Measuring the Similarity and Relatedness of Concepts : A MICAI 2013 Tutorial

Ted Pedersen
Department of Computer Science
University of Minnesota, Duluth
Duluth, MN 55812, USA
tpederse@d.umn.edu

1 Tutorial Description

The ability to quantify the degree to which two concepts are similar or related to each other is a fundamental operation in many Natural Language Processing applications, including Question Answering, Recognizing Textual Entailment, and Word Sense Disambiguation. This tutorial will introduce the theory behind these measures, particularly those based on an underlying ontology or hierarchy of concepts such as WordNet[1]. It will also describe how these measures can be used in practical settings.

1.1 Content

This four hour tutorial will be divided into three sections.

The first will introduce the underlying theory behind measures of semantic similarity and relatedness. The second will show how to take advantage of freely available open source software that implements these measures, paying particular attention to WordNet::Similarity [PPM04][2]. The third will show how to carry out experimental evaluations with gold standard data, and how these measures can be applied in various different NLP tasks.

[1] http://wordnet.princeton.edu
[2] http://wn-similarity.sourceforge.net

1.2 Intended Audience

This tutorial presumes no prior knowledge of similarity and relatedness measures, and so should be accessible to anyone with an interest in the topic.

1.3 Expected Outcomes

Those who attend this tutorial will learn how to :

- understand the distinction between semantic relatedness and semantic similarity,
- measure semantic similarity and relatedness using information from ontologies, definitions, and corpora,
- use these measures from the command line, API, and web interface using the open–source software package WordNet::Similarity,
- conduct experiments using freely available gold standard data, and
- integrate these measures into various NLP applications.

2 Outline of Tutorial Structure

The following describes the overall organization of the tutorial and briefly summarizes the topics to be presented.

2.1 Measures of Similarity and Relatedness

This section will provide a theoretical overview of semantic similarity and relatedness.

2.1.1 Measures of Similarity

Measures of similarity rely on the structure of an is–a hierarchy to establish how much one concept is *like* or is–a *kind–of* another (e.g., *knife* is a kind–of *utensil*).

Path Based In path–based measures simple edge counting is employed, which can then be scaled in various ways by the depth of individual concepts.

Information Content Based Information content [Res99] is a means of augmenting each concept in a hierarchy with a corpus based measure of specificity.

2.1.2 Measures of Relatedness

Relatedness is more general than similarity since two concepts can be related without being similar (e.g., *knife* and *meat*). These measures quantify the degree of relatedness between two concepts without specifying the nature of that relationship.

Lesk Methods based on the Lesk algorithm [Les86, BP03] focus on finding shared content in definitions or descriptions to identify related concepts.

Vector Vector based methods [LD97, Sch98, PP06] generalize the Lesk algorithm and replace the words in definitions with co-occurrence vectors that are compared to establish relatedness between concepts.

2.2 Using Open Source Software

This section will introduce WordNet and the use of the WordNet::Similarity software package. It will also briefly introduce other software that implements these measures in WordNet (like NLTK), and also software that relies on other resources such as Wikipedia and the Unified Medical Language System (UMLS).

2.3 Similarity and Relatedness in the Wild

This section will give the audience specific ideas about how to develop experiments with measures of similarity and relatedness, and how to deploy these measures in NLP applications.

2.3.1 Using Existing Gold Standards

Freely available data sources will be reviewed, as will be methods for conducting experiments that evaluate measures relative to these standards.

2.3.2 Possible Applications

The use of semantic similarity and relatedness in various different applications will be reviewed. Examples will be drawn from Question Answering, Recognizing Textual Entailment, and Word Sense Disambiguation.

3 Instructor Biography

Ted Pedersen is a Professor in the Department of Computer Science at the University of Minnesota, Duluth. His main areas of research are in Natural Language Processing and Computational Linguistics, and focus on identifying the meaning of words and phrases in written text. In particular, he works on both knowledge based and supervised word sense disambiguation, unsupervised word sense induction, identifying collocations in text, and measuring the similarity and relatedness of concepts using both WordNet and the Unified Medical Language System (UMLS). He led the development of WordNet::Similarity, an open source package that measures similarity and relatedness of concepts in general English using the lexical database WordNet. As of September 2013 this software had been cited in more than 850 publications.

References

[BP03] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, Acapulco, August 2003.

[LD97] T. Landauer and S. Dumais. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[Les86] M.E. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM Press, 1986.

[PP06] S. Patwardhan and T. Pedersen. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy, April 2006.

[PPM04] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::Similarity - Measuring the relatedness of concepts. In *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 38–41, Boston, MA, 2004.

[Res99] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[Sch98] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.