WordNet::Similarity - Measuring the Relatedness of Concepts

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

Siddharth Patwardhan
School of Computing
University of Utah
Salt Lake City, UT 84102
sidd@cs.utah.edu

Jason Michelizzi
Department of Computer Science
University of Minnesota
Duluth, MN 55812
mich0212@d.umn.edu

http://search.cpan.org/dist/WordNet-Similarity
http://wn-similarity.sourceforge.net

Abstract

WordNet::Similarity is a freely available software package that makes it possible to measure the semantic similarity or relatedness between a pair of concepts (or word senses). It provides six measures of similarity, and three measures of relatedness, all of which are based on the lexical database WordNet. These measures are implemented as Perl modules which take as input two concepts, and return a numeric value that represents the degree to which they are similar or related.

Introduction

Measures of similarity quantify how much two concepts are alike, based on information contained in an *is–a* hierarchy. For example, an *automobile* might be considered more like a *boat* than a *tree*, if *automobile* and *boat* share *vehicle* as a common ancestor in an *is–a* hierarchy.

The lexical database WordNet is particularly well suited for similarity measures, since it organizes nouns and verbs into hierarchies of *is–a* relations. In version 2.0, there are nine noun hierarchies that include 80,000 concepts, and 554 verb hierarchies that are made up of 13,500 concepts.

Is–a relations in WordNet do not cross part of speech boundaries, so WordNet–based similarity measures are limited to making judgments between noun pairs (e.g., *cat* and *dog*) and verb pairs (e.g., *run* and *walk*). While WordNet includes adjectives and adverbs, these are not organized into *is–a* hierarchies so similarity measures can not be applied.

However, concepts can be related in many ways beyond being similar to each other. For example, a *wheel* is a part of a *car*, *night* is the opposite of *day*, *snow* is made up of *water*, a *knife* is used to cut *bread*, and so forth. As such WordNet provides additional (non–hierarchical) relations such as *has–part*, *is–made–of*, *is–an–attribute–of*, etc. In addition, each concept (or word sense) is described by a short written definition or gloss.

Measures of relatedness are based on these additional source of information, and as such can be applied to a wider range of concept pairs. For example, they can cross part of speech boundaries and assess the degree to which the verb *murder* and the noun *gun* are related. They can even measure the relatedness of concepts that do not reside in any *is–a* hiearchy, such as the adjectives *violent* and *harmful*.

Similarity Measures

Three similarity measures are based on path lengths between concepts: *lch* (Leacock & Chodorow 1998), *wup* (Wu & Palmer 1994), and *path*. The *lch* measure finds the shortest path between two concepts, and scales that value by the maximum path length in the *is–a* hierarchy in which they occur. *wup* finds the path length to the root node from the least common subsumer (LCS) of the two concepts, which is the most specific concept they share as an ancestor. This value is scaled by the sum of the path lengths from the individual concepts to the root. The measure *path* is equal to the inverse of the shortest path length between two concepts.

The three remaining similarity measures are based on *information content*, which is a corpus–based measure of the specificity a concept. These measures include *res* (Resnik 1995), *lin* (Lin 1998), and *jcn* (Jiang & Conrath 1997). The *lin* and *jcn* measures augment the information content of the LCS of two concepts with the sum of the information content of the individual concepts. The *lin* measure scales the information content of the LCS by this sum, while *jcn* subtracts the information content of the LCS from this sum (and then takes the inverse to convert it from a distance to a similarity measure).

By default, the information content of concepts is derived from the sense–tagged corpus SemCor. However, there are utility programs available in WordNet::Similarity that compute information content from untagged corpora such as the Brown Corpus, the Penn Treebank, the British National Corpus, or any given corpus of plain text.

WordNet::Similarity supports two hypothetical root nodes that can be turned on and off. When on, one root node subsumes all of the noun concepts, and another subsumes all of the verb concepts. This allows for similarity measures to be applied to any pair of nouns or verbs. If the hypothetical root nodes are off, then concepts must be in the same *is–a* hierarchy for a similarity measurement to be taken.

Measures of Relatedness

There are three relatedness measures supported in WordNet::Similarity: *hso* (Hirst & St-Onge 1998), *lesk* (Banerjee & Pedersen 2003), and *vector* (Patwardhan 2003). The *hso* measure is path based, and classifies relations in WordNet as having direction. For example, *is-a* relations are upwards, while *has–part* relations are horizontal. It establishes the relatedness between two concepts by trying to find a path between them that is neither too long nor that changes direction too often.

Each concept (or word sense) in WordNet is defined by a short gloss. The *lesk* and *vector* measures use the text of that gloss as a unique representation for the underlying concept. The *lesk* measure assigns relatedness by finding and scoring overlaps between the glosses of the two concepts, as well as concepts that are directly linked to them according to WordNet.

The *vector* measure creates a co–occurrence matrix from a corpus made up of the WordNet glosses. Each content word used in a WordNet gloss has an associated context vector. Each gloss is represented by a *gloss vector* that is the average of all the context vectors of the words found in the gloss. Relatedness between concepts is measured by finding the cosine between a pair of gloss vectors.

Using WordNet::Similarity

WordNet::Similarity is implemented with Perl's object oriented features. It uses the WordNet::QueryData package (Rennie 2000) to create an object representing WordNet. There are a number of methods available that allow for the inclusion of existing measures in Perl source code, and also for the development of new measures.

When an existing measure is to be incorporated into a Perl program, an object of that measure must be created via the *new()* method. Then the *getRelatedness()* method can be called for a pair of word senses that appear in WordNet, and their similarity or relatedness value will be returned.

WordNet::Similarity can also be utilized via a command line interface provided by the utility program *similarity.pl*. This allows a user to run the measures interactively for specific pairs of concepts when given in *word#pos#sense* form. For example, *car#n#3* refers to the third WordNet noun sense of *car*. It also allows for the specification of all the possible senses associated with a *word* or *word#pos* combination. In addition, there is a web interface that is based on this utility.

Regardless of how it is run, WordNet::Similarity supports detailed tracing that shows a variety of diagnostic information specific to each of the different kinds of measures. For example, for the measures that rely on path lengths (*lch*, *wup*, *path*) the tracing shows all the paths found between the concepts. Tracing for the information content measures (*res*, *lin*, *jcn*) includes both the paths between concepts as well as the least common subsumer. Tracing for the *hso* measure shows the actual paths found through WordNet, while the tracing for *lesk* shows the gloss overlaps in WordNet found for the two concepts and their nearby relatives. The *vector* tracing shows the word vectors that are used to create the gloss vector of a concept.

We have incorporated WordNet::Similarity into a generalized approach to word sense disambiguation that is based on semantic relatedness (Patwardhan, Banerjee, & Pedersen 2003). This is implemented in the SenseRelate package (*http://senserelate.sourceforge.net*). The premise of this algorithm is that the sense of a word can be determined by finding which of its senses is most related to the possible senses of its neighbors. We are now exploring the use of similarity and relatedness measures in evaluating the lexical choice component of a text generation system.

Acknowledgments

This work has been partially supported by a National Science Foundation Faculty Early CAREER Development award (#0092784), and by a Grant-in-Aid of Research, Artistry and Scholarship from the Office of the Vice President for Research and the Dean of the Graduate School of the University of Minnesota.

WordNet::Similarity is distributed under the Gnu Public License (GPL). It is freely available from the Comprehensive Perl Archive Network and SourceForge.

References

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

Hirst, G., and St-Onge, D. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum, C., ed., *WordNet: An electronic lexical database*. MIT Press. 305–332.

Jiang, J., and Conrath, D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, 19–33.

Leacock, C., and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed., *WordNet: An electronic lexical database*. MIT Press. 265–283.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the International Conference on Machine Learning*.

Patwardhan, S.; Banerjee, S.; and Pedersen, T. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, 241–257.

Patwardhan, S. 2003. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, Univ. of Minnesota, Duluth.

Rennie, J. 2000. WordNet::QueryData: a Perl module for accessing the WordNet database. http://search.cpan.org/dist/WordNet-QueryData.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 448–453.

Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, 133–138.