

Automated Grammatical Error Detection for Language Learners

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.

Automated Grammatical Error Detection for Language Learners, Second Edition

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault

2014

Ontology-Based Interpretation of Natural Language

Philip Cimiano, Christina Unger, and John McCrae

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 2014

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 Grammatical Error Detection for Language Learners, Second Edition

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault

ISBN: 978-3-031-01025-5 paperback

ISBN: 978-3-031-02153-4 ebook

DOI 10.1007/978-3-031-02153-4

A Publication in the Springer series

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #25

Series Editor: Graeme Hirst, *University of Toronto*

Series ISSN

Synthesis Lectures on Human Language Technologies

Print 1947-4040 Electronic 1947-4059

Automated Grammatical Error Detection for Language Learners

Second Edition

Claudia Leacock
CTB McGraw-Hill

Martin Chodorow
Hunter College and the Graduate Center, City University of New York

Michael Gamon
Microsoft Research

Joel Tetreault
Yahoo! Labs

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #25

ABSTRACT

It has been estimated that over a billion people are using or learning English as a second or foreign language, and the numbers are growing not only for English but for other languages as well. These language learners provide a burgeoning market for tools that help identify and correct learners' writing errors. Unfortunately, the errors targeted by typical commercial proofreading tools do not include those aspects of a second language that are hardest to learn.

This volume describes the types of constructions English language learners find most difficult—constructions containing prepositions, articles, and collocations. It provides an overview of the automated approaches that have been developed to identify and correct these and other classes of learner errors in a number of languages.

Error annotation and system evaluation are particularly important topics in grammatical error detection because there are no commonly accepted standards. Chapters in the book describe the options available to researchers, recommend best practices for reporting results, and present annotation and evaluation schemes.

The final chapters explore recent innovative work that opens new directions for research. It is the authors' hope that this volume will continue to contribute to the growing interest in grammatical error detection by encouraging researchers to take a closer look at the field and its many challenging problems.

KEYWORDS

grammatical error detection, statistical natural language processing, learner corpora, linguistic annotation

*Martin Chodorow: To Mamie,
and to the memory of my parents.*

*Claudia Leacock: To my daughters,
Tess Elspeth Dougherty and Tracy Duva Dougherty.*

Contents

	Acknowledgments	xv
1	Introduction	1
1.1	Introduction to the Second Edition	1
1.2	New to the Second Edition	1
1.3	Working Definition of <i>Grammatical Error</i>	2
1.4	Prominence of Research on English Language Learners	3
1.5	Some Terminology	3
1.6	Automated Grammatical Error Detection: NLP and CALL	4
1.7	Intended Audience	4
1.8	Outline	5
2	Background	7
2.1	In the Beginning	7
2.2	Introduction to Data-Driven and Hybrid Approaches	13
3	Special Problems of Language Learners	17
3.1	Errors Made by English Language Learners	17
3.2	The Influence of L1	21
3.3	Challenges for English Language Learners	22
3.3.1	The English Preposition System	22
3.3.2	The English Article System	25
3.3.3	English Collocations	27
3.4	Summary	28
4	Evaluating Error Detection Systems	31
4.1	Traditional Evaluation Measures	32
4.2	Evaluation Measures for Shared Tasks	36
4.3	Evaluation Using a Corpus of Correct Usage	37
4.4	Evaluation on Learner Writing	38
4.4.1	Verifying Results on Learner Writing	38

4.4.2	Evaluation on Fully Annotated Learner Corpora	40
4.4.3	Using Multiple Annotators and Crowdsourcing for Evaluation	42
4.5	Statistical Significance Testing	43
4.6	Checklist for Consistent Reporting of System Results	44
4.7	Summary	45
5	Data-Driven Approaches to Articles and Prepositions	47
5.1	Extracting Features from Training Data	48
5.2	Types of Training Data	49
5.2.1	Training on Well-Formed Text	50
5.2.2	Artificial Errors	51
5.2.3	Error-Annotated Learner Corpora	52
5.2.4	Comparing Training Paradigms	53
5.3	Methods	54
5.3.1	Classification	54
5.3.2	<i>N</i> -gram Statistics, Language Models, and Web Counts	55
5.3.3	Web-Based Methods	57
5.4	Two End-To-End Systems: <i>Criterion</i> and <i>MSR ESL Assistant</i>	59
5.5	Summary	63
6	Collocation Errors	65
6.1	Defining Collocations	65
6.2	Measuring the Strength of Association between Words	66
6.3	Systems for Detecting and Correcting Collocation Errors	70
7	Different Errors and Different Approaches	75
7.1	Heuristic Rule-Based Approaches	75
7.1.1	Criterion System	76
7.1.2	ESL Assistant	77
7.1.3	Other Heuristic Rule-Based Approaches	77
7.2	More Complex Verb Form Errors	79
7.3	Spelling Errors	80
7.4	Punctuation Errors	82
7.5	Detection of Ungrammatical Sentences	82
7.6	Summary	84

8	Annotating Learner Errors	87
8.1	Issues with Learner Error Annotation	87
8.1.1	Number of Annotators	87
8.1.2	Annotation Schemes	88
8.1.3	How to Correct an Error	89
8.1.4	Annotation Approaches	89
8.1.5	Annotation Tools	91
8.2	Annotation Schemes	91
8.2.1	Examples of Comprehensive Annotation Schemes	91
8.2.2	Example of a Targeted Annotation Scheme	92
8.3	Proposals for Efficient Annotation	93
8.3.1	Sampling Approach with Multiple Annotators	93
8.3.2	Crowdsourcing Annotations	94
8.3.3	Mining Online Community-Driven Revision Logs	97
8.4	Summary	98
9	Emerging Directions	99
9.1	Shared Tasks in Grammatical Error Correction	99
9.1.1	The 2011 HOO Task	100
9.1.2	The 2012 HOO Task	101
9.1.3	The CoNLL 2013 Shared Task	103
9.1.4	Summary	105
9.2	Machine Translation and Error Correction	105
9.2.1	Noisy Channel Model	106
9.2.2	Round Trip Machine Translation (RTMT)	107
9.3	Real-Time Crowdsourcing of Grammatical Error Correction	108
9.4	Does Automated Error Feedback Improve Writing?	109
10	Conclusion	113
A	Appendix A: Learner Corpora	115
A.1	Basque	115
A.2	English	115
A.3	Finnish	120
A.4	French	121
A.5	German	121
A.6	Spanish	121
A.7	Multiple Languages	122

Bibliography	123
Authors' Biographies	153

Acknowledgments

We are grateful to Øistein E. Andersen, Aoife Cahill, Robert Dale, Michael Flor, Jennifer Foster, Ross Israel, John Lee, Nitin Madnani, Hwee Tou Ng, and an anonymous reviewer for their feedback and helpful suggestions.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
January 2014