# Evaluation of Utility of LSA for Word Sense Discrimination

**Esther Levin**
Dept. of Computer Science
City College of New York
NY, NY 10031
esther@cs.ccny.cuny.edu

**Mehrbod Sharifi**
Dept. of Computer Science
City College of New York
NY, NY 10031
mehrbod@yahoo.com

**Jerry Ball**
Air Force Research Laboratory
6030 S Kent Street
Mesa, AZ 85212-6061
Jerry.Ball@mesa.afmc.af.mil

## Abstract

The goal of the on-going project described in this paper is evaluation of the utility of Latent Semantic Analysis (LSA) for unsupervised word sense discrimination. The hypothesis is that LSA can be used to compute context vectors for ambiguous words that can be clustered together – with each cluster corresponding to a different sense of the word. In this paper we report first experimental result on tightness, separation and purity of sense-based clusters as a function of vector space dimensionality and using different distance metrics.

## 1 Introduction

Latent semantic analysis (LSA) is a mathematical technique used in natural language processing for finding complex and hidden relations of meaning among words and the various contexts in which they are found (Landauer and Dumais, 1997; Landauer et al, 1998). LSA is based on the idea of association of elements (words) with contexts and similarity in word meaning is defined by similarity in shared contexts.

The starting point for LSA is the construction of a co-occurrence matrix, where the columns represent the different contexts in the corpus, and the rows the different word tokens. An entry *ij* in the matrix corresponds to the count of the number of times the word token *i* appeared in context *j*. Often the co-occurrence matrix is normalized for document length and word entropy (Dumais, 1994).

The critical step of the LSA algorithm is to compute the singular value decomposition (SVD) of the normalized co-occurrence matrix. If the matrices comprising the SVD are permuted such that the singular values are in decreasing order, they can be truncated to a much lower rank. According to Landauer and Dumais (1997), it is this dimensionality reduction step, the combining of surface information into a deeper abstraction that captures the mutual implications of words and passages and uncovers important structural aspects of a problem while filtering out noise. The singular vectors reflect principal components, or axes of greatest variance in the data, constituting the hidden abstract concepts of the semantic space, and each word and each document is represented as a linear combination of these concepts.

Within the LSA framework discreet entities such as words and documents are mapped into the same continuous low-dimensional parameter space, revealing the underlying semantic structure of these entities and making it especially efficient for variety of machine-learning algorithms. Following successful application of LSA to information retrieval other areas of application of the same methodology have been explored, including language modeling, word and document clustering, call routing and semantic inference for spoken interface control (Bellegarda, 2005).

The ultimate goal of the project described here is to explore the use of LSA for unsupervised identification of word senses and for estimating word sense frequencies from application relevant corpora following Schütze's (1998) context-group discrimination paradigm. In this paper we describe a first set of experiments investigating the tightness, separation and purity properties of sense-based clusters.



*Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 77–80,
New York, June 2006. 

## 2 Experimental Setup

The basic idea of the context-group discrimination paradigm adopted in this investigation is to induce senses of ambiguous word from their contextual similarity. The occurrences of an ambiguous word represented by their context vectors are grouped into clusters, where clusters consist of contextually similar occurrences. The context vectors in our experiments are LSA-based representation of the documents in which the ambiguous word appears. Context vectors from the training portion of the corpus are grouped into clusters and the centroid of the cluster—the sense vector—is computed. Ambiguous words from the test portion of the corpus are disambiguated by finding the closest sense vector (cluster centroid) to its context vector representation. If sense labels are available for the ambiguous words in the corpus, sense vectors are given a label that corresponds to the majority sense in their cluster, and sense discrimination accuracy can be evaluated by computing the percentage of ambiguous words from the test portion that were mapped to the sense vector whose label corresponds to the ambiguous word's sense label.

Our goal is to investigate how well the different senses of ambiguous words are separated in the LSA-based vector space. With an ideal representation the clusters of context vectors would be tight (the vectors in the cluster close to each other and close to centroid of the cluster), and far away from each other, and each cluster would be pure, i.e., consisting of vectors corresponding to words with the same sense. Since we don't want the evaluation of the LSA-based representation to be influenced by the choice of clustering algorithm, or the algorithm's initialization and its parameter settings that determine the resulting grouping, we took an orthogonal approach to the problem: Instead of evaluating the purity of the clusters based on geometrical position of vectors, we evaluate how well-formed the clusters based on sense labels are, how separated from each other and tight they are. As will be discussed below, performance evaluation of such sense-based clusters results in an upper bound on the performance that can be obtained by clustering algorithms such as EM or K-means.

## 3 Results

We used the line-hard-serve-interest corpus(Leacock et al, 1993), with 1151 instances for 3 noun senses of word "Line": cord - 373, division - 374, and text - 404; 752 instances for 2 adjective senses of word "Hard": difficult – 376, not yielding to pressure or easily penetrated – 376; 1292 instances for 2 verb senses of word "Serve": serving a purpose, role or function or acting as – 853, and providing service 439; and 2113 instances for 3 noun senses of word "Interest": readiness to give attention - 361, a share in a company or business – 500, money paid for the use of money -1252.

For all instances of an ambiguous word in the corpus we computed the corresponding LSA context vectors, and grouped them into clusters according to the sense label given in the corpus. To evaluate the inter-cluster tightness and intra-cluster separation for variable-dimensionality LSA representation we used the following measures:

**1. Sense discrimination accuracy**. To compute sense discrimination accuracy the centroid of each sense cluster was computed using 90% of the data. We evaluated the sense discrimination accuracy using the remaining 10% of the data reserved for testing by computing for each test context vector the closest cluster centroid and comparing their sense labels. To increase the robustness of this evaluation we repeated this computation 10 times, each time using a different 10% chunk for test data, round-robin style. The sense discrimination accuracy estimated in this way constitutes an upper bound on the sense discrimination performance of unsupervised clustering such as K-means or EM: The sense-based centroids, by definition, are the points with minimal average distance to all the same-sense points in the training set, while the centroids found by unsupervised clustering are based on geometric properties of all context vectors, regardless of their sense label.

**2. Average Silhouette Value**. The silhouette value (Rousseeuw, 1987) for each point is a measure of how similar that point is to points in its own cluster vs. points in other clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. To construct the silhouette value for each vector i, *S(i)*, the following formula is used:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}},$$

where *a(i)* is an average distance of *i*-object to all other objects in the same cluster and *b(i)* is a minimum of average distance *of i*-object to all objects in other cluster (in other words, it is the average distance to the points in closest cluster among the other clusters). The overall average silhouette value is simply the average of the *S(i)* for all points in the whole dataset.

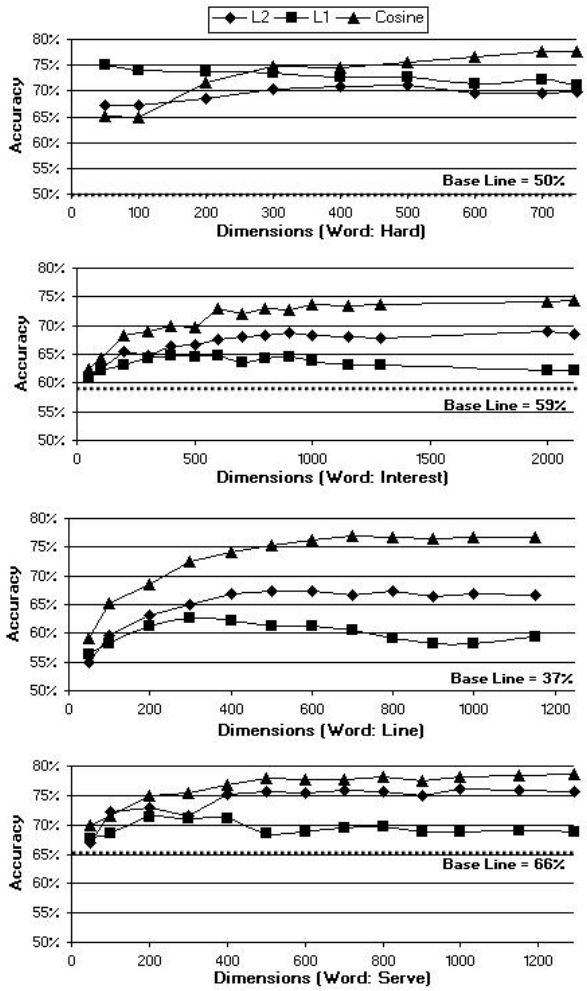


Figure 1: Average discrimination accuracy

Figure 1 plots the average discrimination accuracy as a function of LSA dimensionality for different distance/similarity measures, namely L2, L1 and cosine, for the 4 ambiguous words in the corpus. Note that the distance measure choice affects not only the classification of a point to the cluster, but also the computation of cluster centroid. For L2 and cosine measures the centroid is simply the average of vectors in the cluster, while for L1 it is the median, i.e., the value of *i*-th dimension of the cluster centroid vector is the median of values of the *i*-th dimension of all the vectors in the cluster.

As can be seen from the sense discrimination results in Fig. 1, cosine distance, the most frequently used distance measure in LSA applications, has the best performance in for 3 out of 4 words in the corpus. Only for "Hard" does L1 outperforms cosine for low values of LSA dimension. As to the influence of dimensionality reduction on sense discrimination accuracy, our results show that (at least for the cosine distance) the accuracy does not peak at any reduced dimension, rather it increases monotonically, first rapidly and then reaching saturation as the dimension is increased from its lowest value (50 in our experiments) to the full dimension that corresponds to the number of contexts in the corpus.

These results suggest that the value of dimensionality reduction is not in increasing the sense discrimination power of LSA representation, but in making the subsequent computations more efficient and perhaps enabling working with much larger corpora. For every number of dimensions examined, the average sense discrimination accuracy is significantly better than the baseline that was computed as the relative percentage of the most frequent sense of each ambiguous word in the corpus.

Figure 2 shows the average silhouette values for the sense-based clusters as a function of the dimensionality of the underlying LSA–based vector representation for the 3 different distance metrics and for the 4 words in the corpus. The average silhouette value is close to zero, not varying significantly for the different number of dimensions and distance measures. Although the measured silhouette values indicate that the sense-based clusters are not very tight, the sense-discrimination accuracy results suggest that they are sufficiently far from each other to guarantee relatively high accuracy.

## 4 Summary and Discussion

In this paper we reported on the first in a series of experiments aimed at examining the sense discrimination utility of LSA-based vector representation of ambiguous words' contexts. Our evaluation of average silhouette values indicates that sense-

based clusters in the latent semantic space are not very tight (their silhouette values are mostly positive, but close to zero). However, they are separated enough to result in sense discrimination accuracy significantly higher than the baseline. We also found that the cosine distance measure outperforms L1 and L2, and that dimensionality reduction for sense-based clusters does not improve the sense discrimination accuracy.

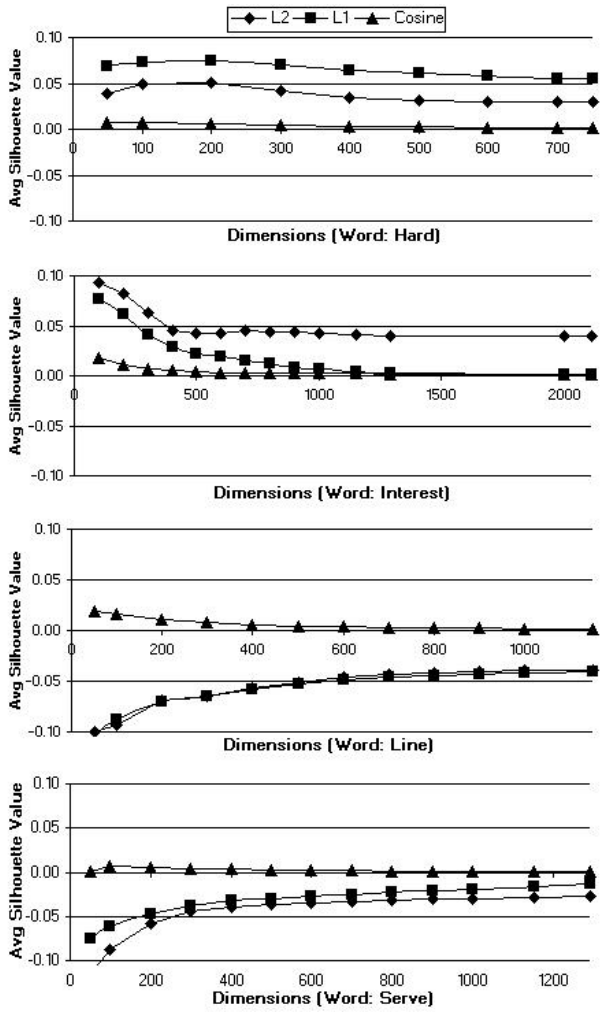

Figure2: Average silhouette values

The clustering examined in this paper is based on pre-established word sense labels, and the measured accuracy constitutes an upper bound on a sense discrimination accuracy that can be obtained by unsupervised clustering such as EM or segmental K-means. In the next phase of this investigation we plan to do a similar evaluation for clustering obtained without supervision by running K-means algorithm on the same corpus. Since such clustering is based on geometric properties of word vectors, we expect it to have a better tightness as measured by average silhouette value, but, at the same time, lower sense discrimination accuracy.

The experiments reported here are based on LSA representation computed using the whole document as a context for the ambiguous word. In the future we plan to investigate the influence of the context size on sense discrimination performance.


### Acknowledgements

The project described above is supported by grant "Using LSA to Compute Word Sense Frequencies" from Air Force Research Lab. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the AFRL.


## 5 References


J.R. Bellegarda. 2005. *Latent Semantic Mapping*, IEEE Signal Processing Magazine, 22(5):70-80.

S.T. Dumais. 1994. *Latent Semantic Indexing (LSI) and TREC-2*, in Proc Second Text Retrieval Conf. (TREC-2), pp 104-105.

T.K. Landauer, S.T. Dumais. 1997. *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge*, Psychological Review, 104(2):211-240.

T.K. Landauer, P. Foltz, and D. Laham. 1998. *Introduction to Latent Semantic Analysis.* Discourse Processes, 25, 259-284.

C. Leacock, G. Towel, E. Voorhees. 1993. *Corpus-Based Statistical Sense Resolution*, Proceedings of the ARPA Workshop on Human Language Technology.

P.J. Rousseeuw. 1987. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics. 20. 53-65.

H. Schütze. 1998. *Automatic Word Sense Discrimination*, Journal of Computational Linguistics, Volume 24, Number 2