

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	xv
List of Tables	xvii
Preface	xix
1. Studying Learning	1
1.1 An Overview of Grammatical Inference	2
1.2 Formal and Empirical Grammatical Inference	3
1.3 Formal Grammatical Inference	5
1.3.1 Language and Grammar	6
1.3.2 Language Families	8
1.3.3 Learning Languages Efficiently	9
1.4 Empirical Grammatical Inference	10
1.4.1 Languages, Grammars, and Language Families	11
1.4.2 Evaluation	12
1.5 Summary	13
1.6 Formal Preliminaries	14
2. Formal Learning	21
2.1 Introduction	21
2.1.1 The Issues of Learning	21
2.1.2 Learning Scenarios	22
2.1.3 Learning Grammars of Languages	23
2.2 Learnability: Definitions and Paradigms	24
2.2.1 Blame the Data, Not the Algorithm	24
2.2.2 A Non-Probabilistic Setting: Identification in the Limit	24
2.2.3 An Active Learning Setting	25
2.2.4 Introducing Complexity	26
2.2.5 A Probabilistic Version of Identification in the Limit	28
2.2.6 Probably Approximately Correct (PAC) Learning	28
2.3 Grammar Formalisms	31
2.3.1 Finite-State Machines Recognizing Strings	31
2.3.2 Probabilistic Finite-State Machines	34

2.3.3	Transducers	37
2.3.4	More Complex Formalisms	41
2.3.5	Dealing with Trees and Graphs	46
2.4	Is Grammatical Inference an Instance of Machine Learning?	47
2.5	Summary	49
3.	Learning Regular Languages	51
3.1	Introduction	51
3.2	Bias Selection Reduces the Problem Space	52
3.3	Regular Grammars	53
3.4	State-Merging Algorithms	56
3.4.1	The Problem of Learning Stress Patterns	56
3.4.2	Merging States	59
3.4.3	Finite-State Representations of Finite Samples	60
3.4.4	The State-Merging Theorem	62
3.5	State-Merging as a Learning Bias	64
3.6	State-Merging as Inference Rules	66
3.7	RPNI	67
3.7.1	How It Works	67
3.7.2	Theoretical Results	68
3.8	Regular Relations	69
3.9	Learning Stochastic Regular Languages	71
3.9.1	Stochastic Languages	72
3.9.2	Structure of the Class Is Deterministic and Known A Priori	73
3.9.3	Structure of the Class Is Deterministic but Not Known A Priori	78
3.9.4	Structure of the Class Is Non-Deterministic and Not Known A Priori	80
3.10	Summary	81
4.	Learning Non-Regular Languages	85
4.1	Substitutability	86
4.1.1	Identifying Structure	86
4.1.2	Learning Using Substitutability	88
4.2	Empirical Approaches	89
4.2.1	Expanding and Reducing Approaches	89
4.2.2	Supervised and Unsupervised Approaches	89
4.2.3	Word-Based and POS-Based Approaches	90
4.2.4	Description of Empirical Systems	90
4.2.5	Comparison of Empirical Systems	102

4.3	Issues for Evaluation	103
4.3.1	Looks-Good-To-Me Approach	104
4.3.2	Rebuilding Known Grammars	105
4.3.3	Compare against a Treebank	107
4.3.4	Language Membership	109
4.4	Formal Approaches	111
4.5	Summary	113
5.	Lessons Learned and Open Problems	115
5.1	Summary	115
5.2	Lessons	116
5.3	Problems	116
5.3.1	Learning Targets	116
5.3.2	Learning Criteria	118
5.4	Resources	120
5.5	Final Words	120
	Bibliography	121
	Author Biographies	137

List of Figures

1.1	General overview of the process of grammatical inference.	3
1.2	The relationship between strings, languages, and language families.	8
1.3	The Chomsky Hierarchy.	9
2.1	Graphical representation of a DFA.	32
2.2	An NFA.	33
2.3	An NFA with λ -transitions.	33
2.4	A PFA.	36
2.5	A PFA with λ -transitions.	36
2.6	An HMM.	37
2.7	An HMM using \sharp to terminate the generation of strings.	38
2.8	A rational transducer.	38
2.9	A semideterministic transducer.	39
2.10	Normalized transducer.	40
2.11	A probabilistic transducer.	41
2.12	A probabilistic semideterministic transducer.	41
2.13	Parse tree for John hit the ball.	43
2.14	Parse tree for the sentence aabaccc for a linear grammar with rules $S \rightarrow aS bS aS^c acc$	44
2.15	Parse tree for an even linear grammar.	45
2.16	Skeleton corresponding to the bracketed string $(a(a(a(b)a)b)(a(a)b))$	47
3.1	A finite-state acceptor which recognizes the language $\sigma\sigma^*$	54
3.2	A minimal, deterministic, finite-state acceptor for Pintupi stress.	57
3.3	A minimal, deterministic, finite-state acceptor for Kwakwala stress.	59
3.4	Machine B represents the machine obtained by merging states 1 and 2 in Machine A.	59
3.5	A prefix tree of Pintupi words.	61
3.6	A suffix tree for Pintupi words.	62
3.7	The result of merging states in the prefix tree for Pintupi words (Figure 3.5) with the same incoming paths of length 2.	65

3.8	A prefix-tree whose states have been enumerated in breadth-first fashion.	68
3.9	A a family of distributions with four parameters.	74
3.10	The count of the parse of $S = \langle bbc \rangle$ and the probabilities obtained after normalization. .	75
3.11	The structure of a bi-gram model.	76
4.1	Substitutability in tree structures.	87
4.2	Schematic overview of phases in ABL.	93
4.3	Alignment of two sentences and the identification of hypotheses.	93
4.4	Initial ADIOS graph for John sees a cat, John walks, and The dog sees Mary.	95
4.5	The initial ADIOS graph, and the two graphs after rewriting the pattern of nodes. . . .	97
4.6	Example dependency parse without terminals, which are visualized as dots.	99
4.7	Elementary tree structures in DOP.	100
4.8	Two derivations using left-most substitution that lead to the same parse.	101
4.9	Schematic representation of the LOOKS-GOOD-TO-ME evaluation approach.	104
4.10	Schematic representation of the REBUILDING KNOWN GRAMMARS evaluation approach. .	105
4.11	Schematic representation of the COMPARE AGAINST A TREEBANK evaluation approach. .	107
4.12	Schematic representation of the LANGUAGE MEMBERSHIP evaluation approach.	110

List of Tables

2.1	Deterministic finite-state automata: DFA	32
2.2	Non-deterministic finite-state automata: NFA	34
2.3	Regular expressions	34
2.4	Probabilistic finite-state automata: PFA	37
2.5	Hidden Markov models: HMM	38
2.6	Transducers	40
2.7	Probabilistic transducers	42
2.8	Context-free grammars	45
2.9	Probabilistic context-free grammars	46
3.1	Pintupi words	57
3.2	Pintupi words with a syllabic representation	57
3.3	All LHOR words up to four syllables in length	58
3.4	Words in Samala with sibilant sounds [ʃ, S]	76
3.5	Impossible Samala words with sibilant sounds [s, ʃ]	77
3.6	SP ₂ probabilities of a sibilant occurring sometime after another one.	78
4.1	Matrix for John walks, John sees Mary and Mary walks	91
4.2	Overview of properties of the two groups of empirical systems	102

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