

## Adding to LSA's Bag For Information Retrieval

Michael Glass

michael.glass@valpo.edu  
Math & CS Dept., Valparaiso University  
Valparaiso, IN 46383 USA

### Introduction

Feature LSA (FLSA) is the practice of creating non-word features and adding them, as synthetic words, into the bags of words that LSA models. Here we report the largely unsuccessful results of two experiments to enhance LSA effectiveness, one adding a syntactically-derived feature and one adding a semantic one.

### Experiment 1: Word Collocations Within Noun Chunks

For our task we imitated part of the TREC-4 information retrieval task, applying full-sentence queries to retrieve wire service articles.

Starting from a chunked corpus, we added synthetic words representing collocations. Each collocation came from two words  $w_1, w_2$ , not necessarily adjacent, that co-occurred within the same noun group. We kept the synthetic words only for pairs that co-occur within a noun group more frequently than chance would predict. These synthetic words thus recognize many word associations that are not phrasal nouns or frozen phrases. For example *rocket, solid*, which occurs 174 times more likely than chance, is derived from *solid fuel booster rocket* and *solid rocket booster motor* as well as a number of other NPs including the shorthand form *solid rocket*.

A 500-dimension LSA semantic space was constructed from the documents in the collection being searched. Both recall and R-Precision, the precision at the moment of the last retrieved relevant document, show a very slight improvement, shown in Table 1. Various versions of this experiment, e.g. different cutoffs for selecting high-probability pairs, synthetic collocations extracted from the verb groups, etc., all yield similarly small, or even negative, changes in R-Precision and recall.

The failure of this method stems from the almost  $4\times$  explosion in the vocabulary from 110,000 to 416,000 words. We can readily decrease performance by adding extra vocabulary words while holding the number of dimensions constant. That we saw a small performance increase could indicate success of the word pairs, masked by the vocabulary increase effect. There being no principled way to increase the semantic space size as a function of vocabulary size, there is no effective way to compare the two conditions using different numbers of dimensions. (Wiemer-Hastings & Zipitria, 2001) achieved greater success in adding structural information to LSA comparisons of sentence-length texts without our problem of increasing the vocabulary.

Table 1: R-Precision with Synthetic Word Pairs

Exp No.	Rel. Docs Retrieved	R-Precision
Control	867	0.12
NP word pairs	879	0.14

### Experiment 2: Artificial Semantic Tag

(Serafin & Di Eugenio, 2004) achieved excellent results at dialogue act annotation of the CallHome Spanish corpus by encoding structural information as synthetic features along with the annotation tags. By contrast, our task used a simple collection of annotated text, adding the annotation as an artificial word into LSA's bag. Using the hand-coded relevance judgment for each document that comes with the TREC-4 data, we added a synthetic feature word \$r (relevant to one of the queries) or \$i (irrelevant to any query) to 90% of our documents. All of the \$r-tagged documents then have a word-the feature annotation-in common, in addition to whatever semantic similarities existed, making their LSA document vectors slightly more similar to each other. The same holds for the \$i-tagged documents.

The experiment injected \$i and \$r tags alternately into an untagged document, retrieved the most relevant documents from the entire collection under each condition, and compared the retrieval sets. We were unable to use these differences to reliably annotate the query document, the differences induced by a single feature word in 440 words (average) of wire service text appeared to be too slight.

### Acknowledgments

Jessica Warnier provided software support and considerable technical assistance to this project.

This work was supported by the Cognitive Science Program, Office of Naval Research, under grant N00014-03-1-0037 to Valparaiso University. The content does not reflect the position or policy of the government and no official endorsement should be inferred.

### References

- Serafin, R. and B. Di Eugenio (2004). FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, ACL-04, Barcelona*.
- Wiemer-Hastings, P. & I. Zipitria (2001). Rules for Syntax, Vectors for Semantics. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.