Georgios Paliouras
Yasubumi Sakakibara (Eds.)

LNAI 3264

# Grammatical Inference: Algorithms and Applications

7th International Colloquium, ICGI 2004
Athens, Greece, October 2004
Proceedings

Springer

Georgios Paliouras Yasubumi Sakakibara (Eds.)

# Grammatical Inference: Algorithms and Applications

7th International Colloquium, ICGI 2004
Athens, Greece, October 11-13, 2004
Proceedings

Springer

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

Lecture Notes in Artificial Intelligence 3264

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

# Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3265: R.E. Frederking, K.B. Taylor (Eds.), Machine Translation: From Real Users to Research. XI, 392 pages. 2004.

Vol. 3264: G. Paliouras, Y. Sakakibara (Eds.), Grammatical Inference: Algorithms and Applications. XI, 291 pages. 2004.

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004.

Vol. 3245: E. Suzuki, S. Arikawa (Eds.), Discovery Science. XIV, 430 pages. 2004.

Vol. 3244: S. Ben-David, J. Case, A. Maruoka (Eds.), Algorithmic Learning Theory. XIV, 505 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004.

Vol. 3215: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVII, 906 pages. 2004.

Vol. 3214: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1302 pages. 2004.

Vol. 3213: M.G.. Negoita, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1280 pages. 2004.

Vol. 3209: B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, G. Stumme (Eds.), Web Mining: From Web to Semantic Web. IX, 201 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004.

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004.

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004.

Vol. 3171: A.L.C. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence – SBIA 2004. XVII, 548 pages. 2004.

Vol. 3159: U. Visser, Intelligent Information Integration for the Semantic Web. XIV, 150 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004.

Vol. 3071: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World. XIII, 409 pages. 2004.

Vol. 3070: L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Artificial Intelligence and Soft Computing - ICAISC 2004. XXV, 1208 pages. 2004.

Vol. 3068: E. André, L. Dybkjær, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems. XII, 324 pages. 2004.

Vol. 3067: M. Dastani, J. Dix, A. El Fallah-Seghrouchni (Eds.), Programming Multi-Agent Systems. X, 221 pages. 2004.

Vol. 3066: S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymała-Busse (Eds.), Rough Sets and Current Trends in Computing. XX, 853 pages. 2004.

Vol. 3065: A. Lomuscio, D. Nute (Eds.), Deontic Logic in Computer Science. X, 275 pages. 2004.

Vol. 3060: A.Y. Tawfik, S.D. Goodwin (Eds.), Advances in Artificial Intelligence. XIII, 582 pages. 2004.

Vol. 3056: H. Dai, R. Srikant, C. Zhang (Eds.), Advances in Knowledge Discovery and Data Mining. XIX, 713 pages. 2004.

Vol. 3055: H. Christiansen, M.-S. Hacid, T. Andreasen, H.L. Larsen (Eds.), Flexible Query Answering Systems. X, 500 pages. 2004.

此为试读,需要完整PDF请访问: www.ertongbook.com

Vol. 3040: R. Conejo, M. Urretavizcaya, J.-L. Pérez-de-la-Cruz (Eds.), Current Topics in Artificial Intelligence. XIV, 689 pages. 2004.

Vol. 3035: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XII, 326 pages. 2004.

Vol. 3034: J. Favela, E. Menasalvas, E. Chávez (Eds.), Advances in Web Intelligence. XIII, 227 pages. 2004.

Vol. 3030: P. Giorgini, B. Henderson-Sellers, M. Winikoff (Eds.), Agent-Oriented Information Systems. XIV, 207 pages. 2004.

Vol. 3029: B. Orchard, C. Yang, M. Ali (Eds.), Innovations in Applied Artificial Intelligence. XXI, 1272 pages. 2004.

Vol. 3025: G.A. Vouros, T. Panayiotopoulos (Eds.), Methods and Applications of Artificial Intelligence. XV, 546 pages. 2004.

Vol. 3020: D. Polani, B. Browning, A. Bonarini, K. Yoshida (Eds.), RoboCup 2003: Robot Soccer World Cup VII. XVI, 767 pages. 2004.

Vol. 3012: K. Kurumatani, S.-H. Chen, A. Ohuchi (Eds.), Multi-Agents for Mass User Support. X, 217 pages. 2004.

Vol. 3010: K.R. Apt, F. Fages, F. Rossi, P. Szeredi, J. Váncza (Eds.), Recent Advances in Constraints. VIII, 285 pages. 2004.

Vol. 2990: J. Leite, A. Omicini, L. Sterling, P. Torroni (Eds.), Declarative Agent Languages and Technologies. XII, 281 pages. 2004.

Vol. 2980: A. Blackwell, K. Marriott, A. Shimojima (Eds.), Diagrammatic Representation and Inference. XV, 448 pages. 2004.

Vol. 2977: G. Di Marzo Serugendo, A. Karageorgos, O.F. Rana, F. Zambonelli (Eds.), Engineering Self-Organising Systems. X, 299 pages. 2004.

Vol. 2972: R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, H. Sossa (Eds.), MICAI 2004: Advances in Artificial Intelligence. XVII, 923 pages. 2004.

Vol. 2969: M. Nickles, M. Rovatsos, G. Weiss (Eds.), Agents and Computational Autonomy. X, 275 pages. 2004.

Vol. 2961: P. Eklund (Ed.), Concept Lattices. IX, 411 pages. 2004.

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. X, 301 pages. 2004.

Vol. 2930: F. Winkler (Ed.), Automated Deduction in Geometry. VII, 231 pages. 2004.

Vol. 2926: L. van Elst, V. Dignum, A. Abecker (Eds.), Agent-Mediated Knowledge Management. XI, 428 pages. 2004.

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004.

Vol. 2915: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction. XIII, 558 pages. 2004.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. XX, 512 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. XVI, 1075 pages. 2003.

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. XV, 504 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003.

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. X, 215 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003.

Vol. 2871: N. Zhong, Z.W. Raś, S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems. XV, 697 pages. 2003.

Vol. 2854: J. Hoffmann, Utilizing Problem Structure in Planing. XIII, 251 pages. 2003.

Vol. 2843: G. Grieser, Y. Tanaka, A. Yamamoto (Eds.), Discovery Science. XII, 504 pages. 2003.

Vol. 2842: R. Gavaldá, K.P. Jantke, E. Takimoto (Eds.), Algorithmic Learning Theory. XI, 313 pages. 2003.

Vol. 2838: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Knowledge Discovery in Databases: PKDD 2003. XVI, 508 pages. 2003.

Vol. 2837: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Machine Learning: ECML 2003. XVI, 504 pages. 2003.

Vol. 2835: T. Horváth, A. Yamamoto (Eds.), Inductive Logic Programming. X, 401 pages. 2003.

Vol. 2821: A. Günter, R. Kruse, B. Neumann (Eds.), KI 2003: Advances in Artificial Intelligence. XII, 662 pages. 2003.

Vol. 2807: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. XIII, 426 pages. 2003.

Vol. 2801: W. Banzhaf, J. Ziegler, T. Christaller, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life. XVI, 905 pages. 2003.

Vol. 2797: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. XII, 281 pages. 2003.

Vol. 2792: T. Rist, R.S. Aylett, D. Ballin, J. Rickel (Eds.), Intelligent Virtual Agents. XV, 364 pages. 2003.

Vol. 2782: M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (Eds.), Cooperative Information Agents VII. XI, 345 pages. 2003.

Vol. 2780: M. Dojat, E. Keravnou, P. Barahona (Eds.), Artificial Intelligence in Medicine. XIII, 388 pages. 2003.

Vol. 2777: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. XIV, 746 pages. 2003.

Vol. 2752: G.A. Kaminka, P.U. Lima, R. Rojas (Eds.), RoboCup 2002: Robot Soccer World Cup VI. XVI, 498 pages. 2003.

Vol. 2741: F. Baader (Ed.), Automated Deduction – CADE-19. XII, 503 pages. 2003.

Vol. 2705: S. Renals, G. Grefenstette (Eds.), Text- and Speech-Triggered Information Access. VII, 197 pages. 2003.

Vol. 2703: O.R. Zaïane, J. Srivastava, M. Spiliopoulou, B. Masand (Eds.), WEBKDD 2002 - MiningWeb Data for Discovering Usage Patterns and Profiles. IX, 181 pages. 2003.

# 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 Zaanen | 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 | Mitsuo 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

PASCAL Network of Excellence


KDnet Network of Excellence


Biovista: Corporate Intelligence in Biotechnology

# Table of Contents

## Invited Papers



## Regular Papers

## Poster Papers



此为试读,需要完整PDF请访问: www.ertongbook.com

# Learning and Mathematics

Dana Angluin

Yale University
P.O. Box 208285, New Haven, CT 06520-8285, USA
angluin@cs.yale.edu
http://www.cs.yale.edu/people/faculty/angluin.html

Our formal models of learning seem to overestimate how hard it is to learn some kinds of things, including grammars. One possible reason for this is that our models generally do not represent learning a concept as an incremental addition to a rich collection of related concepts. This raises the question of how to make a good model of a "rich collection of related concepts." Rather than start by trying to make a general model, or adapting existing formalisms (e.g., logical theories) for the purpose, I have undertaken an extended look at a particular domain, namely mathematics. Mathematics certainly qualifies as a rich collection of related concepts, and has the advantage of thousands of years of effort devoted to improving its representations and clarifying its interconnections. This talk will present some of the issues I have encountered, and will probably consist of more questions than answers.

An anecdote will begin to raise some questions. At a workshop some years ago, a colleague asked me if I was familiar with the following problem. Given a nonempty finite set $U$ of cardinality $n$, and two positive integers $s \leq t \leq n$, find the minimum cardinality of a collection $C$ of subsets of $U$ of size $t$ such that every subset of $U$ of size $s$ is a subset of some element of $C$. Since I was not familiar with the problem, she continued to ask others at the workshop, until finally someone gave her the name of the problem and a pointer to work on it.

The meaning of the problem is clear (to someone with some mathematical training) from a very short description. What kind of representation would it take for us to be able to give something like this description to a search engine and be referred to papers that dealt with it? We already are expected to make our papers available in machine readable form on the web, or risk their irrelevance. Perhaps some enhancement of that representation could make such searches possible?

As another example, students in an elementary discrete mathematics course are often introduced to the concepts of permutations and combinations by means of concrete examples. Liu [1] asks the reader to imagine placing three balls, colored red, blue, and white, into ten boxes, numbered 1 through 10, in such a way that each box holds at most one ball. The problem is to determine the number of ways that this may be done. Lovász, Pelikán and Vesztergombi [2] describe a party with seven participants, each of whom shakes hands once with each of the others, and ask how many handshakes there have been in total. An introductory textbook will typically contain many examples and exercises of this kind.

G. Paliouras and Y. Sakakibara (Eds.): ICGI 2004, LNAI 3264, pp. 1–2, 2004.
© Springer-Verlag Berlin Heidelberg 2004

The situations used involve familiar elements, are easily imagined, and are intended to engage the student's intuitions in helpful ways. However, some students find it quite difficult to get the hang of the implicit rules for these problems. What will *not* help such students is the customary explicit and detailed formalization of the domain as a logical theory. What *might* help would be a somewhat more concrete model in terms of actions and state spaces. This is reminiscent of Piaget's emphasis upon an individual's actions as a basis for more abstract understanding.

These issues provide a window on other questions about mathematical reasoning and representation. It is likely that we will make more and more use of computers to help us create and use mathematics. Questions of how best to do that are far from settled, and will require a deep understanding of the multitude of ways that people actually do mathematics. Ironically, those for whom mathematics is difficult may provide some of the clearest evidence of what is involved.

## References

1. C. L. Liu. *Elements of Discrete Mathematics.* McGraw-Hill, 1977.
2. L. Lovász, J. Pelikán, and K. Vesztergombi. *Discrete Mathematics: Elementary and Beyond.* Springer, 2003.

# Learning Finite-State Models for Machine Translation*

Enrique Vidal and Francisco Casacuberta

Departamento de Sistemas Informáticos y Computación
Instituto Tecnológico de Informática, Universidad Politécnica de Valencia
46071 Valencia, Spain
{evidal,fcn}@iti.upv.es

**Abstract.** In formal language theory finite-state transducers are well-know models for "input-output" rational mappings between two languages. Even if more powerful, recursive models can be used to account for more complex mappings, it has been argued that the input-output relations underlying most usual natural language pairs are essentially rational. Moreover, the relative simplicity of these mappings has recently lead to the development of techniques for learning finite-state transducers from a training set of input-output sentence pairs of the languages considered. Following these arguments, in the last few years a number of machine translation systems have been developed based on stochastic finite-state transducers. Here we review the statistical statement of Machine Translation and how the corresponding modelling, learning and search problems can be solved by using stochastic finite-state transducers. We also review the results achieved by the systems developed under this paradigm. After presenting the traditional approach, where transducer learning is mainly solved under the grammatical inference framework, we propose a new approach where learning is explicitly considered as a statistical estimation problem and the whole stochastic finite-state transducer learning problem is solved by expectation maximisation.

## 1 Introduction

*Machine translation* (MT) is one of the most appealing (and challenging) applications of human language processing technology. Because of its great social and economical interest, in the last 20 years MT has been considered under almost every imaginable point of view: from strictly linguistics-based methods to pure statistical approaches including, of course, formal language theory and the corresponding learning paradigm, *grammatical inference* (GI). Different degrees of success have been achieved so far using these approaches.

Basic MT consists in transforming text from a source language into a target language, but several extensions to this framework have been considered. Among the most interesting of these extensions are *speech-to-speech* MT (STSMT) and

* This work was partially supported by the European Union project TT2 (IST-2001-32091) and by the Spanish project TEFATE (TIC 2003-08681-C02-02).

G. Paliouras and Y. Sakakibara (Eds.): ICGI 2004, LNAI 3264, pp. 3–15, 2004.
© Springer-Verlag Berlin Heidelberg 2004

*computer assisted* (human) *translation* (CAT). In STSMT, which is generally considered significantly harder than pure text MT, the system has to accept a source-language utterance and produce corresponding human-understandable target-language speech. In CAT, on the other hand, the input is source-language text and both the system and the human translator have to collaborate with each other in an attempt to produce high quality target text.

Here we consider MT, STSMT and CAT models that can be be automatically learned through suitable combinations of GI and statistical methods. In particular we are interested in *stochastic finite-state transducers*. Techniques for learning these models have been studied by several authors, in many cases with special motivation for their use in MT applications. [1–12].

## 2 General Statement of MT Problems

The (*text-to-text*) MT problem can be statistically stated as follows. Given a sentence $\mathbf{s}$ from a source language, search for a target-language sentence $\hat{\mathbf{t}}$ which maximises the posterior probability[1]:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t} \mid \mathbf{s}) \ . \tag{1}$$

It is commonly accepted that a convenient way to deal with this equation is to transform it by using the Bayes' theorem:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t}) \cdot \Pr(\mathbf{s} \mid \mathbf{t}) \ , \tag{2}$$

where $\Pr(\mathbf{t})$ is a target *language model* – which gives high probability to well formed target sentences – and $\Pr(\mathbf{s} \mid \mathbf{t})$ accounts for source-target word(-position) relations and is based on *stochastic dictionaries* and *alignment models* [13, 14].

Alternatively the conditional distribution in Eq. 1 can be transformed into a joint distribution:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{s}, \mathbf{t}) \ , \tag{3}$$

which can be adequately modelled by means of *stochastic finite-state transducers* (SFST) [15]. This is the kind of models considered in the present work.

Let us now consider the STSMT problem. Here an acoustic representation of a source-language utterance $\mathbf{x}$ is available and the problem is to search for a target-language sentence $\hat{\mathbf{t}}$ that maximises the posterior probability[2]:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \Pr(\mathbf{t} \mid \mathbf{x}) \ . \tag{4}$$

Every possible decoding of a source utterance $\mathbf{x}$ in the source language can be considered as the value of a hidden variable $\mathbf{s}$ [15] and, assuming $\Pr(\mathbf{x}|\mathbf{s}, \mathbf{t})$ does not depend on $\mathbf{t}$, Eq. 4 can be rewritten as:

[1] For simplicity, $\Pr(X = x)$ and $\Pr(X = x \mid Y = y)$ are denoted as $\Pr(x)$ and $\Pr(x \mid y)$.

[2] From $\hat{\mathbf{t}}$, a target utterance can be produced by using a *text-to-speech* synthesiser.

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \sum_{\mathbf{s}} \Pr(\mathbf{s}, \mathbf{t}) \cdot \Pr(\mathbf{x}|\mathbf{s}) \ . \tag{5}$$

As in plain MT, $\Pr(\mathbf{s}, \mathbf{t})$ can be modelled by a SFST. The term $\Pr(\mathbf{x}|\mathbf{s})$, on the other hand, can be modelled through *hidden Markov models* (HMM) [16], which are the standard acoustic models in automatic speech recognition. Thanks to the homogeneous finite-state nature of both SFST and HMMs, and approximating the sum with a maximisation, Eq. 5 can be easily and efficiently solved by the well-known *Viterbi algorithm* [15].

Finally, let us consider a simple statement of CAT [17]. Given a source text $\mathbf{s}$ and a fixed *prefix* of the target sentence $\mathbf{t}_p$ –previously validated by the human translator–, the problem is to search for a *suffix* of the target sentence $\hat{\mathbf{t}}_s$ that maximises the posterior probability:

$$\hat{\mathbf{t}}_s = \underset{\mathbf{t}_s}{\operatorname{argmax}} \Pr(\mathbf{t}_s|\, \mathbf{s}, \mathbf{t}_p) \ . \tag{6}$$

Taking into account that $\Pr(\mathbf{t}_p|\, \mathbf{s})$ does not depend on $\mathbf{t}_s$, we can write:

$$\hat{\mathbf{t}}_s = \underset{\mathbf{t}_s}{\operatorname{argmax}} \Pr(\mathbf{s}, \mathbf{t}_p \mathbf{t}_s) \ , \tag{7}$$

where $\mathbf{t}_p \mathbf{t}_s$ is the concatenation of the given prefix $\mathbf{t}_p$ and a suffix $\mathbf{t}_s$ suggested by the system. Eq. 7 is similar to Eq. 3, but here the maximisation is constrained to a set of suffixes, rather than full sentences. As in Eq. 3, this joint distribution can be adequately modelled by means of SFSTs [18].

All the above problem statements share the common *learning problem* of estimating $\Pr(\mathbf{s}, \mathbf{t})$, which can be approached by training a SFST from a parallel text corpus.

## 3 Stochastic Finite-State Transducers

Different types of SFSTs have been applied with success in some areas of machine translation and other areas of natural language processing [3, 19, 4, 8, 11, 9, 20]. Here only conventional and *subsequential* SFSTs are considered. A SFST $\mathcal{T}_P$ is a tuple $\langle \Sigma, \Delta, Q, q_0, p_T, f_T \rangle$, where $\Sigma$ is a finite set of *source words*, $\Delta$ is a finite set of *target words*, $Q$ is a finite set of *states*, $q_0$ is the initial state and $p_T$ and $f_T$ are two functions $p_T : Q \times \Sigma \times \Delta^\star \times Q \to [0, 1]$ (*transition probabilities*) and $f_T : Q \to [0, 1]$ (*final-state probability*), that verify:

$$\forall q \in \ Q, \qquad f_T(q) \quad + \sum_{(a,\omega,q') \in \Sigma \times \Delta^\star \times Q} p_T(q, a, \omega, q') = 1.$$

Given $\mathcal{T}_P$, the joint probability of a pair $(\mathbf{s}, \mathbf{t}) \in \Sigma^* \times \Delta^*$ –denoted as $\Pr_{\mathcal{T}_P}(\mathbf{s}, \mathbf{t})$– is the sum of the probabilities of all sequences of states that deal with $(\mathbf{s}, \mathbf{t})$; that is, the concatenation of the source (target) words of the transitions between each pair of adjacent states in the sequence of states is the source sentence $\mathbf{s}$ (target sentence $\mathbf{t}$) [21]. The probability of a particular state sequence is the product of the corresponding transition probabilities, times the final-state probability of the last state in the sequence [21].

此为试读,需要完整PDF请访问: www.ertongbook.com