Communications
in Computer and Information Science 100

Cerstin Mahlow Michael Piotrowski (Eds.)

Systems and Frameworks for Computational Morphology

Second International Workshop, SFCM 2011
Zurich, Switzerland, August 26, 2011
Proceedings

Volume Editors

Cerstin Mahlow
University of Basel
Nadelberg 4, 4051 Basel, Switzerland
E-mail: cerstin.mahlow@unibas.ch

Michael Piotrowski
University of Zurich
Binzmühlestr. 14, 8051 Zurich, Switzerland
E-mail: mxp@cl.uzh.ch

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-23137-7 e-ISBN 978-3-642-23138-4
DOI 10.1007/978-3-642-23138-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011933917

CR Subject Classification (1998): I.2.7

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)

Preface

Morphological resources are the basis for all higher-level natural language processing applications. Morphology components should thus be capable of analyzing single word forms 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 either require morphological analysis and generation, or that can greatly benefit from it, for example in text processing, user interfaces, or information retrieval. 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. In 2011 the workshop took place for the second time. SFCM focuses on actual working systems and frameworks that are based on linguistic principles and that provide linguistically motivated analyses and/or generation on the basis of linguistic categories.

SFCM 2009 focused on systems for a specific language, namely, German. The main theme of SFCM 2011 was phenomena at the interface between morphology and syntax in various languages: Many practical applications have to deal with texts, not just isolated word forms. This requires systems to handle phenomena that cannot be easily classified as either "morphologic" or "syntactic."

The workshop thus had three main goals:

- To stimulate discussion among researchers and developers and to offer an up-to-date overview of available morphological systems for specific purposes.
- To stimulate discussion among developers of general frameworks that can be used to implement morphological components for several languages.
- To discuss aspects of evaluation of morphology systems and possible future competitions or tasks.

Based on the number of submissions and the number of participants at the workshop we can definitely state that the topic of the workshop was met with great interest from the community, both from academia and industry. We received 13 submissions, of which 8 were accepted after a thorough review by the members of the Program Committee and additional reviewers. The peer-review process was double-blind, and each paper received four reviews.

In addition to the regular papers, we had the pleasure of Lauri Karttunen giving an exciting invited talk on new features of the Finite-State Toolkit (FST).

The discussions after the talks and during the demo sessions, as well as the final plenum, showed the interest in and the need and the requirements for further efforts in the field of computational morphology. We will maintain the website for this workshop at `http://sfcm2011.org`.

This book starts with the invited paper by Lauri Karttunen ("Beyond Morphology: Pattern Matching with FST"), reporting on new developments for the Finite-State Toolkit, an enhanced version of XFST. The FST pattern matching algorithm allows applications like tokenizing, named-entity recognition, or even parsing.

Then follows a paper by Mārcis Pinnis and Kārlis Goba ("Maximum Entropy Model for Disambiguation of Rich Morphological Tags"), describing a statistical morphological tagger for Latvian, Lithuanian, and Estonian. The authors explore the use of probabilistic models with maximum entropy weight estimation to cover the rich morphology in these languages.

The paper by Benoît Sagot and Géraldine Walther ("Non-canonical Inflection: Data, Formalisation and Complexity Measures") deals with non-canonical inflection, a popular topic in linguistics, but lacking implementation. Representing inflectional irregularities as morphological rules or as additional information in the lexicon allows the implementation within the Alexina framework. The approach holds for several morphologically rich languages like French, Latin, Italian, Sorani Kurdish, Persian, Croatian, and Slovak.

The following paper of Gertraud Faaß ("A User-Oriented Approach to Evaluation and Documentation of a Morphological Analyser") emphasizes the need for user-centered evaluation of morphological components.

The paper by Krister Lindén, Erik Axelson, Sam Hardwick, Tommi Pirinen, and Miikka Silfverberg ("HFST—Framework for Compiling and Applying Morphologies") reports on the new version of the HFST framework, allowing users to experiment with several finite-state tools for various languages to use in open-source projects.

Then follows a paper by Esmé Manandise and Claudia Gdaniec ("Morphology to the Rescue Redux: Resolving Borrowings and Code-Mixing in Machine Translation") covering morphological issues in machine translation of e-mail messages from Spanish to English when bilingual authors use borrowing, code-mixing, or code-switching.

The last three papers report on morphological systems for specific languages: Arabic, Indonesian, and Swiss German. Mohammed Attia, Pavel Pecina, Antonio Toral, Lamia Tounsi, and Josef Van Genabith ("A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer") report on the creation of resources for modern standard Arabic. The paper by Septina Dian Larasati, Daniel Zeman, and Vladislav Kuboň ("Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus") describes the development of a robust finite state open source morphology tool for Indonesian, motivated by shortcomings of existing resources. The paper by Yves Scherrer ("Morphology Generation for Swiss German Dialects") provides insights into dialectological issues for generation. Although there is a lot of research on Swiss German dialects in the field of linguistics, there is currently only very little related research in the NLP community.

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. Even though other languages are becoming more important, research in computational linguistics still focuses primarily on English, which is well known for its reduced morphology. Morphological analysis and generation are thus often regarded as being required only for the processing of some exotic languages. The papers in this book come from eight countries, discuss a wide variety of languages from many different language families, and illustrate that, in fact, a rich morphology is better described as the norm rather than the exception—proving that for most languages, as we have stated above, morphological resources are indeed the basis for all higher-level natural language processing applications.

The trend toward 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 the development of high-quality real-world applications. Useful applications with carefully conducted evaluation will demonstrate to a broad audience that computational morphology is an actual science with tangible benefits for society.

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 were supported by the EasyChair system. We thank Alfred Hofmann, editor of the series *Communications in Computer and Information Science* (CCIS), and the Springer staff for publishing the proceedings of SFCM 2011. We are grateful for the financial support given by the German Society for Computational Linguistics and Language Technology (GSCL) and the general support of the University of Zurich.

June 2011

Cerstin Mahlow
Michael Piotrowski

Organization

The Second Workshop on Systems and Frameworks for Computational Morphology (SFCM 2011) was organized by Cerstin Mahlow and Michael Piotrowski. The workshop was held at the University of Zurich.

Program Chairs

Cerstin Mahlow	University of Basel, Switzerland
Michael Piotrowski	University of Zurich, Switzerland

Program Committee

Bruno Cartoni	University of Geneva, Switzerland
Simon Clematide	University of Zurich, Switzerland
Axel Fleisch	University of Helsinki, Finland
Piotr Fuglewicz	TiP Sp. z o. o., Katowice, Poland
Thomas Hanneforth	University of Potsdam, Germany
Roland Hausser	Friedrich-Alexander University of Erlangen-Nuremberg, Germany
Lauri Karttunen	Stanford University, USA
Kimmo Koskenniemi	University of Helsinki, Finland
Winfried Lenders	University of Bonn, Germany
Krister Lindén	University of Helsinki, Finland
Anke Lüdeling	Humboldt University Berlin, Germany
Cerstin Mahlow	University of Basel, Switzerland
Günter Neumann	DFKI Saarbrücken, Germany
Michael Piotrowski	University of Zurich, Switzerland
Adam Przepiórkowski	Polish Academy of Sciences, Warsaw, Poland
Christoph Rösener	Institute for Applied Information Science, Saarbrücken, Germany
Helmut Schmid	University of Stuttgart, Germany
Angelika Storrer	University of Dortmund, Germany
Pius ten Hacken	Swansea University, UK
Eric Wehrli	University of Geneva, Switzerland
Andrea Zielinski	FIZ Karlsruhe, Germany

Additional Reviewers

Johannes Handl	Friedrich-Alexander University of Erlangen-Nuremberg, Germany
Besim Kabashi	Friedrich-Alexander University of Erlangen-Nuremberg, Germany

Local Organization

Cerstin Mahlow	University of Basel, Switzerland
Michael Piotrowski	University of Zurich, Switzerland

Sponsoring Institutions

German Society for Computational Linguistics and Language Technology (GSCL)
University of Zurich

Table of Contents