

10th Web as Corpus Workshop (WAC-X 2016) and the EmpiriST Shared Task

Held at ACL 2016

Berlin, Germany
12 August 2016

ISBN: 978-1-5108-2769-1

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2016) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2016)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

| | |
|---|-----|
| <i>Automatic Classification by Topic Domain for Meta Data Generation, Web Corpus Evaluation, and Corpus Comparison</i> | |
| Roland Schäfer and Felix Bildhauer | 1 |
| <i>Efficient construction of metadata-enhanced web corpora</i> | |
| Adrien Barbaresi | 7 |
| <i>Topically-focused Blog Corpora for Multiple Languages</i> | |
| Andrew Salway, Dag Elgesem, Knut Hofland, Øystein Reigem and Lubos Steskal | 17 |
| <i>The Challenges and Joys of Analysing Ongoing Language Change in Web-based Corpora: a Case Study</i> | |
| Anne Krause | 27 |
| <i>Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms. The case of 'rapefugee', 'rapeugee', and 'rapugee'.</i> | |
| Quirin Würschinger, Mohammad Fazleh Elahi, Desislava Zhekova and Hans-Jörg Schmid | 35 |
| <i>EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora</i> | |
| Michael Beißwenger, Sabine Bartsch, Stefan Evert and Kay-Michael Würzner | 44 |
| <i>SoMaJo: State-of-the-art tokenization for German web and social media texts</i> | |
| Thomas Proisl and Peter Uhrig | 57 |
| <i>UdS-(retrain\distributionall\surface): Improving POS Tagging for OOV Words in German CMC and Web Data</i> | |
| Jakob Prange, Andrea Horbach and Stefan Thater | 63 |
| <i>Babler - Data Collection from the Web to Support Speech Recognition and Keyword Search</i> | |
| Gideon Mendels, Erica Cooper and Julia Hirschberg | 72 |
| <i>A Global Analysis of Emoji Usage</i> | |
| Nikola Ljubešić and Darja Fišer | 82 |
| <i>Genre classification for a corpus of academic webpages</i> | |
| Erika Dalan and Serge Sharoff | 90 |
| <i>On Bias-free Crawling and Representative Web Corpora</i> | |
| Roland Schäfer | 99 |
| <i>EmpiriST: AIPHES - Robust Tokenization and POS-Tagging for Different Genres</i> | |
| Steffen Remus, Gerold Hintz, Chris Biemann, Christian M. Meyer, Darina Benikova, Judith Eckle-Kohler, Margot Mieskes and Thomas Arnold | 106 |
| <i>bot.zen @ EmpiriST 2015 - A minimally-deep learning PoS-tagger (trained for German CMC and Web data)</i> | |
| Egon Stemle | 115 |
| <i>LTL-UDE @ EmpiriST 2015: Tokenization and PoS Tagging of Social Media Text</i> | |
| Tobias Horsmann and Torsten Zesch | 120 |