semantic relatedness existing across parts of speech.

Bibliography

Gurevych, Iryna und Hendrik Niederlich. 2005. Computing Semantic Relatedness in German with Revised Information Content Metrics. In *Proceedings of "OntoLex 2005 - Ontologies and Lexical Resources" IJCNLP'05 Workshop*, Jeju Island, Republic of Korea, October 15, 2005. *To appear.*

Gurevych, Iryna. 2005a. Anwendungen des semantischen Wissens über Konzepte im Information Retrieval. In *ZB-Konferenz 2005 "Knowledge eXtended"*, Forschungszentrum Jülich, 02.-04. November 2005.

Gurevych, Iryna. 2005b. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'2005),* Jeju Island, Republic of Korea, October 11–13, 2005.

Hirst, Graeme and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.

Jiang, Jay J. & David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING)*. Tapei, Taiwan.

Lin, Dekang. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning, San Francisco, Cal.*, pages 296-304.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pages 280 – 287.

Resnik, Phil. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montreal, Canada, 20–25 August 1995, volume 1, pages 448–453.