

Automatic Text Selection by Latent Semantic Analysis

Virginie Zampa

L.S.E.

Université Pierre-Mendès-France, BP47

38040 GRENOBLE CEDEX

e-mail : virginie.zampa@upmf-grenoble.fr

Key-words : Latent Semantic Analysis, Intelligent Tutoring System, learning environment, student modeling.

Introduction

This paper investigates how Latent Semantic Analysis (LSA), a model created by Landauer and Dumais [LAN97a], can discriminate texts according to their difficulty just as humans can do this. The first part of this note presents LSA, the second part explains our experiment. Then, results are presented and some educational applications are described.

LSA, a multidimensional factorial analysis method

Description of the model

LSA is a statistical model which automatically extracts semantic information from text. First, it creates a matrix of word occurrences (one line for each word met at least twice and one column by paragraph). Then, this matrix is reduced by a factorial analysis to a large number of dimensions (100 to 300). The proximity between two terms is calculated in term of context (the context of a word is the surrounding words). Therefore, two terms may be close to each other without necessarily occurring in the same paragraph. LSA calculates two kinds of value : semantic centrality (typicality) and semantic proximity between two texts.

Some validations of the model

A large number of tests have been carried out in order to validate LSA in many domains. A survey of these validations is given by Landauer and Dumais [LAN 97a]. Initially (1990), LSA was a tool for document retrieval by key-words. In contrast with other tools, LSA is not restricted to lexical retrieval, but also performs semantic retrieval. Thus LSA takes into account key-words' polysemy, synonymy and inflexions. In 1997, LSA was considered to be a model of human learning. In the learning domain, an experiment based on the TOEFL synonymy test, showed that, after having "learnt" several millions of words, LSA obtained results similar to students admitted to American's universities [FOL 96]. Another experiment proved that this semantic learning is also successful on learning game strategy [LEM 98]. In the evaluation domain, an important correlation was obtained between LSA grading (by proximity's calculation) and teacher's grading. In a first experiment [FOL 96], the students had to write essays, after having learned a corpus of texts. In a second one, the students had to write an essay concerning specific domains [LAN97b]. These experiments show that LSA has a similar behavior to humans, for grading and learning. We will now try to determine whether LSA can classify texts according to their difficulties.

Can LSA evaluate text difficulty ?

Presentation and results of the experiment

The aim of this experiment is to select texts and to study variability between human and LSA categorization. The texts had to be classed by their subjects into two different categories : “easy” and “difficult” (a text is easy if someone who has just begun to learn English, can read and understand it). Ten texts were used for LSA’s grading (five easy ones and five difficult ones) and four for improving LSA's knowledge in the domain. More details about texts are given in the following data set. LSA performs its classification by means of the centrality of each text. The centrality correspond to the typicality : 0 for the typical text, 1 for atypical text.

Type of texts	Mean number of words	Mean number of paragraphs	Mean size (Ko)	Mean centrality
Easy	3975	58	86	0.569542
Difficult	1683	21	54	0.724001
Additional texts	43867	396	255	0.593938

A statistical analysis (Student’s t) shows a significant difference between centrality of easy and difficult texts ($t = 5.06249$, $p < 0,01$).

Application to education

The goal of our thesis is to create an intelligent tutoring system for English learning. The results of the experiment presented below, show that LSA classifies texts into two categories (easy / difficult) as we predicted. However, there is no correlation between LSA and human classifications (0.12), and no correlation between human classifications. In fact, the classifications made by humans or by LSA varied according to their knowledge in English and their knowledge of the domain. In this experiment, the additional texts used for LSA “learning” were four children's stories. Therefor LSA's knowledge corresponds to modelling a young student. An initial way of using use these results would be in helping young English learners in the choice of text according to their level. Our future work can be divided in two parts. Firstly, LSA will be used for user modeling. For that, LSA will learn texts written by a student and examples of texts read by him. Secondly, we will be able to compare the classification made by LSA with the classification made by the student. In this perspective, LSA has two uses : user modeling and personalized help for selecting texts in learning environments.

References

- [FOL 96] FOLTZ, P. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*. 28(2). 197-202.
- [LAN 97a] LANDAUER, T. K & DUMAIS S. T. (1997).A solution to Plato’s problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.
- [LAN97b] LANDAUER, T. K, LAHAM, D., REHDER, B. and SCHREINER, M.E. (1997). How well can passage meaning be derived without using word order ? a comparison of Latent Semantic Analysis and Humans. Shafto, M.G., Langley, P. (Eds), *Proceedings of the 19th annual meeting of the Cognitive Science Society : 412-417, Mahwah, NJ : Erlbaum*.
- [LEM98] LEMAIRE, B., Models of High-Dimensional Semantic Spaces. *Proceedings of the 4th International Workshop on MultiStrategy Learning (MSL’98), June 98*.