Putting the „Wisdom-of-Crowds“ to Use in NLP:

Collaboratively Constructed Semantic Resources on the Web

Prof. Dr. Iryna Gurevych

Ubiquitous Knowledge Processing (UKP) Lab, Computer Science Department

Technical University of Darmstadt, Germany

Since early 90-ies, the Web has served as a **unique corpus** with background knowledge for various NLP tasks. The Web as a corpus has been employed in three principal ways: (i) obtaining Web-based frequencies for specific terms and constructions, (ii) collecting term-specific Web corpora by retrieving the corresponding text snippets, and finally (iii) constructing task- and domain-targeted corpora that are cleaned, linguistically analyzed, and further exploited in a traditional way.

The rise of Web 2.0 and the so called Socio-Semantic technologies in recent years has led to huge amounts of **user generated content** produced by ordinary users on the Web. This content called for user-generated tagging to enable better information navigation and retrieval. Therefore, semantically tagged collaboratively constructed knowledge repositories emerged that represent a novel type of Web-originated resources - we call them **collaboratively constructed semantic resources (CCSR)**. Example instances of CCSR are collaboratively constructed and semantically enriched multilingual online **encyclopedias**, such as Wikipedia, or collaboratively constructed online multilingual **dictionaries**, such as Wiktionary.

Researchers have started to employ Web-based semantic resources as **substitutes** for conventional lexical semantic resources and world knowledge bases, such as thesauri, machine readable dictionaries, or wordnets. In overcoming the limitations of existing resources, such as coverage gaps, significant construction and maintenance costs, and their restricted availability, there is now a hope to **significantly enhance the performance** of numerous algorithms by utilizing the so called “wisdom-of-crowds” in broad coverage NLP systems. Initial research has already shown the high potential of CCSR. **Wikipedia** has been put to use as a background knowledge source to enhance the representation of texts in e.g. text categorization, question answering, or named entity disambiguation. **Wiktionary**, which is a much younger project, has so far been utilized in just a few NLP tasks, such as recognizing opinion orientation of terms in blogs, or diachronic phonology. Moreover, both Wikipedia and Wiktionary have been employed as knowledge sources to measure the relatedness of words. Combining CCSR with statistical measures resulting in the **shallow, approximative semantic knowledge** has thereby brought excellent results in many NLP tasks.

The **Ubiquitous Knowledge Processing Lab** (Technical University of Darmstadt, Germany) has had a significant impact at the research work in this area which will be summarized below.

Analysis of collaboratively constructed resources. In [9], the structure of Wikipedia as a lexical semantic resource is analyzed. The authors compare Wikipedia with conventional linguistic resources and identify sources of lexical semantic information that can be utilized by NLP applications. [5] extends this analysis to a completely new resource, i.e. Wiktionary, and constitutes a pioneering work about Wiktionary as lexical semantic resource. Part of the lexical semantic information is represented in the category and article graphs of Wikipedia and Wiktionary. Therefore, a comparative graph-theoretic analysis of the collaboratively constructed and conventional linguistic resources, such as the German wordnet, is carried out. The analysis reveals similarities in their structures [2]. [8] performs an in-depth comparison specifically for the Wikipedia category graph and the German wordnet.

Utilizing collaboratively constructed resources in NLP applications. In [7], the UKP Lab investigates the performance of a set of semantic relatedness measures on various datasets. Thereby, the Explicit Semantic Analysis measure operating on Wikipedia outperforms the measures operating on the WordNet. The top performing measure is integrated into an information retrieval (IR) algorithm and shows consistent improvements over a baseline IR model. Wiktionary is furthermore utilized as a resource for computing semantic relatedness in [4], and achieves excellent performance for the majority of datasets. Finally, [1] employs the knowledge in Wikipedia and Wiktionary for domain-specific information retrieval and obtains significant improvements. Utilizing CCSR in NLP tasks requires high performance access to the semi-structured knowledge therein. The multilingual Java-based Wikipedia API and the Wiktionary API (English, German) described in [5] are available from the web site of the UKP Lab.

Interoperability of conventional and collaboratively constructed resources. As different resources turn out to be useful for different tasks and often contain complementary information, [3] investigates the representational interoperability of CCSR and

conventional linguistic resources using Wikipedia, Wiktionary, and Wordnet as a case study. They propose a system architecture that translates specific modeling concepts of different knowledge repositories to a set of elementary units. Therefore, NLP algorithms, such as computing semantic relatedness or lexical chaining, tailored toward some conventional knowledge repository, can now be used with any CCSR integrated in the framework.

Current work The ongoing research at the UKP Lab focuses on an in-depth analysis of the content interoperability between conventional resources and CCSR. Thereby, it is essential to link the corresponding word senses in different repositories. The resulting information will be utilized e.g. in paraphrase generation. Furthermore, we investigate the time aspect of CCSR, i.e. the revision history of Wikipedia, as a source of paraphrases to incorporate this information into question answering. The next line of investigation is the construction of lexical semantic graphs as an enhanced text representation for a set of CCSR. Its utility will be evaluated on the task of keyphrase extraction and as a representation for semantic information retrieval, which defines the relevance between topics and documents as a function of similarity between their underlying lexical semantic graphs.

Open issues As there is no editorial control imposed over CCSR, employing them as substitutes for conventional lexical semantic and world knowledge resources entails addressing a set of challenges, such as dealing with the incompleteness of information and the inconsistent structure of entries, uneven coverage, the vagueness and sometimes insufficient quality of the user-contributed information. On the other hand, a lot of information still waits to be exploited, resulting from the multilingual nature of Wikipedia and Wiktionary, additional types of lexical semantic information that can be mined from their content, structure and usage, e.g. time-dependent information in Wikipedia. The impact of CCSR upon the research in “deep” semantics, such as word sense ambiguity, anaphora resolution, discourse structure, and knowledge representation remains to be seen.

Acknowledgements The author would like to thank the members of the UKP Lab who contributed to the ideas and projects described above, in particular Torsten Zesch, Christof Müller, Konstantina Garoufi and Aljoscha Burchardt. This work is funded by the German Research Foundation (DFG GU 798/1-2, GU 798/1-3, and 798/3-1) and the Volkswagen-Foundation (I/82806).

References

[1] C. Müller and I. Gurevych. 2008. Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In: Working Notes for the CLEF 2008 Workshop, Aarhus, Denmark, September 17-19, 2008. (to appear).

[2] K. Garoufi, T. Zesch and I. Gurevych. 2008. Graph-Theoretic Analysis of Collaborative Knowledge Bases in Natural Language Processing In: Poster Proceedings of the 7th International Semantic Web Conference (ISWC 2008), Karlsruhe, Germany, October 2008. (to appear).

[3] K. Garoufi, T. Zesch and I. Gurevych. Representational Interoperability of Linguistic and Collaborative Knowledge Bases. In: Proceedings of KONVENS’08 Workshop on Lexical-Semantic and Ontological Resources Maintenance, Representation, and Standards, Berlin, Germany, September 30 – October 2, 2008.

[4] T. Zesch, C. Müller and I. Gurevych. 2008. Using Wiktionary for Computing Semantic Relatedness. In: Proceedings of Twenty-Third AAAI Conference on Artificial Intelligence (AAAI’08). Chicago, Illinois, July 13–17, 2008.

[5] T. Zesch, C. Müller and I. Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of Language Resources and Evaluation Conference (LREC’08). Marakkech, Morokko, May 28-30, 2008.

[6] I. Gurevych, C. Müller and T. Zesch. 2007. What to be? - Electronic Career Guidance Based on Semantic Relatedness. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Conference (ACL’2007), Prague, Czech Republic, June 23–30, 2007. pp. 1032-1039.

[7] T. Zesch, I. Gurevych and M. Mühlhäuser. 2007. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In: Proceedings of HLT-NAACL’2007, Rochester, NY, April 22-27, Association for Computational Linguistics, ISBN 1-932432-94-9, 2007. pp. 205-208.

[8] T. Zesch and I. Gurevych. 2007. Analysis of the Wikipedia Graph Structure for NLP Applications. In: Proceedings of the HLT-NAACL’2007 Workshop “TextGraphs-2 Graph-based Algorithms for Natural Language Processing”, Rochester, NY, April 26, Association for Computational Linguistics, 2007. pp. 1-8.

[9] T. Zesch, I. Gurevych and M. Mühlhäuser. 2007. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In: Biannual Conference of the Society for Computational Linguistics and Language Technology, Tübingen, Germany, April, pp. 213-221.