On what Latent Semantic Analysis (LSA/LSI) does and doesn't do

Thomas Landauer
University of Colorado
Boulder

Latent Semantic Analysis (LSA) is at once a remarkably simple and remarkably effective model of language. Its foundation is the following extreme simplification: The meaning of a passage is assumed to be the sum of the meanings of its contained words (with, of course a special restricted meaning of "meaning" relative to all that has been said about meaning in philosophy, linguistics, and literature.). This simplification allows observed natural language, for example a large corpus of ordinary text to be treated as a set of simultaneous linear equations that can be solved for the average meaning of the words, and consequently the meaning of any passage. The solution technique used by LSA is Singular Value Decomposition (SVD) followed by empirically optimal dimension reduction. The dimension reduction made possible by SVD has the property of inducing continuous-valued similarity relations between every word and every other, including the $> 98\%$ of pairs that never co-occur in a typical training corpus.

What is remarkable about LSA is that it works very well to emulate human language comprehension across a wide spectrum of linguistic, psycholinguistic, and natural language information handling applications. For example, it closely simulates the rate at which human language learners acquire vocabulary knowledge, improves IR by allowing queries to match relevant documents containing no query words, underwrites automatic scoring of conceptual content of essays that is as reliable as human experts, and supports a direct method of relating the meaning of words and documents in two or more languages.

However, using LSA to model language and to apply it to certain kinds of problems raises a variety of interesting theoretical and computational issues.

For linguistic theory, the fact that LSA does so much so well while ignoring word order within passages (addition being commutative) is extremely surprising, especially given traditional linguistic views of the important factors in semantics.

Psychologically and philosophically, LSA's ability to derive linguistic meaning with no "grounding" in perception and bodily function is puzzling. Similarly, the SVD computation derives all word meanings at once, rather than incrementally as humans must.

The fact that LSA's success is often strongly dependent on an optimal choice of dimensionality — a 250–350 dimensional representation may give two to four times as good results as one with many times fewer or more dimensions-lacks a satisfying formal explanation. Finally, the use of SVD has limited LSA's utility to relatively small document collections.

There has been recent progress, or at least activity, on most of these fronts. Something more can be said about how much is lost by ignoring word order-e.g. an information theoretic conjecture puts it at about 20%. New simulations illuminate the "grounding" problem-e.g., both LSA and humans have been found able to induce perceptual facts, such as the relations in the color circle, indirectly from language experience alone. The unusually strong dependence on optimum dimensionality may reflect a characteristic intrinsic dimensionality of language. The scaling problem, while not being solved well enough to make LSA practical as a general web search engine, has been greatly ameliorated by new algorithms combined with increases in available computational capacities.

Nevertheless, aspects of all these problems remain daunting. Perhaps the most challenging and fundamental is the issue of whether it is possible, and if so how, to automatically learn grammar and syntax and combine them with lexical semantics. A possibly intractable practical problem, if not necessarily a deeply theoretical one, is how to attach the model directly to the "real" world. The incremental learning issue is of both theoretical and practical interest because human language understanding is affected by the order in which words and concepts are encountered and learned-their accumulation appears to follow a course that includes processes that resemble crystal growth more than the solution of simultaneous equations. And finally, because of its superiority in overcoming the "synonymy problem" (there are an unlimited number of ways to say almost the same thing), it would be of considerable value to find ways to scale LSA to www applicability.