EACL 2014

14th Conference of the European Chapter of the Association for Computational Linguistics

Proceedings of the 9th Web as Corpus Workshop (WaC-9)

April 26, 2014
Gothenburg, Sweden

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-937284-83-1

Preface

The World Wide Web has become increasingly popular as a source of linguistic data, not only within the NLP communities, but also with theoretical linguists facing problems of data sparseness or data diversity. Accordingly, web corpora continue to gain importance, given their size and diversity in terms of genres/text types. However, after a decade of activity in the web-as-corpus community, a number of issues in web corpus construction still needs much research.

For instance, questions concerning sampling strategies and their relation to crawling algorithms have not yet been explored in any detail so far. Virtually all existing large web corpora were sampled using breath-first web crawls, which demonstrably yield biased results and make the corpus particularly vulnerable to criticism targeting their sampling frame. In addition, relying on the results of commercial search engines when selecting the seed URLs for such crawls (as has been common practice) introduces an additional bias. This is also an issue for smaller web corpora obtained without web crawling, by simply downloading a number of documents fixed in advance.

Turning to the linguistic post-processing of web corpora, problems may arise, among other things, from the kind of non-copy edited, quasi-spontaneous language typical of numerous genres of computer-mediated communication. Spelling errors and deliberate non-standard spellings are a case in point, and grammatical variation as well as (semi-)graphical elements like emoticons also figure prominently. Technically, all of these present challenges for NLP tools (such as POS-taggers, parsers etc.) that expect "clean", copy-edited standard language. From a conceptual point of view, such variation begs the question whether (and to what extent) web corpora should be normalized and how this can be achieved in a transparent and non-destructive way.

A similar point can be made when it comes to document filtering: Currently available web corpora have usually undergone radical cleaning procedures in order to produce "high-quality" data. However, at least for some uses of the data, aggressive and sometimes arbitrary removal of material in the form of whole documents or parts thereof can be problematic.

Finally, the systematic evaluation of web corpora, for example in the form of task-based comparisons to traditional corpora, has only lately shifted into focus.

Against this backdrop, most of the contributions included in this volume address particular problems related to data collection and normalization, while others offer a broader perspective on the process of constructing a particular web corpus. The papers were selected after a highly competitive review process, and we would like to thank all those who submitted, as well as the program committee who contributed to the review process.

Felix Bildhauer & Roland Schäfer, March 2014

WaC-9 Program Chairs

Felix Bildhauer, Freie Universität Berlin (Germany)
Roland Schäfer, Freie Universität Berlin (Germany)

WaC-9 Program Committee

Adrien Barbaresi, École Normale Supérieure de Lyon (France)
Silvia Bernardini, Università di Bologna (Italy)
Chris Biemann, Technische Universität Darmstadt (Germany)
Jesse Egbert, Northern Arizona University (USA)
Stefan Evert, Friedrich-Alexander Universität Erlangen-Nürnberg (Germany)
Adriano Ferraresi, Università di Bologna (Italy)
William Fletcher, United States Naval Academy (USA)
Dirk Goldhahn, Universität Leipzig (Germany)
Adam Kilgarriff, Lexical Computing Ltd. (UK)
Anke Lüdeling, Humboldt-Universität Berlin (Germany)
Alexander Mehler, Goethe-Universität Frankfurt am Main (Germany)
Uwe Quasthoff, Universität Leipzig (Germany)
Paul Rayson, Lancaster University (UK)
Serge Sharoff, University of Leeds (UK)
Sabine Schulte im Walde, Universität Stuttgart (Germany)
Egon Stemle, European Academy of Bozen/Bolzano (Italy)
Yannick Versley, Universität Heidelberg (Germany)
Stephen Wattam, Lancaster University (UK)
Torsten Zesch, Universität Darmstadt (Germany)

Table of Contents

Conference Program

11:15-11:30 Welcome by Felix Bildhauer, Roland Schäfer

11:30–12:00 *Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources*
Adrien Barbaresi

12:00–12:30 *Focused Web Corpus Crawling*
Roland Schäfer, Adrien Barbaresi and Felix Bildhauer

14:00–14:30 *Less Destructive Cleaning of Web Documents by Using Standoff Annotation*
Maik Stührenberg

14:30–15:00 *Some Issues on the Normalization of a Corpus of Products Reviews in Portuguese*
Magali Sanches Duran, Lucas Avanço, Sandra Aluísio, Thiago Pardo and Maria da Graça Volpe Nunes

15:00–15:30 *bs,hr,srWaC - Web Corpora of Bosnian, Croatian and Serbian*
Nikola Ljubešić and Filip Klubička

16:00–16:30 *The PAISÀ Corpus of Italian Web Texts*
Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci and Vito Pirrelli

16:30–17:00 *Internet Data in a Study of Language Change and a Program Helping to Work with Them*
Varvara Magomedova, Natalia Slioussar and Maria Kholodilova

17:00-18:00 Discussion