Latent Semantic Analysis

Benoît Lemaire (Benoit.Lemaire@imag.fr)
Laboratoire Leibniz-IMAG (CNRS UMR 5522)
46, avenue Félix-Viallet
38031 Grenoble Cedex, FRANCE

Guy Denhière (denhiere@up.univ-mrs.fr)
LPC, University of Aix-Marseille & CNRS
Case 66, 3 place Victor Hugo
13331 Marseille Cedex, FRANCE

Abstract

LSA is a method for extracting and representing the meaning of words, from the statistical analysis of large text corpora. It is both a tool for automatically computing semantic similarities between words or texts and a powerful model of word acquisition, mimicking the process by which children learn the meaning of words through exposure to written material. LSA's semantic representations can also serve as the foundation of other cognitive models, such as text comprehension.
In this tutorial, we will first present how LSA works and what are the main experiments in the literature which tested it. We will explain how the corpus is analyzed, what mathematical procedure is applied to represent each word as a vector in a high-dimensional space, how new sentences can be given a vector representation from the vectors of their words. We will present tests based on word-word similarities, sentence-sentence similarities or text-text similarities.
We will then present some LSA-based cognitive models, namely word acquisition and text comprehension.
We will also present the limits of LSA and various recent cognitive models which aim at overcoming these limits. They are all based on the construction of word similarities from the automatic analysis of huge corpora but they differ from each other in features such as their unit of context (paragraph or moving window), their way of taking into account high-order co-occurrences, the fact that they are incremental or not.

Content

This tutorial is tailored to a cognitive science audience. We will not present LSA as a tool for information retrieval but rather emphasize its cognitive aspects. The tutorial is accessible to participants without mathematical background.

The following is the outline of the tutorial:

- Introduction: representing the meaning of words from the analysis of huge corpora
- LSA technique
  - Building a co-occurrence matrix and reducing it
  - Representing words as vectors
  - Representing texts as vectors
- Tests
  - Word-word similarities: the Landauer & Dumais (1997) TOEFL test; comparison with association norms (Denhière & Lemaire, 2004)
  - Word-sentence similarities: a vocabulary test (Denhière & Lemaire, 2004)
  - Sentence-sentence similarities: measure of coherence (Foltz et al., 1998)
  - Text-text similarities: assessing recall (Foltz, 1996); applications
- Cognitive models
  - Word acquisition: the Landauer & Dumais (1997) simulation
  - Text comprehension: Kintsch (2000) model of predication; Kintsch (2001) and Lemaire & Bianco (2003) models of metaphor comprehension; connecting LSA to Kintsch's constrution-integration model
- Limits
  - The absence of syntax
  - LSA is not incremental
  - Similarities are symetrical
- Competing models
  - Hyperspace Analogue to Language (Burgess, 1998)
  - PMI–IR (Turney, 2001)
  - Random Indexing (Sahlgren, 2001)
  - Word Association Space (Steyvers et al., in press)
  - Incremental Construction of an Association Network (Lemaire & Denhière, 2004)
  - Featural and Unitary Semantic Space (Vigliocco et al., in press)
  - Non-Latent Similarity Algorithm (Cai et al., 2004)
  - Context Dependent Sentence Abstraction Model (Ventura et al., 2004)