Web Corpus Construction

Roland Schäfer and Felix Bildhauer
(Freie Universität Berlin)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 22), 2013, 145 pages, paper-bound, ISBN 9781608459834, doi:10.2200/S00508ED1V01Y201305HLT022

Reviewed by
Serge Sharoff
University of Leeds

The Web is the main source of data in modern computational linguistics. Other volumes in the same series, for example, *Introductions to Opinion Mining* (Liu 2012) and *Semi-supervised Machine Learning* (Søgaard 2013), start their problem statements by referring to data from the Web. This volume starts its own introduction by praising Web corpora for their size, ease of construction, and availability as a source of new text types. A random check of papers from the most recent ACL meeting also shows that the majority of them use Web data in one way or another. Our field definitely needs a comprehensive overview and a DIY manual for the task of constructing a corpus from the Web. This book is, to the best of my knowledge, the first attempt at providing such an overview.

The book consists of an introduction and four chapters outlining the four main steps of Web corpus construction. They include: "Data Collection" (Chapter 2), "Basic Corpus Cleaning" (Chapter 3), "Linguistic Processing" (Chapter 4), and "Corpus Evaluation" (Chapter 5).

Chapter 2 provides a very useful outline of the main properties of the Web and the crawling strategies. The chapter starts with an overview of a large-scale study of Web connectivity from Baeza-Yates, Castillo, and Efthimiadis (2007), listing various parameters of connectivity for a range of Top-Level Domains. However, there is little discussion of the implications for the corpus development task; for example, does the difference of the in-degree parameter of the Web pages from Chile and the UK have any implications for the Web corpora crawled from those domains? The chapter then proceeds to another important topic, which concerns the parameters of crawling; for example, the crawl bias and the number of seeds, and their influence on the final corpus. Section 2.4.1 illustrates the problems with the crawl bias by an example of deWac, a large commonly used corpus of German (Baroni et al. 2009). The second most frequent proper name bigram in this corpus is found to be *Falun Gong*. However, more analysis into the nature of the bias should have been beneficial. It is less likely to be related to the PageRank bias, the main bias discussed in Section 2.4.2. Other most frequent bigrams from deWac are not presented in the book, but it is interesting to note that the fourth place in it is occupied by *Hartz IV*, and the tenth place by *Digital Eyes*. This suggests that the bias comes from frequency spikes (i.e, a large number of instances collected from a small number of Web sites). Another shortcoming of this chapter is that nothing is said specifically about obtaining data from such resources as Twitter or Facebook, which need access via APIs rather than direct crawling.

doi:10.1162/COLI_r_00214

Chapter 3 introduces methods for basic cleaning of the corpus content, such as processing of text formats (primarily HTML tags), language identification, boilerplate removal, and deduplication. Such low-level tasks are not considered to be glamorous from the view of computational linguistics, but they are extremely important for making Web-derived corpora usable (Baroni et al. 2008). The introduction offered in this chapter is reasonably complete, with good explanations of the sources of problems as well as with suggestions for the tools to be used in each task. An important bit which is missing in this chapter concerns the suggestions for choosing a particular cleaning pipeline. Although the choice indeed depends on the purposes of corpus collection, an indication of which pipeline suits which purpose is desirable.

Chapter 4 is devoted to basic steps for linguistic processing of Web corpora, such as tokenization, POS tagging, and lemmatization, as well as orthographic normalization. Even though the processing pipeline is roughly the same for all NLP tasks, it becomes harder for Web corpora because they exhibit greater diversity in comparison with more homogeneous text collections (e.g., WSJ texts). Web texts are also considerably noisier, in the sense of containing nonstandard linguistic expressions, which are likely to be a challenge to the tools trained on more standard texts. The chapter presents some interesting case studies—in particular, the sources of POS tagging errors and nonstandard orthography.

Chapter 5 describes ways for evaluating and comparing corpora. It gives examples of checking for word and sentence length and for sentence-level duplication. It also introduces methods for comparing frequency lists. Like other chapters it includes many interesting observations, such as the methods for extrinsic evaluation of corpora. However, the chapter does not address many issues important for corpus evaluation and comparison. Given that the previous chapters introduced a number of pipelines and corpora, this chapter would have been an ideal place to illustrate all the aspects of the pipelines by evaluating them in a consistent way. There are occasional references to this goal, such as the frequency lists of French nouns in Section 5.3.1, but this particular comparison is fairly impressionistic, and it concludes with a declaration of basic similarity of the underlying corpora. Does this mean that the crawling, cleaning, and linguistic processing pipelines do not matter? In any case, not even an impressionistic comparison of the pipelines is performed for other evaluation methods. Some illustrations are also not informative (e.g., Table 5.1.1 shows two frequency lists with the identical ranks for their words, which leads to the trivial rank correlation value of 1). The chapter contains a single paragraph devoted to composition of Web corpora. Given the size of such corpora, their evaluation crucially depends on understanding what has been crawled. The task has been approached by a number of models, such as supervised and semi-supervised classification, clustering, topic modeling, and so forth, which should have been included in the discussion. The discussion does contain a relevant reference to Mehler, Sharoff, and Santini (2010), which surveys approaches to the genres of the Web, but other aspects of corpus composition need to be addressed, too.

Overall, it is very useful to have a book that introduces all the aspects of Web corpus construction in a single volume with coherent presentation. The volume under review does cover the entire range of topics relevant to Web corpus construction and illustrates them via numerous examples. I would recommend it to students just starting their corpus development experiments. As for the drawbacks of the volume, there is a need to improve the structure of argumentation for the next edition. Bits of information are sometimes introduced in an incomplete way and re-introduced again in subsequent sections. For examples, two tools for crawling are discussed towards the end of Section 2.3.3, while more tools are mentioned as the discussion of crawling strategies progresses. Chapter 1 starts with a fairly random list of non-Web corpora, whereas an overview of the book structure is confined to a short paragraph. Often, frustratingly little information is provided besides an annotated bibliography, rather than a presentation of the relevant methods and issues. In some cases this is accompanied with a statement that "covering this topic is beyond the scope of this volume," even if the nature of the problem and the solutions could have been easily explained in a one-page summary. Another minor concern is an (understandable) emphasis on the tools and corpora developed by the authors, primarily on their German corpus.

I have to admit ambivalence in my final verdict: The book is a useful introduction to an important topic, but it definitely warrants a new edition, which eliminates the shortcomings of the current one.

References

Baeza-Yates, Ricardo, Carlos Castillo, and Efthimis N. Efthimiadis. 2007. Characterization of national Web domains. *ACM Transactions on Internet Technology (TOIT)*, 7(2):9.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, Marco, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. 2008. Cleaneval: A competition for cleaning Web pages. In *Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech.

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, pages 638–643.

Mehler, Alexander, Serge Sharoff, and Marina Santini, editors. 2010. *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Søgaard, Anders. 2013. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

This book review was edited by Pierre Isabelle.

Serge Sharoff is an Associate Professor in the Centre for Translation Studies, University of Leeds. His recent research is in the field of collection, annotation, and use of comparable corpora. He was Chair of ACL SIGWAC, Web as Corpus, and he is one of the organizers of the WAC and BUCC workshops. Sharoff's e-mail address is `s.sharoff@leeds.ac.uk`.