

# Lecture Notes in Artificial Intelligence 4201

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Yasubumi Sakakibara Satoshi Kobayashi  
Kengo Sato Tetsuro Nishino  
Etsuji Tomita (Eds.)

# Grammatical Inference: Algorithms and Applications

8th International Colloquium, ICGI 2006  
Tokyo, Japan, September 20-22, 2006  
Proceedings



Springer

## Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA  
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

## Volume Editors

Yasubumi Sakakibara  
Keio University, Yokohama, Japan  
E-mail: yasu@bio.keio.ac.jp

Satoshi Kobayashi  
University of Electro-Communications, Tokyo, Japan  
E-mail: satoshi@cs.uec.ac.jp

Kengo Sato  
National Institute of Advanced Industrial Science and Technology, Tokyo, Japan  
E-mail: sato-kengo@aist.go.jp

Tetsuro Nishino  
University of Electro-Communications, Tokyo, Japan  
E-mail: nishino@ice.uec.ac.jp

Etsuji Tomita  
University of Electro-Communications, Tokyo, Japan  
E-mail: tomita@ice.uec.ac.jp

Library of Congress Control Number: 2006932583

CR Subject Classification (1998): I.2, F.4, F.3

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-540-45264-8 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-45264-5 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 is a part of Springer Science+Business Media  
springer.com

© Springer-Verlag Berlin Heidelberg 2006  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11872436 06/3142 5 4 3 2 1 0

# Preface

The 8th International Colloquium on Grammatical Inference (ICGI 2006) was held at the University of Electro-Communications (UEC), Tokyo, Japan on September 20-22, 2006. ICGI 2006 was the eighth in a series of successful biennial international conferences in the area of grammatical inference. Previous meetings were held in Essex, UK; Alicante, Spain; Montpellier, France; Ames, Iowa, USA; Lisbon, Portugal; Amsterdam, Netherlands; Athens, Greece. ICGI 2006 was the first conference in this series to be held in Asia. This series of conferences seeks to provide a forum for presentation and discussion of original research papers on all aspects of grammatical inference.

Grammatical inference, the study of learning grammars from data, is an established research field in artificial intelligence, dating back to the 1960s and has been extensively addressed by researchers in automata theory, language acquisition, computational linguistics, machine learning, pattern recognition, computational learning theory and neural networks. ICGI 2006 successively emphasized on the multi-disciplinary nature of the research field and the diverse domains in which grammatical inference is being applied, such as natural language acquisition, computational biology, structural pattern recognition, information retrieval, Web mining, text processing, data compression and adaptive intelligent agents.

We received 44 high-quality papers from 14 countries around the world. The papers were reviewed by three reviewers. Based on the positive comments of the reviewers, 25 full papers were accepted. In addition, we decided to accept 8 short papers for poster presentation. Short papers appear as two-page extended abstracts in a separate section of this volume. The topics of the accepted papers vary from theoretical results of learning algorithms to innovative applications of grammatical inference, and from learning several interesting classes of formal grammars to applications to natural language processing.

In parallel to the submission and reviewing of research papers, a machine translation competition, named Tenjinno, took place. In a separate paper in this volume, the organizers of the competition report on the peculiarities of such an endeavor and some interesting theoretical findings to which they have been led. Last but not least, we were honored by the contributions of our two invited speakers, Yuji Matsumoto, from Nara Institute of Science and Technology, Japan, and Jean-Philippe Vert, from Ecole des Mines de Paris, France. Both invited speakers provided interesting talks on the topics of natural language processing and bioinformatics, and we hope both talks invoked potential applications of grammatical inference.

The editors would like to acknowledge the contribution of the conference's Program Committee and the Additional Reviewers in reviewing the submitted papers and thank the Organizing Committee for their invaluable help in

organizing the conference. Particularly, we would like to thank Colin de la Higuera, Menno van Zaannen, Bradford Starkie, and Dominique Estival for their additional voluntary service to the grammatical inference community, through this conference. We would also like to acknowledge the use of the Cyberchair software, from Borbala online conference services, in the submission and reviewing process. Finally, we are grateful for the generous support and sponsorship of the conference by the University of Electro-Communications, the PASCAL, Inoue foundation for Science, SIG Mathematical Modeling and Problem Solving in Information Processing Society of Japan and New Horizons in Computing (NHC) (Scientific Research on Priority Areas, supported by MEXT, Japan).

September 2006

Yasubumi Sakakibara  
Satoshi Kobayashi  
Kengo Sato  
Tetsuro Nishino  
Etsuji Tomita

# Organization

## Conference Chair

Etsuji Tomita	University of Electro-Communications, Japan
---------------	---

## Technical Program Committee Co-chairs

Yasubumi Sakakibara	Keio University, Japan
Satoshi Kobayashi	University of Electro-Communications, Japan

## Technical Program Committee

Naoki Abe	IBM Thomas J. Watson Research Center, USA
Pieter Adriaans	Perot Systems Corporation/University of Amsterdam, Netherlands
Dana Angluin	Yale University, USA
Hiroki Arimura	Hokkaido University, Japan
Mitra Basu	City University of New York, USA
François Coste	Symbiose, INRIA/IRISA, France
Pierre Dupont	University of Louvain, Belgium
Henning Fernau	University of Hertfordshire, UK
Colin de la Higuera	EURISE, Univ. de St. Etienne, France
Vasant Honavar	Iowa State University, USA
Chih-Jen Lin	National Taiwan University, Taiwan
Laurent Miclet	ENSSAT, Lannion, France
Gopalakrishnaswamy Nagaraja	Indian Institute of Technology, India
Katsuhiko Nakamura	Tokyo Denki University, Japan
Jacques Nicolas	IRISA, France
Tim Oates	University of Maryland Baltimore County, USA
Arlindo Oliveira	Lisbon Technical University, Portugal
Jose Oncina Carratala	Universidad de Alicante, Spain
Georgios Paliouras	Inst. of Informatics and Telecommunications, NCSR, Greece
Rajesh Parekh	Yahoo!, USA
Kengo Sato	CBRC, NAIST, Japan
Giora Slutzki	Iowa State University, USA
Bradford Starkie	Starkie Enterprise, Australia
Eiji Takimoto	Tohoku University, Japan
Menno van Zaanen	Universiteit van Amsterdam, Netherlands
Enrique Vidal	Universidad Politecnica de Valencia, Spain
Osamu Watanabe	Tokyo Institute of Technology, Japan
Thomas Zeugmann	Hokkaido University, Japan

## Additional Reviewers

T. Armstrong	J.-C. Janodet	J. Poland
L. Becerra-Bonache	H.-U. Krieger	J. M. Vilar
M. Bugalho	J. A. Laxminarayana	
D. Eisenstat	A. Martins	

## Organizing Committee Chair

Tetsuro Nishino	University of Electro-Communications, Japan
-----------------	---

## Organizing Committee

Colin de la Higuera	EURISE, Univ. de St. Etienne, France
Kazuhiro Hotta	University of Electro-Communications, Japan
Satoshi Kobayashi	University of Electro-Communications, Japan
Yoichi Motomura	National Institute of Advanced Industrial Science and Technology, Japan
Katsuhiko Nakamura	Tokyo Denki University, Japan
Seiya Okubo	University of Electro-Communications, Japan
Yasuhiro Tajima	Tokyo University of Agriculture and Technology, Japan
Haruhisa Takahashi	University of Electro-Communications, Japan
Jun Tarui	University of Electro-Communications, Japan
Mitsuo Wakatsuki	University of Electro-Communications, Japan

## Sponsoring Institutions

University of Electro-Communications



PASCAL Network

Inoue Foundation for Science

SIG Mathematical Modeling and Problem Solving in Information Processing Society of Japan

New Horizons in Computing (NHC) (Scientific Research on Priority Areas, supported by MEXT, Japan)

# Table of Contents

## Invited Papers

Parsing Without Grammar Rules .....	1
<i>Yuji Matsumoto</i>	
Classification of Biological Sequences with Kernel Methods .....	7
<i>Jean-Philippe Vert</i>	

## Regular Papers

Identification in the Limit of Systematic-Noisy Languages .....	19
<i>Frédéric Tantini, Colin de la Higuera, Jean-Christophe Janodet</i>	
Ten Open Problems in Grammatical Inference .....	32
<i>Colin de la Higuera</i>	
Polynomial-Time Identification of an Extension of Very Simple Grammars from Positive Data .....	45
<i>Ryo Yoshinaka</i>	
PAC-Learning Unambiguous NTS Languages .....	59
<i>Alexander Clark</i>	
Incremental Learning of Context Free Grammars by Bridging Rule Generation and Search for Semi-optimum Rule Sets .....	72
<i>Katsuhiko Nakamura</i>	
Variational Bayesian Grammar Induction for Natural Language .....	84
<i>Kenichi Kurihara, Taisuke Sato</i>	
Stochastic Analysis of Lexical and Semantic Enhanced Structural Language Model .....	97
<i>Shaojun Wang, Shaomin Wang, Li Cheng, Russell Greiner, Dale Schuurmans</i>	
Using Pseudo-stochastic Rational Languages in Probabilistic Grammatical Inference .....	112
<i>Amaury Habrard, François Denis, Yann Esposito</i>	



Learning Analysis by Reduction from Positive Data . . . . .	125
<i>František Mráz, Friedrich Otto, Martin Plátek</i>	
Inferring Grammars for Mildly Context Sensitive Languages in Polynomial-Time . . . . .	137
<i>Tim Oates, Tom Armstrong, Leonor Becerra Bonache, Mike Atamas</i>	
Planar Languages and Learnability . . . . .	148
<i>Alexander Clark, Christophe Costa Florêncio, Chris Watkins, Murielle Serayet</i>	
A Unified Algorithm for Extending Classes of Languages Identifiable in the Limit from Positive Data . . . . .	161
<i>Mitsuo Wakatsuki, Etsuji Tomita, Go Yamada</i>	
Protein Motif Prediction by Grammatical Inference . . . . .	175
<i>Piedachu Peris, Damián López, Marcelino Campos, José M. Sempere</i>	
Grammatical Inference in Practice: A Case Study in the Biomedical Domain . . . . .	188
<i>Sophia Katrenko, Pieter Adriaans</i>	
Inferring Grammar Rules of Programming Language Dialects . . . . .	201
<i>Alpana Dubey, Pankaj Jalote, Sanjeev Kumar Aggarwal</i>	
The Tenjinno Machine Translation Competition . . . . .	214
<i>Bradford Starkie, Menno van Zaanen, Dominique Estival</i>	
Large Scale Inference of Deterministic Transductions: Tenjinno Problem 1 . . . . .	227
<i>Alexander Clark</i>	
A Discriminative Model of Stochastic Edit Distance in the Form of a Conditional Transducer . . . . .	240
<i>Marc Bernard, Jean-Christophe Janodet, Marc Sebban</i>	
Learning n-Ary Node Selecting Tree Transducers from Completely Annotated Examples . . . . .	253
<i>Aurelien Lemay, Joachim Niehren, Remi Gilleron</i>	
Learning Multiplicity Tree Automata . . . . .	268
<i>Amaury Habrard, Jose Oncina</i>	

Learning DFA from Correction and Equivalence Queries . . . . .	281
<i>Leonor Becerra-Bonache, Adrian Horia Dediu, Cristina Tîrnăuică</i>	
Using MDL for Grammar Induction . . . . .	293
<i>Pieter Adriaans, Ciel Jacobs</i>	
Characteristic Sets for Inferring the Unions of the Tree Pattern Languages by the Most Fitting Hypotheses . . . . .	307
<i>Yen Kaow Ng, Takeshi Shinohara</i>	
Learning Deterministic DEC Grammars Is Learning Rational Numbers . . . . .	320
<i>Pieter Adriaans</i>	
Iso-array Acceptors and Learning . . . . .	327
<i>T. Kalyani, V.R. Dare, D.G. Thomas, T. Robinson</i>	
<b>Poster Papers</b>	
A Merging States Algorithm for Inference of <i>RFSAs</i> . . . . .	340
<i>Gloria Alvarez, Pedro García, José Ruiz</i>	
Query-Based Learning of XPath Expressions . . . . .	342
<i>Julien Carme, Michal Ceresna, Max Goebel</i>	
Learning Finite-State Machines from Inexperienced Teachers . . . . .	344
<i>Olga Grinchtein, Martin Leucker</i>	
Suprasymbolic Grammar Induction by Recurrent Self-Organizing Maps . . . . .	346
<i>Fuminori Mizushima, Takashi Toyoshima</i>	
Graph-Based Structural Data Mining in Cognitive Pattern Interpretation . . . . .	349
<i>Lidia Ogiela, Ryszard Tadeusiewicz, Marek R. Ogiela</i>	
Constructing Song Syntax by Automata Induction . . . . .	351
<i>Kazutoshi Sasahara, Yasuki Kakishita, Tetsuro Nishino, Miki Takahasi, Kazuo Okanoya</i>	
Learning Reversible Languages with Terminal Distinguishability . . . . .	354
<i>José M. Sempere</i>	

Grammatical Inference for Syntax-Based Statistical Machine  
Translation..... 356  
    *Menno van Zaanen, Jeroen Geertzen*

**Author Index**..... 359