Cross-Lingual Word Embeddings

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.

Cross-Lingual Word Embeddings
Anders Søgaard, Ivan Vulić, Sebastian Ruder, and 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 2019

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.

Cross-Lingual Word Embeddings

Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui

ISBN: 978-3-031-01043-9 paperback
ISBN: 978-3-031-02171-8 ebook
ISBN: 978-3-031-00182-6 hardcover

DOI 10.1007/978-3-031-02171-8

A Publication in the Springer series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #42
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Cross-Lingual Word Embeddings

Anders Søgaard
University of Copenhagen

Ivan Vulić
University of Cambridge

Sebastian Ruder
DeepMind

Manaal Faruqui
Google Assistant

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #42

ABSTRACT

The majority of natural language processing (NLP) is English language processing, and while there is good language technology support for (standard varieties of) English, support for Albanian, Burmese, or Cebuano—and most other languages—remains limited. Being able to bridge this digital divide is important for scientific and democratic reasons but also represents an enormous growth potential. A key challenge for this to happen is learning to align basic meaning-bearing units of different languages.

In this book, the authors survey and discuss recent and historical work on supervised and unsupervised learning of such alignments. Specifically, the book focuses on so-called cross-lingual word embeddings. The survey is intended to be systematic, using consistent notation and putting the available methods on comparable form, making it easy to compare wildly different approaches. In so doing, the authors establish previously unreported relations between these methods and are able to present a fast-growing literature in a very compact way. Furthermore, the authors discuss how best to evaluate cross-lingual word embedding methods and survey the resources available for students and researchers interested in this topic.

KEYWORDS

natural language processing, machine learning, semantics, cross-lingual learning

Contents

Preface . . . . . **xi**

1 Introduction **1**

2 Monolingual Word Embedding Models **9**

3 Cross-Lingual Word Embedding Models: Typology **13**

4 A Brief History of Cross-Lingual Word Representations **21**

4.1 Cross-Lingual Word Representations using Bilingual Lexicons 24
4.2 Cross-Lingual Word Embeddings and Word Alignments 27
4.3 Representations Based on Latent and Explicit Cross-Lingual Concepts 29
4.4 Summary 32

5 Word-Level Alignment Models **33**

5.1 Word-level Alignment Methods with Parallel Data 33
5.1.1 Mapping-Based Approaches 34
5.1.2 Pseudo-Mixing Approaches 41
5.1.3 Joint Approaches 42
5.1.4 Sometimes Mapping, Joint, and Pseudo-Bilingual Approaches are Equivalent 44
5.2 Word-Level Alignment Methods with Comparable Data 46

6 Sentence-Level Alignment Methods **49**

6.1 Sentence-Level Methods with Parallel Data 49
6.2 Sentence Alignment with Comparable Data 53

7 Document-Level Alignment Models **55**

7.1 Document Alignment with Comparable Data 55

8 From Bilingual to Multilingual Training **59**

8.1 Multilingual Word Embeddings from Word-Level Information 60
8.2 Multilingual Word Embeddings from Sentence-Level and Document-Level Information 64

9 Unsupervised Learning of Cross-Lingual Word Embeddings 67
9.1 Seed Dictionary Induction 68
9.2 Refinement and Heuristics 72
9.3 Limitations of Unsupervised Approaches 74

10 Applications and Evaluation 75
10.1 Intrinsic Evaluation 75
10.2 Extrinsic Evaluation Through Cross-Lingual Transfer 76
10.3 Multi-Modal and Cognitive Approaches to Evaluation 80

11 Useful Data and Software 83
11.1 Monolingual Resources 83
11.2 Cross-Lingual Data 84
11.3 Cross-Lingual Word Embedding Models 84
11.4 Evaluation and Application 86

12 General Challenges and Future Directions 89

Bibliography 93

Authors' Biographies 119

Preface

In recent decades, we have witnessed an increasing *digital language divide*, that is, large differences in what information and services are accessible to people around the world, depending on their mother tongue. Learning cross-lingual word embeddings facilitates transfer of technologies between languages, narrowing this divide. The last five years has seen a surge in cross-lingual word embedding algorithms, many of which lend intuitions, algorithms, and terminology from neighboring fields, including computer vision, graph theory, and information retrieval. Since these algorithms bring different baggage, they often seem very different at surface level. This book surveys cross-lingual word embedding algorithms and aims to show that many of these algorithms are, in fact, quite similar.

Thanks to our editors and anonymous reviewers for providing helpful feedback on our first draft of this book. Its shortcomings are entirely our fault: The book is intended to be *both* readable by first-year M.Sc. students *and* interesting to an expert audience. Our intention is to introduce what is necessary to appreciate the major challenges we face in cross-lingual NLP, without wasting too much time on fundamentals in linguistic typology, linear algebra, graph theory, or machine learning. In cross-lingual NLP, data is so noisy and biased that few theoretical guarantees can be established, and we are typically left with our gut feelings and a catalog of crazy ideas. We hope this book will provide its readers with both.

Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui
May 2019