An Application of Latent Semantic Analysis to Word Sense Discrimination for Words with Related and Unrelated Meanings

Juan Pino and **Maxine Eskenazi**
(jmpino, max)@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

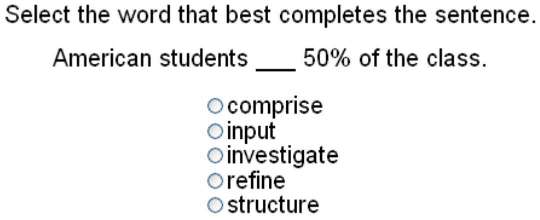
Abstract

We present an application of Latent Semantic Analysis to word sense discrimination within a tutor for English vocabulary learning. We attempt to match the meaning of a word in a document with the meaning of the same word in a fill-in-the-blank question. We compare the performance of the Lesk algorithm to Latent Semantic Analysis. We also compare the performance of Latent Semantic Analysis on a set of words with several unrelated meanings and on a set of words having both related and unrelated meanings.

1 Introduction

In this paper, we present an application of Latent Semantic Analysis (LSA) to word sense discrimination (WSD) within a tutor for English vocabulary learning for non-native speakers. This tutor retrieves documents from the Web that contain target words a student needs to learn and that are at an appropriate reading level (Collins-Thompson and Callan, 2005). It presents a document to the student and then follows the document reading with practice questions that measure how the student's knowledge has evolved. It is important that the fill-in-the-blank questions (also known as cloze questions) that we ask to the students allow us to determine their vocabulary knowledge accurately. An example of cloze question is shown in Figure 1.

Some words have more than one meaning and so the cloze question we give could be about a different meaning than the one that the student learned in the document. This is something that can lead to confusion and must be avoided. To do this, we need to use some automatic measure of semantic similarity.



Select the word that best completes the sentence.

American students ___ 50% of the class.

○ comprise
○ input
○ investigate
○ refine
○ structure

Figure 1: Example of cloze question.

To define the problem formally, given a target word w, a string r (the reading) containing w and n strings q_1, ..., q_n (the sentences used for the questions) each containing w, find the strings q_i where the meaning of w is closest to its meaning in r. We make the problem simpler by selecting only one question.

This problem is challenging because the context defined by cloze questions is short. Furthermore, a word can have only slight variations in meaning that even humans find sometimes difficult to distinguish. LSA was originally applied to Information Retrieval (Dumais et al., 1988). It was shown to be able to match short queries to relevant documents even when there were no exact matches between the words. Therefore LSA would seem to be an appropriate technique for matching a short context, such as a question, with a whole document.

So we are looking to first discriminate between the meanings of words, such as "compound", that have several very different meanings (a chemical compound or a set of buildings) and then to disambiguate words that have senses that are closely related such as "comprise" ("be composed of" or "compose"). In the following sections, we present

LSA and some of its applications, then we present some experimental results that compare a baseline to the use of LSA for both tasks we have just described. We expect the task to be easier on words with unrelated meanings. In addition, we expect that LSA will perform better when we use context selection on the documents.

2 Related Work

LSA was originally applied to Information Retrieval (Dumais et al., 1988) and called Latent Semantic Indexing (LSI). It is based on the singular value decomposition (SVD) theorem. A $m \times n$ matrix X with $m \geq n$ can be written as $X = U \cdot S \cdot V^T$ where U is a $m \times n$ matrix such that $U^T \cdot U = I_m$; S is a $n \times n$ diagonal matrix whose diagonal coefficients are in decreasing order; and V is a $n \times n$ matrix such that $V^T \cdot V = I_n$.

X is typically a term-document matrix that represents the occurrences of vocabulary words in a set of documents. LSI uses truncated SVD, that is it considers the first r columns of U (written U_r), the r highest coefficients in S (S_r) and the first r columns of V (V_r). Similarity between a query and a document represented by vectors $\mathbf{d}$ and $\mathbf{q}$ is performed by computing the cosine similarity between $S_r^{-1} \cdot U_r^T \cdot \mathbf{d}$ and $S_r^{-1} \cdot U_r^T \cdot \mathbf{q}$. The motivation for computing similarity in a different space is to cope with the sparsity of the vectors in the original space. The motivation for truncating SVD is that only the most meaningful semantic components of the document and the query are represented after this transformation and that noise is discarded.

LSA was subsequently applied to number of problems, such as synonym detection (Landauer et al., 1998), document clustering (Song and Park, 2007), vocabulary acquisition simulation (Landauer and Dumais, 1997), etc.

Levin and colleagues (2006) applied LSA to word sense discrimination. They clustered documents containing ambiguous words and for a test instance of a document, they assigned the document to its closest cluster. Our approach is to assign to a document the question that is closest. In addition, we examine the cases where a word has several unrelated meanings and where a word has several closely related meanings.

3 Experimental Setup

We used a database of 62 manually generated cloze questions covering 16 target words[1]. We manually annotated the senses of the target words in these questions using WordNet senses (Fellbaum, 1998). For each word and for each sense, we manually gathered documents from the Web containing the target word with the corresponding sense. There were 84 documents in total. We added 97 documents extracted from the tutor database of documents that contained at least one target word but we did not annotate their meaning.

We wanted to evaluate the performances of LSA for WSD for words with unrelated meanings and for words with both related and unrelated meanings. For the first type of evaluation, we retained four target words. For the second type of evaluation, all 16 words were included. We also wanted to evaluate the influence of the size of the context of the target words. We therefore considered two matrices: a term-document matrix and a term-context matrix where context designates five sentences around the target word in the document. In both cases each cell of the matrix had a *tf-idf* weight. Finally, we wanted to investigate the influence of the dimension reduction on performance. In our experiments, we explored these three directions.

4 Results

4.1 Baseline

We first used a variant of the Lesk algorithm (Lesk, 1986), which is based on word exact match. This algorithm seems well suited for the unsupervised approach we took here since we were dealing with discrimination rather than disambiguation. Given a document $\mathbf{d}$ and a question $\mathbf{q}$, we computed the number of word tokens that were shared between $\mathbf{d}$ and $\mathbf{q}$, excluding the target word. The words were lower cased and stemmed using the Porter stemmer. Stop words and punctuation were discarded; we used the standard English stopword list. Finally, we selected a window of nw words around the target word in the question $\mathbf{q}$ and a window of ns sentences around the target word in the document $\mathbf{d}$. In order to detect sentence boundaries, we used

[1] available at: www.cs.cmu.edu/ jmpino/questions.xls

the OpenNLP toolkit (Baldridge et al., 2002). With $nw = 10$ and $ns = 2$, we obtained an accuracy of 61% for the Lesk algorithm. This can be compared to a random baseline of 44% accuracy.

4.2 LSA

We indexed the document database using the Lemur toolkit (Allan et al., 2003). The database contained both the manually annotated documents and the documents used by the tutor and containing the target words. The Colt package (Binko et al.,) was used to perform singular value decomposition and matrix operations because it supports sparse matrix operations. We explored three directions in our analysis. We investigated how LSA performs for words with related meanings and for words with unrelated meanings. We also explored the influence of the truncation parameter r. Finally, we examined if reducing the document to a selected context of the target word improved performance.

Figures 2 and 3 plot accuracy versus dimension reduction in different cases. In all cases, LSA outperforms the baseline for certain values of the truncation parameter and when context selection was used. This shows that LSA is well suited for measuring semantic similarity between two contexts when at least one of them is short. In general, using the full dimension in SVD hurts the performances. Dimension reduction indeed helps discarding noise and noise is certainly present in our experiments since we do not perform stemming and do not use a stopword list. One could argue that filling the matrix cells with *tf-idf* weights already gives less importance to noisy words.

Figure 2 shows that selecting context in documents does not give much improvement in accuracy. It might be that the amount of context selected depends on each document. Here we had a fixed size context of five sentences around the target word.

In Figure 3, selecting context gives some improvement, although not statistically significant, over the case with the whole document as context. The best performance obtained for words with unrelated meanings and context selection is also better than the performance for words with related and unrelated meanings.

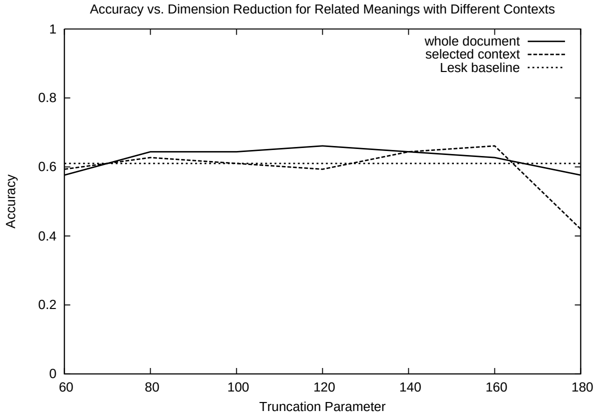


Figure 2: Accuracy vs. r, the truncation parameter, for words with related and unrelated meanings and with whole document or selected context (95% confidence for whole document: [0.59; 0.65], 95% confidence for selected context: [0.52; 0.67])

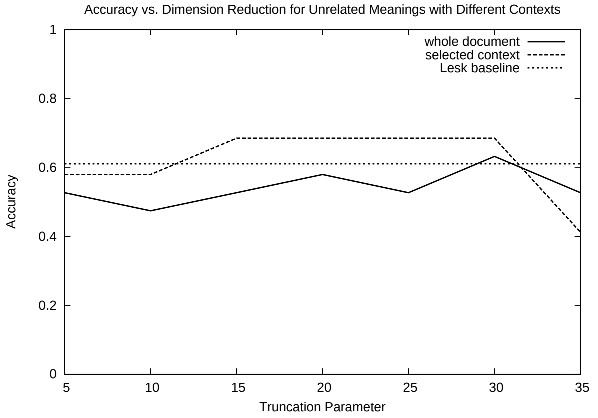


Figure 3: Accuracy vs. r, the truncation parameter, for words with unrelated meanings only and with whole documents or selected context ((95% confidence for whole document: [0.50; 0.59], 95% confidence for selected context: [0.52; 0.71]))

5 Discussion

LSA helps overcome sparsity of short contexts such as questions and gives an improvement over the exact match baseline. However, reducing the context of the documents to five sentences around the target word does not seem to give significant improvement. This might be due to the fact that capturing the right context for a meaning is a difficult task and that a fixed size context does not always represent a relevant context. It is yet unclear how to set the truncation parameter. Although dimension reduction seems to help, better results are sometimes obtained when the truncation parameter is close to full dimension or when the truncation parameter is farther from the full dimension.

6 Conclusion and Future Work

We have shown that LSA, which can be considered as a second-order representation of the documents and question vectors, is better suited than the Lesk algorithm, which is a first-order representation of vectors, for measuring semantic similarity between a short context such as a question and a longer context such as a document. Dimension reduction was shown to play an important role in the performances. However, LSA is relatively difficult to apply to large amounts of data because SVD is computationally intensive when the vocabulary size is not limited. In the context of tutoring systems, LSA could not be applied on the fly, the documents would need to be preprocessed and annotated beforehand.

We would like to further apply this promising technique for WSD. Our tutor is able to provide definitions when a student is reading a document. We currently provide all available definitions. It would be more beneficial to present only the definitions that are relevant to the meaning of the word in the document or at least to order them according to their semantic similarity with the context. We would also like to investigate how the size of the selected context in a document can affect performance. Finally, we would like to compare LSA performance to other second-order vector representations such as vectors induced from co-occurrence statistics.

Acknowledgments

Thanks Mehrbod Sharifi, David Klahr and Ken Koedinger for fruitful discussions. This research is supported by NSF grant SBE-0354420. Any conclusions expressed here are those of the authors.

References

James Allan, Jamie Callan, Kevin Collins-Thompson, Bruce Croft, Fangfang Feng, David Fisher, John Lafferty, Leah Larkey, Thi N. Truong, Paul Ogilvie, et al. 2003. The lemur toolkit for language modeling and information retrieval.

Jason Baldridge, Thomas Morton, and Gann Bierner. 2002. The opennlp maximum entropy package. Technical report, Technical report, SourceForge.

Pavel Binko, Dino Ferrero Merlino, Wolfgang Hoschek, Tony Johnson, Andreas Pfeiffer, et al. Open source libraries for high performance scientific and technical computing in java.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Susane T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological review*, 104:211–240.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25:259–284.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems Documentation*, pages 24–26.

Esther Levin, Mehrbod Sharifi, and Jerry Ball. 2006. Evaluation of utility of lsa for word sense discrimination. In *Proceedings of HLT/NAACL*, pages 77–80.

Wei Song and Soon Cheol Park. 2007. A novel document clustering model based on latent semantic analysis. In *Proceedings of the Third International Conference on Semantics, Knowledge and Grid*, pages 539–542.