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

Volume Editors
Georgios Paliouras
National Center of Scientific Research (NCSR), "Demokritos"
Institute of Informatics and Telecommunications
P.O.Box 60228, Ag. Paraskevi, Attiki, 15310, Greece
E-mail: paliourg@iit.demokritos.gr

Yasubumi Sakakibara
Keio University
Dept. of Biosciences and Informatics
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan
E-mail: yasu@bio.keio.ac.jp

Library of Congress Control Number: 2004112896

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

ISSN 0302-9743
ISBN 3-540-23410-1 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 is a part of Springer Science+Business Media
springeronline.com
© Springer-Verlag Berlin Heidelberg 2004
Printed in Germany
Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11332824 06/3142 5 4 3 2 1 0
Preface

The 7th International Colloquium on Grammatical Inference (ICGI 2004) was held in the National Centre for Scientific Research “Demokritos”, Athens, Greece on October 11–13, 2004. ICGI 2004 was the seventh 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; and Amsterdam, The Netherlands. This series of conferences seeks to provide a forum for the 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 2004 emphasized the multidisciplinary 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 45 high-quality papers from 19 countries. The papers were reviewed by at least two – in most cases three – reviewers. In addition to the 20 full papers, 8 short papers that received positive comments from the reviewers were accepted, and they appear 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 estimations of probabilistic grammars.

In conjunction with ICGI 2004, a context-free grammar learning competition, named Omphalos, took place. In an invited paper in this volume, the organizers of the competition report on the peculiarities of such an endeavor and some interesting theoretical findings. Last but not least, we are honored by the contributions of our invited speakers Prof. Dana Angluin, from Yale University, USA, and Prof. Enrique Vidal, from Universidade Politecnica de Valencia, Spain.

The editors would like to acknowledge the contribution of the 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, Georgios Petasis, Georgios Sigletos and Evangelia Alexopoulou 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 NCSR “Demokritos”, the PASCAL and KD-net European Networks of Excellence, and Biovista: Corporate Intelligence in Biotechnology.

October 2004
Georgios Paliouras and Yasubumi Sakakibara
Technical Program Committee

Pieter Adriaans
Perot Systems Corporation,
and University of Amsterdam,
The Netherlands

Dana Angluin
Yale University, USA

Walter Daelemans
University of Antwerp, The Netherlands

Pierre Dupont
Université Catholique de Louvain, Belgium

Dominique Estival
Defence Science and Technology Organisation
(DSTO), Australia

Colin de la Higuera
EURISE, Univ. de St. Etienne, France

Vasant Honavar
Iowa State University, USA

Makoto Kanazawa
University of Tokyo, Japan

Laurent Miclet
ENSSAT, Lannion, France

Gopalakrishnaswamy Nagaraja
Indian Institute of Technology, Bombay, India

Jacques Nicolas
IRISA, France

Tim Oates
University of Maryland Baltimore County,
USA

Arlindo Oliveira
Lisbon Technical University, Portugal

Jose Oncina Carratala
Universidade de Alicante, Spain

Rajesh Parekh
Blue Martini, USA

Lenny Pitt
University of Illinois at Urbana-Champaign,
USA

Arun Sharma
University of New South Wales, Australia

Giora Slutzki
Iowa State University, USA

Etsuji Tomita
University of Electro-Communications, Japan

Esko Ukkonen
University of Helsinki, Finland

Menno van Zaalen
Tilburg University, The Netherlands

Enrique Vidal
Universidade Politécnica de Valencia, Spain

Takashi Yokomori
Waseda University, Japan

Thomas Zeugmann
University at Lübeck, Germany

Additional Reviewers

Cláudia Antunes
Christine Largeron

O. Boeffard
J.A. Laxminarayana

Miguel Bugalho
Thierry Murgue

Rafael Carrasco
Katsuhiko Nakamura

François Coste
Takeshi Shinohara

Daniel Fredouille
Yasuhiro Tajima

Akio Fujiyoshi
Franck Thollard

Rebecca Hwa
Mitsu Wakatsuki

François Jacquenet
Satoshi Kobayashi
Organizing Committee

Georgios Paliouras, NCSR “Demokritos”, Greece
Colin de la Higuera, Univ. de St. Etienne, France
Georgios Petasis, NCSR “Demokritos”, Greece
Georgios Sigletos, NCSR “Demokritos”, Greece

Sponsoring Organizations

National Center for Scientific Research
(NCSR) “Demokritos”

PASCAL Network of Excellence

KDnet Network of Excellence

Biovista: Corporate Intelligence in Biotechnology
Table of Contents

Invited Papers

Learning and Mathematics .............................................. 1
  Dana Angluin

Learning Finite-State Models for Machine Translation ............... 3
  Enrique Vidal and Francisco Casacuberta

The Omphalos Context-Free Grammar Learning Competition .......... 16
  Bradford Starkie, François Coste, and Menno van Zaanen

Regular Papers

Mutually Compatible and Incompatible Merges
for the Search of the Smallest Consistent DFA ........................ 28
  John Abela, François Coste, and Sandro Spina

Faster Gradient Descent Training of Hidden Markov Models,
Using Individual Learning Rate Adaptation ........................... 40
  Pantelis G. Bagos, Theodore D. Liakopoulos, and Stavros J. Hamodrakas

Learning Mild Context-Sensitivity:
Toward Understanding Children’s Language Learning ................ 53
  Leonor Becerra-Bonache and Takashi Yokomori

Learnability of Pregroup Grammars .................................... 65
  Denis Béchet, Annie Foret, and Isabelle Tellier

A Markovian Approach to the Induction of Regular String Distributions .. 77
  Jérôme Callut and Pierre Dupont

Learning Node Selecting Tree Transducer
from Completely Annotated Examples .................................... 91
  Julien Carme, Aurélien Lemay, and Joachim Niehren

Identifying Clusters from Positive Data ................................ 103
  John Case, Sanjay Jain, Eric Martin, Arun Sharma, and Frank Stephan

Introducing Domain and Typing Bias in Automata Inference .......... 115
  François Coste, Daniel Fredouille, Christopher Kermorvant,
  and Colin de la Higuera

Analogical Equations in Sequences: Definition and Resolution ........ 127
  Arnaud Delhay and Laurent Miclet
Representing Languages by Learnable Rewriting Systems ................. 139  
Rémi Eyraud, Colin de la Higuera, and Jean-Christophe Janodet

A Divide-and-Conquer Approach to Acquire Syntactic Categories .......... 151  
Pablo Gamallo, Gabriel P. Lopes, and Joaquim F. Da Silva

Grammatical Inference Using Suffix Trees ..................................... 163  
Jeroen Geertzen and Menno van Zaanen

Learning Stochastic Finite Automata ............................................. 175  
Colin de la Higuera and Jose Oncina

Navigation Pattern Discovery Using Grammatical Inference ............... 187  
Nikolaos Karampatziakis, Georgios Paliouras, Dimitrios Pierrakos,  
and Panagiotis Stamatopoulos

A Corpus-Driven Context-Free Approximation  
of Head-Driven Phrase Structure Grammar ..................................... 199  
Hans-Ulrich Krieger

Partial Learning Using Link Grammars Data ..................................... 211  
Erwan Moreau

deg-GRIDS: Context-Free Grammatical Inference  
from Positive Examples Using Genetic Search .................................. 223  
Georgios Petasis, Georgios Paliouras, Constantine D. Spyropoulos,  
and Constantine Halatsis

The Boisdale Algorithm – An Induction Method  
for a Subclass of Unification Grammar from Positive Data ............... 235  
Bradford Starkie and Henning Fernau

Learning Stochastic Deterministic Regular Languages  
................................................. 248  
Franck Thollard and Alexander Clark

Polynomial Time Identification of Strict Deterministic Restricted  
One-Counter Automata in Some Class from Positive Data ............... 260  
Mitsuo Wakatsuki, Kiyoshi Teraguchi, and Etsuji Tomita

Poster Papers

Learning Syntax from Function Words ......................................... 273  
David Brooks and Mark Lee

Running FCRPNI in Efficient Time for Piecewise  
and Right Piecewise Testable Languages ..................................... 275  
Antonio Cano, José Ruiz, and Pedro García

Extracting Minimum Length Document Type Definitions Is NP-Hard ...... 277  
Henning Fernau
Table of Contents

Learning Distinguishable Linear Grammars from Positive Data ................. 279
J.A. Laxminarayana, José M. Sempere, and G. Nagaraja

Extending Incremental Learning of Context Free Grammars in Synapse ... 281
Katsuhiko Nakamura

Identifying Left-Right Deterministic Linear Languages ....................... 283
Jorge Calera-Rubio and Jose Oncina

Efficient Learning of $k$-Reversible Context-Free Grammars from Positive Structural Examples .................................. 285
Shinnosuke Seki and Satoshi Kobayashi

An Analysis of Examples and a Search Space for PAC Learning of Simple Deterministic Languages with Membership Queries .......... 288
Yasuhiro Tajima, Yoshiyuki Kotani, and Matsuaki Terada

Author Index ............................................................ 291