Lecture Notes in Artificial Intelligence 6062

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Anssi Yli-Jyrä András Kornai
Jacques Sakarovitch Bruce Watson (Eds.)

Finite-State Methods and Natural Language Processing

8th International Workshop, FSMNLP 2009
Pretoria, South Africa, July 21-24, 2009
Revised Selected Papers

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Anssi Yli-Jyrä
University of Helsinki, Department of Modern Languages
00014 University of Helsinki, Finland
E-mail: anssi.yli-jyra@helsinki.fi

András Kornai
Harvard University, Institute for Quantitative Social Science
1737 Cambridge St, Cambridge MA 02138, USA
and: Computer and Animation Research Institute, Hungarian Academy
of Sciences, Kende u 13-17, Budapest 1111, Hungary
E-mail: andras@kornai.com

Jacques Sakarovitch
CNRS and Telecom ParisTech
Laboratoire Traitement et Communication de l'Information
46, rue Barrault, 75634 Paris Cedex 13, France
E-mail: sakarovitch@enst.fr

Bruce Watson
University of Pretoria, FASTAR Research Group
Petoria 0002, South Africa
E-mail: bw@bruce-watson.com

Library of Congress Control Number: 2010931643

CR Subject Classification (1998): I.2, H.3, F.4.1, I.2.7, F.3, H.4
LNCS Sublibrary: SL 7 – Artificial Intelligence
ISSN 0302-9743
ISBN-10 3-642-14683-X Springer Berlin Heidelberg New York
ISBN-13 978-3-642-14683-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

This volume of *Lecture Notes in Artificial Intelligence* is a collection of revised versions of the papers and lectures presented at the 8th International Workshop on Finite-State Methods and Natural Language Processing, FSMNLP 2009.

The workshop was held at the University of Pretoria, South Africa, during July 21–24, 2009. This was the first time in the history of the FSMNLP series that the event was not located in Europe.

As its predecessors, the scope of FSMNLP 2009 included a range of topics around computational morphology, natural language processing, finite-state methods, automata, and related formal language theory. However, a special theme, *Finite-State Methods for Under-Resourced Languages*, was adopted as recognition of the event's location on the African continent. The Program Committee was composed of internationally leading researchers and practitioners selected from academia, research labs, and companies.

In total, 21 papers underwent a blind refereeing process in which each paper submitted to the workshop was reviewed by at least three Program Committee members, with the help of external referees. Of those papers, the workshop accepted 13 as regular papers and a further 6 as extended abstracts. The papers came from Croatia, Finland, France, Georgia, Germany, Poland, South Africa, Spain and the United States. In addition, the workshop program contained tutorials by Colin de la Higuera, Kemal Oflazer, and Johan Schalkwyk, inivited talks by Kenneth R. Beesley, Thomas Hanneforth, André Kempe, Jackson Muhirwe, and Johan Schalkwyk, as well as the Zulu competition announcement presented by Colin de la Higuera.

After the FSMNLP 2009 workshop, the papers were re-reviewed for inclusion in this collection, with the assistance of more referees. As the result, four regular papers and four extended abstracts were selected in their final, revised format. The collection also includes the abstracts and extended papers of the competition announcement and of almost all invited lectures and tutorials presented at the workshop.

It is a pleasure for the editors to thank the members of the Program Committee and the external referees for reviewing the papers and maintaining the high standard of the FSMNLP workshops. We are grateful to all the contributors to the conference, in particular to the invited speakers and sponsors, for making FSMNLP 2009 a scientific success despite the challenges in the global economical situation and with the long flight distances. Last, but not least, we wish to express our sincere appreciation to the local organizers for their tireless efforts.

14 April 2010

A. Yli-Jyrä
A. Kornai
J. Sakarovitch
B. Watson

Organization

FSMNLP 2010 was organized by the Department of Computer Science, University of Pretoria.

Conference Chair

Bruce Watson	University of Pretoria, South Africa

Organizing Committee

Loek Cleophas	University of Pretoria, South Africa (OC Chair)
Derrick Kourie	University of Pretoria, South Africa
Jakub Piskorski	Polish Academy of Sciences, Warsaw, Poland
Pierre Rautenbach	University of Pretoria, South Africa
Bruce Watson	University of Pretoria, South Africa
Anssi Yli-Jyrä	Department of General Linguistics, University of Helsinki, Finland

Program Committee Chairs

Andras Kornai	Budapest Institute of Technology, Hungary and MetaCarta, Cambridge, USA
Jacques Sakarovitch	Ecole nationale supérieure des Télécommunications, Paris, France
Anssi Yli-Jyrä	Department of General Linguistics, University of Helsinki, Finland

Program Committee

Cyril Allauzen	Google Research, New York, USA
Sonja Bosch	University of South Africa, South Africa
Francisco Casacuberta	Instituto Tecnologico De Informática, Valencia, Spain
Damir Cavar	University of Zadar, Croatia
Jean-Marc Champarnaud	Université de Rouen, France
Loek Cleophas	University of Pretoria, South Africa
Maxime Crochemore	King's College London, UK
Jan Daciuk	Gdańsk University of Technology, Poland
Frank Drewes	Umea University, Sweden
Dafydd Gibbon	University of Bielefeld, Germany
John Goldsmith	University of Chicago, USA

Karin Haenelt	Fraunhofer Gesellschaft and University of Heidelberg, Germany
Thomas Hanneforth	University of Potsdam, Germany
Colin de la Higuera	Jean Monnet University, Saint-Etienne, France
Johanna Högberg	Umea University, Sweden
Arvi Hurskainen	University of Helsinki, Finland
Lauri Karttunen	Palo Alto Research Center and Stanford University, USA
André Kempe	Cadege Technologies, Paris, France
Kevin Knight	University of Southern California, USA
Derrick Kourie	University of Pretoria, South Africa
Marcus Kracht	University of California, Los Angeles, USA
Hans-Ulrich Krieger	DFKI GmbH, Saarbrücken, Germany
Eric Laporte	Université de Marne-la-Vallée, France
Andreas Maletti	Universitat Rovira i Virgili, Spain
Michael Maxwell	University of Maryland, USA
Stoyan Mihov	Bulgarian Academy of Sciences, Sofia, Bulgaria
Kemal Oflazer	Sabanci University, Turkey
Jakub Piskorski	Polish Academy of Sciences, Warsaw, Poland
Laurette Pretorius	University of South Africa, South Africa
Michael Riley	Google Research, New York, USA
Strahil Ristov	Ruder Boskovic Institute, Zagreb, Croatia
James Rogers	Earlham College, USA
Max Silberztein	Université de Franche-Comté, France
Bruce Watson	University of Pretoria, South Africa
Sheng Yu	University of Western Ontario, Canada
Menno van Zaanen	Tilburg University, The Netherlands
Lynette van Zijl	Stellenbosch University, South Africa

Additional Referees

S. Amsalu	S. Gerdjikov	S. Pissis
F. Barthelemy	H. Liang	P. Prochazka
M. Constant	P. Mitankin	M. Silfverberg
B. Daille	M.-J. Nederhof	N. Smith

Sponsors

FASTAR Research Group - University of Pretoria
University of Pretoria, Faculty of Engineering, Built Environment & IT
Google Research
Microsoft Research
University of South Africa (UNISA)

Table of Contents

Tutorials

Learning Finite State Machines 1
Colin de la Higuera

Special Theme Tutorials

Developing Computational Morphology for Low- and Middle-Density Languages 11
Kemal Oflazer

Invited Papers

fsm2 – A Scripting Language Interpreter for Manipulating Weighted Finite-State Automata 13
Thomas Hanneforth

Selected Operations and Applications of n-Tape Weighted Finite-State Machines 31
André Kempe

OpenFst 47
Johan Schalkwyk

Special Theme Invited Talks

Morphological Analysis of Tone Marked Kinyarwanda Text 48
Jackson Muhirwe

Regular Papers

Minimizing Weighted Tree Grammars Using Simulation 56
Andreas Maletti

Compositions of Top-Down Tree Transducers with ε-Rules 69
Andreas Maletti and Heiko Vogler

Reducing Nondeterministic Finite Automata with SAT Solvers 81
Jaco Geldenhuys, Brink van der Merwe, and Lynette van Zijl

Joining Composition and Trimming of Finite-State Transducers 93
Johannes Bubenzer and Kay-Michael Würzner

Special Theme Extended Abstracts

Porting Basque Morphological Grammars to *foma*, an Open-Source Tool 105
Iñaki Alegria, Izaskun Etxeberria, Mans Hulden, and Montserrat Maritxalar

Describing Georgian Morphology with a Finite-State System 114
Oleg Kapanadze

Finite State Morphology of the Nguni Language Cluster: Modelling and Implementation Issues 123
Laurette Pretorius and Sonja Bosch

A Finite State Approach to Setswana Verb Morphology 131
Laurette Pretorius, Biffie Viljoen, Rigardt Pretorius, and Ansu Berg

Competition Announcements

Zulu: An Interactive Learning Competition 139
David Combe, Colin de la Higuera, and Jean-Christophe Janodet

Author Index 147