ACL 2016

The 54th Annual Meeting of the Association for Computational Linguistics

Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task

August 12, 2016
Berlin, Germany

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 978-1-945626-15-9

Introduction

These proceedings contain the papers presented at the 10th Web as Corpus Workshop (WAC-X 2016) organized by the ACL Special Interest Group on Web as Corpus (SIGWAC), co-located with the ACL conference 2016. It took place on August 12, 2016.

With WAC-X, the series of WAC workshops continues its successful tradition going back to 2005. Thematically, the WAC workshops have always been positioned between computational linguistics and theoretically oriented empirical linguistics, and this year is no exception. A majority of the accepted papers relates in some way to the **construction of web corpora** (Barbaresi, Salway et al., Krause, Würschinger et al., Mendels et al., Ljubešić and Fišer, Schäfer) with a clear tendency towards specialized corpora collected for individual research questions and towards data sources similar to but not identical to the web (such as Twitter). The extraction and generation of **meta data** for web-derived (or similar) corpora has also been a recurring theme in Web as Corpus workshops (Schäfer and Bildhauer, Barbaresi, Dalan and Sharoff). A lot of the accepted papers also deal with **research based on web data** (Krause, Würschinger et al., Mendels et al., Ljubešić and Fišer), demonstrating that web corpora are a unique source of data in (computational) linguistics and related fields.

We received a total of 15 full paper submissions for the main workshop (5 short, 10 long) of which 9 were accepted (2 short, 7 long), resulting in an overall acceptance rate of 60% as the result of a double-blind peer review process (three independent reviews per paper).

Adding to the success of the WAC-X event was the inclusion of the final presentations for the shared task on **Automatic Linguistic Annotation of Computer-Mediated Communication/Social Media (EmpiriST)**. The papers by the five competing teams and the introductory paper by the organizers are also included in these proceedings. System descriptions were reviewed non-anonymously by the task organizers and participants. Each submitted paper received two reviews. All papers meeting our formal requirements and quality standards after revisions were accepted for publication, regardless of whether they make a novel research contribution.

In these proceedings, the WAC-X papers are printed before the EmpiriST papers. In both groups, the papers are printed in the order of the corresponding presentations.

We would like to thank all authors for submitting their research to WAC-X and the members of the program committee for their hard work reviewing the papers and making valuable suggestions.

Paul C. Cook
Stefan Evert
Roland Schäfer
Egon Stemle

SIGWAC web page: `https://www.sigwac.org.uk`

Organizers:

Paul Cook, University of New Brunswick
Stefan Evert, Friedrich-Alexander Universität Erlangen-Nürnberg
Roland Schäfer, Freie Universität Berlin
Egon Stemle, European Academy of Bozen/Bolzano

Program Committee:

Adrien Barbaresi, Österreichische Akademie der Wissenschaften, Wien
Silvia Bernardini, Università di Bologna
Douglas Biber, Northern Arizona University, Flagstaff
Felix Bildhauer, Institut für Deutsche Sprache Mannheim
Katrien Depuydt, Instituut voor Nederlandse Lexicologie, Leiden
Jesse de Does, Instituut voor Nederlandse Lexicologie, Leiden
Cédrick Fairon, Université catholique de Louvain
William H. Fletcher, U.S. Naval Academy, Annapolis
Iztok Kosem, Trojina Institute for Applied Slovene Studies, Ljubljana
Simon Krek, Jožef Stefan Institute, Ljubljana
Lothar Lemnitzer, Berlin-Brandenburgische Akademie der Wissenschaften
Nikola Ljubešić, Sveučilišta u Zagrebu
Siva Reddy, University of Edinburgh
Steffen Remus, Technische Universität Darmstadt
Pavel Rychly, Masaryk University, Brno
Kevin Scannell, Saint Louis University
Serge Sharoff, University of Leeds
Klaus Schulz, Ludwig-Maximilians-Universität München
Kay-Michael Würzner, Berlin-Brandenburgische Akademie der Wissenschaften
Torsten Zesch, Universität Duisburg-Essen
Pierre Zweigenbaum, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Orsay

Table of Contents

Conference Program

9:30–10:30 WAC-X morning session

9:30–9:40 *Welcome and Introduction*

9:40–10:00 *Automatic Classification by Topic Domain for Meta Data Generation, Web Corpus Evaluation, and Corpus Comparison*
Roland Schäfer and Felix Bildhauer

10:00–10:30 *Efficient construction of metadata-enhanced web corpora*
Adrien Barbaresi

11:00–12:30 WAC-X noon session

11:00–11:30 *Topically-focused Blog Corpora for Multiple Languages*
Andrew Salway, Dag Elgesem, Knut Hofland, Øystein Reigem and Lubos Steskal

11:30–12:00 *The Challenges and Joys of Analysing Ongoing Language Change in Web-based Corpora: a Case Study*
Anne Krause

12:00–12:30 *Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms. The case of 'rapefugee', 'rapeugee', and 'rapugee'.*
Quirin Würschinger, Mohammad Fazleh Elahi, Desislava Zhekova and Hans-Jörg Schmid

13:30–14:30 EmpiriST session

13:30–13:50 *EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora*
Michael Beißwenger, Sabine Bartsch, Stefan Evert and Kay-Michael Würzner

13:50–14:10 *SoMaJo: State-of-the-art tokenization for German web and social media texts*
Thomas Proisl and Peter Uhrig

14:10–14:30 *UdS-(retrain|distributional|surface): Improving POS Tagging for OOV Words in German CMC and Web Data*
Jakob Prange, Andrea Horbach and Stefan Thater

14:30–15:10 WAC-X and EmpiriST Teaser Talks

14:30–14:35 *Babler - Data Collection from the Web to Support Speech Recognition and Keyword Search*
Gideon Mendels, Erica Cooper and Julia Hirschberg

14:35–14:40 *A Global Analysis of Emoji Usage*
Nikola Ljubešić and Darja Fišer

14:40–14:45 *Genre classification for a corpus of academic webpages*
Erika Dalan and Serge Sharoff

14:45–14:50 *On Bias-free Crawling and Representative Web Corpora*
Roland Schäfer

14:55–15:00 *EmpiriST: AIPHES - Robust Tokenization and POS-Tagging for Different Genres*
Steffen Remus, Gerold Hintz, Chris Biemann, Christian M. Meyer, Darina Benikova, Judith Eckle-Kohler, Margot Mieskes and Thomas Arnold

15:00–15:05 *bot.zen @ EmpiriST 2015 - A minimally-deep learning PoS-tagger (trained for German CMC and Web data)*
Egon Stemle

15:05–15:10 *LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text*
Tobias Horsmann and Torsten Zesch

15:10–16:30 WAC-X and EmpiriST Poster Session