Communications
in Computer and Information Science 41

Cerstin Mahlow Michael Piotrowski (Eds.)

State of the Art in Computational Morphology

Workshop on Systems and Frameworks
for Computational Morphology, SFCM 2009
Zurich, Switzerland, September 4, 2009
Proceedings

Springer

Volume Editors

Cerstin Mahlow
University of Zurich
Zurich, Switzerland
E-mail: mahlow@cl.uzh.ch

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

Library of Congress Control Number: Applied for

CR Subject Classification (1998): J.5, H.3.1, F.4.2, F.4.3, I.2.7, H.5.2

ISSN 1865-0929
ISBN-10 3-642-04130-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-04130-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12745241 06/3180 5 4 3 2 1 0

Preface

From the point of view of computational linguistics, morphological resources are the basis for all higher-level applications. This is especially true for languages with a rich morphology, such as German or Finnish. A morphology component 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 also 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.

In 1994, the First Morpholympics took place at the University of Erlangen-Nuremberg, a competition between several systems for the analysis and generation of German word forms. Eight systems participated in the First Morpholympics; the conference proceedings [1] thus give a very good overview of the state of the art in computational morphology for German as of 1994.

Today, 15 years later, some of the systems that participated in the Morpholympics still exist and are being maintained. However, there are also new developments in the field of computational morphology, for German and for other languages. Unfortunately, publications on morphologic analysis and generation are spread over many different conferences and journals, so that it is difficult to get an overview of the current state of the art and of the available systems. One of our goals for the Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009) was therefore to bring together researchers, developers, and maintainers of morphology systems for German and of frameworks for computational morphology from academia and industry, in order to produce an up-to-date overview of available systems for German.

As in many areas of computational linguistics, there are rule-based and statistical approaches to morphological analysis; SFCM 2009 focused on systems and frameworks based on linguistic principles and providing linguistically motivated analyses and/or generation on the basis of linguistic categories.

One factor for this decision was our own experience in implementing rule-based morphological analyzers, so we know that they are able to deliver detailed, structured analyses (see [2]). The possibility to draw upon the morphological processes of inflection, derivation, and compounding involved when analyzing or generating word forms, the respective parse and generation trees, and certain elements of the category, is, in our view, a potential advantage when compared with statistically created results.

The second and deciding factor is that, based on the results of Morpho Challenge[1], we have come to the conclusion that statistical morphological analyzers are not yet able to deliver the quality of results required for practical applications. In the Morpho Challenge morpheme analysis task, the analyses proposed by the participants' algorithms are compared against a linguistic gold standard. At Morpho Challenge 2008 [3], the best system for German achieved an F-measure of 54.06%. The best recall value was 59.51% (this system achieved 49.53% precision), the best result for precision was 87.92% (with 7.44% recall).[2] These figures are much too low to consider the systems as suitable for use in most types of applications, and in interactive applications in particular. If we compare the results of Morpho Challenge with the figures reported from the First Morpholympics [4], the decision to focus on rule-based systems suggests itself.

In the call for papers for this workshop we asked for contributions on actual, working systems and frameworks of at least prototype quality. To ensure fruitful discussions among workshop participants, we asked that submissions on concrete morphology systems should be for German; submissions on morphological frameworks were considered relevant if the framework can be used to implement components for different languages.

The workshop thus had three main goals:

- To stimulate discussion among researchers and developers and to offer an up-to-date overview of available systems for German morphology which provide deep analyses and are suitable for generating specific word forms.
- 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, such as a new edition of the Morpholympics.

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. We received 16 submissions, of which 9 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 at least three reviews.

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://sfcm2009.org`. Here you can find additional material not included in the proceedings. If there is a follow-up to this workshop—whether in a similar format or in the form of a competition—it will also be announced on this site.

[1] Morpho Challenge is a shared task and conference for the evaluation of statistical morphological components based on unsupervised machine-learning.

[2] See `http://www.cis.hut.fi/morphochallenge2008/` for details. The results of Morpho Challenge 2009 were not yet available at the time of this writing.

This book starts with a theory-oriented paper by Thomas Hanneforth ("Using Ranked Semirings for Representing Morphology Automata"), proposing complex weight structures to represent morphological analyzers.

The following three papers report on frameworks: Johannes Handl, Besim Kabashi, Thomas Proisl, and Carsten Weber ("JSLIM – Computational Morphology in the Framework of the SLIM Theory of Language") present a recent implementation of the SLIM theory, based on Left-Associative Grammar [5]. This framework allows the implementation of morphological analyzers and generators for different languages. Krister Lindén, Miikka Silfverberg, and Tommi Pirinen ("HFST Tools for Morphology – An Efficient Open Source Package for Construction of Morphological Analyzers") present a recent implementation of Two-Level Morphology [6]. Both frameworks (JSLIM and HFST) are intended to be distributed as open-source software, and both papers report on actually implemented systems for several languages with a coverage permitting the analysis of real-world texts. This is a very positive trend, which—we hope—will encourage the development of concrete analyzers and generators using the knowledge of the community and coordinating the efforts. Thomas Hanneforth ("fsm2 – A Scripting Language for Creating Weighted Finite-State Morphologies") reports on a scripting language for creating and manipulating morphological analyzers for different languages based on weighted semirings.

The following three papers report on morphological systems for German. Andrea Zielinski, Christian Simon, and Tilman Wittl ("Morphisto – Service-Oriented Open Source Morphology for German") present their recent efforts in developing an open-source analyzer and generator for German. It was initially developed within the TextGrid project—a modular platform for collaborative textual editing for researchers in philology, linguistics, and related fields. Morphisto is based on SMOR and the SFST tools [7], offering a comprehensive free lexicon. Heinz Dieter Maas, Christoph Rösener, and Axel Theofilidis ("Morphosyntactic and Semantic Analysis of Text: The MPRO Tagging Procedure") present further details of a system which already participated in the first Morpholympics in 1994. MPRO is still maintained and being used in several applications. Similarly, the system presented by Pius ten Hacken ("Word Manager") has been maintained for more than 15 years and is being used in various environments.

The last two papers deal with unknown words. The first one, by Stephan Bopp and Sandro Pedrazzini ("Morphological Analysis Using Linguistically Motivated Decomposition of Unknown Words") is a practical application of Word Manager described in the preceding paper. While Bopp and Pedrazzini concentrate on the decomposition of complex German compounds using an existing morphological system, the paper of Krister Lindén and Jussi Tuovila ("Corpus-based Lexeme Ranking for Morphological Guessers") is concerned with adding new words to the lexicon of a morphological component. Their approach is intended to help the lexicographer by providing automatically generated suggestions for lexemes.

In summary, these contributions show that high-quality research is being conducted in the area of rule-based computational morphology, and that there are further developments of mature systems, new implementations based on established theoretical frameworks, and new approaches to the problems of morphology. We also see a trend towards open-source developments, which we find very promising. Open-source projects allow

the collaboration of researchers interested in morphological systems which fulfill high demands on performance and quality. It should not be necessary to develop a morphological analyzer from scratch if one is needed for a project. 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 will demonstrate to a broad audience that computational morphology (and natural language processing in general) is not an esoteric boondoggle, but an actual science with tangible benefits for society, which are good reason for publicly funding research in this area.

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 Stefan Göller, the editor of the series ("Communications in Computer and Information Science") (CCIS), and Springer for publishing the proceedings of SFCM 2009. We are grateful for the financial support given by the Institute of Computational Linguistics at the University of Zurich and by the German Society for Computational Linguistics and Language Technology (GSCL). Last, but not least, we thank Roland Hausser for encouraging us to organize the SFCM 2009 workshop as an almost successor of the First Morpholympics in 1994, and Norbert Fuchs for providing many helpful hints during the entire organization process.

July 2009

Cerstin Mahlow
Michael Piotrowski

References

1. Hausser, R.: Linguistische Verifikation. Dokumentation zur Ersten Morpholympics. Niemeyer, Tübingen (1996)
2. Mahlow, C., Piotrowski, M.: SMM: Detailed, structured morphological analysis for Spanish. Polibits. Computer science and computer engineering with applications (39) (2009)
3. Kurimo, M., Varjokallio, M.: Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard – Morpho Challenge 2008. In: Workshop of the Cross-Language Evaluation Forum (CLEF 2008). (2008)
4. Lenders, W., Bátori, I., Dogil, G., Görz, G., Seewald, U.: Stellungnahme der Jury für die Morpholympics 94. In Hausser, R., ed.: Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994. Niemeyer, Tübingen (1996) 15–24
5. Hausser, R.: Foundations of Computational Linguistics: Human-Computer Communication in Natural Language. 2nd rev. and ext. edn. Springer, Heidelberg (2001)
6. Koskenniemi, K.: Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. PhD thesis, University of Helsinki (1983)
7. Schmid, H., Fitschen, A., Heid, U.: A German computational morphology covering derivation, composition, and inflection. In: IVth International Conference on Language Resources and Evaluation (LREC 2004). (2004) 1263–1266

Organization

The Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009) was organized by Cerstin Mahlow and Michael Piotrowski. The workshop was held at the University of Zurich.

Program Chairs

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

Program Committee

Simon Clematide	University of Zurich, Switzerland
Thomas Hanneforth	University of Potsdam, Germany
Roland Hausser	Friedrich-Alexander University of Erlangen-Nuremberg, Germany
Lauri Karttunen	PARC, Palo Alto, 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 Zurich, Switzerland
Günter Neumann	DFKI Saarbrücken, Germany
Michael Piotrowski	University of Zurich, Switzerland
Helmut Schmid	University of Stuttgart, Germany
Angelika Storrer	University of Dortmund, Germany
Martin Volk	University of Zurich, Switzerland
Shuly Wintner	University of Haifa, Israel
Andrea Zielinski	FIZ Karlsruhe, Germany

Additional Reviewers

Bruno Cartoni	University of Geneva, Switzerland
Johannes Handl	Friedrich-Alexander University of Erlangen-Nuremberg, Germany
Besim Kabashi	Friedrich-Alexander University of Erlangen-Nuremberg, Germany
Thomas Proisl	Friedrich-Alexander University of Erlangen-Nuremberg, Germany

Luzius Thöny	University of Zurich, Switzerland
Carsten Weber	Friedrich-Alexander University of Erlangen-Nuremberg, Germany

Local Organization

Cerstin Mahlow	University of Zurich, Switzerland
Michael Piotrowski	University of Zurich, Switzerland
Nancy Renning	University of Zurich, Switzerland

Sponsoring Institutions

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

Table of Contents