Communications in Computer and Information Science 537

Editorial Board

Simone Diniz Junqueira Barbosa
Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil

Phoebe Chen
La Trobe University, Melbourne, Australia

Alfredo Cuzzocrea
ICAR-CNR and University of Calabria, Cosenza, Italy

Xiaoyong Du
Renmin University of China, Beijing, China

Joaquim Filipe
Polytechnic Institute of Setúbal, Setúbal, Portugal

Orhun Kara
TÜBİTAK BİLGEM and Middle East Technical University, Ankara, Turkey

Igor Kotenko
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia

Ting Liu
Harbin Institute of Technology (HIT), Harbin, China

Krishna M. Sivalingam
Indian Institute of Technology Madras, Chennai, India

Dominik Ślęzak
University of Warsaw and Infobright, Warsaw, Poland

Takashi Washio
Osaka University, Osaka, Japan

Xiaokang Yang
Shanghai Jiao Tong University, Shangai, China

More information about this series at http://www.springer.com/series/7899

Cerstin Mahlow · Michael Piotrowski (Eds.)

Systems and Frameworks for Computational Morphology

Fourth International Workshop, SFCM 2015 Stuttgart, Germany, September 17–18, 2015 Proceedings

Springer

Editors
Cerstin Mahlow
Institut für Deutsche Sprache
Mannheim
Germany

Michael Piotrowski
Leibniz Institute of European History
Mainz
Germany

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-23978-1 ISBN 978-3-319-23980-4 (eBook)
DOI 10.1007/978-3-319-23980-4

Library of Congress Control Number: 2015947943

Springer Cham Heidelberg New York Dordrecht London
© Springer International Publishing Switzerland 2015
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at SFCM 2015: The Fourth International Workshop on Systems and Frameworks for Computational Morphology, held on September 17 and 18, 2015, at the University of Stuttgart, Germany.

From the point of view of computational linguistics, morphological resources form the basis for all higher-level applications. This is especially true for languages with a rich morphology like Czech, German, Finnish, Italian, Latin, Pali, Polish, Sanskrit, and Serbian, to name some of the languages targeted in this volume. A morphology component should thus be capable of analyzing single wordforms as well as whole corpora. For many practical applications not only morphological analysis but also generation is required, i.e., the production of surfaces corresponding to specific categories.

Apart from uses in computational linguistics, there are numerous practical applications that can benefit from morphological analysis and/or generation or even require it, for example in textual analysis, word processing, information retrieval, or dialog systems. These applications have specific requirements for morphological components, including requirements from software engineering, such as programming interfaces or robustness.

With the workshop on Systems and Frameworks for Computational Morphology (SFCM) we have established a place for presenting and discussing recent advances in the field of computational morphology. SFCM focuses on linguistically motivated morphological analysis and generation, computational frameworks for implementing such systems, and linguistic frameworks suitable for computational implementation. In 2015 the workshop took place for the fourth time. The main theme for SFCM 2009 was systems for a specific language, namely, German; SFCM 2011 looked at phenomena at the interface between morphology and syntax in various languages; SFCM 2013 discussed the role of morphological analysis and generation to improve the rather disappointing situation with respect to language technology for languages other than English. All workshop programs are accessible via the series website at http://www.sfcm.eu.

SFCM 2015 aimed at broadening the scope to include research on very under-resourced languages, interactions between computational morphology and formal, quantitative, and descriptive morphology, as well as applications of computational morphology in the Digital Humanities. For the first time, it was a two-day workshop and featured a special session dedicated to CLARIN ("Common Language Resources and Technology Infrastructure"). Dörte de Kok from CLARIN-D and Krister Lindén from FIN-CLARIN gave insights on CLARIN in general and how the German and Finnish CLARIN centers support computational morphology.

Based on the number of submissions and the number of participants at the workshop we can definitely state that the topic of the workshop has met with great interest from the community, both from academia and industry. The broader scope of this workshop

as outlined in the call for papers is reflected in the broader variety of topics discussed and use cases described. We received 16 submissions describing complete works of research as well as novel challenges and visions, of which 10 were accepted after a thorough review by the members of the program committee. The peer review process was double-blind, and each paper received three independent reviews.

In addition to the regular papers, we had the pleasure of Magda Ševčíková from Charles University Prague giving an invited talk on the role of morphology in the Prague Dependency Treebank.

The discussions after the talks and during the demo session, as well as the final plenum, showed the interest in and the need and requirements for further efforts in the field of computational morphology. During the last years, we see more workshops and conferences following the lead idea of SFCM to bring together researchers with different perspectives interested in the common topic of morphological phenomena. We also see more events actively supporting the idea of a "workshop" by offering ample opportunity to demonstrate and discuss ongoing research in an informal atmosphere, where participants can get critical but supportive feedback, in addition to the traditional presentation of complete research by talks in front of a plenum. We are encouraged to continue the series of SFCM workshops by the advent of similar morphology-oriented events targeting specific languages (such as the "Greek Workshop on Frameworks and Systems for Computational Morphology" in 2013 on Rhodes) or addressing computational aspects in morphology from the linguistic point of view (such as the workshop on "Computational Methods for Descriptive and Theoretical Morphology" in 2015 in Vienna).

Topics of this Book

This book starts with the invited paper by Magda Ševčíková ("Morphology within the Multi-layered Annotation Scenario of the Prague Dependency Treebank"), presenting morphological annotation as an element in a large multi-layered treebank. Following the approach of relations between form and function, morphological information is represented as attributes at the tectogrammatical layer. This allows the use for practical applications like dependency-based machine translation and the creation of lexical databases.

The following paper, "Designing and Comparing G2P-Type Lemmatizers for a Morphology-Rich Language" by Steffen Egger, presents work on lemmatization of ancient Latin. He finds that general-purpose string-to-string transduction models as used for grapheme-to-phoneme conversion perform better than techniques based on suffix transformation. The lemmatizer is aimed to complement lexicon-based systems.

In the paper "Morphological Disambiguation of Classical Sanskrit," Oliver Hellwig targets another ancient language. He describes a system for tokenization and morphological analyzation of Sanskrit combining a morphological rule-base with statistical selection of the most probable analysis.

The third paper aiming at ancient languages is "Morphological Analysis and Generation for Pali" by David Alfter and Jürgen Knauth. They introduce a system for

analyzing and generating Pali word forms. The system can be integrated into a general technical infrastructure and supports linguistic research on Pali.

The paper "A Universal Feature Scheme for Rich Morphological Annotation" by John Sylak-Glassman, Christo Kirov, David Yarowski, and Roger Que introduces a general set of features that represent fine distinctions in meaning expressed by inflectional morphology across languages. For evaluation, the texts of the Bible are used as a large parallel corpus. This work is in the field of typology and cross-linguistic morphology to improve NLP applications such as machine translation and information extraction.

In the paper "Dsolve—Morphological Segmentation for German using Conditional Random Fields," Kay-Michael Würzner and Bryan Jurish present a system for the segmentation of complex German word forms. Segmentation is handled as a classification task using conditional random fields. Unlike previous segmentation approaches, Dsolve also predicts types of morph boundaries, which boosts performance.

Maciej Janicki's paper "A Multi-purpose Bayesian Model for Word-Based Morphology" presents morphology as a systematic correspondence between full word forms without segmenting word forms into smaller units. The Bayesian models trained this way perform very well when evaluated for lexicon expansion and the generation of inflected forms in German and Polish.

In their paper "Using HFST—Helsinki Finite-State Technology for Recognizing Semantic frames," Krister Lindén, Sam Hardwick, Miikka Silfverberg, and Erik Axelson show the use of HFST as a comprehensive framework using the example of recognizing semantic frames. HFST is a toolkit for text analysis covering all steps from tokenization over morphological analysis up to semantic tagging. This paper emphasizes the usefulness of such toolkits for text analysis in the Digital Humanities.

The next paper, "Morpho-SLaWS: An API for Morphosyntactic Annotation of the Serbian Language" by Toma Tasovac, Saša Rudan, and Siniša Rudan, gives another insight into the use of NLP tools in Digital Humanities. The Serbian Lexical Web Service (SLaWS) offers a broad range of functions to be used as a resource-oriented web service. Morpho-SLaWS is the morphological component of this infrastructure and can be combined with other linguistic resources and tools.

Next, in their paper "Morphological Analysis and Generation of Monolingual and Bilingual Medical Lexicons," Serena Pelosi, Annibale Elia, and Alessandro Maisto describe the automatic creation of Italian–English medical lexical resources. They use finite-state transducers to analyze combinations of prefixes, confixes, and suffixes used in medical terms. This approach allows also for recognition of relevant neologisms and multi-word expressions.

Finally, the paper "Grammar Debugging" by Michael Maxwell argues for the representation of morphological and phonological features in a linguistic way that allows for automatic conversion into parsers. For debugging, he presents a tool that enables the linguist to follow each step during analysis and generation. Here again, linguists do not need programming skills but can adjust the parser on the linguistic level.

The contributions show that high-quality research is being conducted in the area of computational morphology: Mature systems are further developed and new systems and applications are emerging. Other languages than English are becoming more

important. The papers in this book come from eight countries and two continents, discuss a wide variety of living and ancient languages, and illustrate that, in fact, morphological resources are indeed the basis for higher-level natural language processing applications.

The trend towards open-source developments still goes on and evaluation is considered an important issue. Making high-quality morphological resources freely available will help to advance the state of the art and allow for the development of high-quality real-world applications. Useful applications shown as use cases here with carefully conducted evaluation demonstrate to a broad audience that computational morphology might not be a solved problem but is mature enough to be used in research settings in the Digital Humanities. It also shows that computational morphology is an actual science with tangible benefits for society.

July 2015

Cerstin Mahlow
Michael Piotrowski

Acknowledgments

We would like to thank the authors for their contributions to the workshop and to this book. We also thank the reviewers for their effort and for their constructive feedback, encouraging and helping the authors to improve their papers. The submission and reviewing process and the compilation of the proceedings was supported by the EasyChair system. We thank Aliaksandr Birukou, the editor of the series *Communications in Computer and Information Science* (CCIS), and the Springer staff for publishing the proceedings of SFCM 2015. We are grateful for the financial support given by the German Society for Computational Linguistics and Language Technology (GSCL), the Institut für Deutsche Sprache, and CLARIN-D. We thank Jonas Kuhn and the staff from the Institute for Natural Language Processing (IMS) at the University of Stuttgart for hosting the workshop and for helping with local organization.

Organization

The Fourth International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2015) was organized and chaired by Cerstin Mahlow and Michael Piotrowski. The workshop was held at the Institute for Natural Language Processing (IMS) at the University of Stuttgart, Germany.

Program Chairs

Cerstin Mahlow	Institut für Deutsche Sprache, Mannheim, Germany
Michael Piotrowski	Leibniz Institute of European History, Mainz, Germany

Program Committee

Delphine Bernhard	Université de Strasbourg, France
Bruno Cartoni	Google, Switzerland
Simon Clematide	University of Zurich, Switzerland
Thomas Hanneforth	University of Potsdam, Germany
Lauri Karttunen	Stanford University, USA
Kimmo Koskenniemi	University of Helsinki, Finland
Krister Lindén	University of Helsinki, Finland
Anke Lüdeling	Humboldt-Universität zu Berlin, Germany
Cerstin Mahlow	Institut für Deutsche Sprache, Germany
Günter Neumann	DFKI Saarbrücken, Germany
Michael Piotrowski	Leibniz Institute of European History, Germany
Yves Scherrer	University of Geneva, Switzerland
Helmut Schmid	Ludwig-Maximilians-Universität München, Germany
Angelika Storrer	University of Mannheim, Germany
Marcin Wolinski	Polish Academy of Science, Poland
Andrea Zielinski	Fraunhofer IOSB, Germany

Host

Jonas Kuhn	University of Stuttgart, Germany

Sponsoring Institutions

German Society for Computational Linguistics and Language Technology (GSCL)
University of Stuttgart, Germany
Institut für Deutsche Sprache, Mannheim, Germany
CLARIN-D

Contents

Morphology Within the Multi-layered Annotation Scenario of the Prague Dependency Treebank 1
Magda Ševčíková

Designing and Comparing G2P-Type Lemmatizers for a Morphology-Rich Language 27
Steffen Eger

Morphological Disambiguation of Classical Sanskrit 41
Oliver Hellwig

Morphological Analysis and Generation for Pali 60
David Alfter and Jürgen Knauth

A Universal Feature Schema for Rich Morphological Annotation and Fine-Grained Cross-Lingual Part-of-Speech Tagging 72
John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky

Dsolve—Morphological Segmentation for German Using Conditional Random Fields 94
Kay-Michael Würzner and Bryan Jurish

A Multi-purpose Bayesian Model for Word-Based Morphology 104
Maciej Janicki

Using HFST—Helsinki Finite-State Technology for Recognizing Semantic Frames 124
Krister Lindén, Sam Hardwick, Miikka Silfverberg, and Erik Axelson

Developing Morpho-SLaWS: An API for the Morphosyntactic Annotation of the Serbian Language 137
Toma Tasovac, Saša Rudan, and Siniša Rudan

Morphological Analysis and Generation of Monolingual and Bilingual Medical Lexicons 148
Annibale Elia, Alessandro Maisto, and Serena Pelosi

Grammar Debugging 166
Michael Maxwell

Author Index 185