Web Corpus Graphs

Lea Helmers Roland Schäfer
lea.helmers@fu-berlin.de roland.schaefer@fu-berlin.de
Deutsche Grammtik/Linguistik, Freie Universität Berlin

Web corpora play an important role in theoretical linguistics as well as computational linguistics (Biemann et al., 2013; Schäfer and Bildhauer, 2013). The size of crawled web corpora has reached the order of ten to a hundred billion tokens (Pomikálek et al., 2012; Schäfer and Bildhauer, 2012). A central issue in the construction of web corpora of such size is automated quality control, which is first of all understood as the process of making sure that noisy data from various sources is detected and removed (Eckart et al., 2012). Secondly, the suitability of web corpora for computational linguistics tasks (like collocation extraction) has been used as a means of assessing the quality of web corpora (Biemann et al., 2013; Kilgarriff et al., in prep.).

A third kind of quality control is related to web crawling, and it has so far received virtually no attention at all. Web crawling is an act of sampling, and many parameters of the crawling algorithm and the post-crawl selection from among the documents influence the representativeness of the sample w. r. t. the population, regardless of the concrete definition of the sampling frame. We suggest that quality control for web corpus crawling can be dealt with by, among other things, analyses of the topological makeup of the corpus. We present a first approximation by treating the documents (or rather their hosts) in a web corpus (the 9.1 billion token DECOW2012) as nodes in a web corpus graph. The edges (= hyperlinks) between those hosts are extracted from the whole crawl data (not just the final corpus documents), which contains approximately 1.1 billion raw hyperlinks. The final DECOW2012 corpus graph itself contains 393,528 nodes (= hosts) and 7,636,127 edges (= hyperlink connections of different strengths between the hosts).

The properties of this graph are compared to known properties of the web graph (cf. Broder et al., 2000; Kumar et al., 2000, and much subsequent work). We focus on the analysis of the number of documents which the hosts contribute to the corpus, and the distribution of degrees of the hyperlinks between the hosts (including standard measures such as centrality, PageRank, etc.). We argue that analyses of this kind are essential for our understanding of the relation between the data on the web and the data in a web corpus. Thus, they help to answer the difficult question of whether the corpus is representative – and of what. Although quite technical at first sight, web corpus graphs are therefore directly relevant for linguistic evaluations of web corpora.

C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski, and T. Zesch. Scalable construction of high-quality web corpora. *Special issue of JLCL*, 2013.

A. Broder, R. Kumar, F. Maghoul, P. Raghavan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. In *Proceedings of the 9th WWW conference*, pages 309–320. North-Holland Publishing Co, 2000.

T. Eckart, U. Quasthoff, and D. Goldhahn. The influence of corpus quality on statistical measurements of language resources. In *Proceedings of LREC 08*, pages 2318–2321, Istanbul, 2012.

A. Kilgarriff, V. Baisa, M. Jakubicek, V. Kovara, and P. Rychly. How to evaluate a corpus. ms., in prep.

R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal. The web as a graph. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–10, New York, NY, USA, 2000. ACM.

J. Pomikálek, M. Jakubíček, and P. Rychlý. Building a 70 billion word corpus of English from ClueWeb. In *Proceedings of LREC 08*, pages 502–506, 2012.

R. Schäfer and F. Bildhauer. Building large corpora from the web using a new efficient tool chain. In *Proceedings of LREC 2012*, pages 486–493, Istanbul, ELRA 2012.

R. Schäfer and F. Bildhauer. *Web Corpus Construction.* Synthesis Lectures on Human Language Technologies. Morgan and Claypool, San Francisco, 2013.