Computing Semantic Representations: A Comparative Analysis

Xiaowei Zhao (xzhao2@richmond.edu)
Ping Li (pli@richmond.edu)
Department of Psychology, University of Richmond
Richmond, VA 23173 USA

How can we formally capture the complex semantic relationships of the human lexicon? This question has been the focus of much recent computational studies. The ability to represent semantics faithfully in formal mechanisms not only is important for understanding the nature of the lexical system of natural languages, but also has significant implications for understanding the mental representation of meaning and its processing and acquisition.

Two best-known models in this regard are the Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and the Hyperspace Analog to Language (HAL; Burgess & Lund, 1997). Both of them are based on large-scale computational analyses of human speech corpora. The LSA model represents the corpora as a high-dimensional co-occurrence matrix of words in texts, and reduces its dimensions using singular value decomposition. The HAL model builds a semantic word co-occurrence matrix, which is weighted according to co-occurrence frequency. In contrast to these two models that automatically extract meanings by computational algorithms, a third model, the WordNet (Miller, 1990), is a computational thesaurus that provides semantic classification of the English lexicon in terms of hyponyms, synonyms, and antonyms, as well as searchable word entries with semantic definitions. Harm (2002) developed a system to extract the semantic features of the WordNet definitions so that lexical entries can be represented as feature-based vectors. In this study, we examine the virtues and drawbacks of the three models with respect to their ability to represent semantics accurately.

Because of our interest in modeling a developmental lexicon, we selected as our test vocabulary 600 words from the MacArthur Communicative Development Inventories (CDI; Dale & Fenson, 1996). The vocabulary can be divided into four major grammatical categories (nouns, verbs, adjectives, and closed-class words). The nouns can be further divided into 12 subcategories according to their meanings (e.g., clothes, toys, food, etc). The LSA, HAL and WordNet matrices used in our analyses were made available either by the authors or by their electronic distributions.

To examine the accuracy of word classification and representations of the three models, we used a simple k-nearest neighbor (kNN) classifier (Duda, Hart & Stork, 2000). The average classification rates of 4 grammatical categories and the 12 noun subcategories were treated respectively with a 5NN classifier. Figure 1 presents the results. It shows that the WordNet vectors give the best classification rates overall, followed by HAL and then LSA for the 4 grammatical categories, and by LSA and then HAL for the 12 noun subcategories. The best performance of the WordNet model indicates that the lexicographic and psycholinguistic analyses of words can yield accurate lexical-semantic representations, although it comes with a price: a significant amount of work is required to hand-code the features of words by human researchers. The better performance of HAL for the major grammatical categories indicates that HAL captures important information about grammatical relationships of words because of its representation and weighting of word sequences (word-to-word co-occurrence matrices). Finally, the better performance of LSA for the noun subcategories indicates that LSA is able to capture more subtle semantic differences and relationships among words, because a word's representation in this model involves a large number of other words in text (word-to-text co-occurrence matrices).

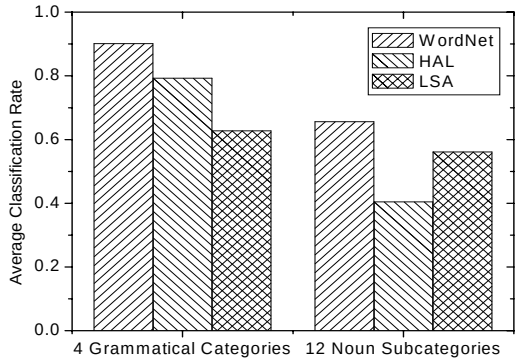

Figure 1: Average classification rates by a 5NN classifier

Acknowledgments

This research was supported by a grant from the National Science Foundation (BCS-0131829).

References

Burgess, C. & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 1-34.

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28,* 125-127.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (2nd ed.). John Wiley and Sons.

Harm, M. (2002). Building large scale distributed semantic feature sets with WordNet. *Technical Report PDP-CNS-02-1*, Carnegie Mellon University.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Miller, G.A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3,* 235-312.