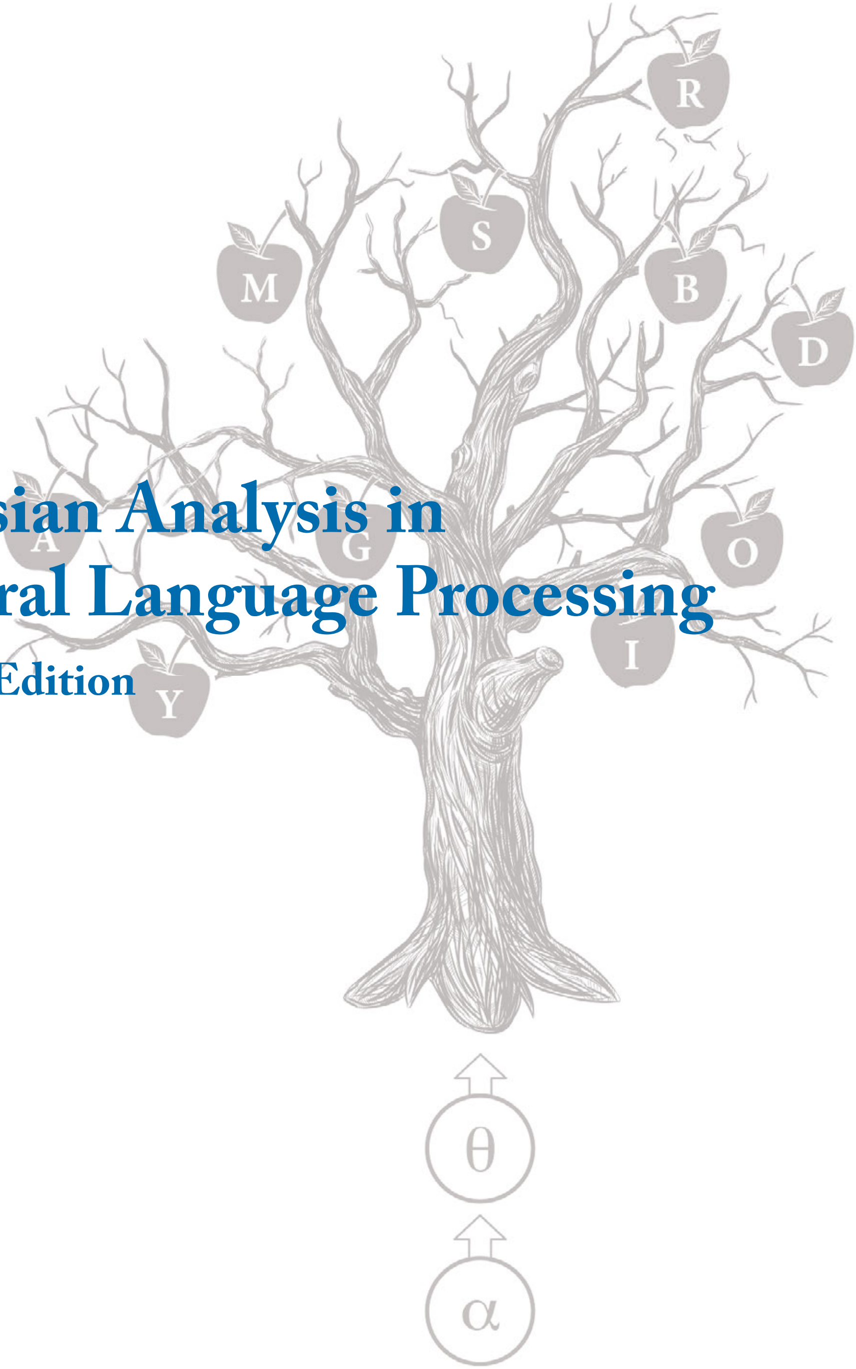
Bayesian Analysis in Natural Language Processing

Second Edition



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.

Bayesian Analysis in Natural Language Processing, Second Edition
Shay Cohen
2019

Argumentation Mining
Manfred Stede and Jodi Schneider
2018

Quality Estimation for Machine Translation
Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold
2018

Natural Language Processing for Social Media, Second Edition
Atefeh Farzindar and Diana Inkpen
2017

Automatic Text Simplification
Horacio Saggion
2017

Neural Network Methods for Natural Language Processing
Yoav Goldberg
2017

Syntax-based Statistical Machine Translation
Philip Williams, Rico Sennrich, Matt Post, and Philipp Koehn
2016

Domain-Sensitive Temporal Tagging
Jannik Strötgen and Michael Gertz
2016

Linked Lexical Knowledge Bases: Foundations and Applications
Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek
2016

Bayesian Analysis in Natural Language Processing
Shay Cohen
2016

Metaphor: A Computational Perspective
Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov
2016

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 2019

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.

Bayesian Analysis in Natural Language Processing, Second Edition

Shay Cohen

ISBN: 978-3-031-01042-2 paperback
ISBN: 978-3-031-02170-1 ebook
ISBN: 978-3-031-00181-9 hardcover

DOI 10.1007/978-3-031-02170-1

A Publication in the Springer series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #41
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Bayesian Analysis in Natural Language Processing

Second Edition

Shay Cohen
University of Edinburgh

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #41

ABSTRACT

Natural language processing (NLP) went through a profound transformation in the mid-1980s when it shifted to make heavy use of corpora and data-driven techniques to analyze language. Since then, the use of statistical techniques in NLP has evolved in several ways. One such example of evolution took place in the late 1990s or early 2000s, when full-fledged Bayesian machinery was introduced to NLP. This Bayesian approach to NLP has come to accommodate various shortcomings in the frequentist approach and to enrich it, especially in the unsupervised setting, where statistical learning is done without target prediction examples.

In this book, we cover the methods and algorithms that are needed to fluently read Bayesian learning papers in NLP and to do research in the area. These methods and algorithms are partially borrowed from both machine learning and statistics and are partially developed "in-house" in NLP. We cover inference techniques such as Markov chain Monte Carlo sampling and variational inference, Bayesian estimation, and nonparametric modeling. In response to rapid changes in the field, this second edition of the book includes a new chapter on representation learning and neural networks in the Bayesian context. We also cover fundamental concepts in Bayesian statistics such as prior distributions, conjugacy, and generative modeling. Finally, we review some of the fundamental modeling techniques in NLP, such as grammar modeling, neural networks and representation learning, and their use with Bayesian analysis.

KEYWORDS

natural language processing, computational linguistics, Bayesian statistics, Bayesian NLP, statistical learning, inference in NLP, grammar modeling in NLP, neural networks, representation learning

Dedicated to Mia

Contents

List of Figures

List of Algorithms

List of Generative Stories

Preface (First Edition)

When writing about a topic that intersects two areas (in this case, Bayesian Statistics and Natural Language Processing), the focus and the perspective need to be considered. I took a rather practical one in writing this book, aiming to write it for those in the same position as myself during my graduate studies. At that time, I already had a reasonable grasp of the problems in natural language processing and knowledge of the basic principles in machine learning. I wanted to learn more about Bayesian statistics—in a rather abstract way—particularly the parts that are most relevant to NLP. Thus, this book is written from that perspective, providing abstract information about the key techniques, terminology and models that a computational linguist would need in order to apply the Bayesian approach to his or her work.

Most chapters in this book, therefore, are rather general and have relevance to other uses of Bayesian statistics. The last chapter in this book, though, presents some specific NLP applications for grammar models that are mostly (but not exclusively) used in NLP.

Ideally, this book should be read by a person who already has some idea about statistical modeling in NLP, and wants to gain more depth about the specific application of Bayesian techniques to NLP. The motivation for this decision to focus more on the mathematical aspects of Bayesian NLP is simple. Most computational linguists get exposed quite early in their graduate career or otherwise to the basic core terminology in NLP, the linguistic structures it predicts and perhaps some of the linguistic motivation behind them. Ideas from Bayesian statistics and other statistical tools are often "picked up" on the way. As such, there are sometimes misconceptions and a missing global picture. This book tries to provide some of these missing details to the reader.

There are several approaches to doing statistics, two of which are the frequentist approach and the Bayesian approach. The frequentist approach is also sometimes referred to as "classic statistics." One of the things that motivated me to learn more about Bayesian statistics is the rich and colorful history it has. To this day, the famous frequentist-Bayesian divide still exists. This kind of divide regarding the philosophy that statistical analysis should follow is even more persistently and more ardently argued about than theories of grammar were in the famous "linguistics war" between generative semanticians and generative grammarians. It does not end there—even within the Bayesian camp there are those who support a subjective interpretation of probability and those who support an objective one.

Although I was captivated by the mathematical elegance of Bayesian statistics when I was first exposed to the core ideas (in principle, Bayesian statistics relies on one basic principle of applying Bayes' rule to invert the relationship between data and parameters), I take a pragmatic approach and do not try to present Bayesian statistics as the ultimate theory for doing statistical

NLP. I also do not provide the philosophical arguments that support Bayesian statistics in this monograph. Instead, I provide the technical mechanisms behind Bayesian statistics, and advise the reader to determine whether the techniques work well for him or her in the problems they work on. Here and there, I also describe some connections that Bayesian statistics have to the frequentist approach, and other points of confluence. If the reader is interested in learning more about the philosophy behind Bayesian statistics, I suggest reading Jaynes (2003) and also looking at Barnett (1999). To better understand the history and the people behind Bayesian statistics, I suggest reading the book by McGrayne (2011). This book consists of eight chapters as following:

Chapter 1 is a refresher about probability and statistics as they relate to Bayesian NLP. We cover basic concepts such as random variables, independence between random variables, conditional independence, and random variable expectations; we also briefly discuss Bayesian statistics and how it differs from frequentist statistics. Most of this chapter can probably be skipped if one already has some basic background in computer science or statistics.

Chapter 2 introduces Bayesian analysis in NLP with two examples (the latent Dirichlet allocation model and Bayesian text regression), and also provides a high-level overview of the topic.

Chapter 3 covers an important component in Bayesian statistical modeling—the prior. We discuss the priors that are most commonly used in Bayesian NLP, such as the Dirichlet distribution, non-informative priors, the normal distribution and others.

Chapter 4 covers ideas that bring together frequentist statistics and Bayesian statistics through the summarization of the posterior distribution. It details approaches to calculate a point estimate for a set of parameters while maintaining a Bayesian mindset.

Chapter 5 covers one of the main inference approaches in Bayesian statistics: Markov chain Monte Carlo. It details the most common sampling algorithms used in Bayesian NLP, such as Gibbs sampling and Metropolis-Hastings sampling.

Chapter 6 covers another important inference approach in Bayesian NLP, variational inference. It describes mean-field variational inference and the variational expectation-maximization algorithm.

Chapter 7 covers an important modeling technique in Bayesian NLP, nonparametric modeling. We discuss nonparametric models such as the Dirichlet process and the Pitman-Yor process.

Chapter 8 covers basic grammar models in NLP (such as probabilistic context-free grammars and synchronous grammars), and the way to frame them in a Bayesian context (using models such as adaptor grammars, hierarchical Dirichlet process PCFGs and others).

In addition, the book includes two appendices that provide background information that offers additional context for reading this book. Each chapter is accompanied by at least five exercises. This book (perhaps with the exercises) could be used as teaching material. Specifically, it could be used to teach a number of lectures about Bayesian analysis in NLP. If a significant amount of time is devoted to Bayesian NLP in class (say, four lectures), I would suggest devoting one lecture to chapter 3, one lecture to chapter 4, one lecture to chapters 5 and 6 together, and one lecture to chapter 7. Topics from chapter 8 (such as adaptor grammars or Bayesian PCFGs) can be injected into the various lectures as examples.

Acknowledgments (First Edition)

I am indebted to all the people who helped me to write this book. First, I would like to especially thank Lea Frermann, Trevor Cohn, and Jacob Eisenstein, who carefully read a draft of this book, and left detailed feedback. I would also like to thank all other people who gave feedback in one form or another: Omri Abend, Apoorv Agarwal, Anahita Bhiwandiwalla, Jordan Boyd-Graber, Daniel Gildea, Sharon Goldwater, Mark Johnson, Mirella Lapata, Shalom Lappin, Adam Lopez, Brendan O'Connor, Mohammad Sadegh Rasooli, Siva Reddy, Stefan Riezler, Giorgio Satta, Stuart Shieber, Mark Steedman, Karl Stratos, Swabha Swayamdipta, Bonnie Webber and Dani Yogatama. Thanks also to Sharon Rosenfeld, who proofread this book to help make it more readable. Thanks also to Samantha Draper, Graeme Hirst, Michael Morgan and CL Tondo for helping with the publication of this book.

I would also like to thank all of the great students who attended my class at the Department of Computer Science in Columbia University in Spring 2013 ("Bayesian analysis in NLP") and indirectly helped me better understand what is the level of coverage needed from this book for young researchers making their first steps in the area of Bayesian NLP (students such as Jessica Forde, Daniel Perlmutter, and others whom I have already mentioned). Thanks also go to my collaborators on projects in the area of Bayesian NLP, who helped me shape my understanding of it: David Blei, Jordan Boyd-Graber, Kevin Gimpel, and Ke Zhai.

I want to also thank my mentors throughout all the years, and especially my advisor, Noah Smith, with whom I first studied Bayesian NLP, Michael Collins, my post-doc advisor, who supported me in spending time writing this book during my post-doctoral fellowship and in teaching the Bayesian NLP class at Columbia, and Mark Johnson, whose work, as well as our conversations and email exchanges, influenced me in writing this book.

Also, thanks to Sylvia Cohen, my spouse, who has always been there for me during the writing of this book. Similar thanks to Sylvia's family, who always made me feel at home in Pittsburgh, while I was studying topics such as Bayesian analysis. Finally, I would like to thank my parents and siblings—whose prior beliefs in me never changed, no matter what the evidence was.

Shay Cohen
Edinburgh
May 2016

Preface (Second Edition)

I did not expect to release a second edition for this book so quickly, but the last few years of fast-paced and exciting developments in the world of Natural Language Processing (NLP) have called for various updates, leading to this second edition.

The main addition to this book is Chapter 9, which focuses on representation learning and neural networks in NLP, particularly in a Bayesian context. This chapter was written based on the observation that in the past five years or so, NLP literature has been dominated by the use of neural networks; and as such, I believe the fundamentals needed to be addressed in this book. Adapting the content to the Bayesian "mission" of this book (coupled with the NLP context) was not always easy, and I will let the reader be the judge of whether I have accomplished my mission.

In addition to introducing this new chapter in this edition, several typographical errors have been fixed, and some additional content was integrated into various chapters.

There are several people who have helped with this edition. I want to thank Trevor Cohn, Marco Damonte, Jacob Eisenstein, Lea Frermann, Annie Louis, Chunchuan Lyu, Nikos Papasarantopoulos, Shashi Narayan, Mark Steedman, Rico Sennrich, and Ivan Titov for their help and comments. I also want to thank my students and postdocs who have taught me more than I have taught them in some areas of the new material in this book.

Shay Cohen
Edinburgh
February 2019