Lecture Notes in Artificial Intelligence 8105

Subseries of Lecture Notes in Computer Science

Iryna Gurevych Chris Biemann
Torsten Zesch (Eds.)

Language Processing and Knowledge in the Web

25th International Conference, GSCL 2013
Darmstadt, Germany, September 25-27, 2013
Proceedings

Springer

Volume Editors

Iryna Gurevych
Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technische Universität Darmstadt, Darmstadt, Germany and Ubiquitous Knowledge Processing Lab, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany
E-mail: gurevych@ukp.informatik.tu-darmstadt.de

Chris Biemann
FG Language Technology
Department of Computer Science, Technische Universität Darmstadt
Darmstadt, Germany
E-mail: biem@cs.tu-darmstadt.de

Torsten Zesch
Ubiquitous Knowledge Processing Lab
Department of Computer Science, Technische Universität Darmstadt, Darmstadt, Germany and Ubiquitous Knowledge Processing Lab, German Institute for International Educational Research (DIPF) Frankfurt am Main, Germany
E-mail: zesch@ukp.informatik.tu-darmstadt.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-40721-5 e-ISBN 978-3-642-40722-2
DOI 10.1007/978-3-642-40722-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013946726

CR Subject Classification (1998): I.2, H.3, H.4, F.1, I.5, I.4, C.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedication

This volume is dedicated to the memory of Prof. Wolfgang Hoeppner, who passed away on June 4, 2012.

For many years, Wolfgang Hoeppner was a member and coordinator of the Scientific Board of the German Society for Computational Linguistics and Language Technology (GSCL). He contributed greatly to GSCL and the interdisciplinary development of GSCL at the intersection of knowledge discovery, human–computer interaction, and language technology.

Iryna Gurevych, Chris Biemann, and Torsten Zesch
Co-editors

Preface

The International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2013) was held in Darmstadt, Germany, during September 25–27, 2013. The meeting brought together an international audience from Germany and other countries. The conference's main theme was "Language Processing and Knowledge in the Web."

Language processing and knowledge in the Web has been an area of great and steadily increasing interest within the language processing and related communities over the past years. Both in terms of academic research and commercial applications, the Web has stimulated and influenced research directions, yielding significant results with impact beyond the Web itself. Thus, the conference turned out to be a very useful forum in which to highlight the most recent advances within this domain and to consolidate the individual research outcomes.

The papers accepted for publication in the present Springer volume address language processing and knowledge in the Web on several important dimensions, such as computational linguistics, language technology, and processing of unstructured textual content in the Web.

About one third of the papers are dedicated to fundamental computational linguistics research in multilingual settings. On the one hand, the work deals with different languages, such as German, Manipuri, or Chinese. On the other hand, it deals with a wide range of computational linguistics tasks, such as word segmentation, modeling compounds, coreference resolution, word sense annotation, named entity recognition, or lexical-semantic processing.

The second third of papers address a wide range of language technology tasks, such as construction of a new error tagset for an Arabic learning corpus and prediction of cause of death from verbal autopsy text. Two papers deal with different aspects of machine translation. An evaluation of several approaches to sentiment analysis is the subject of another contribution. Last but not least, one article deals with dependency-based algorithms in question answering for Russian.

The third portion of the papers presented in this volume deals with processing of unstructured textual content in the Web. An important issue is the construction of Web corpora for computational research. One paper presents a tool for creating tailored Twitter corpora, while another describes the construction of a corpus of parsable sentences from the Web. Optimizing language processing components to work on noisy Web content is the subject of several papers. Finally, one contribution exploits Wikipedia as a knowledge resource for topic modeling, and another presents a novel summarization algorithm for community-based question-answering services.

In summary, the GSCL 2013 conference clearly demonstrated the recent advances in language processing research for processing the textual content in the

Web. It also showed that Web corpora can be effectively employed as a resource in language processing. A particular property of the Web is its multilinguality, which is reflected in a significant number of papers dealing with languages other than English and German published in the present volume.

We would like to sincerely thank the Organizing Committee of GSCL 2013 and the reviewers for their hard work, the invited speakers for their inspiring contributions to the program, the sponsors and funding agencies for their financial contributions, and Tristan Miller for his technical assistance in compiling the final volume. We also express our gratitude to the Hessian LOEWE research excellence program and to the Volkswagen Foundation for funding the conference organizers as part of the research center "Digital Humanities" (Chris Biemann) and the Lichtenberg Professorship Program under grant № I/8280 (Iryna Gurevych).

Iryna Gurevych
Chris Biemann
Torsten Zesch

Organization

GSCL 2013 was organized by the Ubiquitous Knowledge Processing (UKP) Lab of the Technische Universität Darmstadt's Department of Computer Science.

Organizing Committee

General Chair

Iryna Gurevych	Technische Universität Darmstadt and German Institute for International Educational Research (DIPF), Germany

Program Chairs

Chris Biemann	Technische Universität Darmstadt, Germany
Torsten Zesch	Technische Universität Darmstadt and German Institute for International Educational Research (DIPF), Germany

Workshops/Tutorials Chair

György Szarvas	Nuance Communications, Germany

Local Chairs

Christian M. Meyer	Technische Universität Darmstadt, Germany
Wolfgang Stille	Technische Universität Darmstadt, Germany

Program Committee

Abend, Omri	The Hebrew University of Jerusalem, Israel
Auer, Sören	Universität Leipzig, Germany
Bernhard, Delphine	Université de Strasbourg, France
Buitelaar, Paul	Digital Enterprise Research Institute, National University of Ireland, Ireland
Choudhury, Monojit	Microsoft Research, India
Cimiano, Philipp	Bielefeld University, Germany
Cysouw, Michael	Philipps-Universität Marburg, Germany
Dagan, Ido	Bar-Ilan University, Israel
De Luca, Ernesto William	University of Applied Sciences, Potsdam, Germany
de Melo, Gerard	International Computer Science Institute, USA
Dipper, Stefanie	Ruhr-Universität Bochum, Germany
Fellbaum, Christiane	Princeton University, USA

Frank, Annette	Heidelberg University, Germany
Girju, Roxana	University of Illinois at Urbana-Champaign, USA
Heid, Ulrich	University of Hildesheim, Germany
Heyer, Gerhard	Universität Leipzig, Germany
Hirst, Graeme	University of Toronto, Canada
Hoeppner, Wolfgang (†)	Universität Duisburg-Essen, Germany
Kozareva, Zornitsa	Information Sciences Institute, University of Southern California, USA
Lobin, Henning	Justus Liebig University Giessen, Germany
Lüdeling, Anke	Humboldt-Universität zu Berlin, Germany
Magnini, Bernardo	ITC-irst, Italy
Mahlow, Cerstin	University of Zurich, Switzerland
Manandhar, Suresh	University of York, UK
McCarthy, Diana	University of Sussex, UK
Mehler, Alexander	Goethe University Frankfurt am Main, Germany
Mihalcea, Rada	University of North Texas, USA
Miller, Tristan	Technische Universität Darmstadt, Germany
Mohammad, Saif	National Research Council Canada, Canada
Navigli, Roberto	Sapienza University of Rome, Italy
Nenkova, Ani	University of Pennsylvania, USA
Neumann, Günter	German Research Center for Artificial Intelligence (DFKI), Germany
Ng, Vincent	University of Texas at Dallas, USA
Padó, Sebastian	Heidelberg University, Germany
Palmer, Alexis	Saarland University, Germany
Poesio, Massimo	University of Essex, UK
Quasthoff, Uwe	Universität Leipzig, Germany
Rehm, Georg	German Research Center for Artificial Intelligence (DFKI), Germany
Riezler, Stefan	Heidelberg University, Germany
Schlangen, David	Bielefeld University, Germany
Schmidt, Thomas	Institut für Deutsche Sprache, Germany
Schmitz, Ulrich	University of Duisburg-Essen, Germany
Schröder, Bernhard	University of Duisburg-Essen, Germany
Stein, Benno	Bauhaus-Universität Weimar, Germany
Storrer, Angelika	Technische Universität Dortmund, Germany
Søgaard, Anders	University of Copenhagen, Denmark
Teich, Elke	Saarland University, Germany
Temnikova, Irina	University of Wolverhampton, UK
Wandmacher, Tonio	SYSTRAN, France
Witt, Andreas	Institut für Deutsche Sprache, Germany
Witte, René	Concordia University, Canada
Wolff, Christian	Universität Regensburg, Germany

Big Data and Text Analytics

Hans Uszkoreit

Saarland University, Germany
German Research Center for Artificial Intelligence (DFKI),
Germany

Abstract. Text analytics is faced with rapidly increasing volumes of language data. In our talk we will show that big language data are not only a challenge for language technology but also an opportunity for obtaining application-specific language models that can cope with the long tail of linguistic creativity. Such models range from statistical models to large rule systems. Using examples from relation/event extraction we will illustrate the exploitation of large-scale learning data for the acquisition of application specific syntactic and semantic knowledge and discuss the achieved improvements of recall and precision.

Biography: Hans Uszkoreit is Professor of Computational Linguistics and—by cooptation—of Computer Science at Saarland University. At the same time he serves as Scientific Director at the German Research Center for Artificial Intelligence (DFKI) where he heads the DFKI Language Technology Lab. He has more than 30 years of experience in language technology which are documented in more than 180 international publications. Uszkoreit is Coordinator of the European Network of Excellence META-NET with 60 research centers in 34 countries and he leads several national and international research projects. His current research interests are information extraction, atomatic translation and other advanced applications of language and knowledge technologies as well as computer models of human language understanding and production.

Distributed Wikipedia LDA

Massimiliano Ciaramita

Google Research
Zurich, Switzerland

Abstract. When someone mentions Mercury, are they talking about the planet, the god, the car, the element, Freddie, or one of some 89 other possibilities? This problem is called disambiguation, and while it's necessary for communication, and humans are amazingly good at it, computers need help. Automatic disambiguation is a long standing problem and is the focus of much recent work in natural language processing, web search and data mining. The surge in interest is due primarily to the availability of large scale knowledge bases such as Wikipedia and Freebase which offer enough coverage and structured information to support algorithmic solutions and web-scale applications. In this talk I will present recent work on the disambiguation problem based on a novel distributed inference and representation framework that builds on Wikipedia, Latent Dirichlet Allocation and pipelines of MapReduce.

Biography: Massimiliano Ciaramita is a research scientist at Google Zurich. Previously he has worked as a researcher at Yahoo! Research and the Italian National Research Council. He did his undergraduate studies at the University of Rome "La Sapienza" and obtained ScM and PhD degrees from Brown University. His main research interests involve language understanding and its applications to search technologies. He has worked on a wide range of topics in natural language processing and information retrieval, including disambiguation, acquisition, information extraction, syntactic and semantic parsing, query analysis, computational advertising and question answering. He co-teaches (with Enrique Alfonseca) "Introduction to Natural Language Processing" at ETH Zurich.

Multimodal Sentiment Analysis

Rada Mihalcea

Department of Computer Science and Engineering
University of North Texas, USA

Abstract. During real-life interactions, people are naturally gesturing and modulating their voice to emphasize specific points or to express their emotions. With the recent growth of social websites such as YouTube, Facebook, and Amazon, video reviews are emerging as a new source of multimodal and natural opinions that has been left almost untapped by automatic opinion analysis techniques. One crucial challenge for the coming decade is to be able to harvest relevant information from this constant flow of multimodal data. In this talk, I will introduce the task of multimodal sentiment analysis, and present a method that integrates linguistic, audio, and visual features for the purpose of identifying sentiment in online videos. I will first describe a novel dataset consisting of videos collected from the social media website YouTube and annotated for sentiment polarity at both video and utterance level. I will then show, through comparative experiments, that the joint use of visual, audio, and textual features greatly improves over the use of only one modality at a time. Finally, by running evaluations on datasets in English and Spanish, I will show that the method is portable and works equally well when applied to different languages.

Biography: Rada Mihalcea is an Associate Professor in the Department of Computer Science and Engineering at the University of North Texas. Her research interests are in computational linguistics, with a focus on lexical semantics, graph-based algorithms for natural language processing, and multilingual natural language processing. She serves or has served on the editorial boards of the journals of *Computational Linguistics*, *Language Resources and Evaluation*, *Natural Language Engineering*, *Research in Language in Computation*, *IEEE Transations on Affective Computing*, and *Transactions of the Association for Computational Linguistics*. She was a program co-chair for the Conference of the Association for Computational Linguistics (2011), and the Conference on Empirical Methods in Natural Language Processing (2009). She is the recipient of a National Science Foundation CAREER award (2008) and a Presidential Early Career Award for Scientists and Engineers (2009).

Table of Contents