Applications of Latent Semantic Analysis

Thomas K Landauer (landauer@psych.colorado.edu)
University of Colorado at Boulder, CB 344 Boulder, CO 80309
and Knowledge Analysis Technologies, LLC

Latent Semantic Analysis (LSA) treats language learning and representation as a problem in mathematical induction. It casts the passages of a large and representative text corpus as a system of simultaneous linear equations in which passage meaning equals the sum of word meanings. Solution by Singular Value Decomposition (SVD) and dimension reduction produces a high-dimensional vector representing the average contribution to passage meanings of every word, and thus of the similarity between any two passages. LSA simulates human language understanding with surprising fidelity. Combining LSA with other statistical language modeling methods increases its practical scope. A variety of tests and applications illustrate its power, limits, and raise interesting theoretical issues.

Examples from Previously Published Results

LSA Improved IR 10-30% by recognizing documents of similar meaning but different words (Dumais, 1991); powered automatically constructed cross-language information retrieval (Landauer and Littman, 1990); mimicked the 10 words/ day vocabulary acquisition rate of children (Landauer & Dumais, 1997), and college student learning of psychology from textbooks (Landauer, Foltz & Laham, 1998) as measured by multiple-choice tests; simulated human categorization and similarity ratings (Laham, 2000), enabled simulations of predication and metaphor. (W. Kintsch, 2001); predicted paragraph comprehension differences caused by variation in S-S coherence; predicted which texts students would learn most from as a function of their prior knowledge (Rehder et al.; Landauer, Foltz & Laham, 1998); and improved summarizing skills by automatic componential feedback (E. Kintsch & Steinhart, 2000).

New Tests, Advances and Applications

LSA now scales to ca. 100 million word corpora by larger computer memory and new algorithms. Systems based on LSA measure the quality of sentences written to contextually define a word, r = .81 with expert ratings; connect by conceptual meaning each of a million paragraphs of an e-library; power collaborative learning environments that automatically alert participants to relevant contributions of others and assess contributions; enhance technical manuals to improve learning and speed performance; from text about tasks, occupational histories, etc., help guide career choice, fill jobs, and assemble optimal teams; combined with other statistical language models, score essays as accurately as expert human readers and provide componential feedback and plagiarism detection.

Some Implications, Limitations, and Issues

Successes to date disprove the poverty of the stimulus argument for lexical meaning and recast the problem of syntax learning, but leave much room for improvement. Size matters. The largest text corpora used in these applications equals one student's reading through high school; spoken language experience is an order of magnitude greater. Semantic atoms are not only single words; idioms need lexicalization. Syntax surely matters; LSA ignores word order. LSA's knowledge resembles intuition; people also use language for logic. Relations to other input matter. Perceptual and intentional experience contribute to meaning representation. (However, whether these bases are essential, more fundamental or involve different representational mechanisms is an open question. LSA represents perceptual phenomena vicariously, e.g. color relations. Demonstrations that people think in other modes, or that LSA does not exhaust linguistic meaning do not question LSA's validity, but call for more modeling, testing, and integration.

Possible Avenues for Research and Resolution

Similar inductive methods have been applied in perception (e.g. by S. Edelman, 1999)., opening a road to integrating language and perception. New models with learning of sentential order based meaning are needed. Simon Dennis's new unpublished model is a serious contender.

Acknowledgments

Funding support from ARI, AFRL, ONR, NSF, DoEd.

References

Landauer, T. K. (in press). On the computational basis of learning and cognition: Arguments from LSA. In B. H. Ross (Ed) *The Psychology of Learning and Motivation.* New York: Academic Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.

For all other references cited plus demonstrations and public LSA research tools, see http://LSA.colorado.edu and http://www.Knowledge-Technologies.com/