

Validity, Reliability, and Significance

Empirical Methods for NLP and Data Science

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.

Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science

Stefan Riezler and Michael Hagmann

2021

Pretrained Transformers for Text Ranking: BERT and Beyond

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates

2021

Automated Essay Scoring

Beata Beigman Klebanov and Nitin Madnani

2021

Explainable Natural Language Processing

Anders Søgaard

2021

Finite-State Text Processing

Kyle Gorman and Richard Sproat

2021

Semantic Relations Between Nominals, Second Edition

Vivi Nastase, Stan Szpakowicz, Preslav Nakov, and Diarmuid Ó Séaghdha

2021

Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning

Mohammad Taher Pilehvar and Jose Camacho-Collados

2020

Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots

Michael McTear

2020

Natural Language Processing for Social Media, Third Edition

Anna Atefeh Farzindar and Diana Inkpen

2020

Statistical Significance Testing for Natural Language Processing

Rotem Dror, Lotem Peled, 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 and Pragmatics

Emily M. Bender and Alex Lascarides

2019

Cross-Lingual Word Embeddings

Anders Søgaard, Ivan Vulić, Sebastian Ruder, 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

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

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.

Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science
Stefan Riezler and Michael Haggmann

ISBN: 978-3-031-01051-4 paperback

ISBN: 978-3-031-02179-4 ebook

ISBN: 978-3-031-00190-1 hardcover

DOI 10.1007/978-3-031-02179-4

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

Lecture #55

Series Editor: Graeme Hirst, *University of Toronto*

Series ISSN

Print 1947-4040 Electronic 1947-4059

Validity, Reliability, and Significance

Empirical Methods for NLP and Data Science

Stefan Riezler

Department of Computational Linguistics
& Interdisciplinary Center for Scientific Computing
Heidelberg University, Heidelberg, Germany

Michael Hagmann

Department of Computational Linguistics
Heidelberg University, Heidelberg, Germany

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #55

ABSTRACT

Empirical methods are means to answering methodological questions of empirical sciences by statistical techniques. The methodological questions addressed in this book include the problems of validity, reliability, and significance. In the case of machine learning, these correspond to the questions of whether a model predicts what it purports to predict, whether a model's performance is consistent across replications, and whether a performance difference between two models is due to chance, respectively. The goal of this book is to answer these questions by concrete statistical tests that can be applied to assess validity, reliability, and significance of data annotation and machine learning prediction in the fields of NLP and data science.

Our focus is on model-based empirical methods where data annotations and model predictions are treated as training data for interpretable probabilistic models from the well-understood families of generalized additive models (GAMs) and linear mixed effects models (LMEMs). Based on the interpretable parameters of the trained GAMs or LMEMs, the book presents model-based statistical tests such as a validity test that allows detecting circular features that circumvent learning. Furthermore, the book discusses a reliability coefficient using variance decomposition based on random effect parameters of LMEMs. Last, a significance test based on the likelihood ratio of nested LMEMs trained on the performance scores of two machine learning models is shown to naturally allow the inclusion of variations in meta-parameter settings into hypothesis testing, and further facilitates a refined system comparison conditional on properties of input data.

This book can be used as an introduction to empirical methods for machine learning in general, with a special focus on applications in NLP and data science. The book is self-contained, with an appendix on the mathematical background on GAMs and LMEMs, and with an accompanying webpage including R code to replicate experiments presented in the book.

KEYWORDS

empirical methods, measurement theory, validity, bias features, circularity, generalized additive models, deviance, nullification, reliability, experimental design, variance components, linear mixed models, orthogonal estimators, significance, likelihood ratio

To Sabine and Janna & Ida.

Contents

	Preface	xv
	Acknowledgments	xvii
1	Introduction	1
	1.1 Empirical Methods in Machine Learning	1
	1.2 Scope and Outline of this Book	3
	1.3 Intended Readership	6
2	Validity	9
	2.1 Validity Problems in NLP and Data Science	9
	2.1.1 Bias Features	9
	2.1.2 Illegitimate Features	10
	2.1.3 Circular Features	11
	2.2 Theories of Measurement and Validity	12
	2.2.1 The Concept of Validity in Psychometrics	12
	2.2.2 The Theory of Scales of Measurement	14
	2.2.3 Theories of Measurement in Philosophy of Science	15
	2.3 Prediction as Measurement	16
	2.3.1 Feature Representations	17
	2.3.2 Measurement Data	18
	2.4 Descriptive and Model-Based Validity Tests	19
	2.4.1 Dataset Bias Test	20
	2.4.2 Transformation Invariance Test	25
	2.4.3 A Model-Based Test for Circularity	28
	2.5 Notes on Practical Usage	53
3	Reliability	55
	3.1 Untangling Terminology: Reliability, Agreement, and Others	55
	3.2 Performance Evaluation as Measurement	56
	3.3 Descriptive and Model-Based Reliability Tests	57
	3.3.1 Agreement Coefficients for Data Annotation	57

	3.3.2	Bootstrap Confidence Intervals for Model Evaluation	61
	3.3.3	Model-Based Reliability Testing	66
	3.4	Notes on Practical Usage	88
4		Significance	91
	4.1	Parametric Significance Tests	93
	4.2	Sampling-Based Significance Tests	97
	4.2.1	Bootstrap Resampling	97
	4.2.2	Permutation Tests	99
	4.3	Model-Based Significance Testing	101
	4.3.1	The Generalized Likelihood Ratio Test	102
	4.3.2	Likelihood Ratio Tests using LMEMs	104
	4.4	Notes on Practical Usage	113
A		Mathematical Background	115
	A.1	Generalized Additive Models	115
	A.1.1	General Form of Model	115
	A.1.2	Example	116
	A.1.3	Parameter Estimation	117
	A.2	Linear Mixed Effects Models	120
	A.2.1	General Form of Model	120
	A.2.2	Example	121
	A.2.3	Parameter Optimization	125
	A.3	The Distribution of the Likelihood Ratio Statistic	126
	A.3.1	Score Function and Fisher Information	126
	A.3.2	Taylor Expansion and Asymptotic Distribution	127
		Bibliography	129
		Authors' Biographies	147

Preface

There is a particular book that accompanied the first author since his days as doctoral student: Paul R. Cohen’s textbook *Empirical Methods for Artificial Intelligence* [Cohen, 1995]. The book was introduced to him by Mark Johnson, with the recommendation that it contained essential information for an empirical researcher that is not easily available in a comparably concise form anywhere else. This assessment of Cohen’s book is still valid today.

Myriad books on machine learning, deep learning, and artificial intelligence have been published since Cohen’s book appeared in 1995. With rare exceptions such as [Hardt and Recht \[2021\]](#), however, questions about data practices, the concepts of validity and reliability, or techniques of exploratory data analysis are not mentioned in contemporary books on machine learning. A discussion of confirmatory techniques for statistical hypothesis testing and their relevance for practical machine learning research is also not integrated in most machine learning textbooks. For these topics, Cohen’s exposition of exploratory and confirmatory techniques of empirical science is still the to-go textbook. However, Cohen’s book has not been updated since its publication date.

The goal of our book is to extend and update Cohen’s book using model-based techniques to address the questions of validity, reliability, and significance in empirical machine learning research. In our book, these techniques are based on interpretable probabilistic models as described in [Wood \[2017\]](#). These models are not necessarily more recent than Cohen’s book, but they possess the necessary expressiveness to model experimental data from data annotation and machine learning prediction experiments, and they are associated with proven statistical properties for drawing inferences about the parameters and models. The goal of our book is to provide the reader with an instrument in the form of model-based statistical tests that enables assessing the methodological questions of validity, reliability, and significance. We showcase our techniques on examples from the authors’ areas of expertise—NLP and medical data science—and hope that the proposed techniques will also be of use to readers from other areas of machine learning and artificial intelligence.

Stefan Riezler and Michael Haggmann
November 2021

Acknowledgments

This book would not have been possible without the help of several people who are actively involved in empirical machine learning research, and who were willing to read early drafts of our book and comment on its relevance to their own work.

Firstly, we would like to thank the students at the departments of computational linguistics and computer science who participated in two iterations of a seminar class on the topics of the book, and who detected and corrected many mistakes in earlier versions of the book. We are indebted to Nathan Berger for proofreading our writing as a native speaker of American English and for comments on the intelligibility of the contents as a Ph.D. student in computer science. We would like to thank Michael Staniek for his critical comments on the coherence of our argumentation and the usefulness of the presented methods for a Ph.D. student in computational linguistics. We are indebted to Mayumi Ohta for testing our R scripts and for contributing to the experimental material that illustrates our statistical tests.

We thank Artem Sokolov for patiently going over several early versions of the book, for a great many discussions on all aspects of our work, and for serving as an endless source of recommendations on related work.

Last, we would like to thank Graeme Hirst and Michael Morgan for giving us the chance to publish this book in the first place and for selecting two excellent reviewers. We thank the anonymous reviewers for providing feedback on all levels of detail of the book and for giving invaluable guidance on how to present our material in a clear and appealing form.

Clearly, various errors and shortcomings remain, and we would be grateful if readers could point them out to us so that they can be corrected.

Stefan Riezler and Michael Haggmann
November 2021