Wikipedia-based Corpora for Analyzing Revisions, Discussions and Text Quality in Collaborative Writing

Johannes Daxenberger[†], Oliver Ferschke[†] and Iryna Gurevych[†‡]

[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information

[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt

http://www.ukp.tu-darmstadt.de

Wikipedia is both a valuable resource and a remarkable example of a non-standard data source for computational linguistics research. It is a unique source of authentic corpus material to study collaborative writing in the Web and offers terabytes of collaboratively constructed content in numerous domains. Due to the magnitude, diversity and the collaborative construction of its content, creating task-specific corpora for computational linguistics from Wikipedia is a non-trivial task. For example, the semi- and sometimes unstructured data makes its segmentation difficult.

In this presentation, we report about our ongoing work on several Wikipedia-based corpora tailored to support computational linguistics research with respect to:

- classifying the user edits in Wikipedia revisions (`Wikipedia Revision Act Corpus`);
- discourse analysis of the Wikipedia discussions (`Simple English Wikipedia Discussion Corpus`, `English Wikipedia Discussion Corpus`);
- analyzing the text quality of the Wikipedia articles (`Wikipedia Quality Assessment Corpus`, `Wikipedia Quality Feedback Dataset`).

The long-term goal of this research is a comprehensive analysis framework of collaborative writing based on user revisions, discussions and quality features of the resulting texts. Beyond the text level, this also involves the temporal and social analysis dimensions. The corpora presented below originate from the English Wikipedia and carry expert- or community-based annotations to support the respective analysis tasks.

Wikipedia Revision Act Corpus (WPRAC) is a subset of 20 articles from the Wikipedia Quality Assessment Corpus described below. The articles have been selected according to their age, quality and edit frequency. Since articles go through different stages of maturity and quality over time, we selected revisions from particular time spans to be annotated. We extracted revisions using the Wikipedia Revision Toolkit [2, 3]. The obtained adjacent revision pairs were automatically segmented, resulting in a list of 1,995 local edits. Currently, trained annotators are classifying them according to the twenty types of user edits defined in the annotation schema. Example user edits are spelling corrections, adding content, or reverting vandalism.

Simple English Wikipedia Discussion Corpus (SEWD) consists of 100 article Talk pages extracted from the Simple English Wikipedia [1]. Talk pages are used to report article deficiencies, announce intended changes and discuss the future development of an article. We annotated 1,400 automatically segmented discussion turns with dialog acts capturing the coordination efforts for article improvement. Example dialog acts contain problem statements, or action announcements.

Subsequently, we also built the English Wikipedia Discussion Corpus (EWPD). It is based on a revised version of the SEWD annotation schema and consists of 5,225 annotated turns from 200 Talk pages extracted from the English Wikipedia.

Wikipedia Quality Assessment Corpus contains 15,000 Wikipedia article pairs consisting of a distinguished article (*good* or *featured*) and a non-distinguished counterpart of comparable length. We determine distinguished articles based on the label *good* or *featured* as defined by the Wikipedia community[1]. As the criteria applied by the community may change over time, inconsistencies in the actual text quality of *featured* or *good* articles should be kept in mind.

Wikipedia Quality Feedback Dataset contains 7.9 million article judgments by the users for more than 740,000 articles gathered between June and September 2011. Each user judgment characterizes an article on the five-point Likert scale according to four dimensions: trustworthiness, objectivity, completeness and readability. Additionally, it contains information about the user expertise in the article's subject area, which can be used as a proxy for the user confidence. We built a database-driven framework to access this data and to enable large-scale analysis of the judgments in combination with the Wikipedia article revisions.

[1] http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the Hessian research excellence program “Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz” (LOEWE) as part of the research center “Digital Humanities”.

References

[1] Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, April 2012.

[2] Oliver Ferschke, Torsten Zesch, and Iryna Gurevych. Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia’s Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, Portland, OR, June 2011.

[3] Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.