

Guest editorial to the special issue on grammatical inference

Georgios Paliouras · Yasubumi Sakakibara

Published online: 14 August 2006
Springer Science + Business Media, LLC 2007

Grammars are a powerful representation of sequential patterns of various types, ranging from speech and other audio signals, to biological sequences and user navigation on the Web. The long study of grammars, especially in natural languages, has resulted in a rich repertoire of grammar variants and flavors that provide different representation power and consequently different complexity of analysis. Beyond the issues of representation complexity, most of the work on grammars has focused on their use by parsing programs that analyze strings of a given “language”, in order to prove their grammaticality and/or identify interesting elements of the language.

The use of grammars by parsers, especially the more expressive and complex grammars, remains a challenging and interesting research issue on its own. However, grammatical inference goes a step further to study methods that learn grammars from data. Grammatical inference is an established research field in Artificial Intelligence, dating back to the 60s and has been extensively addressed by researchers in automata theory, language acquisition, computational linguistics, machine learning, pattern recognition, computational learning theory and neural networks. From a theoretical perspective, the main aim of this work is to study the learnability of different types of grammars from different types of data and propose efficient algorithms for learning. In parallel, a significant amount of work focuses on innovative applications of grammatical inference algorithms to various knowledge discovery tasks.

The main forum for presenting this type of work in the past 15 years has been the International Colloquium on Grammatical Inference (ICGI) which takes place in different countries and different continents every two years. The seventh ICGI was held in the National Centre for Scientific Research “Demokritos”, Greece on October 11–13th, 2004. The topics of the

G. Paliouras (✉)
Institute of Informatics and Telecommunications,
National Center for Scientific Research “Demokritos”,
15310, Ag. Paraskevi, Attiki, Greece
e-mail: paliourg@iit.demokritos.gr

Y. Sakakibara
Department of Biosciences and Informatics,
Keio University, Yokohama, 223-8522, Japan
e-mail: yasu@bio.keio.ac.jp

papers accepted in ICGI-2004 varied 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 parallel to the submission and reviewing of research papers, a context-free grammar learning competition, named Omphalos, took place. The competition was created in an effort to promote the development of new and better grammatical inference algorithms for context-free languages, to provide a forum for the comparison of different grammatical inference algorithms and to gain insight into the current state of the art of context-free grammatical inference algorithms. The proceedings of the conference are available in Paliouras and Sakakibara (2004), while the presentations and related material can be found in the Web site of the conference:

<http://www.iit.demokritos.gr/icgci04>

From the 23 full papers that were presented in the conference, we have selected six and asked the authors to present an improved and significantly extended version of their work. The submitted papers underwent a very rigorous and thorough refereeing process in which each paper received three reviews. Finally, four papers were accepted and appear here.

The paper by Remi Eyraud, Colin de la Higuera and Jean-Christophe Janodet presents a theoretical result on the issue of grammar learnability. The paper proposes to learn a class of languages using a different representation class from formal grammars, that is, representing and manipulating grammars as string re-writing systems (SRS), which are expressive enough to define the class of regular languages and an important subset of context-free languages. The paper further introduces a concept of characteristic sample for a new class of SRS, called closed DSRS, and presents a polynomial-time algorithm, called LARS (Learning Algorithm for Rewriting Systems), for learning the class (closed DSRS).

The paper by Julien Carme, Remi Gilleron, Aurelien Lemay and Joachim Niehren develops new algorithms for learning monadic node selection queries in unranked trees from annotated examples with the aim of applying them to the difficult task of Web information extraction. The paper introduces Node Selecting Tree Transducers (NSTT), an adaptation of tree automata which enables one to express node selecting queries with tree automata, and discusses their relationship with queries definable in monadic second order logic (MSO).

The paper by Francisco Casacuberta and Enrique Vidal summarizes works of the authors on machine translation for a period of about 15 years. The paper explains how modeling, learning and search problems can be solved by using stochastic finite-state transducers, and reviews the results achieved by the machine translation systems that the authors have developed under this paradigm. The paper can be interpreted as describing a paradigm shift from a deterministic learning approach to more statistically oriented techniques in this domain.

Finally, the paper by Alexander Clark presents the method behind the winning entry to the Omphalos context-free grammar learning competition held at ICGI-2004. The paper introduces a new method of learning grammars, which combines several existing approaches. It provides an overview of the basic properties of context-free grammars, focusing on a subclass, called Non-terminally Separated (NTS) grammars. The proposed method not only solved several Omphalos problems, but also shows how the innovative integration of existing methods can lead to significant progress in the state-of-the-art of grammatical inference.

We believe that the four papers presented in this special issue are indicative of the state-of-the-art in the area of grammatical inference, covering both theoretical and practical strands of work. We therefore hope that this issue will not only contribute to a better understanding

of grammatical inference, but also increase the reader's interest in the many challenges that still lie ahead of us.

Reference

Paliouras, G., & Sakakibara, Y. (Eds.) (2004) Grammatical inference: algorithms and applications. Lecture notes in artificial intelligence, no. 3264, Springer-Verlag.