

Automated Essay Scoring

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.

Automated Essay Scoring

Beata Beigman Klebanov and Nitin Madnani
2021

Pretrained Transformers for Text Ranking: BERT and Beyond

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates
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 Faruqi

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.

Automated Essay Scoring

Beata Beigman Klebanov and Nitin Madnani

ISBN: 978-3-031-01054-5 paperback

ISBN: 978-3-031-02182-4 PDF

ISBN: 978-3-031-00193-2 hardcover

DOI 10.1007/978-3-031-02182-4

A Publication in the Springer series

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #52

Series Editor: Graeme Hirst, *University of Toronto*

Series ISSN

Print 1947-4040 Electronic 1947-4059

Automated Essay Scoring

Beata Beigman Klebanov and Nitin Madnani
Educational Testing Service

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #52

ABSTRACT

This book discusses the state of the art of automated essay scoring, its challenges and its potential. One of the earliest applications of artificial intelligence to language data (along with machine translation and speech recognition), automated essay scoring has evolved to become both a revenue-generating industry and a vast field of research, with many subfields and connections to other NLP tasks. In this book, we review the developments in this field against the backdrop of Elias Page’s seminal 1966 paper titled “The Imminence of Grading Essays by Computer”.

Part I establishes what automated essay scoring is about, why it exists, where the technology stands, and what are some of the main issues. In Part II, the book presents guided exercises to illustrate how one would go about building and evaluating a simple automated scoring system, while Part III offers readers a survey of the literature on different types of scoring models, the aspects of essay quality studied in prior research, and the implementation and evaluation of a scoring engine. Part IV offers a broader view of the field inclusive of some neighboring areas, and Part V closes with summary and discussion.

This book grew out of a week-long course on automated evaluation of language production at the North American Summer School for Logic, Language, and Information (NASSLLI), attended by advanced undergraduates and early-stage graduate students from a variety of disciplines. Teachers of natural language processing, in particular, will find that the book offers a useful foundation for a supplemental module on automated scoring. Professionals and students in linguistics, applied linguistics, educational technology, and other related disciplines will also find the material here useful.

KEYWORDS

automated essay scoring, automated writing evaluation, natural language processing, educational technology, artificial intelligence, AES, AWE, NLP, EdTech, AI

*To my Dad, a physicist who likes a good story, and
To my Mom, a doctor who likes a well-written one,
With love,
Beata*

*To my family for their enduring support ... and patience!
Nitin*

Contents

	Preface	xix
	PART I Introduction	1
1	Should We Do It? Can We Do It?	3
1.1	The Case for Automated Scoring of Essays	3
1.1.1	Argument from Need and Positive Consequence	4
1.1.2	Argument from Feasibility I: Computers are Smart	4
1.1.3	Arguments from Feasibility II: Define the Goal	5
1.1.4	Argument from Quality and Utility: High-Quality Low-Cost Large-Scale Scoring	8
1.2	Challenges to Automated Scoring of Essays	9
1.2.1	Anticipated Objection #1: Originality	9
1.2.2	Anticipated Objection #2: Content	10
1.2.3	Anticipated Objection #3: Gaming	11
1.2.4	Anticipated Objection #4: Feedback	12
1.3	Summary	13
	PART II Getting Hands-On	15
2	Building an Automated Essay Scoring System	17
2.1	Introduction	17
2.2	Setting Up	17
2.2.1	Data	17
2.2.2	Model and Features	18
2.2.3	Evaluation Metrics	24
2.2.4	Software	28
2.3	Building the System	28
2.3.1	Experiment 0: Use All Features	29

2.3.2	Experiment 1: Feature Fairness	33
2.3.3	Experiment 2: Feature Collinearity	34
2.3.4	Experiment 3: Additional Evidence for Feature Contributions	34
2.3.5	Experiment 4: Feature Transformations	35
2.3.6	Experiment 5: Negative Feature Contributions	36
2.3.7	Experiment 6: Test on Held-out Data	38
2.3.8	Experiment 7: Cross-Task Evaluation	39
2.3.9	Experiment 8: Task-Specific Models	39
2.3.10	Experiments 9a and 9b: More Reliable Human Ratings	41
2.3.11	Experiment 10: A More Sophisticated Learner	43
2.4	Conclusions	44
3	From Lessons to Guidelines	45
3.1	Introduction	45
3.2	Perspectives on Automated Scoring	45
3.3	Case Studies	49
3.3.1	Adding Automated Scoring to an Existing Assessment	49
3.3.2	Creating a New Assessment that Includes Automated Scoring	51
3.3.3	Including Automated Scoring in a Classroom Setting	52
3.4	Summary	52
3.5	Looking Ahead	53
 PART III A Deep Dive: Models, Features, Architecture, and Evaluation		55
4	Models	57
4.1	Introduction	57
4.2	Linear Regression	57
4.3	Latent Semantic Analysis	60
4.4	Other Non-Neural Models	62
4.5	Neural Networks	63
4.5.1	Deep Learning	65
4.5.2	Interlude: Word Embeddings	66
4.5.3	Deep Learning for Automated Essay Scoring	67
4.5.4	Discussion	73

5	Generic Features	75
5.1	Introduction	75
5.2	Discourse-Level Features	77
5.2.1	Essay Organization	77
5.2.2	Essay Development	78
5.2.3	Coherence	80
5.3	Selection of Content: Vocabulary and Topicality	86
5.3.1	Vocabulary	86
5.3.2	Topicality	88
5.4	Conventions	92
5.4.1	Early Approaches	93
5.4.2	Feature-Driven Supervised Learning	94
5.4.3	Neural Approaches to Detection of Grammatical Errors	98
5.4.4	Grammaticality on a Scale	99
6	Genre- and Task-Specific Features	101
6.1	What's in an Essay?	101
6.2	Persuasive/Argumentative Writing	104
6.2.1	Persuasion vs. Argumentation	104
6.2.2	Features Based on Use of Evaluative Language	106
6.2.3	Features Based on Use of Figurative Language	107
6.2.4	Features Based on Argument Structure	109
6.2.5	Features Based on Argument Content	115
6.2.6	Discussion: Between Content and Structure	122
6.3	Narrative Writing/Convey Experience	124
6.3.1	Scoring Narrative Essays	124
6.3.2	Scoring Transcripts of Oral Narratives	127
6.4	Expository Writing Based on Sources	128
6.4.1	Scoring Human Summaries—An Overview	129
6.4.2	Approaches that Use Source Document as the Sole Reference	130
6.4.3	Approaches that Use Source Document and Expert/Peer Summaries as References	132
6.4.4	Approaches that Use Additional Expert-Provided Materials as References	136
6.4.5	Approaches that Use Transformed Text and/or Expert/Peer Summaries as Reference	138
6.5	Reflective Writing	143

6.6	Other Tasks/Genres	151
6.7	Summary	152
7	Automated Scoring Systems: From Prototype to Production	155
7.1	Introduction	155
7.2	Criteria	156
7.3	Example Architecture	156
7.3.1	Prelude: Apache Storm	156
7.3.2	Architecture Details	157
7.3.3	Evaluation	159
7.3.4	Illustrating the Architecture	160
7.4	Conclusions	163
8	Evaluating for Real-World Use	165
8.1	Introduction	165
8.2	Validity	165
8.3	Fairness	167
8.4	Fairness for Essay Scoring	168
8.5	RSMTTool	169
8.6	Postscript: Connections to FATML	171
 PART IV Further Afield: Feedback, Content, Speech, and Gaming		 173
9	Automated Feedback	175
9.1	What is Feedback?	175
9.2	Feedback Systems	176
9.3	Evaluation of Feedback Systems	179
10	Automated Scoring of Content	181
10.1	Introduction	181
10.2	Approaches	184
10.3	Response-Based Scoring	185
10.3.1	Features	185
10.3.2	Model	186

10.4	Emerging Trend: Deep Neural Networks	186
10.5	Summary	187
11	Automated Scoring of Speech	189
11.1	Introduction	189
11.2	Automated Speech Recognition for Speech Scoring	190
11.3	Features for Assessing Spontaneous Speech	191
11.3.1	Delivery: Pronunciation, Fluency	191
11.3.2	Language Use: Vocabulary, Grammar	193
11.3.3	Topic Development: Content, Discourse	195
11.4	Scoring Models	197
12	Fooling the System: Gaming Strategies	199
12.1	Introduction	199
12.2	Shell Language	200
12.3	Artificially Generated Essays	201
12.4	Off-Topic Responses	203
12.5	Plagiarism	204
12.6	Other Related Work	205
12.7	Summary	207
	PART V Summary and Discussion	209
13	Looking Back, Looking Ahead	211
13.1	Report Card: Where are We Now?	211
13.1.1	Accomplishments	211
13.1.2	Needs Improvement	212
13.2	Going off the Page	214
13.2.1	Assessing Writing in Multiple Languages	214
13.2.2	Standardized Testing	214
13.2.3	Increased Attention to Fairness	215
13.2.4	Pervasiveness of Technology	216
13.3	Discussion	216
13.3.1	Support Consequential Decision Making	216
13.3.2	Create a Better Written Product	217

13.3.3 Help the User Learn to Write Better	218
13.3.4 Relationships Between Types of Use	218
13.4 Conclusion	219
Definitions-in-Context	221
Index	223
References	229
Authors' Biographies	293

Preface

This book grew out of a week-long course on automated evaluation of language production that we gave at the North American Summer School for Logic, Language, and Information in 2016. As anyone who has given a natural language processing (NLP) course for NASSLLI or ESSLLI knows, one generally expects early- or middle-stage graduate students from a variety of disciplines, or advanced undergraduates; some students are looking to merely familiarize themselves with a sub-field, others want to understand how things are done in the sub-field in a hands-on fashion, and still others who already have substantial computer science training want some help in understanding the current research frontier.

In an attempt to cater to all these different types of audiences, we wrote a book that can be read in different ways by different readers. In Part I, we provide an introduction addressing the first set of questions on anybody's mind when discovering a new field of endeavor—What is it about? Why does it exist? Where does it stand? What are some of the main issues? We aimed for an introduction that isn't too technical for people without much technical background, yet that would offer a reader with such background a new perspective that would not be commonly taught in a computer science, linguistics, education, or natural language processing course—a historical perspective that highlights the hopes and promise of this line of endeavor, along with long-known and more recently discovered challenges.

The paths of different readers through the book might diverge from this point on. In Part II, we address the eager newcomer who is ready to give this sub-field some more attention; we provide a guided set of exercises to illustrate how one would go about building and evaluating an automated scoring system, and the kinds of issues one is likely to encounter (Chapter 2), followed by a broader discussion of lessons learned (Chapter 3). A reader who wanted a glimpse and a taste, without having background in either computer science or linguistics, and without the intention of engaging with the research frontier in the area, might be best served by skimming Part III (Chapters 4–8) and focusing next on Part IV, in order to get a broader view of the field—inclusive of some neighboring areas—and moving to Part V for the overall summary and discussion.

If you are a reader who is already reasonably well versed in the standard machine learning material, is perhaps taking advanced undergraduate- and graduate-level courses in NLP, and is looking for a more technical orientation into the state of the field and the relevant literature—Part III is written for you. We consider different types of scoring models (Chapter 4), describe in detail the specific aspects of essay quality that have been studied in prior research—both those that cover a relatively generic construct of a good quality essay (Chapter 5) and those that were built to address more specific requirements of task or genre (Chapter 6). Chapters 7

and 8 discuss the implementation of a scoring engine and the evaluation of its performance, respectively. The subsequent Part IV, where a variety of neighboring fields are discussed briefly, might be familiar to this type of reader, and so, perhaps, are to be skimmed on the way to the summary and discussion in Part V.

This book is not intended as a stand-alone textbook that would take the reader with no knowledge of computer science, linguistics, NLP and psychometrics to the cutting edge of current research. We believe that some of the perspectives offered in this book would help teachers of courses in natural language processing who consider including automated scoring as one of the modules/exercises to supplement instruction. We also hope that teachers of courses in linguistics, applied linguistics, educational technology, and other related disciplines would find the material here useful for supplementing their instruction. The large majority of the work discussed in this book has been done on essays in English; we provide some pointers to work on automated essay scoring in other languages in Part V.

Finally, we want to say a few words on the typographical emphases we will use in the book. We will use small caps—like THIS—when an important concept is most fully explained; these locations are linked from the Definitions-in-Context section. You can track mentions of these concepts throughout the book using the index. We will use italics for local emphasis, and boldface to draw the reader’s attention to an important statement or finding, as well as for tracking some chapter-level organizational elements.

Beata Beigman Klebanov and Nitin Madnani
August 2021