Grammatical Inference for Computational Linguistics

Synthesis Lectures on Human Language Technologies

Editor
Graeme Hirst, *University of Toronto*

Synthesis Lectures on Human Language Technologies is edited by Graeme Hirst of the University of Toronto. The series consists of 50- to 150-page monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Grammatical Inference for Computational Linguistics
Jeffrey Heinz, Colin de la Higuera, and Menno van Zaanen
2015

Automatic Detection of Verbal Deception
Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari
2015

Natural Language Processing for Social Media
Atefeh Farzindar and Diana Inkpen
2015

Semantic Similarity from Natural Language and Ontology Analysis
Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain
2015

Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition
Hang Li
2014

Ontology-Based Interpretation of Natural Language
Philipp Cimiano, Christina Unger, and John McCrae
2014

Automated Grammatical Error Detection for Language Learners, Second Edition
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2014

Web Corpus Construction
Roland Schäfer and Felix Bildhauer
2013

Recognizing Textual Entailment: Models and Applications
Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto
2013

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax
Emily M. Bender
2013

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing
Anders Søgaard
2013

Semantic Relations Between Nominals
Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz
2013

Computational Modeling of Narrative
Inderjeet Mani
2012

Natural Language Processing for Historical Texts
Michael Piotrowski
2012

Sentiment Analysis and Opinion Mining
Bing Liu
2012

Discourse Processing
Manfred Stede
2011

Bitext Alignment
Jörg Tiedemann
2011

Linguistic Structure Prediction
Noah A. Smith
2011

Learning to Rank for Information Retrieval and Natural Language Processing
Hang Li
2011

Computational Modeling of Human Language Acquisition
Afra Alishahi
2010

Introduction to Arabic Natural Language Processing
Nizar Y. Habash
2010

Cross-Language Information Retrieval
Jian-Yun Nie
2010

Automated Grammatical Error Detection for Language Learners
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2010

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer
2010

Semantic Role Labeling
Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems
Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing
Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics
Graham Wilcock
2009

Dependency Parsing
Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval
ChengXiang Zhai
2008

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2016

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews—without the prior permission of the publisher.

Grammatical Inference for Computational Linguistics

Jeffrey Heinz, Colin de la Higuera, Menno van Zaanen

ISBN: 978-3-031-01031-6 paperback
ISBN: 978-3-031-02159-6 ebook

DOI 10.1007/978-3-031-02159-6

A Publication in the Springer series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES
Series ISSN: 1947-4040 print 1947-4059 ebook

Lecture #28
Series Editor: Graeme Hirst, *University of Toronto*

First Edition
10 9 8 7 6 5 4 3 2 1

Grammatical Inference for Computational Linguistics

Jeffrey Heinz
University of Delaware

Colin de la Higuera
Nantes University

Menno van Zaanen
Tilburg University

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #28

ABSTRACT

This book provides a thorough introduction to the subfield of theoretical computer science known as grammatical inference from a computational linguistic perspective. Grammatical inference provides principled methods for developing computationally sound algorithms that learn structure from strings of symbols. The relationship to computational linguistics is natural because many research problems in computational linguistics are learning problems on words, phrases, and sentences: What algorithm can take as input some finite amount of data (for instance a corpus, annotated or otherwise) and output a system that behaves "correctly" on specific tasks?

Throughout the text, the key concepts of grammatical inference are interleaved with illustrative examples drawn from problems in computational linguistics. Special attention is paid to the notion of "learning bias." In the context of computational linguistics, such bias can be thought to reflect common (ideally universal) properties of natural languages. This bias can be incorporated either by identifying a learnable class of languages which contains the language to be learned or by using particular strategies for optimizing parameter values. Examples are drawn largely from two linguistic domains (phonology and syntax) which span major regions of the Chomsky Hierarchy (from regular to context-sensitive classes). The conclusion summarizes the major lessons and open questions that grammatical inference brings to computational linguistics.

KEYWORDS

grammatical inference, language learning, natural languages, formal languages

To our families, friends, and automata

Contents

List of Figures

List of Tables

Preface

Once authors have written the words "The End," they realize how much has been left out. And this rule holds if instead of one author, there are three. This is one reason why prefaces are important. They let readers know (and remind the authors) what the book achieves and what it does not.

The tasks addressed in this book become more formidable with each passing day. It is becoming more complex and intricate because there are more and more cases where one is delivered a huge amount of strings, words, or sentences, or has access to some such data, and one is asked how to build a model summarizing or explaining this information. Furthermore, for many reasons—for example, the fact that most computer scientists have taken courses on graph theory and formal languages—the types of models people are seeking will be very often linked with grammars and automata.

That is why, today, there are people attempting to build or infer grammars or finite-state machines in fields as different as verification, pattern recognition, bioinformatics, and linguistics. That is why techniques of all sorts are being used to infer these models: some rely on statistics, others on linear algebra, some on formal language theory, and many quite often on a combination of these. And finally, that is why certain choices have been made in this book, and therefore some readers might be frustrated.

Before we explain why some of our choices may frustrate readers, let us state who we think our readers are. One reason we embarked on this project was because there is no text which introduces grammatical inference to people working in computational linguistics and natural language processing. Our hope is that this book helps bridge the gap between the needs of *these* researchers and a particular way of thinking about the problems of learning automata and grammars in machine learning. We sincerely believe grammatical inference can help address problems in computational linguistics and that problems in computational linguistics can inform and lead to new developments in grammatical inference (in fact, such mutual benefits exist and are ongoing).

We also have in mind readers who are not encountering automata and grammars for the first time. The kind of background knowledge we expect readers to have is of the type that could be found in standard textbooks on formal language theory that one might take as an advanced undergraduate student or a beginning graduate student. We also expect readers to have some familiarity with topics in computational linguistics and natural language processing, like the kinds discussed in the books by Jurafsky and Martin [2008] or Manning and Schütze [1999] (or their more recent editions).

So what are the choices that may frustrate readers? The first choice we made was to concentrate on *only some* tools and techniques, and not attempt to be exhaustive.

The second choice was to cover the tools and techniques which have been developed in what may informally be called the school of grammatical inference, as represented, over the past 30 years, by the papers published in the series of conferences called ICGI—International Conference on Grammatical Inference. These share a certain number of aspects.

- They build upon well-understood formal language formalisms and avoid, whenever possible, technical complications in the definitions of the objects themselves.
- They either attempt to deliver formal learnability results, independent of some particular corpus, or, on the other hand, aim to produce a very general algorithm whose proof of concept will be given by its results on particular corpora without corpus-specific tweaks.

Consequently, this means the knowledge we present builds from formal language theory and concentrates on those techniques whose intricate theoretical backbone comes from that field.

A third choice is that the book is not self-contained, in the sense that not every algorithm discussed is presented and proved correct in full detail. Instead, we have chosen to focus on ideas, and to include only the notation, definitions, and theorems that we felt important because they support those main points. We do not include proofs, but we try to point to them and further material which helps readers find detailed descriptions and explanations of the algorithms or formalisms. For instance, we often refer to de la Higuera [2010], a book on grammatical inference with a general orientation, which is self-contained.

Together all of this means leaving out certain results, which no doubt deserve closer attention.

For finite-state machines, one notable area left out is *spectral methods*, which identify finite-state machines with sets of matrices and therefore transforms the learning problem into one which searches for an optimal set of parameters which fits those matrices. The techniques here are attractive: they allow the learning of very rich classes of finite-state machines, rely on linear algebra's vast literature, and can be redefined as global optimization problems, for which a large number of researchers are bettering the algorithms all the time.

For formal grammars, a number of results (sometimes grouped under the name *grammar induction*) are based on starting with a backbone grammar, either extremely general or devised from using data for which the structure is known, and adjusting the parameters by just observing relative frequencies. The types of grammars will themselves be adapted to better fit the knowledge we have of natural language.

We do not argue here that the techniques covered in this book work better, just that they correspond to a uniform set of ideas which, when understood, can allow a number of problems to be solved.

Perhaps one argument which we would like to put forward is that of intelligibility. Albeit informal in most cases, the idea is that the types of techniques proposed in this book rely on wanting to understand the machines and grammars learned. An undeclared goal is that one should be able to run a grammatical inference algorithm, obtain perhaps a large automaton or grammar, and nevertheless be able to observe it and understand it, not just its effects. This helps to explain why we believe that the issue of learning the *structure* of the grammar is essential, and why this theme recurs throughout the book.

One may wonder if this is necessary, as the grammar will often be evaluated through a success or error rate, not through its capacity to speak to us. On the other hand, there are increasingly many applications where the user wants more than a black box.

All of this, and the idea of making the book useful to as many readers as possible, was what the authors had in mind when they launched this adventure.

ACKNOWLEDGMENTS

The authors would like to thank Dion Bot, Rémi Eyraud, Thomas Graf, Ákos Kádár, Esmée Mertens, Jim Rogers, Eva Ummelen, Merel van de Wiel, and the anonymous reviewers for their useful feedback.

Jeffrey Heinz gladly thanks Mika, Emma, and Maya for their support.

Colin de la Higuera is grateful to Lindsey for her support.

Menno van Zaanen would like to thank Tanja, Colwin, and Lejla for their support.

September 2015