Defining the Dimensions of the Human Semantic Space

Vladislav D. Veksler (vekslv@rpi.edu)
Ryan Z. Govostes (govosr@rpi.edu)
Wayne D. Gray (grayw@rpi.edu)

Cognitive Science Department, 110 8th Street
Troy, NY 12180 USA

Abstract

We describe VGEM, a technique for converting probability-based measures of semantic relatedness (e.g. Normalized Google Distance, Pointwise Mutual Information) into a vector-based form to allow these measures to evaluate relatedness of multi-word terms (documents, paragraphs). We use a genetic algorithm to derive a set of 300 dimensions to represent the human semantic space. With the resulting dimension sets, VGEM matches or outperforms the probability-based measure, while adding the multi-word term functionality. We test VGEM's performance on multi-word terms against Latent Semantic Analysis and find no significant difference between the two measures. We conclude that VGEM is more useful than probability-based measures because it affords better performance, and provides relatedness between multi-word terms; and that VGEM is more useful than other vector-based measures because it is more computationally feasible for large, dynamic corpora (e.g. WWW), and thus affords a larger, dynamic lexicon.

Keywords: measures of semantic relatedness, computational linguistics, natural language processing, vector generation, multidimensional semantic space, semantic dimensions, VGEM, Normalized Google Distance, NGD, Latent Semantic Analysis, LSA

Introduction

Measures of Semantic Relatedness (MSRs) are statistical methods for extracting word associations from text corpora. Among the many applications of MSRs are cognitive modeling applications (e.g. Fu & Pirolli, 2007), augmented search engine technology (e.g. Dumais, 2003), and essay grading algorithms used by Educational Testing Services (e.g. Landauer & Dumais, 1997).

Two of the varieties of MSRs are vector-based and probability-based. Probability-based MSRs, such as PMI (Pointwise Mutual Information; Turney, 2001) and NGD (Normalized Google Distance; Cilibrasi & Vitanyi, 2007), are easily implemented on top of search engines (like Google™ search) and thus have a virtually unlimited vocabulary. However, these techniques cannot measure relatedness between multi-word terms (e.g. sentences, paragraphs, documents).

Vector-based MSRs, such as LSA (Latent Semantic Analysis; Landauer & Dumais, 1997) and GLSA (Generalized Latent Semantic Analysis; Matveeva, Levow, Farahat, & Royer, 2005), have the capability to measure relatedness between multi-word terms. However, these MSRs have non-incremental vocabularies based on limited corpora. The general problem with vector-based MSRs is that these measures traditionally require preprocessing steps of creating a word-by-document or a word-by-word matrix, and dimensionality reduction. For example, LSA creates a word-by-document matrix, and reduces dimensions using Singular Value Decomposition. Since Singular Value Decomposition is computationally infeasible for sufficiently large matrices, LSA is limited by both the size of the lexicon, and the size of the corpus. GLSA avoids the corpus-size problem by using a word-by-word matrix, but it is still very computationally expensive to create a GLSA vector-space for all possible words. These techniques require complete corpus re-processing when even a single document is added to the corpus, and thus cannot be used in combination with large dynamic corpora such as the World Wide Web.

There have been some attempts to reduce the vector-MSR preprocessing procedure by avoiding complex dimensionality reduction techniques such as Singular Value Decomposition. HAL (Hyperspace Analogue to Language, Lund & Burgess, 1996) is a vector-based MSR that uses a subset of words from the word-by-word matrix as the vector dimensions. However, HAL is limited in several ways. HAL uses a very simple co-occurrence measure to determine the values in the word-by-word matrix, whereas measures like PMI and NGD may work better. Second, HAL still requires a major preprocessing step, creating the word-by-word matrix. Thus, unlike PMI and NGD, HAL still cannot be used in combination with large dynamic corpora.

In this paper, we describe VGEM (Vector Generation from Explicitly-defined Multidimensional semantic space), a technique to convert probability-based MSR output into vector form by explicitly defining the dimensions of the human semantic space. Unlike other vector-based MSRs, given a standard set of dimensions, VGEM requires no preprocessing, and can be easily implemented on top of search engines (like Google™ search). Unlike probability-based MSRs, VGEM can measure relatedness between paragraphs and documents. We derive high-fidelity dimensions for VGEM via a genetic algorithm, and test VGEM's performance against a probability-based measure using human free-association norms. Finally, we test VGEM's performance on measuring relatedness of multi-word terms against that of LSA.

VGEM

To convert a probability-based measure, *M*, to vector form we first need to define VGEM's semantic space. VGEM's semantic space is explicitly defined with a set of dimensions *d = {d1, d2, ..., dn}*, where each dimension is a word. To compute a vector for some target word, *w*, in this semantic space, VGEM uses *M* to calculate the semantic relatedness between *w* and each dimension-word in *d*:

$$\mathrm{v}(M,w,d) = [\ M(w,d1), M(w,d2), ..., M(w,dn)\ ]$$

For example, if *d* = {"animal", "friend"}, the vector for the word "dog" would be [*M*("dog","animal"), *M*("dog","friend")]. If *M*("dog", "animal") is 0.81 and *M*("dog","friend") is 0.84, then the vector is [0.81, 0.84]. See Table 1, Figure 1.

Table 1: Sample two-dimensional VGEM semantic space.

Words	Dimensions	
	Animal	Friend
Dog	0.81	0.84
Cat	0.81	0.67
Tiger	0.79	0.13
Robot	0.02	0.60

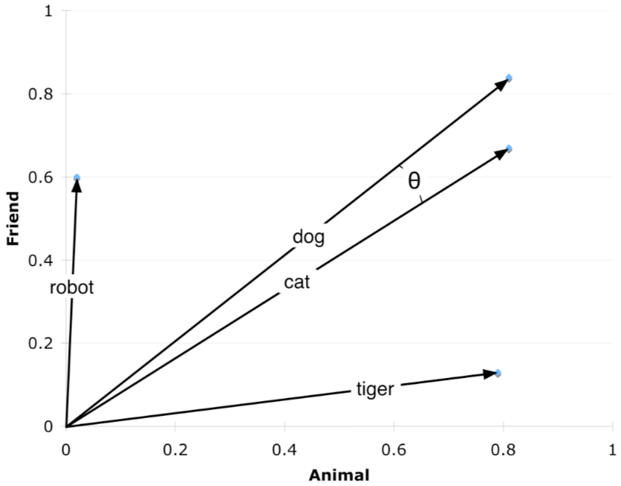

Figure 1. Graphical representation of the sample two-dimensional VGEM semantic space in Table 1.

Like other vector-based measures (e.g. LSA, GLSA), VGEM defines the relatedness between two words to be the cosine of the angle between the vectors that represent those words. As the angle becomes smaller, and the cosine approaches 1.0, the words are considered more related. A value of 1.0 means that the two words are identical in meaning. For example, in Figure 1 the angle between “dog” and “cat” is relatively small, so the cosine of that angle will be close to 1.0 (.994), and the two words will be considered to be more related than any other pair of words shown.

Using this vector-based approach allows VGEM to represent a group of words as a vector sum of the words that make up the group. To compute a vector for a paragraph, VGEM creates a vector representation for each word in that paragraph and adds those vectors component by component. This vector sum represents the meaning of the whole paragraph, and its relatedness to another vector may be measured as the cosine of the angle between them. Continuing with the example in Table 1/Figure 1, the vector to represent the words "dog cat tiger" would be the sum of first three vectors in Table 1, v=[2.41, 1.64].

Advantages of VGEM

The main advantage of VGEM over probability-based MSRs is that it can compute relatedness between multi-word terms. A probability-based MSR cannot find the relatedness between two paragraphs because the probability of any two paragraphs co-occurring (word for word) in any context is virtually zero. VGEM, like other vector-based measures, can represent a paragraph or a document as a vector, and then compare that vector to other vectors within its semantic space.

The main advantage of VGEM over other vector-based MSRs is that VGEM does not require extensive preprocessing. Among other things, this affords VGEM a larger dynamic lexicon. Other vector-based MSRs cannot handle corpora that are very large or corpora that change often, as adding even a single word may require reprocessing the corpora from scratch.

An additional advantage of this technique is the potential to model human expertise and knowledge biases. Explicit specification of the dimensions allows us to bias the semantic space. That is, we can use specific medical terms in the dimension set to represent the semantic space of a medical doctor, and specific programming terminology to represent the semantic space of a software engineer. Unfortunately, this last point is beyond the scope of this paper.

Dimensions

Picking the right set of dimensions is key in obtaining high performance using VGEM. We use a genetic algorithm to pick a good set of dimensions.

Method

NSS In this report, we base VGEM on the probability-based MSR – Normalized Similarity Score (NSS). NSS is an MSR that is derived from NGD. To be more precise, the relatedness between two words *x* and *y* is derived as follows:

$$\mathrm{NSS}(x, y) = 1 - \mathrm{NGD}(x, y)$$

where NGD is a formula derived by Cilibrasi & Vitanyi (2007):

$$\mathrm{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

f(*x*) is the frequency with which *x* may be found in the corpus, and *M* is the total number of searchable texts in the corpus. It is not necessary to use NSS with VGEM, as PMI and other similar metrics may be used. We chose NSS because some previous testing has revealed that overall it is a better model of language than PMI (Lindsey, Veksler, Grintsvayg, & Gray, 2007).

Corpora We trained our technique on two separate corpora. The first corpus was a subset of about 2.3 million paragraphs from the Factiva™ database ("Factiva," 2004) as a training corpus for NSS. The corpus consisted of archived news articles. We also used a standard TASA (Touchstone Applied Science Associates) corpus, used for LSA by Landauer & Dumais (1997). The TASA corpus is a collection of text representative of the reading material that a person is supposed to have been exposed to by their first year in college (Zeno, Ivens, Millard, & Duvvuri, 1995). As corpus choice makes a great difference (Lindsey et al., 2007), future research will explore other corpora (e.g. WWW, Wikipedia, and Project Gutenberg).

MSR Evaluation Function We use the Nelson-McEvoy free association norms (Nelson, McEvoy, & Schreiber, 1998) to compare VGEM's associative ratings with human ones. The association norms consist of 5019 *cue-targets* sets gathered from more than 6000 participants, where target words were participants' free-association responses to the respective cue words (e.g. *cue*='old', *targets*={'new', 'young', 'ancient', 'man', 'wrinkle', 'age', 'grandparent', 'house', 'wise'}). Number of targets per cue in the Nelson-McEvoy set varies from 1 to 34 (M=14.38, SE=.07). For each set of *cue-targets* we picked *n* random distracter words, where *n* is equal to the number of *targets*. Distracter words were chosen from the same dataset (e.g. 'buyer', which is a target word for *cue*='owner', could have been chosen as a distracter word for *cue*='old'). We took a random sample of 500 *cue-targets-distracters* test cases to evaluate a given MSR, *M* (for our purposes, each evaluated MSR is actually $\text{VGEM}_{\textbf{NSS-Factiva}}$ with a different set of dimensions). In our chosen set of 500 test cases the number of targets per cue varies from 2 to 28 (M=14.18, SE=.23).

To evaluate each *cue-targets-distracters* test case, each of the *targets-distracters* lists is sorted according to the *M*-derived relatedness values, *M(cue,word)*, where *word*∈{*targets,distracters*}. The score for each *cue-targets-distracters* test case for a given MSR, *M*, was calculated as follows:

$$Score_{\text{case}} = \frac{\textit{Number of targets in top n words}}{n}$$

where "*top n words*" are the top half of the sorted *targets-distracters* list. If, according to *M*, all target words are more related to *cue* than any of the distracter words, the score for that test case is *100%*. If none of the target words are picked by *M* to be more related to the *cue* than any of the distracter words, the score is *0*. The overall score for *M* is the average of all test case scores.

Candidate Dimensions Candidate dimensions were chosen from a subset of Alan Beale's 2+2gfreq wordlist (Beale, 2007), which is a list of 32,638 English words plus variations thereof. Beale's word list is arranged by groups, based on word-frequency data supplied by Google™. We used words from groups 11 through 13 (a total of 5,066 words), which contain words that are widely used, but are not overly common (e.g., the, of, and) nor overly specific (chemical compounds, medical jargon, etc.).

Genetic Algorithm (GA) We used a genetic algorithm to derive a set of 300 'good' dimensions from the 5,066 candidate dimensions described above. 300 is the number of dimensions used in the popular implementation of LSA, derived by Landauer and Dumais (1997). It also seems to be a reasonable, computationally-inexpensive size of the VGEM semantic space for beginning to explore VGEM capabilities.

The genetic algorithm was set up as follows. We chose a population size of 100 individuals with chromosomes of 300 dimensions. The initial population's genetic makeup was chosen uniformly at random from the set of candidate dimensions. The population was split up into two arbitrary designations of male and female, where males are intended to be more variable (prone to mutations) while females are more stable.

For each iteration of the genetic algorithm, the population was ranked in accordance with the MSR evaluation function described above. The lower-ranked half of the individuals of each gender was then removed from the population.

Next, each remaining female was crossed with two males, where the males were chosen uniformly at random, with replacement, from the remaining population. Each pair produced one new female and one new male. An offspring received 150 unique dimensions from each parent's chromosome. To increase variance, the chromosome was then subjected to mutation. We replaced *n* dimensions in each chromosome with dimensions chosen uniformly at random from the set of candidate dimensions, $0 \leq n < 60$ for males and $0 \leq n < 6$ for females.

Due to constraints on time and computational resources we ran only 8,000 generations for each of the training corpora. Future work will involve a more rigorous search for better dimension sets.

Hill-climbing (HC) After the genetic algorithm completed, the highest scoring set of VGEM dimensions was passed through a hill-climbing algorithm to refine the results. Here we try replacing each dimension in the winning set with an unused dimension from the candidate set. If such a replacement increases the score, it is incorporated in the final solution.

Results

Hill-climbing significantly improved the best results of the genetic algorithm, proving that the GA had not converged yet. Improvements for the VGEM-NSS-Factiva were 1.77%, $t(499)_{\text{two-tail}}$=6.68, p<.01, and for VGEM-NSS-TASA were .65%, $t(499)_{\text{two-tail}}$=4.15, p<.01. See Figure 2.

VGEM-NSS-Factiva performed significantly better than NSS-Factiva, $t(499)_{\text{two-tail}}$=12.03, p<.01. VGEM-NSS-TASA matched the performance of NSS-TASA, showing no significant difference, $t(499)_{\text{two-tail}}$=.16, p=.87. See Figure 2.

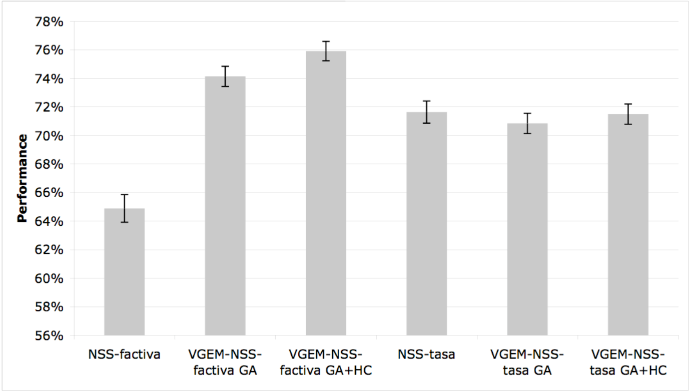

Figure 2. Performance of NSS trained on Factiva and TASA corpora, and VGEM-NSS on same corpora with GA- and HC- derived dimensions. Error bars signify standard error.

Discussion

The current report is a demonstration of VGEM and should not be interpreted as a test of VGEM-NSS-TASA versus VGEM-NSS-Factiva. Our demonstration of VGEM is limited by our use of only 300 dimensional VGEM spaces, the relatively small number of candidate dimensions (5,066), GA generations/population size (8,000 and 100), and the use of only two relatively small training corpora. It may well be the case that VGEM-NSS-TASA would perform better than VGEM-NSS-Factiva with higher dimensionality, more GA iterations, or a different set of candidate dimensions.

However, we can safely conclude that using VGEM on probability-based MSRs like NSS does not hinder and can even improve performance. We speculate that VGEM can improve the performance of probability-based MSRs because VGEM dimensions provide context. That is, relatedness between two words is not merely defined as, "how often do these words co-occur," but rather, "how often do these words occur in the same context." This especially helps in the case of synonyms, as synonyms are rarely found near each other in text.

Additionally, translating NSS into vector form provides the capability to measure relatedness of multi-word terms.

Multi-word terms

Our motivation for translating probability-based MSRs into vector form is to add the capability to measure relatedness between documents, not merely single words. NSS (and similar measures) can be used to find relatedness between documents based on the relatedness of all the words in these documents. However, there are two problems with this. First, these computations are computationally expensive (e.g. to relate two 10,000 word documents would require 100 million NSS computations). Second, and perhaps more importantly, some terms are more important than others, and a simple average of word-to-word relatedness values would not be an accurate approximation of document relatedness.

Vector-based measures like VGEM and LSA resolve both of these problems. The computational complexity of measuring relatedness between two documents is linear (to relate two 10,000 word documents would require 20,000 vector additions). As for term weighting, in vector addition the vector lengths represent term weights.

Method

To test VGEM's capability of measuring the relatedness of multi-word terms we compare VGEM performance (based on the dimension sets that we had derived) to that of LSA. LSA is a powerful technique that has been used with success for automatically grading student essays (Landauer & Dumais, 1997), to model human language learning (Landauer & Dumais, 1997), to model language comprehension (Lemaire, Denhiere, Bellissens, & Jhean-Iarose, 2006), and more.

We test VGEM and LSA on how well these measures can pick out 2-sentence blocks from within a given text passage when presented with distracters from other passages. The assumption here is that text extracts from the same document should be found more related than text extracts from different documents.

We used thirty-two short text passages to make *cue-targets-distracters* test cases. The text passages were chosen randomly from TOEFL-like reading comprehension sections on the Web ("Reading Comprehension," 2005; Reading comprehension and vocabulary," 2007). We created 32 test cases, one for each passage. The *cue* for each test case was the concatenation of the first two sentences of the passage. The remaining pairs of sentences from that passage were *targets* for that test case. The set of distracters consisted of *n* pairs of random sentences from the other thirty-one text passages, where *n* is equal to the number of *targets*. The number of *targets* per *cue* varied from 1 to 10 (M=4.38, SE=.57). The overall performance for each MSR on these 32 test cases is evaluated using the same procedure as for the free association test cases, described in the MSR Evaluation Function section above.

Results

Performance difference between VGEM-NSS based on the TASA corpus (M=89.1%, SE=2.6%) and LSA trained on the same corpus (M=87.7%, SE=2.8%) was not significant, $t(31)_{\text{two-tail}}$=.58, p=.568. The difference between VGEM-NSS based on the Factiva corpus (M=82.5%, SE=2.9%) and LSA based on the TASA corpus was not significant, $t(31)_{\text{two-tail}}$=1.68, p=.103. See Figure 3.

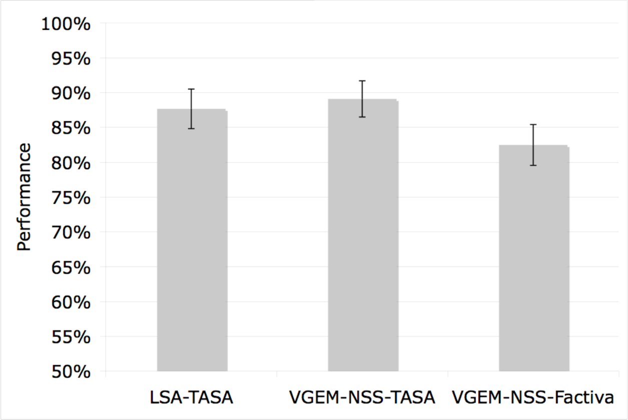

Figure 3. Performance of VGEM-NSS and LSA on within document text identification. Error bars signify standard error.

Discussion

Given the derived sets of dimensions, we find VGEM's multi-word text relatedness measurement capabilities to be on par with those of LSA, without the need for computationally expensive preprocessing. The genetic algorithm may be viewed as a necessary preprocessing step for VGEM, but it only needs to be done once, whereas LSA has to be retrained every time the corpus changes. It may even be possible to avoid the genetic algorithm altogether, and simply use an established ontology as a set of dimensions for VGEM.

It should be noted that it is possible for text extracts from the same document to be less related to each other than text extracts from different documents, so 100% performance may not be possible or wanted in the above evaluation. It is, however, encouraging that both LSA and VGEM trained on the same corpus (TASA) picked almost 90% of text extracts within documents to be more closely related than across documents.

The real power of these vector-based techniques is that multi-word terms provide more context than individual words. For example, we used VGEM-NSS-Factiva with the best-found set of dimensions to compare the sentence (cue) "when was the last time you scored a goal" to (target) "my favorite sports are basketball and football" and to (distracter) "the goal of this paper is to provide a technical description of our measure." Although the distracter sentence has a word in common with the cue ("goal"), whereas the target sentence has no words in common with the cue, VGEM can actually tease out the context and score the cue-target pair as more related than the cue-distracter pair.

As with LSA (Landauer & Dumais, 1997), we speculate that the larger the body of text, the more accurate VGEM's judgments should be. In future studies we plan to examine VGEM's ability to judge document similarity and compare it to human judgments.

Summary

In this paper we described a technique for converting probability-based MSRs like NSS and PMI into vector form to allow these computationally inexpensive measures to compare multi-word terms (documents, paragraphs). The proposed technique, VGEM, was used to convert a probability-based measure, NSS, into vector-based form. We used a genetic algorithm to derive a set of dimensions for VGEM-NSS, and tested VGEM-NSS against NSS on human free-association norms. Finally, we tested VGEM-NSS capability to measure relatedness of multi-word terms against another vector-based measure, LSA.

The results are promising. On tests of human free-association norms VGEM did not hinder and even improved the performance of NSS. We speculate that VGEM can improve the performance of probability-based measures because VGEM calculates word relatedness based on whether words occur in similar contexts, as opposed to how often words co-occur. Additionally, VGEM-NSS was able to measure multi-word term relatedness. On tests of multi-word term relatedness VGEM-NSS did as well as the more computationally expensive measure, LSA.

We conclude that VGEM is more useful than probability-based MSRs because it affords better performance and provides relatedness between multi-word terms. We conclude that VGEM is more useful than other vector-based measures because it is more computationally feasible for large, dynamic corpora (e.g. WWW), and thus affords a larger, dynamic lexicon. Although the use of a genetic algorithm may be viewed as a computationally expensive step, we hold that it needs to be done only once to establish a universal set of dimensions, whereas other vector-based MSRs have to be reprocessed every time the corpus changes.

Future studies will include a more rigorous search for a universal set of dimensions for VGEM, explore the use of established taxonomies as dimension sets, and evaluate VGEM across different corpora (e.g. the World Wide Web, Wikipedia, etc.). Additionally, we plan to examine practical applications of VGEM, such as semantic search indexing and text-summary generation.

Acknowledgements

We would like to thank TASA for making their corpus available for academic research. We would also like to thank Thomas Landauer for giving us permission to use the TASA corpus, and for making the corpus available, and for providing access to the LSA engine online at http://lsa.colorado.edu/.

This work was supported, in part, by grants from the Office of Naval Research (N000140710033) and the Air Force Office of Scientific Research (FA9550-06-1-0074) to Wayne Gray.

References

Beale, A. (2007). Release 5 of the 12dicts word lists. 2007, from http://wordlist.sourceforge.net/12dicts-readme-r5.html

Cilibrasi, R., & Vitanyi, P. M. B. (2007). The Google similarity distance. [Article]. *Ieee Transactions on Knowledge and Data Engineering, 19*(3), 370-383.

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science, 27*(3), 491-524.

Factiva. (2004). 2007, from http://www.factiva.com

Fu, W. T., & Pirolli, P. (2007). SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web. *Human Computer Interaction*.

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*(2), 211-240.

Lemaire, B., Denhiere, G., Bellissens, C., & Jhean-Iarose, S. (2006). A computational model for simulating text comprehension. [Article]. *Behavior Research Methods, 38*(4), 628-637.

Lindsey, R., Veksler, V. D., Grintsvayg, A., & Gray, W. D. (2007). *Be Wary of What Your Computer Reads: The Effects of Corpus Selection on Measuring Semantic Relatedness.* Presented at the 8th International Conference of Cognitive Modeling, ICCM 2007, Ann Arobor, MI.

Lund, K., & Burgess, C. (1996). Hyperspace analogue to language (HAL): A general model semantic representation. *Brain and Cognition, 30*(3), 5-5.

Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). *Term representation with generalized latent semantic analysis.* Presented at the 2005 Conference on Recent Advances in Natural Language Processing.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. (Publication. Retrieved 12/06/2006, from http://www.usf.edu/FreeAssociation/.

Reading Comprehension. (2005). 2007, from http://www.majortests.com/sat/reading-comprehension.php

Reading comprehension and vocabulary. (2007). 2007, from http://education.kulichki.com/lang/toefl1.html

Turney, P. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)* (pp. 491-502). Freiburg, Germany.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*: Touchstone Applied Science Associates (TASA), Inc.