Semantic Similarity from Natural Language and Ontology 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.

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 iv

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 2015

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.

Semantic Similarity from Natural Language and Ontology Analysis Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain

ISBN: 978-3-031-01028-6 paperback ISBN: 978-3-031-02156-5 ebook

DOI 10.1007/978-3-031-02156-5

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

Lecture #27 Series Editor: Graeme Hirst, *University of Toronto* Series ISSN Print 1947-4040 Electronic 1947-4059

Semantic Similarity from Natural Language and Ontology Analysis

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain École des mines d'Alès – LGI2P

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #27

ABSTRACT

Artificial Intelligence federates numerous scientific fields in the aim of developing machines able to assist human operators performing complex treatments—most of which demand high cognitive skills (e.g. learning or decision processes). Central to this quest is to give machines the ability to estimate the *likeness* or *similarity* between *things* in the way human beings estimate the similarity between stimuli.

In this context, this book focuses on semantic measures: approaches designed for comparing semantic entities such as units of language, e.g. words, sentences, or concepts and instances defined into knowledge bases. The aim of these measures is to assess the similarity or relatedness of such semantic entities by taking into account their semantics, i.e. their meaning—intuitively, the words *tea* and *coffee*, which both refer to stimulating beverage, will be estimated to be more semantically similar than the words *toffee* (confection) and *coffee*, despite that the last pair has a higher syntactic similarity. The two state-of-the-art approaches for estimating and quantifying semantic similarities/relatedness of semantic entities are presented in detail: the first one relies on corpora analysis and is based on Natural Language Processing techniques and semantic models while the second is based on more or less formal, computer-readable and workable forms of knowledge such as semantic networks, thesauri or ontologies.

Semantic measures are widely used today to compare units of language, concepts, instances or even resources indexed by them (e.g., documents, genes). They are central elements of a large variety of Natural Language Processing applications and knowledge-based treatments, and have therefore naturally been subject to intensive and interdisciplinary research efforts during last decades. Beyond a simple inventory and categorization of existing measures, the aim of this monograph is to convey novices as well as researchers of these domains toward a better understanding of semantic similarity estimation and more generally semantic measures. To this end, we propose an in-depth characterization of existing proposals by discussing their features, the assumptions on which they are based and empirical results regarding their performance in particular applications. By answering these questions and by providing a detailed discussion on the foundations of semantic measures, our aim is to give the reader key knowledge required to: (i) select the more relevant methods according to a particular usage context, (ii) understand the challenges offered to this field of study, (iii) distinguish room of improvements for state-of-the-art approaches and (iv) stimulate creativity toward the development of new approaches. In this aim, several definitions, theoretical and practical details, as well as concrete applications are presented.

KEYWORDS

semantic similarity, semantic relatedness, semantic measures, distributional measures, domain ontology, knowledge-based semantic measure

Contents

	Prefa	ace			
	Ackı	nowledgments xv			
1	Introduction to Semantic Measures 1				
	1.1	Semantic Measures in Action			
		1.1.1 Natural Language Processing			
		1.1.2 Knowledge Engineering, Semantic Web, and Linked Data			
		1.1.3 Biomedical Informatics and Bioinformatics			
		1.1.4 Other applications			
	1.2	From Similarity toward Semantic Measures			
		1.2.1 Human Cognition, Similarity, and Existing Models			
		1.2.2 Definitions of Semantic Measures and Related Vocabulary			
		1.2.3 From Distance and Similarities to Semantic Measures			
	1.3	Classification of Semantic Measures			
		1.3.1 How to Classify Semantic Measures			
		1.3.2 A General Classification of Semantic Measures			
2	Corpus-Based Semantic Measures				
	2.1	From Text Analysis to Semantic Measures			
	2.2	Semantic Evidence of Word Similarity in Natural Language			
		2.2.1 The Meaning of Words			
		2.2.2 Structural Relationships: Paradigmatic and Syntagmatic			
		2.2.3 The Notion of Context			
		2.2.4 Distributional Semantics			
	2.3	Distributional Measures			
		2.3.1 Implementation of the Distributional Hypothesis			
		2.3.2 From Distributional Model to Word Similarity			
		2.3.3 Capturing Deeper Co-Occurrences			
	2.4	Other Corpus-Based Measures			
	2.5	Advantages and Limits of Corpus-Based Measures			
		2.5.1 Advantages of Corpus-Based Measures			

		2.5.2 Limits of Corpus-Based Measures
	2.6	Conclusion
3	Know	vledge-Based Semantic Measures 57
	3.1	Ontologies as Graphs and Formal Notations
		3.1.1 Ontologies as Graphs
		3.1.2 Relationships
		3.1.3 Graph Traversals
		3.1.4 Notations for Taxonomies
	3.2	Types of Semantic Measures and Graph Properties
		3.2.1 Semantic Measures on Cyclic Semantic Graphs
		3.2.2 Semantic Measures on Acyclic Graphs
	3.3	Semantic Evidence in Semantic Graphs and their Interpretations
		3.3.1 Semantic Evidence in Taxonomies
		3.3.2 Concept Specificity
		3.3.3 Strength of Connotations between Concepts
	3.4	Semantic Similarity between a Pair of Concepts
		3.4.1 Structural Approach
		3.4.2 Feature-Based Approach
		3.4.3 Information Theoretical Approach94
		3.4.4 Hybrid Approach
		3.4.5 Considerations when Comparing Concepts in Semantic Graphs96
		3.4.6 List of Pairwise Semantic Similarity Measures
	3.5	Semantic Similarity between Groups of Concepts
		3.5.1 Direct Approach
		3.5.2 Indirect Approach 109
		3.5.3 List of Groupwise Semantic Similarity Measures
	3.6	Other knowledge-Based Measures114
		3.6.1 Semantic Measures Based on Logic-Based Semantics
		3.6.2 Semantic Measures for Multiple Ontologies
	3.7	Advantages and Limits of Knowledge-Based Measures
	3.8	Mixing Knowledge-Based and Corpus-Based Approaches
		3.8.1 Generalities
		3.8.2 Wikipedia-Based Measure: How to Benefit from Structured
		Encyclopedia Knowledge 119
	3.9	Conclusion

x

4	Met	hods and Datasets for the Evaluation of Semantic Measures 131		
	4.1	A General Introduction to Semantic Measure Evaluation		
	4.2	Criteria for Semantic Measure Evaluation		
		4.2.1 Accuracy, Precision, and Robustness		
		4.2.2 Computational Complexity		
		4.2.3 Mathematical Properties		
		4.2.4 Semantics		
		4.2.5 Technical Details		
	4.3	Existing Protocols and Datasets		
		4.3.1 Protocols Used to Compare Measures		
		4.3.2 Datasets		
	4.4	Discussions		
5	Conclusion and Research Directions 159			
A	Examples of Syntagmatic Contexts			
B	A Brief Introduction to Singular Value Decomposition			
C	A Br	A Brief Overview of Other Models for Representing Units of Language 179		
D	Software Tools and Source Code Libraries 187			
	Bibliography 197			
	Authors' Biographies			

xi

Preface

In the last decades, numerous researchers from different domains have developed and studied the notion of semantic measure and more specifically the notions of semantic similarity and semantic relatedness. Indeed, from the biomedical domain, where ontologies and conceptual annotations abound—e.g., genes are characterised by concepts from the Gene Ontology, scientific articles are indexed by terms defined into the Medical Subject Heading thesaurus (MeSH)—to Natural Language Processing (NLP) where text mining requires the semantics of units of language to be compared, researchers provided a vast body of research related to semantic measures: algorithms and approaches designed in the aim of comparing concepts, instances characterised by concepts and units of language w.r.t their meaning. Despite the vast literature dedicated to the domain, most of which is related to the definition of new measures, no extensive introduction proposes to highlight the large diversity of contributions which have been proposed so far. In this context, understanding the foundations of these measures, knowing the numerous approaches which have been proposed and distinguishing those to use in particular application contexts is challenging.

This book proposes an extended introduction to semantic measures targeting both students and domain experts. The aim is to provide a general introduction to the diversity of semantic measures in order to distinguish the central notions and the key concepts of the domain. In a second step, we present the two main families of measures to further discuss technical details related to specific implementations. By organizing information about measures and by providing references to key research papers, our aim is to improve semantic measure understanding, to facilitate their use and to provide a condensed overview of state-of-the-art contributions related to the domain.

The first chapter introduces the motivations which highlight the importance of studying semantic measures. Starting by presenting various applications that benefit from semantic measures in different usage contexts, it then guides the reader toward a deeper understanding of those measures. Intuitive notions and the vocabulary commonly used in the literature are introduced. We present in particular the central notions of semantic relatedness, semantic similarity, semantic distance. More formal definitions and properties used for studying semantic measures are also proposed. Next, these definitions and properties are used to characterise the broad diversity of measures which have been introduced in the literature. A classification of semantic measures is then proposed; it distinguishes the two main approaches corresponding to corpus-based and knowledge-based semantic measures. These two families of semantic measures are further presented in detail in Chapter 2 and Chapter 3, respectively. The foundations of these measures and several implementations which have been proposed in the literature are discussed—software tools enabling practical use of measures are also presented in appendix. Chapter 4 is dedicated to semantic measures evaluation and selection. It presents several aspects of measures that can

xiv PREFACE

be studied for their comparison, as well as state-of-the-art protocols and datasets used for their evaluation. Finally, Chapter 5 concludes by summarizing important notions which are introduced in this book, and by highlighting several important research directions which must be studied for improving both semantic measures and their understanding.

By following this progression, we hope that the reader will find a detailed and stimulating introduction to semantic measures. Our aim is to give the reader access to an extensive state-of-the-art of this field, as well as key knowledge required to: (i) select the more relevant methods according to a particular usage context, (ii) understand the challenges offered to this field of study, (iii) distinguish room of improvements for state-of-the-art approaches and (iv) stimulate creativity toward the development of new approaches.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain Nîmes – France, May 2015

Acknowledgments

The authors would like to express their sincere gratitude towards Mike Morgan and Graeme Hirst, respectively the publisher and editor of the Synthesis Lectures on Human Language Technologies, for giving us the opportunity to share our work and expertise on the topic of semantic similarity through this book. We have really appreciated their involvement and professionalism in orchestrating this project. We also warmly thank the associated publisher collaborators, CL Tondo and Samantha Draper, for their involvement and their support during the editing process. We also wish to express a special thank you to Jane Hayward for her remarkable work and enthusiasm in making this book more pleasant to read by improving our English, within the realms of possibility.

This book has significantly benefited from the relevant and constructive remarks and suggestions provided by the anonymous reviewers. We are deeply grateful to their work. We would also like to thank David Sánchez, Montserrat Batet, Jérôme Euzenat, and Pascale Kuntz for their numerous comments and insightful suggestions on early versions of technical contents introduced in this book.

Finally, we would like to thank our families for their constant encouragement and support during the preparation of this book, as well as their understanding regarding the long working time.

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain May 2015