Matuschek, Michael; Gurevych, Iryna
Technische Universität Darmstadt, www.ukp.tu-darmstadt.de

Beyond the Synset: Synonyms in Collaboratively Constructed Semantic Resources

We present a comparative analysis of synonyms in collaboratively constructed and linguistic lexical semantic resources and its implications for NLP research. Our focus is on the Wiki-based resources constructed mostly by non-experts on the Web which rely on user collaboration for quality management, as opposed to conventional sources of synonyms such as WordNet or thesauri. The most prominent examples are Wikipedia[1] (a free Encyclopedia) and its dictionary spin-offs Wiktionary[2] and OmegaWiki[3], where the latter has a strong focus on crosslinguality. We will examine three major ways how synonyms emerge in these resources, all of which imply a different operational definition of synonymy. We will then show how these synonyms can be mined and used building upon previous research in this field ((Zesch, Müller & Gurevych, 2008), (Wolf & Gurevych, 2010)), and we will also examine what theoretical conclusions about the notion of synonymy can possibly be drawn from our examinations.

The first part is the explicit encoding of synonymy, for example a link between word senses in Wiktionary, where it can be argued that the user community agrees that they are synonymous; this gives rise to a new notion of cognitive synonymy which is anchored in the “collective mind”. Following earlier work considering German resources (Meyer & Gurevych, 2010), we analyze in detail how synonyms are dealt with in different English resources, what problems arise for their exploitation (e.g. due to inconsistencies) and how this compares to conventional lexical resources.

The second part is the implicit encoding of synonyms, e.g. deducing synonymy through a transitive relation between two senses in Wiktionary. Another example is the redirect/link anchor structure in Wikipedia. Here, it can be claimed that synonyms link to the same article. Contrary to previous work (Nakayama et al., 2008), we show that this claim does not really hold, but some interesting observations can be made regarding the link structure and how it relates to the idea of synonymy. We also examine how links lead to insights about capital or subordinate traits of “distant relatives” (cf. (Cruse, 1986)) which might give us a better idea of why words are perceived as similar.

The third part does not rely on the structure of the resources but on the inference of synonymy from context. Two examples are mining synonyms from the Wikipedia revision history (cf. (Nelken & Yamangil, 2008)) as well as from an aligned corpus of Wikipedia and Simple Wikipedia. The hypothesis is that (apart from spelling corrections) terms could be synonyms if they have been replaced by each other in an article’s history (Kulessa, 2008) or if they are used interchangeably in the “normal” and “simple” versions of an article (Zhu, Bernhard, & Gurevych 2010). Both examples relate to the notion of propositional synonymy, but replacement of terms might also imply that they were deemed invalid somehow; this observation could be another path for future research.

To substantiate our work, we give illustrative examples of synonyms collected from the examined resources, and we provide statistical evidence about their structure and content.

[1] http://www.wikipedia.org/
[2] http://www.wiktionary.org/
[3] http://www.omegawiki.org/

Bibliography

Cruse, D.A. 1986: *Lexical Semantics.* Cambridge Textbooks in Linguistics, Cambridge University Press.

Kulessa, S. 2008: *Mining Wikipedia's Revision History for Paraphrase Extraction (Master Thesis).* Technische Universität Darmstadt.

Meyer, C., & Gurevych, I. 2010: Worth its Weight in Gold or Yet Another Resource — A Comparative Study of Wiktionary, OpenThesaurus and GermaNet. *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 38-49). Iaşi, Romania.

Nakayama, K., Pei, M., Erdmann, M., Ito, M., Shirakawa, M., Hara, T., 2008: Wikipedia Mining: Wikipedia as a Corpus for Knowledge Extraction. *Proceedings of Annual Wikipedia Conference (Wikimania).*

Nelken, R., & Yamangil, E. 2008: Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms. *Proceedings of the Wikipedia and AI Workshop at the AAAI Conference.* Chicago, USA.

Wolf, E., & Gurevych, I. 2010: Expert-Built and Collaboratively Constructed Lexical Semantic Resources for Natural Language Processing. *Language and Linguistics Compass.* (to appear)

Zesch, T., Müller, C. & Gurevych, I. 2008: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech, Morocco.

Zhu, Z., Bernhard, D. & Gurevych, I. 2010: A Monolingual Tree-based Translation Model for Sentence Simplification. *Proceedings of The 23rd International Conference on Computational Linguistics,* Bei Jing, China. (to appear)