A Word-Net Vector Space Frequency Semantic Link Distance Model of Word-Meaning Equivalence

Kunal Patel (kunaliitb@hotmail.com, golden@utdallas.edu)
School of Computer Science and Electrical Engineering and
School of Behavioral and Brain Sciences, GR 4.1
University of Texas at Dallas, Box 830688
Richardson, TX 75083-0688

Richard M. Golden (golden@utdallas.edu)
School of Behavioral and Brain Sciences, GR 4.1
University of Texas at Dallas, Box 830688
Richardson, TX 75083-0688

Problem

The ability to identify indices of semantic similarity among words is an important problem which has important applications in the fields of both artificial intelligence and experimental psychology. Unfortunately, the problem of rank-ordering words according to their semantic similarity measure is not a straightforward one even with powerful semantic network models of the lexicon such as WORDNET (Fellbaum, 1998). The problem arises because "word-meaning equivalence" is not some objective quantity but is a subjective context-dependent property. Two words that seem very different in one context might be judged semantically similar within another context (and vice-versa). For example, a model of the lexicon such as WORDNET would predict the superordinates of CUP: "CROCKERY" or "DISHWARE" to be more consistent with the meaning of the word "CUP" than the subordinates of "CUP": "TEA CUP" or "COFFEE CUP". But such predictions might not be consistent with human performance. That is, people might simply be more likely to use the phrase "COFFEE CUP" instead of the word "CROCKERY" when they want to express the meaning "CUP". The goal of this research is to compare the standard semantic similarity measure of distance in WORDNET which is based upon the number of links separating two words in WORDNET with two new algorithms for computing semantic distance.

Algorithms

Semantic Link Distance

The *semantic link distance (SLD) algorithm* is a WORDNET-based algorithm which identifies for a given target word the set of all synonyms of that word, the set of all superordinates of the target word which were one or two links from the target word and the set of all subordinates of the target word. These words are then rank-ordered according to their semantic similarity to the target word by counting the number of links separating each word from the target word.

High Frequency Word/ Semantic Link Distance

The *high frequency word semantic link distance* (HFW-SLD) algorithm takes the results of the semantic link distance algorithm and removes the low frequency words.

Vector Space Frequency Semantic Link Distance

The *vector space frequency frequency semantic link* distance (VSF-SLD) algorithm represents each word generated by the SLD algorithm as a point in a two-dimensional vector space where the first component was the SLD and the second component was derived from the word's frequency of usage. The word with the largest vector magnitude is chosen as the most semantically similar word. The algorithm chooses the *k*th most semantically similar word as the word with the the *k*th closest distance to the algorithm's first choice.

Results

The above three algorithms were used to rank-order words and short word phrases according to their semantic similarity to 18 commonly target words. The VSF-SLD algorithm was superior to the HFW-SLD algorithm. The performance of the SLD algorithm was noticeably worse than both the VSF-SLD and HFW-SLD algorithms.

Acknowledgments

The order of the authors is arbitrary. This research is supported in part by the NSF ITR Initiative Award 0113369.

References

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Alexander, B, and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Workshop on WordNet and Other Lexical Resources, 2nd meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.