Modeling Students' Knowledge Representation with Latent Semantic Analysis

Peter W. Foltz (pfoltz@nmsu.edu)
Amber Wells (ambwells@nmsu.edu)
Department of Psychology, New Mexico State University
Las Cruces, NM, 88003-8001, USA

Latent Semantic Analysis (LSA) serves as both a theory and method for representing the meaning of words based on a statistical analysis of their contextual usage (e.g., Landauer & Dumais, in press; Foltz, Britt, & Perfetti 1996). Based on an analysis of large amounts of textual information, LSA generates a high-dimensional semantic space in which terms are represented as vectors in the space. The space captures effects due to the pattern of word usage across many contexts. This permits a comparison of semantic similarity between terms and larger units of text, even if they are used in different contexts

LSA has been tested using several different approaches to demonstrate that its representation of meaning corresponds well to humans' knowledge representations. These tests have included: representing and evaluating student's knowledge of history as expressed in their essays, predicting the results of lexical priming and word sorting tasks, and comparing performance on synonym tests to those of humans. In this paper, we compared the representation of readers' knowledge structures of information learned from an introductory psychology text to the representation generated by LSA.

Students enrolled in an introductory psychology class read a textbook chapter on memory. After reading the chapter, they were given a list of 16 concepts mentioned in the chapter and rated their knowledge of each of the concepts. They then were presented with all possible pairs of these 16 concepts and rated their relatedness on a 7 point Likert scale. Students also took a Nelson-Denny reading comprehension test and a test for knowledge about psychological concepts related to memory. In addition, graduate students, who were highly familiar with the domain, read the same chapter and performed the same ratings tasks. This provided a comparison between novice and expert representations of the topic.

The LSA representation of the same knowledge was generated through an analysis of the complete text of the textbook. The textbook was separated into paragraphs and the matrix of 4903 paragraphs by 19160 unique terms was analyzed with LSA, retaining 300 factors. Concept-concept comparisons were then made by calculating the cosine between the vectors for each pair of concepts. These cosines were then compared to the subjects' ratings of relatedness between concepts.

Correlations were computed between the subjects' ratings and LSA's predictions. Overall, the novice subjects correlated significantly with LSA ($r(119)=.27$, $p<.01$). Using a median split, those scoring low on the Nelson-Denny comprehension test had lower correlations to the LSA predictions ($r(119)=.18$, $p<.05$) than those who had high comprehension scores ($r(119)=.28$, $p<.01$). This indicates that LSA captures the effects of reading ability that result in a better knowledge representation. Based on preliminary data of experts' ratings, LSA correlates more strongly to the expert's knowledge representation, ($r(119)=0.40$ $p<.01$). Thus, the representation generated by LSA is much closer to that of a domain expert than that of a novice.

Overall, the results indicate that LSA does provide an accurate representation of the semantic distances between concepts that are expressed by readers of a text, even when those concepts do not co-occur within similar contexts. LSA's representation is more similar to that of a domain expert and it captures effects of the reader's reading ability.

Since LSA can model readers' knowledge representations, it also has practical applications for knowledge assessment. For example, the method can be used for performing automatic evaluations of students' knowledge of particular topics within a training system. Through an LSA analysis of a text or sets of texts, LSA's representation of distances between concepts should correspond well to that of a domain expert. Comparisons can then be made between human performance on these concepts and the model. Using this approach, we can generate measures that determine the overall comprehension of information in a domain, as well as methods for diagnosing knowledge deficits on specific topics within a domain. This information can then be used to determine what additional specific information must be provided to a student in order to improve their knowledge of the domain.

References

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1996) Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. W. Cottrell (Ed.) Proceedings of the 18th Annual Cognitive Science Conference. (pp. 110-115), Lawrence Erlbaum, NJ.

Landauer, T. K. & Dumais, S. T. (in press) A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. Psychological Review.