

Semantic Similarity Frontiers: From Concepts to Documents

David Jurgens

School of Computer Science
McGill University
jurgens@cs.mcgill.ca

Mohammad Taher Pilehvar

Department of Computer Science
Sapienza University of Rome
pilehvar@di.uniroma1.it

1 Introduction

Semantic similarity forms a central component in many NLP systems, from lexical semantics, to part of speech tagging, to social media analysis. Recent years have seen a renewed interest in developing new similarity techniques, buoyed in part by work on embeddings and by SemEval tasks in Semantic Textual Similarity and Cross-Level Semantic Similarity. The increased interest has led to hundreds of techniques for measuring semantic similarity, which makes it difficult for practitioners to identify which state-of-the-art techniques are applicable and easily integrated into projects and for researchers to identify which aspects of the problem require future research.

This tutorial synthesizes the current state of the art for measuring semantic similarity for all types of conceptual or textual pairs and presents a broad overview of current techniques, what resources they use, and the particular inputs or domains to which the methods are most applicable. We survey methods ranging from corpus-based approaches operating on massive or domains-specific corpora to those leveraging structural information from expert-based or collaboratively-constructed lexical resources. Furthermore, we review work on multiple similarity tasks from sense-based comparisons to word, sentence, and document-sized comparisons and highlight general-purpose methods capable of comparing multiple types of inputs. Where possible, we also identify techniques that have been demonstrated to successfully operate in multilingual or cross-lingual settings.

Our tutorial provides a clear overview of currently-available tools and their strengths for practitioners who need out of the box solutions and provides researchers with an understanding of the limitations of current state of the art and what open problems remain in the field. Given the breadth of available approaches, participants will

also receive a detailed bibliography of approaches (including those not directly covered in the tutorial), annotated according to the approaches abilities, and pointers to when open-source implementations of the algorithms may be obtained.

2 Outline

1. Foundations (20 min.)
2. Semantic Similarity: State of the Art (90 min.)
 - (a) Concept and Word Similarity
 - (b) Phrase and Sentence Similarity
 - (c) Paragraph and Document Similarity
3. Cross-level semantic similarity (20 min.)
4. Open source tools for measuring semantic similarity (15 min.)
5. Open Problems and Future Work in Semantic Similarity (15 min.)



3 Instructors

David Jurgens is a postdoctoral scholar at McGill University. He received his PhD from the University of California, Los Angeles. His research interests include lexical semantics, word sense disambiguation, latent attribute inference, and the relationship between language and location. He is currently co-chairing the 2015 and 2016 International Workshops on Semantic Evaluation (SemEval). His research has been featured in the MIT Technology Review, Forbes, Business Insider, and Schneier on Security.

Mohammad Taher Pilehvar is a postdoctoral scholar at Sapienza University of Rome. He received his PhD from the same university under the supervision of Roberto Navigli. He does research in multiple areas of Lexical Semantics such as semantic similarity and Word Sense Disambiguation (WSD). His main focus is on unified graph-based semantic similarity measures and large-scale frameworks for the evaluation of WSD systems. He has co-organized a task on Cross

Level Semantic Similarity in SemEval-2014 (Jurgens et al., 2014) and is the first author of a paper on semantic similarity that was nominated for the best paper award at ACL 2013 (Pilehvar et al., 2013).

Acknowledgments

 The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234. 

References

- David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. Semeval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), in conjunction with COLING 2014*, pages 17–26.
- Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. In *Proceedings of ACL*, pages 1341–1351.