Book Reviews

Semantic Similarity from Natural Language and Ontology Analysis

Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain
(École des Mines d'Alès - LGI2P)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 27), 2015, xv+238 pp; paperback, ISBN 978-1-62705-446-1; e-book, ISBN 978-1-62705-447-8; doi:10.2200/S00639ED1V01Y201504HLT027, $75.00

Reviewed by
Deyi Xiong
Soochow University

Downloaded from http://direct.mit.edu/coli/article-pdf/42/4/829/1807507/coli_r_00269.pdf by guest on 11 August 2024

Learning semantic similarity for units of language or concepts is crucial not only for numerous tasks in computational linguistics, but also for language understanding and reasoning in the broad context of artificial intelligence. With growing interests and efforts in modeling and computing semantic measures in recent years, we have witnessed much progress in the following two strands of research that approach semantic similarity: corpus-based statistical methods and knowledge-enhanced methods with human knowledge defined in ontologies. This book by Harispe, Ranwez, Janaqi, and Montmain provides a detailed introduction to state-of-the-art research in these two lines of work.

The book is clearly organized into five chapters and four appendixes, covering the motivation, notion, and classification of semantic measures, corpus-based and knowledge-based methodologies in semantic measure modeling, evaluations, data sets, tools, challenges, and future directions.

The book begins with a solid introduction of the notion of semantic measure. It highlights the importance of semantic measures from a broad perspective of artificial intelligence and provides the mathematical and cognitive foundation of semantic measures. In order to give general definitions of semantic measures, including semantic relatedness and semantic similarity, related concepts such as semantic proxy and differences from traditional distance functions are introduced in Chapter 1. This chapter also provides a short description of applications of semantic measures in various fields (natural language processing, information retrieval, semantic web, linked data, biomedical informatics, etc.). It is very useful for the reader that the chapter distinguishes different types of semantic measures (e.g., semantic relatedness/unrelatedness, semantic similarity/dissimilarity, semantic/taxonomic distance) with a graph showing their differences and relations (e.g., semantic similarity is a part of semantic relatedness). A general profile that demonstrates the landscape of different semantic measures is offered at the end of this chapter.

Chapter 2 provides a detailed introduction of corpus-based semantic measures defined at the word level. In order to develop a conceptual understanding of these measures, this chapter first presents a general pipeline normally used for defining such measures, which consists of five steps: collecting a corpus of texts as the semantic proxy, extracting the vocabulary, building a raw semantic model, transforming the raw model

doi:10.1162/COLI_r_00269

© 2017 Association for Computational Linguistics

into a refined model, and finally computing semantic similarity based on word representations from the refined model. Subsequently, it discusses notions of word meaning, context, and distributional semantics that are essential elements for the definition of semantic measures, though there are some debates on these notions. A variety of distributional models based on the distributional hypothesis are then described with details on both the strategies (e.g., frequency weighting and dimension reduction) used to build these models and the approaches used to learn word representations and estimate semantic similarity of word representations, which are classified into the geometric approach (e.g., Latent Semantic Analysis), the set-based approach (e.g., Dice index), and the probabilistic approach (e.g., Pointwise Mutual Information). Finally, the chapter concludes with advantages and limits of corpus-based semantic measures.

Compared with corpus-based semantic measures, knowledge-based semantic measures are not extensively and systematically surveyed and discussed in the literature. Chapter 3, the longest chapter in the book, is dedicated to filling this gap. This chapter starts with a general introduction to ontologies that can be built on graph representations as well as graph-based ontology processing technologies that can be used to calculate ontology-based semantic measures. It presents two different types of knowledge-based semantic measures, namely, semantic measures on cyclic and acyclic semantic graphs, with details on methods defining these measures, such as the shortest path and random walk approach. The chapter proceeds with the description of the core elements of semantic measures: different kinds of semantic evidence in semantic graph including the depth/width of a taxonomy, and concept specificity. It also elaborates on two particular ontology-based semantic measures: concept-to-concept semantic similarity and its extension groupwise concept similarity, listing papers where these measures are proposed and defined. Other knowledge-based measures, such as logic-based measures and measures defined on multiple ontologies, are also briefly introduced. Similar to Chapter 2, this chapter discusses the advantages and limits of knowledge-based semantic measures. Finally, a very useful and interesting section of this chapter is Section 3.8, which introduces recent efforts in hybrid measures combining both knowledge-based and corpus-based semantic measures.

Chapter 4 is dedicated to the evaluation of the various corpus-based and knowledge-based semantic measures presented in the preceding two chapters. It discusses two important topics, objective evaluation and task-oriented measure selection, and presents a variety of evaluation criteria for semantic measures (e.g., accuracy, precision, and computational complexity). Intrinsic and extrinsic evaluation strategies are also introduced. Perhaps the strongest part of this chapter is the extensive treatment of data sets that are widely used to evaluate semantic measures in the literature (e.g., WordSim353 for word relatedness, TOEFL for word similarity).

The last chapter concludes the book, demonstrates the challenges in semantic measures, and offers several suggestions for future research, for example, providing generic data sets and tools, formalizing and standardizing ontology-to-semantic-graph transformation, and promoting interdisciplinary studies among cognitive sciences, logic, natural language processing, and so on. There are four appendixes in the book, which provide examples, specific introduction to important algorithms, and useful tools and open-source software tools for both the computation and analysis of corpus-based and knowledge-based semantic measures.

Overall, the book provides a useful introduction and overview of research on semantic measures from natural language where both corpus-based and knowledge-based semantic measures are clearly presented. The book will serve as a good reference for graduate students who are eager to enter this area and natural language processing

Downloaded from http://direct.mit.edu/coli/article-pdf/42/4/829/1807507/coli_r_00269.pdf by guest on 11 August 2024

researchers and practitioners who want to quickly find a suitable measure for their tasks. Not only does it provide technical details of state-of-the-art semantic measure systems and algorithms, but it also offers a very useful list of tools and open-source software tools for modeling and computing various semantic measures and a list of widely used data sets for semantic measure evaluation.

There are a few aspects of this book with which I am not quite satisfied as a fastidious reader. First, the book title appears not to be fully consistent with the content covered by the book, since semantic similarity is only one particular semantic measure and the book discusses more. Second, although the introductory chapter gives an introduction to applications of semantic measures in various fields, I still think that this introduction deserves a complete chapter rather than a section, because in the current form it is not sufficient for NLP beginners who are going to use semantic measures. Finally, I am a little disappointed that the very recent research efforts in computing semantic similarity based on word, phrase, or even sentence embeddings learned with neural networks are not covered in Chapter 2 (corpus-based semantic measures). As such, I would like to encourage and welcome a new edition of this book to discuss these topics.

Deyi Xiong is a professor at Soochow University. Previously, he was a research scientist at the Institute for Infocomm Research of Singapore. His primary research interests are in the areas of natural language processing and statistical machine translation. He is particularly interested in semantics-driven machine translation. Xiong's e-mail address is dyxiong@suda.edu.cn.

Downloaded from http://direct.mit.edu/coli/article-pdf/42/4/829/1807507/coli_r_00269.pdf by guest on 11 August 2024

Downloaded from http://direct.mit.edu/coli/article-pdf/42/4/829/1807507/coli_r_00269.pdf by guest on 11 August 2024