

Statistical Methods for Annotation Analysis



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 Methods for Annotation Analysis](#)

Silviu Paun, Ron Artstein, and Massimo Poesio
2021

[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.

Statistical Methods for Annotation Analysis

Silviu Paun, Ron Artstein, and Massimo Poesio

ISBN: 978-3-031-03753-5 paperback

ISBN: 978-3-031-03763-4 PDF

DOI 10.1007/978-3-031-03763-4

A Publication in the Springer series

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #54

Series Editor: Graeme Hirst, *University of Toronto*

Series ISSN

Print 1947-4040 Electronic 1947-4059

Statistical Methods for Annotation Analysis

Silviu Paun

Queen Mary University of London

Ron Artstein

University of Southern California

Massimo Poesio

Queen Mary University of London and Turing Institute

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #54

ABSTRACT

Labelling data is one of the most fundamental activities in science, and has underpinned practice, particularly in medicine, for decades, as well as research in corpus linguistics since at least the development of the Brown corpus. With the shift towards Machine Learning in Artificial Intelligence (AI), the creation of datasets to be used for training and evaluating AI systems, also known in AI as corpora, has become a central activity in the field as well.

Early AI datasets were created on an *ad-hoc* basis to tackle specific problems. As larger and more reusable datasets were created, requiring greater investment, the need for a more systematic approach to dataset creation arose to ensure increased quality. A range of statistical methods were adopted, often but not exclusively from the medical sciences, to ensure that the labels used were not subjective, or to choose among different labels provided by the coders. A wide variety of such methods is now in regular use. This book is meant to provide a survey of the most widely used among these statistical methods supporting annotation practice.

As far as the authors know, this is the first book attempting to cover the two families of methods in wider use. The first family of methods is concerned with the development of labelling schemes and, in particular, ensuring that such schemes are such that sufficient agreement can be observed among the coders. The second family includes methods developed to analyze the output of coders once the scheme has been agreed upon, particularly although not exclusively to identify the most likely label for an item among those provided by the coders.

The focus of this book is primarily on Natural Language Processing, the area of AI devoted to the development of models of language interpretation and production, but many if not most of the methods discussed here are also applicable to other areas of AI, or indeed, to other areas of Data Science.

KEYWORDS

statistics, corpus annotation, agreement, coefficients of agreement, probabilistic annotation models, variational autoencoders, latent models, neural models for learning from the crowd

In memory of Janyce Wiebe

Contents

	Preface	xvii
	Acknowledgements	xix
1	Introduction	1
	1.1 Reliability and Validity, and Other Issues	2
	1.2 A Very Short Guide to the Probabilistically More Advanced Content in the Book	3
	1.3 The Companion Website	4
	 PART I Analysing Agreement	 7
2	Coefficient of Agreement	9
	2.1 Introduction and Motivations	9
	2.2 Coefficients of Agreement	11
	2.2.1 Agreement, Reliability, and Validity	11
	2.2.2 A Common Notation	11
	2.2.3 The Need for Dedicated Measures of Agreement	12
	2.2.4 Chance-Corrected Coefficients for Measuring Agreement Between Two Coders	16
	2.2.5 More than Two Coders	22
	2.2.6 Krippendorff's Alpha and Other Weighted Agreement Coefficients .	25
	2.2.7 Relations Among Coefficients	30
	2.2.8 An Integrated Example	30
	2.3 Missing Data	33
	2.4 Unitizing or Markable Identification	35
	2.5 Bias and Prevalence	37
	2.5.1 Annotator Bias	37
	2.5.2 Prevalence	40
	2.A Appendix: Proofs for Theorems Presented in This Chapter	42

	2.A.1	Annotator Bias and Variance with Multiple Coders	42
	2.A.2	Annotator Bias for Weighted Measures	43
3		Using Agreement Measures for CL Annotation Tasks	47
	3.1	General Methodological Recommendations	47
	3.1.1	Generating Data to Measure Reproducibility	48
	3.1.2	Establishing Significance	49
	3.1.3	Interpreting the Value of Kappa-Like Coefficients	49
	3.1.4	Agreement and Machine Learning	51
	3.2	Labelling Units with a Common and Predefined Set of Categories	51
	3.2.1	Part-of-Speech Tagging	51
	3.2.2	Dialogue Act Tagging	53
	3.2.3	Named Entities	59
	3.2.4	Other Labelling Tasks	60
	3.3	Marking Boundaries and Unitizing	62
	3.3.1	Segmentation and Topic Marking	62
	3.3.2	Prosody	65
	3.4	Set-Based Labels	66
	3.4.1	Anaphora	66
	3.4.2	Discourse Deixis	69
	3.4.3	Summarization	70
	3.5	Word Senses	72
	3.6	Summary	75
	3.6.1	Methodology	75
	3.6.2	Choosing a Coefficient	76
	3.6.3	Interpreting the Values	77
4		Probabilistic Models of Agreement	79
	4.1	Introduction	79
	4.2	Easy Items, Difficult Items, and Agreement	81
	4.2.1	Aickin's α	82
	4.2.2	Modelling Stability	89
	4.2.3	Coder Stability: A Discussion	90
	4.3	Latent Class Analysis of Agreement Patterns	92
	4.3.1	Varying Panel of Coders	94
	4.3.2	Fixed Panel of Coders	97
	4.3.3	An NLP Case Study	99
	4.4	Summary	101

	PART II	Analysing and Using Crowd Annotations . . .	103
5		Probabilistic Models of Annotation	105
	5.1	Introduction	105
	5.2	Terminology and a Simple Annotation Model	106
	5.3	Modelling Annotator Behaviour	110
	5.4	Modelling Item Difficulty	115
	5.5	Hierarchical Structures	122
	5.6	Adding Features	126
	5.7	Modelling Sequence Labelling Tasks	130
	5.8	Aggregating Anaphoric Annotations	133
	5.9	Aggregation with Variational Autoencoders	137
	5.10	Modelling Complex Annotations	142
6		Learning from Multi-Annotated Corpora	147
	6.1	Introduction	147
	6.2	Learning with Soft Labels	149
	6.3	Learning Individual Coder Models	152
	6.4	Dealing with Noise	155
	6.5	Pooling Coder Confusions	160
	6.6	Summary	163
		Bibliography	167
		Authors' Biographies	197

Preface

When, almost 15 years ago, two of us (Massimo and Ron) completed what was to become our (Artstein and Poesio, 2008) paper, we felt elated at reaching a milestone after three years of hard work. But we were also aware that the stage we had reached, on the one hand, left a number of open questions, particularly about the interpretability of coefficients of agreement. On the other hand, it failed to cover important areas of research within the field of statistical methods for annotation analysis, such as latent models of agreement or probabilistic annotation models, which we felt could make important contributions to data creation and use within Computational Linguistics even though, at the time, it had only begun being experimented with in such pioneering work as Beigman Klebanov and Beigman (2009), Bruce and Wiebe (1999), Carpenter (2008), and Reidsma and Carletta (2008). 2008 was also the year of the seminal paper by Snow et al. (2008) that started the crowdsourcing revolution in Natural Language Processing (NLP), in which such methods were to play an essential role. We were therefore delighted when, many years later, Graeme Hirst and Mike Morgan offered us the opportunity to make further progress along the path begun in that paper. This book is the result of that progress.

As the 2008 paper covered the material in its scope—coefficients of agreement—in a, we thought, reasonably thorough way; and, as even more thorough works covering that material have appeared since, such as Gwet (2014), we didn't attempt to expand our coverage of that topic much in this book. We merely aimed to incorporate some material that had been left out in the original paper and to update the presentation to take into account more recent proposals, e.g., on unitizing. Our effort focused instead primarily on covering methods that had not been covered at all in the 2008 paper, also because meanwhile those techniques have become much more widespread in NLP. We hope this book offers as accessible an introduction to latent models of agreement, probabilistic models of aggregation, and learning directly from multiple coders, as the 2008 paper did for coefficients of agreement.

Silviu Paun, Ron Artstein, and Massimo Poesio
September 2021

Acknowledgements

We are extremely grateful to Graeme Hirst and Mike Morgan for offering us the opportunity to continue our work in this area, and to the many colleagues who collaborated with us on this effort at various stages and/or offered their comments and support. We would especially like to thank Bob Carpenter, Jon Chamberlain, Janosch Haber, Dirk Hovy, Tommaso Fornaciari, Udo Kruschwitz, Chris Madge, Becky Passonneau, Barbara Plank, Edwin Simpson, Alexandra Uma, and Juntao Yu, for much inspiration, feedback, and help over the years. We would also like to thank Jacopo Amidei, Lora Aroyo, Valerio Basile, Beata Beigman-Klebanov, Raffaella Bernardi, Alex Braylan, Chris Callison-Burch, Jean Carletta, Chris Cieri, Barbara di Eugenio, Anca Dumitrache, James Fiumara, Iryna Gurevych, Nancy Ide, Klaus Krippendorff, Walter Lasecki, Matt Lease, Mark Liberman, Yann Mathet, Sabine Schulte im Walde, Yannick Versley, Renata Vieira, Aline Villavicencio, and Janyce Wiebe.

Silviu Paun and Massimo Poesio were supported by the DALI project, ERC Advanced Grant 695662 to Massimo Poesio.

Ron Artstein was sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

Silviu Paun, Ron Artstein, and Massimo Poesio
September 2021