

Automatic Text Simplification

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.

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 2017

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.

Automatic Text Simplification

Horacio Saggion

ISBN: 978-3-031-01038-5 paperback

ISBN: 978-3-031-02166-4 ebook

DOI 10.1007/978-3-031-02166-4

A Publication in the Springer series

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #32

Series Editor: Graeme Hirst, *University of Toronto*

Series ISSN

Print 1947-4040 Electronic 1947-4059

Automatic Text Simplification

Horacio Saggion

Department of Information and Communication Technologies
Universitat Pompeu Fabra

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #32

ABSTRACT

Thanks to the availability of texts on the Web in recent years, increased knowledge and information have been made available to broader audiences. However, the way in which a text is written—its vocabulary, its syntax—can be difficult to read and understand for many people, especially those with poor literacy, cognitive or linguistic impairment, or those with limited knowledge of the language of the text. Texts containing uncommon words or long and complicated sentences can be difficult to read and understand by people as well as difficult to analyze by machines. Automatic text simplification is the process of transforming a text into another text which, ideally conveying the same message, will be easier to read and understand by a broader audience. The process usually involves the replacement of difficult or unknown phrases with simpler equivalents and the transformation of long and syntactically complex sentences into shorter and less complex ones. Automatic text simplification, a research topic which started 20 years ago, now has taken on a central role in natural language processing research not only because of the interesting challenges it possesses but also because of its social implications. This book presents past and current research in text simplification, exploring key issues including automatic readability assessment, lexical simplification, and syntactic simplification. It also provides a detailed account of machine learning techniques currently used in simplification, describes full systems designed for specific languages and target audiences, and offers available resources for research and development together with text simplification evaluation techniques.

KEYWORDS

syntactic simplification, lexical simplification, readability measures, text simplification systems, text simplification evaluation, text simplification resources

To Sandra, Jonas, Noah, and Isabella

Contents

	Acknowledgments	xv
1	Introduction	1
1.1	Text Simplification Tasks	1
1.2	How are Texts Simplified?	2
1.3	The Need for Text Simplification	3
1.4	Easy-to-read Material on the Web	5
1.5	Structure of the Book	6
2	Readability and Text Simplification	7
2.1	Introduction	7
2.2	Readability Formulas	8
2.3	Advanced Natural Language Processing for Readability Assessment	9
2.3.1	Language Models	10
2.3.2	Readability as Classification	10
2.3.3	Discourse, Semantics, and Cohesion in Assessing Readability	12
2.4	Readability on the Web	14
2.5	Are Classic Readability Formulas Correlated?	15
2.6	Sentence-level Readability Assessment	16
2.7	Readability and Autism	18
2.8	Conclusion	19
2.9	Further Reading	19
3	Lexical Simplification	21
3.1	A First Approach	21
3.2	Lexical Simplification in LexSiS	22
3.3	Assessing Word Difficulty	24
3.4	Using Comparable Corpora	25
3.4.1	Using Simple English Wikipedia Edit History	25
3.4.2	Using Wikipedia and Simple Wikipedia	25
3.5	Language Modeling for Lexical Simplification	26

3.6	Lexical Simplification Challenge	28
3.7	Simplifying Numerical Expressions in Text	29
3.8	Conclusion	30
3.9	Further Reading	31
4	Syntactic Simplification	33
4.1	First Steps in Syntactic Simplification	33
4.2	Syntactic Simplification and Cohesion	34
4.3	Rule-based Syntactic Simplification using Syntactic Dependencies	36
4.4	Pattern Matching over Dependencies with JAPE	37
4.5	Simplifying Complex Sentences by Extracting Key Events	40
4.6	Conclusion	43
4.7	Further Reading	44
5	Learning to Simplify	47
5.1	Simplification as Translation	47
5.1.1	Learning Simple English	48
5.1.2	Facing Strong Simplifications	49
5.2	Learning Sentence Transformations	49
5.3	Optimizing Rule Application	55
5.4	Learning from a Semantic Representation	57
5.5	Conclusion	58
5.6	Further Reading	58
6	Full Text Simplification Systems	59
6.1	Text Simplification in PSET	59
6.2	Text Simplification in Simplext	60
6.2.1	Rule-based “Lexical” Simplification	63
6.2.2	Computational Grammars for Simplification	64
6.2.3	Evaluating Simplext	67
6.3	Text Simplification in PorSimples	67
6.3.1	An Authoring Tool with Simplification Capabilities	69
6.4	Conclusion	70
6.5	Further Reading	70

7	Applications of Automatic Text Simplification	71
7.1	Simplification for Specific Target Populations	71
7.1.1	Automatic Text Simplification for Reading Assistance	71
7.1.2	Simplification for Dyslexic Readers	72
7.1.3	Simplification-related Techniques for People with Autism Spectrum Disorder	72
7.1.4	Natural Language Generation for Poor Readers	73
7.2	Text Simplification as NLP Facilitator	73
7.2.1	Simplification for Parsing	73
7.2.2	Simplification for Information Extraction	74
7.2.3	Simplification in and for Text Summarization	74
7.2.4	Simplifying Medical Literature	75
7.2.5	Retrieving Facts from Simplified Sentences	75
7.2.6	Simplifying Patent Documents	76
7.3	Conclusion	76
7.4	Further Reading	77
8	Text Simplification Resources and Evaluation	79
8.1	Lexical Resources for Simplification Applications	79
8.2	Lexical Simplification Resources	80
8.3	Corpora	83
8.4	Non-English Text Simplification Datasets	86
8.5	Evaluation	90
8.6	Toward Automatically Measuring the Quality of Simplified Output	92
8.7	Conclusion	93
8.8	Further Reading	93
9	Conclusion	95
	Bibliography	97
	Author's Biography	121

Acknowledgments

I am indebted to my fellow colleagues Stefan, Sanja, Biljana, Susana, Luz, Daniel, Simon, and Montserrat for sharing their knowledge and expertise with me.

Horacio Saggion
January 2017