

# Lecture Notes in Artificial Intelligence 1891

Subseries of Lecture Notes in Computer Science

Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science

Edited by G. Goos, J. Hartmanis and J. van Leeuwen

**Springer**

*Berlin*

*Heidelberg*

*New York*

*Barcelona*

*Hong Kong*

*London*

*Milan*

*Paris*

*Singapore*

*Tokyo*

Arlindo L. Oliveira (Ed.)

# Grammatical Inference: Algorithms and Applications

5th International Colloquium, ICGI 2000  
Lisbon, Portugal, September 11-13, 2000  
Proceedings



Springer

Series Editors

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

Volume Editor

Arlindo L. Oliveira  
INESC / IST  
R. Alves Redol 9  
1000 Lisbon, Portugal  
E-mail: aml@inesc.pt

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Grammatical inference: algorithms and applications : 5th international colloquium ; proceedings / ICGI 2000, Lisbon, Portugal, September 11 - 13, 2000. Arlindo L. Oliveira (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; HongKong ; London ; Milan ; Paris ; Singapore ; Tokyo : Springer, 2000

(Lecture notes in computer science ; Vol. 1891 : Lecture notes in artificial intelligence)  
ISBN 3-540-41011-2

CR Subject Classification (1998): I.2, F.4.2-3, I.5.1, I.5.4, J.5, F.2  
ISBN 3-540-41011-2 Springer-Verlag 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-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York  
a member of BertelsmannSpringer Science+Business Media GmbH  
© Springer-Verlag Berlin Heidelberg 2000  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Stefan Sossna  
Printed on acid-free paper      SPIN 10722523      06/3142      5 4 3 2 1 0

# Preface

The Fifth International Colloquium on Grammatical Inference (ICGI-2000) was held in Lisbon on September 11–13th, 2000. ICGI-2000 was the fifth in a series of successful biennial international conferences in the area of grammatical inference. Previous conferences were held in Essex, U.K.; Alicante, Spain; Montpellier, France; and Ames, Iowa, USA.

This series of meetings seeks to provide a forum for the presentation and discussion of original research on all aspects of grammatical inference. Grammatical inference, the process of inferring grammar from given data, is a field that is not only challenging from a purely scientific standpoint but also finds many applications in real world problems.

Despite the fact that grammatical inference addresses problems in a relatively narrow area, it uses techniques from many domains, and intersects a number of different disciplines. Researchers in grammatical inference come from fields as diverse as machine learning, theoretical computer science, computational linguistics, pattern recognition and artificial neural networks.

From a practical standpoint, applications in areas such as natural language acquisition, computational biology, structural pattern recognition, information retrieval, text processing and adaptive intelligent agents have either been demonstrated or proposed in the literature.

ICGI-2000 was held jointly with CoNLL-2000, the Computational Natural Language Learning Workshop and LLL-2000, the Second Learning Language in Logic Workshop. The technical program included the presentation of 24 accepted papers (out of 35 submitted) as well as joint sessions with CoNLL and LLL. A tutorial program organized by Gabriel Pereira Lopes took place after the meetings and included tutorials by Raymond Mooney, Gregory Grefenstette, Walter Daelemans, António Ribeiro, Joaquim Ferreira da Silva, Gael Dias, Nuno Marques, Vitor Rossio, João Balsa and Alexandre Agostini. The joint realization of these events represents a unique opportunity for researchers in these related fields to interact and exchange ideas.

I would like to thank Claire Nédellec, Claire Cardie, Walter Daelemans, Colin de la Higuera and Vasant Honavar for their help in several aspects of the organization; the members of the technical program committee and the reviewers for their careful evaluation of the submissions; the members of the local organizing committee, Ana Teresa Freitas and Ana Fred, for their help in setting up the event; and Ana de Jesus for her invaluable secretarial support.

September 2000

Arlindo Oliveira  
Technical Program Chair

## Technical Program Committee

Pieter Adriaans	Syllogic/University of Amsterdam, The Netherlands
Michael Brent	Johns Hopkins University, USA
Walter Daelemans	Tilburg University, The Netherlands
Pierre Dupont	University de St. Etienne, France
Dominique Estival	Syrinx Speech Systems, Australia
Ana Fred	Lisbon Technical University, Portugal
Jerry Feldman	ICSI, Berkeley, USA
Lee Giles	NEC Research Institute, USA
Colin de la Higuera	EURISE, University de St. Etienne, France
Vasant Honavar	Iowa State University, USA
Laurent Miclet	ENSSAT, France
G. Nagaraja	Indian Institute of Technology, India
Jacques Nicolas	IRISA, France
Arlindo Oliveira	INESC/IST, Portugal
Jose Oncina Carratala	Universidade de Alicante, Spain
Rajesh Parekh	Allstate Research and Planning Center, USA
Lenny Pitt	University of Illinois at Urbana-Champaign, USA
Yasubumi Sakakibara	Tokyo Denki University, Japan
Arun Sharma	University of New South Wales, Australia
Giora Slutzki	Iowa State University, USA
Esko Ukkonen	University of Helsinki, Finland
Stefan Wermter	University of Sunderland, UK
Enrique Vidal	University Politecnica de Valencia, Spain
Thomas Zeugmann	Kyushu University, Japan

## Organizing Committee

Conference Chair:	Arlindo Oliveira, INESC/IST
Tutorials:	Gabriel Pereira Lopes, Universidade Nova de Lisboa
Local Arrangements:	Ana Fred, Lisbon Technical University
Social Program:	Ana Teresa Freitas, INESC/IST
Secretariat:	Ana de Jesus, INESC

## Additional Reviewers

Daniel Gildea  
Mitch Harris  
Satoshi Kobayashi  
Eric Martin  
Franck Thollard  
Takashi Yokomori

# Table of Contents

Inference of Finite-State Transducers by Using Regular Grammars and Morphisms .....	1
<i>Francisco Casacuberta</i>	
Computational Complexity of Problems on Probabilistic Grammars and Transducers .....	15
<i>Francisco Casacuberta and Colin de la Higuera</i>	
Efficient Ambiguity Detection in C-NFA, a Step Towards the Inference of Non Deterministic Automata .....	25
<i>François Coste and Daniel Fredouille</i>	
Learning Regular Languages Using Non Deterministic Finite Automata ...	39
<i>François Denis, Aurélien Lemay, and Alain Terlutte</i>	
Smoothing Probabilistic Automata: An Error-Correcting Approach .....	51
<i>Pierre Dupont and Juan-Carlos Amengual</i>	
Inferring Subclasses of Contextual Languages .....	65
<i>J.D. Emerald, K.G. Subramanian, and D.G. Thomas</i>	
Permutations and Control Sets for Learning Non-regular Language Families	75
<i>Henning Fernau and José M. Sempere</i>	
On the Complexity of Consistent Identification of Some Classes of Structure Languages .....	89
<i>Christophe Costa Florêncio</i>	
Computation of Substring Probabilities in Stochastic Grammars .....	103
<i>Ana L. N. Fred</i>	
A Comparative Study of Two Algorithms for Automata Identification ...	115
<i>P. García, A. Cano, and J. Ruiz</i>	
The Induction of Temporal Grammatical Rules from Multivariate Time Series .....	127
<i>Gabriela Guimarães</i>	
Identification in the Limit with Probability One of Stochastic Deterministic Finite Automata .....	141
<i>Colin de la Higuera and Franck Thollard</i>	
Iterated Transductions and Efficient Learning from Positive Data: A Unifying View .....	157
<i>Satoshi Kobayashi</i>	

An Inverse Limit of Context-Free Grammars - A New Approach to Identifiability in the Limit . . . . .	171
<i>Pavel, Martinek</i>	
Synthesizing Context Free Grammars from Sample Strings Based on Inductive CYK Algorithm . . . . .	186
<i>Katsuhiko Nakamura and Takashi Ishiwata</i>	
Combination of Estimation Algorithms and Grammatical Inference Techniques to Learn Stochastic Context-Free Grammars . . . . .	196
<i>Francisco Nevado, Joan-Andreu Sánchez, and José-Miguel Benedí</i>	
On the Relationship between Models for Learning in Helpful Environments	207
<i>Rajesh Parekh and Vasant Honavar</i>	
Probabilistic $k$ -Testable Tree Languages . . . . .	221
<i>Juan Ramón Rico-Juan, Jorge Calera-Rubio, and Rafael C. Carrasco</i>	
Learning Context-Free Grammars from Partially Structured Examples . . . .	229
<i>Yasubumi Sakakibara and Hidenori Muramatsu</i>	
Identification of Tree Translation Rules from Examples . . . . .	241
<i>Hiroshi Sakamoto, Hiroki Arimura, and Setsuo Arikawa</i>	
Counting Extensional Differences in BC-Learning . . . . .	256
<i>Frank Stephan and Sebastiaan A. Terwijn</i>	
Constructive Learning of Context-Free Languages with a Subpansive Tree .	270
<i>Noriko Sugimoto, Takashi Toyoshima, Shinichi Shimozone, and Kouichi Hirata</i>	
A Polynomial Time Learning Algorithm of Simple Deterministic Languages via Membership Queries and a Representative Sample . . . . .	284
<i>Yasuhiro Tajima and Etsuji Tomita</i>	
Improve the Learning of Subsequential Transducers by Using Alignments and Dictionaries . . . . .	298
<i>Juan Miguel Vilar</i>	
<b>Author Index . . . . .</b>	<b>313</b>