# Latent Concepts versus Latent Topics

Vasile Rus, Rajendra Banjade, Nobal Niraula, Dan Stefanescu
Department of Computer Science/Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
USA

**Abstract:** We present in this paper a comparison study between two fully automated methods to derive meaning representations: Latent Dirichlet Allocation and Latent Semantic Analysis. The comparison is conducted in terms of performance with respect to the task of paraphrase identification. In particular, we investigate which of the two methods can best capture word meaning in two different settings: (1) same number of concepts and topics are used in LSA and LDA, respectively, and (2) an optimal number of topics is chosen for LDA.

## Introduction

Representing and acquiring meaning is an important topic in text and discourse processing. It can help, for instance, to address the problem of student input assessment in dialogue-based Intelligent Tutoring Systems (Rus & Graesser, 2006). Fully automated methods for meaning derivation, such as Latent Semantic Analysis (LSA; Landauer et al, 2006) or Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan, 2003), are desirable as they enable the derivation of meaning from very large collections of texts without human intervention.

LDA is a generative model of text in which the assumption is that documents are random mixtures of topics, i.e. generated according to some topic distribution, whereas topics are distributions over words in the vocabulary.

Like LDA, LSA is fully automated for meaning derivation and representation. LSA represents meanings as vectors in a reduced dimensionality space (300-500 dimensions), called the LSA semantic space. The dimensions of the reduced LSA space are called latent concepts. Because LSA has a unique representation for each word, it does not represent different senses of the same word explicitly. Some argue that the LSA vector represents an average meaning of all the senses of the word while some others believe that LSA represents the dominant meaning of the word.

This paper offers a comparison between the two methods, LSA and LDA, in the context of the paraphrase identification task.

## Method

We compare how well word meanings derived with LSA and LDA, respectively, can be used to solve the problem of paraphrase identification in two different ways. First, we specify the same number of latent topics and latent concepts when deriving the LDA and LSA representations. That is, we compare how well the same number of latent topics or latent concepts can represent word meanings derived automatically from large corpora. Second, we use an optimal number of topics and concepts to derive word meanings. For LSA, we use the widely accepted number of 300 dimensions or concepts. For LDA, we select the optimal number of topics based on a measure of topic coherence. That is, we vary the number of topics from 10 to 300 (with an increment of 10), measure the coherence of the derived topics, and then choose the LDA model corresponding to the number of topics with the highest topic coherence score. The number of

topics with the highest topic coherence is 100. Topic coherence was measured using average Pointwise Mutual Information (PMI) between top 20 words in each topic. PMI was derived from English Wikipedia (downloaded in December, 2012). LDA and LSA models were derived from TASA corpus (compiled by Touchstone Applied Science Associates), a balanced collection of representative texts from various genres (science, language arts, health, economics, social studies, business, and others).

### Paraphrase Identification and Paraphrase Corpus

We measure how well LDA and LSA can derive the meaning of words by evaluating how well the derived representations can help solve the task of paraphrase identification between two sentences. Paraphrase identification is about making a judgment with respect to how semantically similar two sentences are. If the two sentences have the same meaning we say they are paraphrase otherwise they are not.

To decide whether two sentences are paraphrases or not, we combined LDA and LSA with two sentence-to-sentence similarity methods: a greedy matching method and an optimal matching method. In the greedy matching method, each word in one sentence is greedily matched with the highest matching word in the other sentence based on an LDA or LSA-based word-to-word similarity score. The sentence-to-sentence similarity score is the normalized average of the individual word-to-word scores. In optimal matching, the overall matching score between two sentences is optimized by with respect to the overall sum of individual word-to-word similarity scores. For LDA, we defined a word-to-word similarity score as the cosine between the vectors that represent the contributions of the words to each topic. For LSA, we use the cosine between two words' LSA vectors as a measure of word-to-word similarity.

The MSRP corpus is our testbed to evaluate both LSA and LDA based representations. The MSRP corpus is the largest public annotated paraphrase corpus and has been used in most of the recent studies that addressed the problem of paraphrase identification. The corpus consists of 5,801 sentence pairs collected from newswire articles, 3,900 of which were labeled as paraphrases by human annotators. The whole set is divided into a training subset (4,076 sentences of which 2,753, or 67.5%, are true paraphrases), and a test subset (1,725 pairs of which 1,147, or 66.5%, are true paraphrases). The average number of words per sentence is 17.

## Analytic Strategy

We followed a training-testing methodology according to which we first trained the proposed methods on a set of training data after which we used the learned models on testing data. In our case, we learned a threshold for the text-to-text similarity score above which a pair of sentences was deemed a paraphrase and any score below the threshold means the sentences were not paraphrases. We report performance of the various methods using accuracy (percentage of correct predictions) and kappa statistics (a measure of agreement between our method's output and experts' labels while accounting for chance agreement).

## Results and Discussion

A summary of results is shown in Table 1. The first raw presents results obtained with a simple baseline method of predicting all the time the dominant class (true paraphrases) in the training data set. As one can notice from the table, the LDA based results are better than LSA results for

both Greedy and Optimal semantic matching methods. Given these results and the fact that LDA models various senses of a word, we are inclined to recommend the use of LDA over LSA.

**Table 1.** Results on the MSRP test data (for LDA we report results for 300 topics and the optimal number of 100 topics).

| ***Method*** | ***Accuracy*** | ***Kappa*** |
|---|---|---|
| Baseline | 66.55 | 0.22 |
| LSA Greedy | 72.86 | 33.89 |
| LSA Optimal | 73.04 | 35.95 |
| LDA-Greedy | 73.04/72.33 | 35.01/34.68 |
| LDA-Optimal | **73.27/74.68** | **36.74/37.01** |

## References


Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation, The Journal of Machine Learning Research 3, 993-1022.

Landauer, T.; McNamara, D. S.; Dennis, S.; and Kintsch, W. (2007). Handbook of Latent Semantic Analysis. Mahwah, NJ: Erlbaum.

Dolan, B.; Quirk, C.; and Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources, COLING 2004.

Rus, V. & Graesser, A.C. (2006). Deeper natural language processing for evaluating student answers in intelligent tutoring systems, Paper presented at the Annual Meeting of the American Association of Artificial Intelligence (AAAI-06), July 16-20, 2006, Boston, MA.