Statistical Semantic Representations

Walter Kintsch
Department of Psychology
Institute of Cognitive Science
University of Colorado

(Walter.Kintsch@Colorado.edu)

Statistical semantics attempts to infer semantic knowledge from the analysis of linguistic corpora. For example, Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer et al., 2007) constructs a map of meaning that allows the ready computation of similarities between word meanings as well as text meanings. LSA takes as input word co-occurrence counts in a large number of documents and infers from that a high-dimensional semantic space. The meaning of words and texts can be expressed as vectors in this space, which allows the ready computation of the semantic similarity between words and texts. The LSA algorithm is not the only one that can be used for this purpose, however. Jones and Mewhort (2007) represent word meanings as holographs that record the context in which each was experienced. The holograph records not only word co-occurrences, but also information about the order in which the words appeared in each sentence. Griffiths, Steyvers, and Tenenbaum (2007) take yet a different approach, expressing word and document meaning as probability mixtures of a set of semantic topics. Griffiths et al. use a Bayesian learning algorithm to compute these probability mixtures. All of these models have been evaluated against data from the experimental literature, with considerable success. At some level, the predictions and implications of these different models are similar, although they differ in important ways and have different strengths and weaknesses. I shall focus on two limitations of the statistical approach to semantics.

Typically, semantic representations are generated from data that consist only of word co-occurrences in documents, neglecting information about word order, syntax, or discourse structure. Several ways to include word order as well as syntactic information in the construction of corpus-based semantic representations have been proposed. In our work, dependency grammar is used to guide the construction of semantic representations and comparisons. A parser derives the structural relations between words in a sentence, such as Noun-Verb dependencies, Verb-Noun dependencies, and Modifier-Noun dependencies. which are the building blocks of propositions. Thus, semantic relatedness between sentences can be computed taking into account their propositional structure, and not merely the words used and their order.

A second limitation of statistical models of semantics is that it is based solely upon verbal information, whereas human semantics integrates perception and action with the symbolic aspects of meaning. A map of meaning that considers only its verbal basis can nevertheless be useful, in that language mirrors real world phenomena. Furthermore, it is argued that meaning, while based on perception and action, transcends this basis and includes a symbolic level, which we attempt to model by statistical semantics.

Statistical models of semantics, particularly LSA, have been used widely and successfully in a number of practical applications (e.g., Landauer et al., 2007). I describe the use of an LSA-based software that helps middle-school students to learn how to write summaries. The software has been employed in a large number of schools and has been shown to be effective in teaching summarization.

Acknowledgements

This research was supported by grants from NSF and the J. S. McDonnell Foundation. The collaboration of Praful Mangalath and the *Summary Street* research group is gratefully acknowledged.

References

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211-244.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1-37.

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

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. Eds. (2007). *The Handbook of LSA*. Mahwah, NJ: Erlbaum