Similarity Between Semantic Spaces

Xiangen Hu (xhu@memphis.edu)
Department of Psychology; The University of Memphis
Memphis, TN 38152

Zhiqiang Cai (zcai@memphis.edu)
Department of Psychology; The University of Memphis
Memphis, TN 38152

Arthur C. Graesser (agraesser@memphis.edu)
Department of Psychology; The University of Memphis
Memphis, TN 38152

Matthew Ventura (mventura@memphis.edu)
Department of Psychology; The University of Memphis
Memphis, TN 38152

Abstract

One of the challenges in Latent Semantic Analysis (LSA) is deciding which corpus is best for a specific application.Important factors of LSA influence the generation of high quality LSA space including the size of the corpus, the weight (local or global) functions, number of dimensions to keep, etc. These factors are often difficult to determine and as a result hard to control for. In this paper, we provide a general method to measure similarity between semantic spaces. Using this method, one can evaluate semantic spaces (such as LSA spaces) that are generated from different sets of parameters or different corpora. The method we have developed is generic enough to evaluate differing types of semantic spaces.

Keywords: Semantic Space, Latent Semantic Analysis, Similarity Measures Between Texts

Introduction

The use of higher dimensional semantic spaces, such as Latent Semantic Analysis (LSA) (Landauer & Littman, 1990; Dumais, 1990; Laham, 1997; Landauer, Laham, Rehder, & Schreiner, 1997; Landauer, Foltz, & Laham, 1998), Hyperspace Analogue to Language (HAL) (Burgess, Livesay, & Lund, 1996; Burgess & Lund, 1997; Burgess, 1998), Non-Latent Similarity (NLS) algorithm (Cai et al., 2004), is very common in computational linguistics. The semantic spaces have been used in applications that involve information retrieval(Dumais, 1990), essay grading, and text comparison (Foltz, Laham, & Landauer, 1999). The scope and depth of the applications are so diverse that different semantic spaces are needed for different purposes (Franceschetti et al., 2001). Furthermore, the process of generating semantic spaces is very complicated (Deerwester et al., 1990). Some of them involve careful selection of corpora (Franceschetti et al., 2001). From all the previous studies and applications of semantic spaces such as LSA, HAL, and NLS, we observed that there are several key parameters (such as dimensions, domain, corpus size, etc.) that need to be set before generating an appropriate space for an application. There are different methods that can be used to compute similarities between documents (Hu et al., 2003, 2003; Hu, Cai, Wiemer-Hasting, Graesser, & McNamara, 2005). The issue of evaluating the quality of semantic space is very important, not only at the level of theoretical importance, but also at the level of specific applications.

Currently, the quality of semantic spaces is usually evaluated by human experts. This is done by comparing performances between applications that use a specific semantic space and have human experts perform some benchmark tests. For example, to evaluate an LSA space with a given set of parameters (e.g., number of dimensions), an LSA similarity measure between texts is compared with experts' judgement of the similarity of those texts (Olde, Graesser, & Tutoring Research Group, 2002). There are many possible variables that are involved in creating semantic spaces, which makes it impractical for human experts to evaluate all semantic spaces.

In this paper, we present a systematic method to automatically evaluate semantic spaces. This method allows us to measure similarity between semantic spaces that are created from different sets of parameters (domain, corpus size, dimensions, etc.). Furthermore, this method can even be used to find differences between semantic spaces that are created using entirely different methods, such as LSA, HAL, and NLS.

Observations

We first provide a mathematical model for Semantic Space. This model is simply an abstraction of some commonly used semantic spaces such as LSA , HAL, and NLS. Based on this model, we provide a measure of similarity between semantic spaces. At the end of the paper, we outline procedures of how to use the similarity measures to evaluate semantic spaces. The method presented in this paper is based on the following observations from semantic spaces, such as LSA , HAL, and NLS:

1. Semantics is a property that applies to five differ-

ent levels of language entities: *word*, *phrase*, *sentence*, *paragraph*, and *document*. In any given language,

- the smallest semantic units are words. For example, "*this*", "*is*", "*a*", "*big*", "*table*".
- a phrase is an *ordered array* of words. For example, "*big table*".
- a sentence is an *ordered array* of words and phrases. For example, "*This is a big table.*"
- a paragraph is an *ordered array* of sentences. For example, "*This is a big table. It was broken.*"
- a document is an *ordered array* of paragraphs.

2. Semantics of any level of the language entities can be represented numerically or algebraically
   - Semantics and the numerical or algebraic *representation* are synonymous.
3. Semantics of different levels of the language entities may be represented differently, but
   - Semantics of a higher level language entity is computed as a function of semantics of its lower level language entities.
   - Semantic relations between any two entities at the same level can be numerically measured as a function of the semantics of the entities.
4. The meaning of any word is represented by its (*numerically measurable*) *relations* with other words in the same semantic space. We call such a relation *induced semantic structure* of the word in the given semantic space.

Definitions

To formalize the above assumptions and the concept of semantic structure, we have a formal definition of vector-based semantic spaces.

Definition 1 *A vector-based semantic space contains five components:*

1. *A set of words* $X_0 = \{x_1, x_2, ..., x_N\}$;
2. *A hierarchy of layers,* $X_1, ..., X_M$, *where an element in the set* X_i *is a finite ordered array of elements in* X_{i-1} $(i = 1, ..., M)$;
3. *Vector representation for elements in each of the layers.*
4. *Measure of similarity between elements within each of the layers.*
5. *Maps from lower level representations to higher level representations*

For vector representations of elements in each layer, i. e., $\forall x \in X_i$, there exists a vector $E_i(x) \in R^\infty$ with only finite non-zero entries, $i = 1, ..., M$;

For measure of similarity for each layer: Assume $S_i : R^\infty \times R^\infty \longrightarrow R, i = 1, ..., M.$ such that

- $S_i(\mathbf{x}_1, \mathbf{x}_2) < \infty$ if $\mathbf{x}_1$ and $\mathbf{x}_2$ have only finite non-zero elements,
- $S_i(\mathbf{x}_1, \mathbf{x}_2) > 0$ if $\mathbf{x}_1 = \mathbf{x}_2 \neq \mathbf{0}$,
- $S_i(\mathbf{x}_1, \mathbf{x}_2) = 0$ if $\mathbf{x}_1 = \mathbf{0}$ or $\mathbf{x}_2 = \mathbf{0}$.

The similarity measure $s_i(x_1, x_2)$ between $x_1, x_2 \in X_i$ is defined in as

$$\frac{S_i(E_i(x_1), E_i(x_2))}{\sqrt{S_i(E_i(x_1), E_i(x_1))}\sqrt{S_i(E_i(x_2), E_i(x_2))}},$$

where $E_i(x_1)$ and $E_i(x_2)$ are not zero vectors.

For maps from lower level representations to higher level representations follow the following constraints:

- if $x = (y_1, ..., y_k) \in X_i$, $y_1, ..., y_k \in X_{i-1}$. for some $k > 0$, then

  $$E_i(x) = H_i(E_{i-1}(y_1), ..., E_{i-1}(y_k)),$$

  where H_i is a function $H_i : [R^\infty]^k \longrightarrow R^\infty$;
- For $x_1 = (y_{11}, ...y_{1u}) \in X_i$ and $x_2 = (y_{21}, ...y_{2v}) \in X_i$, where $y_{11} \in X_{i-1}$, $S_i(E_i(x_1), E_i(x_2))$ is in the form of Eqn. (1), where U_i is a function $U_i : [R^\infty]^u \times [R^\infty]^v \longrightarrow R$, for some $u, v > 0$.

  $$S_i(E_i(x_1), E_i(x_2)) = U_i(\mathbf{U}, \mathbf{V}) \quad (1)$$

  where $\mathbf{U} = (E_{i-1}(y_{11}), E_{i-1}(y_{12}), ..., E_{i-1}(y_{1k_1}))$ and $\mathbf{V} = (E_{i-1}(y_{21}), E_{i-1}(y_{22}), ..., E_{i-1}(y_{2k_2}))$ and $x_1, x_2 \in X_i, y_{11}, ..., y_{1k_1}; y_{21}, ..., y_{2k_2} \in X_{i-1}$

Definition 1 is similar with the four components model of Lowe (2001). The difference that arises between Lowe and out definition is that this definition considers not only the word level, but also all other levels with assumed mapping from lower layers to higher layers. To understand the above definition, consider the five language entities, namely, word, phrase, sentence, paragraph, and document. Each corresponds to a different layer. X_1 is a set of phrases, X_2 is a set of sentences, etc. For every element, there is a vector representation in R^∞. In Definition 1, we do not specify a limited dimensionality for the vector representation. Instead we assume there is an infinite dimensional vector with only a finite number of non-zero entries. To understand 5 and 4 of Definition 1, one can take LSA as a simple example, where the semantic vector of a sentence is simply a vector summation of the vectors of the words in the sentence. Furthermore, the similarity between two words (or two sentences) is a function of the two word (sentence) vectors. 5 of Definition 1 emphasizes the relations between different layers. The similar relations can be seen from LSA, where the computation of similarity between documents is a function of the vectors of the words.

For the purpose of this paper, we next generalize the idea of "near neighbor" of LSA in the new framework of semantic space. From this concept, we further introduce the idea of induced semantic structure. These two concepts will serve as the foundation for the remainder of the paper.

Definition 2 *Given a semantic space with layers* $X_0, ..., X_M$, $\forall x \in X_i$, *the neighbor of* x *is* $\{(y, s_i(x,y)) | \, y \in X_i\}$.

The neighbor of any element in any of the layer X_i is simply a partial ordered set. We call such an ordered set *induced semantic structure*.

Definition 3 *Given a semantic space with layers* $X_0, ..., X_M$, $\forall x \in X_i$, *the induced semantic structure* $\mathcal{S}_{x,i} \subset X_i \times X_i$ *is a partial order defined in Eqn. (2).*

$$\left\{ (x_1, x_2) \,\middle|\, \begin{array}{c} (x_1, x_2 \in X_i) \text{ and} \\ (s_i(E_i(x), E_i(x_1)) \geq S_i(E_i(x), E_i(x_2))) \end{array} \right\}. \quad (2)$$

Assumptions

With the above definitions, we have the following assumptions. These assumptions serve as the theoretical foundation for our similarity measure of vector based semantic spaces.

Assumption 1 *The meaning of a word is embedded in its relations with other words.*

As an illustrative example (see Fig 1), the word "life" has different near neighbors for different LSA spaces.

Assumption 2 *If a given word is shared in different semantic spaces, the relation between the semantics of the word in different semantic spaces is a function of the corresponding induced semantic structures.*

In Assumption 2, we consider only the algebraic (ordering) nature (as in Eqn. (2)) of the near neighbors.

Assumption 3 *The relations between any two semantic spaces are a function of the relations of the semantic structures of all the shared words*

Assumption 3 extends Assumption 2 from the level of the word to the entire semantic space.

Similarity Measures

With the above assumptions, we are able to measure similarity between semantic spaces at three different levels: *Combinatorial Similarity, Permutational Similarity* , and *Quantitative Similarity.* From 5 of Definition 1, we see that all layers $X_i, ..., X_M$ of a semantic space actually depend on X_0 and the mappings from lower layers to higher layers. This makes it easier to introduce the general measure of semantic similarity between semantic spaces.. In this paper, we only consider similarity measures that are derived at the layer of the basic items, namely, X_0.

Combinatorial Similarity

Based on Assumption 1, the meaning of a word is determined by its relations with all other words in a semantic space. Using Assumption 2, we first have the *Combinatorial Similarity* at the level of individual word. Applying Assumption 3, we will have the *Combinatorial Similarity* at the level of semantic spaces.

Assume X_0 and Y_0 are layers in two semantic spaces. For any given item $x \in X_0 \cap Y_0$, there are two induced semantic structures in each of the semantic spaces. Denote them as $\mathcal{S}_x^1$ and $\mathcal{S}_x^2$. Assume N_1 and N_2 are the number of words in the two semantic spaces (defined in 1 of Definition 1), respectively, where $T \leq \min(N_1, N_2)$. Furthermore, assume[1] $S_{x,T}^1$ and $S_{x,T}^2$ are the top T nearest neighbor of word x. The combinatorial similarity for word x between the two semantic spaces is defined as

$$C_x^T = \frac{\left\| S_{x,T}^1 \cap S_{x,T}^2 \right\|}{\left\| S_{x,T}^1 \cup S_{x,T}^2 \right\|} \quad (3)$$

where $\|X\|$ is the number of items in set X. Given $T \leq \min(N_1, N_2)$, with such a definition of semantic similarity for any word x. One can obtain similarity for any collection of words, $W \subset X_0 \cap Y_0$, as statistical properties of $\left\{ C_x^T \middle| \, x \in W \subset X_0 \cap Y_0 \right\}$. For simplicity, we only consider mean and standard derivation of $\left\{ C_x^T \middle| \, x \in W \subset X_0 \cap Y_0 \right\}$, although we may consider other characteristics. Furthermore, we have the similarity defined as a function of the value T. In fact

$$\left\{ C_x^T \middle| \, x \in W \subset X_0 \cap Y_0, 1 \leq T \leq \min(N_1, N_2) \right\} \quad (4)$$

contains all information between the two semantic spaces at the "combinatorial sense". Statistical properties of (4) can be used to measure the *Combinatorial Similarity* between two spaces. For example, if W is a collection of physics glossory terms, then statistical properties of (4), namely, mean and standard deviation, would be a measure of semantic similarity of these terms between the two spaces.

Permutational Similarity

Permutational similarity is defined in the same way as *Combinatorial Similarity*, except the comparison of the top T nearest neighbors of x in the two semantic spaces is not only combinatorial, but also permutational. Consider $\left(S_{x,T}^1 \cap S_{x,T}^2\right) = \left\{x_1^{'}, ..., x_\tau^{'}\right\} = \left\{x_1^{"}, ..., x_\tau^{"}\right\}$ and the orders of the nearest neighbors for x: $\left(x_1^{'}, ..., x_\tau^{'}\right)$ and $\left(x_1^{"}, ..., x_\tau^{"}\right)$ for the two semantic spaces, respectively. d is a function that measures the permutational distance between two orders. It is assumed that

$$d\left((x_1, ..., x_\tau), (x_1, ..., x_\tau)\right) = 0,$$

and

$$d\left(\left(x_1^{'}, ..., x_\tau^{'}\right), \left(x_1^{"}, ..., x_\tau^{"}\right)\right) \leq d\left(\left(x_1^{'}, ..., x_\tau^{'}\right), \left(x_\tau^{'}, ..., x_1^{'}\right)\right).$$

We define the quantity

$$P_x^T = \left(1 - \frac{d\left(\left(x_1^{'}, ..., x_\tau^{'}\right), \left(x_1^{"}, ..., x_\tau^{"}\right)\right)}{d\left(\left(x_1^{'}, ..., x_\tau^{'}\right), \left(x_\tau^{'}, ..., x_1^{'}\right)\right)}\right) C_x^T \quad (5)$$

[1] In some cases, there is no unique T top neighbours, because the induced semantic structure is only a partial order. We only consider the simplest case here. We will not consider the cases where no unique top T nearest neighbours in this paper.

as permutational similarity of x in two semantic spaces for a given T. Similarly, permutational similarity can be defined at the level of semantic spaces for any given set of words, $W \subset X_0 \cap Y_0$,

$$\left\{ P_x^T \middle| \, x \in W \subset X_0 \cap Y_0, 1 \leq T \leq \min\left(N_1, N_2\right) \right\} \quad (6)$$

contains similarity between the two semantic spaces at the permutational level. Consequently, statistical properties of (6), such as mean and standard deviation can be used for such purposes.

Quantitative Similarity

Combinatorial Similarity and *Permutational similarity* are based on algebraic properties of the induced semantic structure as a partial order. *Quantitative Similarity* is based on quantitative property of the nearest neighbor (Definition 2). For any $x \in X_0 \cap Y_0$, and $T \leq \min\left(N_1, N_2\right)$, there is a simple quantitative relation between $\left\{ \left(y, s_0^1\left(x,y\right)\right) \middle| \, y \in S_{x,T}^1 \cap S_{x,T}^2 \right\}$ and $\left\{ \left(y, s_0^2\left(x,y\right)\right) \middle| \, y \in S_{x,T}^1 \cap S_{x,T}^2 \right\}$. For example, one could use Pearson correlation r between the two set of quantities. In the same way,

$$Q_x^T = \left(1 - \frac{\sum s_0^1\left(x,y\right) s_0^2\left(x,y\right)}{\sqrt{\sum \left[s_0^1\left(x,y\right)\right]^2 \sum \left[s_0^2\left(x,y\right)\right]^2}} \right) C_x^T \quad (7)$$

where the sum is obtained for all $y \in S_{x,T}^1 \cap S_{x,T}^2$. Furthermore, a set of quantities can be obtained for $W \subset X_0 \cap Y_0$,

$$\left\{ Q_x^T \middle| \, x \in W \subset X_0 \cap Y_0, 1 \leq T \leq \min\left(N_1, N_2\right) \right\}. \quad (8)$$

Quantitative similarity is defined as a set of statistical properties of (8). As usual, mean and standard deviation can be used for such purposes.

An Example

In this section, we apply the definitions, assumptions, and similarity measures to LSA spaces. The following is true for LSA spaces:

1. LSA contains a set of words.
2. At the layer corresponding to phrases, sentences, and documents, LSA does not consider ordering of lower layer items. One may view the "bag of words" as a equivalent class of the ordered arrays containing the same set of items.
3. The representation of LSA is a finite dimensional vector. It can be viewed as infinite vector with finite non-zero entries.
4. The similarity measure $S_i\left(\mathbf{x}_1, \mathbf{x}_2\right)$ is simply the dot-product of the vectors
5. The maps between lower layers to higher layers is a pool of the items from the lower layers. The vector representation of higher layer items is simply a vector summation of the vectors from the lower layers.

Assume that we have two LSA spaces[2] $L_1 = \left(X_0, X_2, X_3, X_4, X_5\right)$ and $L_2 = \left(Y_0, Y_2, Y_3, Y_4, Y_5\right)$, and a common set of words $W = \left\{x_1, x_2, ..., x_N\right\} = X_0 \cap Y_0$. Two matrices can be obtained by considering near neighbors (Definition 2) for all words in W : $S_1 = \left(s_{1ij}\right)$ and $S_2 = \left(s_{2ij}\right)$, where $s_{kij} = \cos_k\left(x_i, x_j\right)$ is the cosine match between x_i and x_j within the LSA space k, $i, j = 1, ..., N$; $k = 1, 2$. Notice that such two matrices contain all necessary information needed for all three different levels of similarities.

For the purpose of illustration, and due to space limitation of the paper, we compute C_x^T, P_x^T, and Q_x^T for the word "life".

Table 1 lists near neighbors for several LSA spaces. We computed C_x^T, P_x^T, and Q_x^T (as defined in Eqn. (3), (5) and (7)) with the value $T = 50$ (see Tables 2, 3, and 4), We observed that the meaning of "life" is most similar between 6th grade and 9th grade and between 9th grade and 12th grade in the Touchstone Applied Science Associates (TASA) corpus.

In this example, we have five different spaces generated from TASA corpus. The same method can be used to find semantic similarities for a group of words between different corpora. When using a group of words, instead of using quantities C_x^T, P_x^T, and Q_x^T as similarity measures, one needs to use statistical properties of coresponding sets of quantities (defined in (4),(6),(8)) such as mean and standard deviation.

Table 1: Top 20 nearest neighbours of "life" for different LSA spaces (only showing 6th–12th, due to the space limitation of the paper). The rank is taking from http://lsa.colorado.com.

6th	9th	12th
life	life	life
reincarnation	contemplated	death
premiums	reincarnation	lifetime
policyholder	sai	hamlin
premium	pipal	pipal
sai	nirvana	nirvana
cycles	lifetime	zarathustra
holdover	death	ahuramazda
condemning	hinduism	ahriman
chekhov	afterlife	policyholder
captial	excerpted	romantics
pipal	ribman	essayists
nirvana	reaffirm	sai
hinduism	militarily	beaumarchais
span	kindless	1658
priori	condemning	pseudonym
maturity	premiums	poquelin
immoral	premium	ribman
humane	policyholder	kindless

[2]In applications of LSA, only X_0 and very small portion of other layers are meaningful. Due to the simple algorithm of combining word vectors to sentence or document vectors, items in other layers can be computed very easily.

Table 2: Combinatorial similarity of the word "life" between several LSA spaces.

	3rd	6th	9th	12th	College
3ed	1				
6th	0.0526	1			
9th	0.0204	0.2987	1		
12th	0.0000	0.0989	0.2500	1	
College	0.0000	0.0989	0.1111	0.2048	1

Table 3: Permutational Similarity of the word "life" between several LSA spaces.

	3rd	6th	9th	12th	College
3ed	1				
6th	0.0491	1			
9th	0.0136	0.2078	1		
12th	0.0000	0.0478	0.2130	1	
College	0.0000	0.0462	0.0626	0.1450	1

Table 4: Quantitative Similarity of the word "life" between several LSA spaces.

	3rd	6th	9th	12th	College
3ed	1				
6th	0.0524	1			
9th	0.0202	0.2975	1		
12th	0.0000	0.0979	0.2495	1	
College	0.0000	0.0975	0.1101	0.2043	1

Possible applications of the method

Although we only have shown a simple application of the semantic similarity measures, namely, only at the level of single word ("*life*"), we argue that the method can be easily applied to evaluate similarities between semantic spaces. For example, several parameters need to be set for any give LSA space. The method introduced in this paper can be used to measure differences due to different values of those parameters.

- *Size of Documents*: LSA created a word-document matrix as the original matrix for later SVD. It is not very clear what the ideal document size is. Using the similarity measure we have here, we can systematically vary the document size and measure the similarity among LSA spaces with different document sizes.
- *Selection of number of dimensions to keep*: After SVD, only dimensions corresponding to the largest singular values are kept. The number of dimensions kept is only about 1~3% of the total number of dimensions. One question is to address the robustness of the selection of the dimensions. Using the similarity measure, we can compare LSA spaces with different dimensions.
- *Type of corpus*: LSA has been used in different domains. For example, LSA has been used in tutoring systems that teach computer literacy and qualitative physics(Graesser et al., 2002). Usually, different corpora are used for different target application domains. There are also general purpose corpora, such as the TASA corpus. One question is how different the LSA spaces are that are created from these different corpora. The similarity measure offered here can be used to evaluate the LSA spaces, such as LSA spaces generated from physics text, computer literacy texts, and TASA.

Summary

In this paper, we provide a general approach to measure similarity between semantic spaces. We first offer some observations of commonly used semantic spaces. From these observations, we introduce a set of general assumptions. Finally we have a mathematical model of semantic spaces. Based on this model, we are able to derive three quantitative measures between semantic spaces. Finally, we have used the similarity measures to examine semantic similarity of a single word ('life") in several LSA spaces.

The space limitation of this paper does not permit us to offer more, elaborated applications of these similarity measures. However, we have offered several possible applications of the method. We argue that the method introduced in this paper will help researchers to select parameters in the process of creating semantic spaces.

Acknowledgments

The Institute for Intelligent Systems (IIS) is an interdisciplinary research group comprised of approximately 100 researchers from psychology, computer science, physics, engineering, linguistics, education and other disciplines (visit http://www.iismemphis.org). This research, conducted by the authors and the IIS, was supported by the National Science Foundation (REC 0106965) and the DoD Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or NSF. We would like to thank Tom Landauer and Walter Kintsch by supplying a number of corpora used in this study.

References

Burgess, C. (1998). From simple associations to the building blocks of language modeling meaning in memory with the hal model. *Behavior Research Methods, Instruments, and Computers*, *30*, 188-198.

Burgess, C., Livesay, K., & Lund, K. (1996). Modeling parsing constraints in high-dimensional semantic space: On the use of proper names. In *Proceedings of the cognitive science society.* Hillsdale, N.J.: Erlbaum.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*, 177-210.

Cai, Z., McNamara, D. S., Louwerse, M., Hu, X., Rowe, M., & Graesser, A. C. (2004). Nls: Non-latent similarity algorithm. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual meeting of the cognitive science society* (p. 180-185). Mahwah, NJ: Erlbaum.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, K., T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society For Information*, *141*, 391-407.

Dumais, S. (1990). *Enhancing performance in latent semantic indexing (lsi) retrieval* (TM-ARH-017527 Technical Report). Bellcore.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *In proceedings of EdMedia '99*.

Franceschetti, D., Karnavat, A., Marineau, J., McCallie, G. L., Olde, B. A., Terry, B. L., Graesser, A., & C. (2001). Development of physics text corpora for latent semantic analysis. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the cognitive science society* (p. 297-300). Mahwah, NJ: Erlbaum.

Graesser, A. C., Hu, X., Olde, B. A., Ventura, M., Olney, A., Louwerse, M., Franceschetti, D. R., & Person, N. (2002). Implementing latent semantic analysis in learning environments with conversational agents and tutorial dialog. In W. G. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual meeting of the cognitive science society* (p. 37). Mahwah, NJ: Erlbaum.

Hu, X., Cai, Z., Franceschetti, D., Penumatsa, P., Graesser, A., Louwerse, M., McNamara, D., & TRG. (2003). Lsa: The first dimension and dimensional weighting. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the 25rd annual conference of the cognitive science society* (p. 1-6). Boston, MA: Cognitive Science Society.

Hu, X., Cai, Z., Graesser, A. C., Louwerse, M. M., Penumatsa, P., Olney, A., & TRG. (2003). An improved lsa algorithm to evaluate student contributions in tutoring dialogue. In G. Gottlob & T. Walsh (Eds.), *Proceedings of the eighteenth international joint conference on artificial intelligence* (p. 1489-1491). San Francisco: Morgan Kaufmann.

Hu, X., Cai, Z., Wiemer-Hasting, P., Graesser, A., & McNamara, D. S. (2005). Strengths, limitations, and extensions of lsa. In D. S. S. W. Landauer T.;McNamara (Ed.), *Lsa: A road to meaning.* Mahwah, NJ: Erlbaum. (in press)

Laham, D. (1997). Latent semantic analysis approaches to categorization. *In M. G. Shafto and P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society (p. 979). Mawhwah NJ: Erlbaum.*

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. *In M. G. Shafto and P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society*, 412-417.

Landauer, T. K., & Littman, M. L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. *In Proceedings of the 6th Annual Conferenceof the Centre for the New Oxford English Dictionaryand Text Research*, 31-38.

Lowe, W. (2001). Towards a theory of semantic space. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the twenty-third annual conference of the cognitive science society* (pp. 576–581).Mahwah NJ: Lawrence Erlbaum Associates.

Olde, B. A., Graesser, A. C., & Tutoring Research Group the. (2002). Latent semantic analysis: What is it and how can it improve and assess student learning? *Paper presented at the North East Regional Conference on Excellence in Learning and Teaching, Oswego, NY.*