Chapter 9
LSA in the Classroom

Walter Kintsch
University of Colorado, USA

Eileen Kintsch
University of Colorado, USA

ABSTRACT

LSA is a machine learning method that constructs a map of meaning that permits one to calculate the semantic similarity between words and texts. We describe an educational application of LSA that provides immediate, individualized content feedback to middle school students writing summaries.

INTRODUCTION

The development of ever more efficient machine learning systems during the past decades has the potential to revolutionize computer applications in education. These systems are capable of learning, without supervision, the meaning of words from a large linguistic corpus, as well as the meaning of sentences and texts composed with these words. As we shall show, certain restrictions apply, but this work has already reached a sufficient level of maturity with several educational applications currently in use. Examples of the kind of systems we have in mind are Latent Semantic Analysis (LSA) (Landauer, McNamara, Dennis, & Kintsch, 2007), the topics model (Griffiths, Steyvers, & Tenenbaum, 2007), and the holograph model (Jones & Mewhort, 2007). We shall limit our discussion here to LSA, the method most widely used in education. The following section briefly summarizes the LSA method, but an example of an educational application of LSA will be the main focus of this chapter, concluding with a brief discussion of the limitations of this approach.

LSA was introduced by Landauer and Dumais in a seminal paper in 1997 (Landauer & Dumais, 1997). LSA was originally developed in the context of information retrieval, but Landauer and Dumais realized the potential of the method for modeling a wide variety of semantic phenomena. LSA infers word meanings from analyzing a large linguistic corpus. An example of a widely used corpus is the TASA corpus that consists of 44k documents a high-school graduate might have

DOI: 10.4018/978-1-60960-741-8.ch009

been exposed to during his or her lifetime. The total corpus comprises 11M word tokens, about 90k different words. This is a rich corpus, but the only information LSA actually uses consists of which words co-occurred which other words in each document. Sentence structure, syntax, discourse structure and so on are all neglected. Nevertheless, there is a great deal of information remaining, which LSA makes good use of.

The input to LSA consists of a huge matrix, listing the frequencies with which each word occurs in each document. This is an extremely sparse matrix, with most cells filled with 0's, because most words co-occur with only relatively few other words. The problem with such a matrix is that words whose meanings are quite unrelated do co-occur in the same document. Thus, although the raw word vector has all the right information in it, it is drowned in a sea of irrelevancies. What we want is the latent structure underlying the co-occurrence data, disregarding the noise inherent in the data. This latent structure is what LSA computes. LSA first uses a weighting scheme that de-emphasizes semantically uninformative words. For instance function words like "the," "of," or "but" play a very important role in comprehension in that they allow us to construct the syntactic structure of a sentence, specifying which role each word plays in a sentence. But since these high-frequency function words occur with many different words, they carry little weight semantically. LSA then uses a well-known mathematical technique called singular-value decomposition to reduce the dimensionality of the original matrix to about 300 dimensions. Dimensionality reduction achieves a two-fold purpose: It gets rid of much of the irrelevant noise in the corpus data, revealing its latent structure, and it fills in the original, sparse matrix, relating the main meaning-bearing words to each other, whether they had co-occurred in the corpus or not. As a result, in LSA each word in the corpus and each document is represented by a vector of 300 numbers. These numbers have no meaning by themselves, but together they define a semantic space – a high-dimensional map of meanings. Just as in a familiar two-dimensional map we can locate any two points with respect to each other and measure the distance between them, we can locate word meanings and document meanings in this 300-dimensional space and measure their distance. A useful measure of the similarity of two words, or of a word and a document, is the cosine between their vectors. Words that are unrelated have a cosine of 0 (or even a small negative value), and the more similar they are, the higher their cosine; identical words have a cosine of 1. Introductions to how LSA actually works can be found in Landauer and Dumais (1997) and Landauer et al. (2007).

What makes this semantic map so useful is the ability to calculate similarity measures (cosines) between any two points in the space, even though they may have never co-occurred in the corpus from which the space has been derived. Most importantly, we can represent the meaning of new documents in this space. A basic assumption upon which LSA is built is that the meaning of a document is the sum of the meaning of the words in that document. Thus, for any new, arbitrary document we can just add up all the word vectors to obtain the LSA vector for that document. Obviously, the assumption that the meaning of a document is the sum of the word meanings cannot be strictly true. "The hunter killed the deer" and "The deer killed the hunter" have the same words but have very different meanings. Nevertheless, for many purposes this simplifying assumption provides a very good approximation. The neglect of syntax limits the usefulness of LSA, as we shall see below, but there are many cases, both of theoretical and practical interest, where LSA has proven to be a powerful tool for the analysis of semantic phenomena.

We shall not discuss here the uses of LSA for the purpose of modeling human language understanding. The interested reader will find discussions of that approach in Landauer et al. (2007). Instead, we will describe an educational application of