Analyzing and Accessing Wikipedia as a Lexical Semantic Resource

Torsten Zesch, Iryna Gurevych, Max Mühlhäuser

Department of Telecooperation
Ubiquitous Knowledge Processing Group
Darmstadt University of Technology, Hochschulstraße 10
D-64289 Darmstadt, Germany
{zesch,gurevych,max} (at) tk.informatik.tu-darmstadt.de

Analyzing and Accessing Wikipedia as a Lexical Semantic Resource

In this paper, we analyze Wikipedia as an emerging lexical semantic resource that is growing exponentially. Recent research has shown that Wikipedia can be successfully employed for NLP tasks, e.g. question answering (Ahn et al., 2004), text classification (Gabrilovich and Markovitch, 2006) or named entity disambiguation (Bunescu and Pasca, 2006). We extend this work by focusing on the analysis of Wikipedia content, in particular its category structure, and present a highly efficient Java-based API to use Wikipedia in large scale NLP.

At first, we compare Wikipedia with conventional lexical semantic resources such as dictionaries, thesauri, semantic wordnets or paper bound encyclopedias. We show that different parts of Wikipedia reflect different aspects of these resources. Wikipedia articles form a heavily linked encyclopedia, whereas the category system is a collaboratively constructed thesaurus used for collaborative tagging (Voss, 2006). Our analysis reveals that Wikipedia additionally contains a vast amount of knowledge about named entities, domain specific terms or specific word senses that cannot be easily found in other freely available lexical semantic resources. We also show that the redirect system of Wikipedia can be used as a dictionary for synonyms, spelling variations and abbreviations.

Next, we perform a detailed analysis of the category graph of Wikipedia employing graph-theoretic methods. Previous studies (Holloway et al., 2005; Buriol et al., 2006; Zlatic et al., 2006) have focused on the Wikipedia article graph. We analyze the category graph, as it can be regarded as an important lexical semantic resource of its own. From this analysis, we draw some conclusions for adapting algorithms from semantic wordnets to the Wikipedia category graph.

If Wikipedia is to be used as a lexical-semantic resource in large-scale NLP tasks, efficient programmatic access to the knowledge therein is required. We review existing access mechanisms (Riddle, 2006; Sigurbjörnsson et al., 2006; Wikimedia Foundation, 2006) and show that they are limited with respect to either their performance or the access functions provided. Therefore, we introduce a general purpose, high performance Java-based API for Wikipedia that overcomes these limitations and is specifically designed to turn Wikipedia into a lexical semantic resource. We transform the Wikipedia database dump into an optimized representation that can be more efficiently accessed. The information nuggets, such as redirects, are explicitly stored there, whereas they are scattered in the original dump. As a case study, the API is used to perform the graph-theoretic analysis of Wikipedia mentioned above.

The first release of the Wikipedia-API provides access to the full set of explicit information encoded in Wikipedia including articles, links, categories and redirects. In particular, the category system is converted into a graph representation, on which the whole range of standard graph algorithms can be applied, e.g. finding the shortest path between two given categories. We plan to make the API freely available for research purposes. Our ongoing work investigates the use of knowledge in Wikipedia to compute semantic relatedness of words and extensions of the API for mining the lexical semantic knowledge represented in Wikipedia articles.

References

David Ahn, Valentin Jijkoun, Gilad Mishne, Karin Müller, Maarten de Rijke, and Stefan Schlobach. 2004. Using Wikipedia at the TREC QA Track. In *Proceedings of TREC 2004*.

Razvan Bunescu and Marius Pasca. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the EACL*, pages 9–16, Trento, Italy.

Luciana Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Stefano Millozzi. 2006. Temporal Analysis of the Wikigraph. In *Proceedings of Web Intelligence*, Hong Kong.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *AAAI*, pages 1301–1306, Boston, MA.

Todd Holloway, Miran Bozicevic, and Katy Börner. 2005. Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. *ArXiv Computer Science e-prints*, cs/0512085.

Tyler Riddle. 2006. Parse::mediawikidump. URL `http://search.cpan.org/~triddle/Parse-MediaWikiDump-0.40/`.

Börkur Sigurbjörnsson, Jaap Kamps, and Maarten de Rijke. 2006. Focused Access to Wikipedia. In *Proceedings DIR-2006*.

Jakob Voss. 2006. Collaborative thesaurus tagging the Wikipedia way. *ArXiv Computer Science e-prints*, cs/0604036.

Wikimedia Foundation. 2006. Wikipedia. URL `http://en.wikipedia.org/wiki/Wikipedia:Searching`.

Vinko Zlatic, Miran Bozicevic, Hrvoje Stefancic, and Mladen Domazet. 2006. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74:016115.