Statistical Significance Testing for Natural Language Processing

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.

Statistical Significance Testing for Natural Language Processing
Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart
2020

Deep Learning Approaches to Text Production
Shashi Narayan and Claire Gardent
2020

Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics to Pragmatics
Emily M. Bender and Alex Lascarides
2019

Cross-Lingual Word Embeddings
Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui
2019

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

Statistical Significance Testing for Natural Language Processing

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart

ISBN: 978-3-031-01046-0 paperback
ISBN: 978-3-031-02174-9 ebook
ISBN: 978-3-031-00185-7 hardcover

DOI 10.1007/978-3-031-02174-9

A Publication in the Springer series
SYNTHESIS LECTURES ON ADVANCES IN AUTOMOTIVE TECHNOLOGY

Lecture #45
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Statistical Significance Testing for Natural Language Processing

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart
Technion – Israel Institute of Technology

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #45

ABSTRACT

Data-driven experimental analysis has become the main evaluation tool of Natural Language Processing (NLP) algorithms. In fact, in the last decade, it has become rare to see an NLP paper, particularly one that proposes a new algorithm, that does not include extensive experimental analysis, and the number of involved tasks, datasets, domains, and languages is constantly growing. This emphasis on empirical results highlights the role of statistical significance testing in NLP research: If we, as a community, rely on empirical evaluation to validate our hypotheses and reveal the correct language processing mechanisms, we better be sure that our results are not coincidental.

The goal of this book is to discuss the main aspects of statistical significance testing in NLP. Our guiding assumption throughout the book is that the basic question NLP researchers and engineers deal with is whether or not one algorithm can be considered better than another one. This question drives the field forward as it allows the constant progress of developing better technology for language processing challenges. In practice, researchers and engineers would like to draw the right conclusion from a limited set of experiments, and this conclusion should hold for other experiments with datasets they do not have at their disposal or that they cannot perform due to limited time and resources. The book hence discusses the opportunities and challenges in using statistical significance testing in NLP, from the point of view of experimental comparison between two algorithms. We cover topics such as choosing an appropriate significance test for the major NLP tasks, dealing with the unique aspects of significance testing for non-convex deep neural networks, accounting for a large number of comparisons between two NLP algorithms in a statistically valid manner (multiple hypothesis testing), and, finally, the unique challenges yielded by the nature of the data and practices of the field.

KEYWORDS

Natural Language Processing, statistics, statistical significance, hypothesis testing, algorithm comparison, deep neural network models, replicability analysis

Contents

Preface

The field of Natural Language Processing (NLP) has made substantial progress in the last two decades. This progress stems from multiple sources: the data revolution that has made abundant amounts of textual data from a variety of languages and linguistic domains available, the development of increasingly effective predictive statistical models, and the availability of hardware that can apply these models to large datasets. This dramatic improvement in the capabilities of NLP algorithms carries the potential for a great impact.

The extended reach of NLP algorithms has also resulted in NLP papers giving more and more emphasis to the experiment and result sections by showing comparisons between multiple algorithms on various datasets from different languages and domains. It can be safely argued that the ultimate test for the quality of an NLP algorithm is its performance on well-accepted datasets, sometimes referred to as "leader-boards". This emphasis on empirical results highlights the role of statistical significance testing in NLP research: If we rely on empirical evaluation to validate our hypotheses and reveal the correct language processing mechanisms, we better be sure that our results are not coincidental.

The goal of this book is to discuss the main aspects of statistical significance testing in NLP. Particularly, we aim to briefly summarize the main concepts so that they are readily available to the interested researcher, address the key challenges of hypothesis testing in the context of NLP tasks and data, and discuss open issues and the main directions for future work.

We start with two introductory chapters that present the basic concepts of statistical significance testing: Chapter 2 provides a brief presentation of the hypothesis testing framework, and Chapter 3 introduces common statistical significance tests. Then, Chapter 4 discusses the application of statistical significance testing to NLP. In Chapter 4, we assume that two algorithms are compared on a single dataset, based on a single output that each of them produces, and discuss the relevant significance tests for various NLP tasks and evaluation measures. The chapter puts an emphasis on the aspects in which NLP tasks and data differ from common examples in the statistical literature, e.g., the non–Gaussian distribution of the data and the dependence between the participating examples, e.g., sentences in the same corpus. This chapter, which extends our ACL 2018 paper [Dror et al, 2018], provides our recommended matching between NLP tasks with their evaluation measures and statistical significance tests.

The next two chapters relax two of the basic assumptions of Chapter 4: (a) that each of the compared algorithms produces a single output for each test example (e.g., a single parse tree for a given input sentence), and (b) that the comparison between the two algorithms is performed on a single dataset. Particularly, Chapter 5 addresses the comparison between two algorithms

based on multiple solutions where each of them produces for a single dataset, while Chapter 6 addresses the comparison between two algorithms across several datasets.

The first challenge stems from the recent emergence of Deep Neural Networks (DNNs), which has made data-driven performance comparison much more complicated. This is because these models are non-deterministic due to their non-convex objective functions, complex hyper-parameter tuning process and training heuristics such as random dropouts, that are often applied in their implementation. Chapter 5, therefore, defines a framework for a statistically valid comparison between two DNNs based on multiple solutions each of them produces for a given dataset. The chapter summarizes previous attempts in the NLP literature to perform this comparison task and evaluates them in light of the proposed framework. Then, it presents a new comparison method that is better fitted to the pre-defined framework. This chapter is based on our ACL 2019 paper [Dror et al., 2019].

The second challenge is crucial for the efforts to extend the reach of NLP technology to multiple domains and languages. These well-justified efforts result in a large number of comparisons between algorithms, across corpora from a large number of languages and domains. The goal of this chapter is to provide the NLP community with a statistical analysis framework, termed Replicability Analysis, which will allow us to draw statistically sound conclusions in evaluation setups that involve multiple comparisons. The classical goal of replicability analysis is to examine the consistency of findings across studies in order to address the basic dogma of science, namely that a finding is more convincingly true if it is replicated in at least one more study [Heller et al., 2014, Patil et al., 2016]. We adapt this goal to NLP, where we wish to ascertain the superiority of one algorithm over another across multiple datasets, which may come from different languages, domains, and genres. This chapter is based on our TACL paper [Dror et al., 2017].

Finally, while this book aims to provide a basic framework for proper statistical significance testing in NLP research, it is by no means the final word on this topic. Indeed, Chapter 7 presents a list of open questions that are still to be addressed in future research. We hope that this book will contribute to the evaluation practices in our community and eventually to the development of more effective NLP technology.

INTENDED READERSHIP

The book is intended for researchers and practitioners in NLP who would like to analyze their experimental results in a statistically sound manner. Hence, we assume technical background in computer science and related areas such as statistics and probability, mostly at the undergraduate level. Moreover, while in Chapter 4 we discuss various NLP tasks and their proposed significance tests, our discussion of these tasks is quite shallow. Furthermore, when we analyze experimental results with NLP tasks in Chapters 5 and 6 we do not provide the details of the tasks because we assume the reader is familiar with the basic tasks of NLP. Despite these assumptions about the reader's background, we are trying as much as possible to be self-contained when it comes

to statistical hypothesis testing and the derived concepts and methodology, as presenting these ideas to the NLP audience is a core objective of this book.

Further Reading For broader and more in-depth reading on the fundamental concepts of statistics, we refer the reader to other existing resources such as Montgomery and Runger [2007] (which provides an engineering perspective) and Johnson and Bhattacharyya [2019]. For further reading on the topic of multiple comparisons in statistics, we recommend the book by Bretz et al. [2016] which demonstrates the basic concepts and provides examples with R code.

This book evolved from a series of conference and journal papers—Dror et al. [2017], Dror et al [2018], Dror et al. [2019]—which have been greatly expanded in order to form this book. First, we added background chapters that discuss the foundations of statistical hypothesis testing and provide the details of the statistical significance tests that we find most relevant for NLP. Then, we take the handbook approach and provide the pseudocode of the various methods discussed throughout the book, along with concrete recommendations and guidelines—our goal is to allow the practitioner to directly and easily implement the methods described in this book. Finally, in Chapter 7, we critically discuss the ideas presented in this book and point to challenges that are yet to be addressed in order to perform statistically sound analysis of NLP experimental results.

FOCUS OF THIS BOOK

This book is intended to be self-contained, presenting the framework of statistical hypothesis testing and its derived concepts and methodology in the context of NLP research. However, the main focus of the book is on this statistical framework and its application to the analysis of NLP experimental results, rather than on providing in-depth coverage of the NLP field.

Most of the book takes the handbook approach and aims to provide concrete solutions to practical problems. As such, it does not provide in-depth technical coverage of statistical hypothesis testing to a level that will allow the reader to propose alternative solutions to those proposed here, or to solve some of the open challenges we point to. Yet, our hope is that highlighting the challenges of statistically sound evaluation of NLP experiments, both those that already have decent solutions and those that are still open, will attract the attention of the community to these issues and facilitate future development of additional methods and techniques.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart
April 2020

Acknowledgments

This book is an outcome of three years of exploration. The journey started with a course by Dr. Marina Bogomolov on multiple hypothesis testing, which was given in the fall of 2017 at the Faculty of Industrial Engineering and Management (IE&M) of the Technion. Marina, as well as Gili Baumer, her M.Sc. student and the tutor of the course at the time, were instrumental in the research that resulted in Chapter 6 of this book.

Many people commented on the ideas we discuss in the book, read drafts of the papers that were eventually extended into this book as well as versions of the book itself, and provided valuable feedback. Among these are David Azriel, Eustasio Del Barrio, Yuval Pinter, David Traum (who, as the program chair of ACL 2019, made a substantial contribution to the shaping of our ideas in Chapter 5), Or Zuk, and the members of the Natural Language Processing Group of the IE&M Faculty of the Technion: Reut Apel, Chen Badler, Eyal Ben David, Amichay Doitch, Ofer Givoli, Amir Feder, Ira Leviant, Rivka (Riki) Malka, Nadav Oved, Guy Rotman, Ram Yasdi, Yftah Ziser, and Dor Zohar.

The anonymous reviewers of the book and original papers provided detailed comments on various aspects of this work, from minor technical details to valuable suggestions on the structure, that dramatically improved its quality. Graeme Hirst, Michael Morgan, and Christine Kiilerich orchestrated the book-writing effort and provided valuable guidance.

Finally, we would like to thank the generous support of the Technion Graduate School. Rotem Dror has also been supported by a generous Google Ph.D. fellowship.

Needless to say that all the mistakes and shortcomings of the book are ours. Please let us know if you find any.

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart
April 2020