Document Similarity Misjudgment by LSA: Misses vs. False Positives

Kyung Hun Jung (kjung2@kennesaw.edu)
Department of Psychology, Kennesaw State University (Kennesaw Campus)
402 Bartow Ave Kennesaw (Rm 4030), GA 30144 USA

Eric Ruthruff (ruthruff@unm.edu)
Department of Psychology, Logan Hall MSC03-2020
1 University of New Mexico, Albuquerque, NM 87131-0001 USA

Timothy Goldsmith (gold@unm.edu)
Department of Psychology, Logan Hall MSC03-2020
1 University of New Mexico, Albuquerque, NM 87131-0001 USA

Abstract

Modeling text document similarity is an important yet challenging task. Even the most advanced computational linguistic models often misjudge document similarity relative to humans. Regarding the pattern of misjudgment between models and humans, Lee and colleagues (2005) suggested that the models' primary failure is occasional underestimation of strong similarity between documents. According to this suggestion, there should be more extreme misses (i.e., models failing to pick up on strong document similarity) than extreme false positives (i.e., models falsely detecting document similarity that does not exist). We tested this claim by comparing document similarity ratings generated by humans and latent semantic analysis (LSA). Notably, we implemented LSA with 441 unique parameter settings, determined optimal parameters that yielded high correlations with human ratings, and finally identified misses and false positives under the optimal parameter settings. The results showed that, as Lee et al. predicted, large errors were predominantly misses rather than false positives. Potential causes of the misses and false positives are discussed.

Keywords: text document relatedness; semantic similarity; latent semantic analysis (LSA)

Introduction

Modeling how humans judge the semantic similarity of text documents is an interesting topic in cognitive science with numerous practical implications. In an effort to better model human document similarity judgments, Lee, Pincombe, and Welsh (2005) compared several models of document similarity, including latent semantic analysis (LSA). They found that LSA's cosine similarity scores yielded higher agreement (r = .60) with aggregate human ratings than other models, such as Tversky's (1977) ratio model (r = .50). Considering that the inter-rater correlation among human raters is also about .60, LSA seems to judge about as well as a single human rater. However, the moderate correlation between humans and LSA also suggests that LSA does not fully capture human similarity judgments[1].

To better understand the weaknesses of LSA, and thereby improve models of text document similarity, this study investigated the pattern of discrepancy between LSA and humans with respect to their document similarity ratings. Specifically, we examined the frequency and degree of underestimation (misses) and overestimation (false positives) made by LSA relative to humans under favorable parameter settings of LSA.

Misses vs. False Positives

Regarding the nature of the misjudgment by models, Lee et al. (2005) suggested that extreme misses would be a stronger cause than extreme false positives. They made this suggestion based on an observation that the *common features model* (Lee & Navarro, 2002) occasionally misses the high similarity between documents that is readily apparent to humans. In a scatterplot of the model ratings against human ratings for each document pair, the authors found a cluster of points (document pairs) with low model ratings but high human ratings. That is, the model missed some of the strong document similarities that humans detected.

Lee et al.'s (2005) analysis above was based on the common features model, but it might apply to LSA as well. The common features model judges document similarity primarily based on the proportion of common features (words) shared by two documents. Notably, LSA's underlying model, the vector space model, determines document similarity in a similar manner. Therefore, Lee et al.'s findings suggesting more extreme misses over extreme false positives may also apply to LSA. This interesting hypothesis, if validated, would provide a valuable clue to improving models of text document similarity. However, it has not yet been rigorously tested.

Testing the hypothesis seems straightforward at first glance: compare human and LSA's document similarity ratings for a set of documents pairs. Then, document pairs with especially low LSA ratings compared to human ratings should be considered as misses and the reverse as false

[1] For the document pairs used in Lee et al. (2005), the highest reported correlation between a model and humans was .77 (Yeh, Ramage, Manning, Agirre, & Soroa, 2009).

positives. However, LSA's document representation depends on the parameters used such as the quantity and quality of the background documents (Bullinaria & Levy, 2006), the dimensionality (Dumais, 1991; Landauer & Dumais, 1997), and the local-global weighting schemes (Lintean, Moldovan, Rus, & Mcnamara, 2010; Nakov, Popova, & Mateev, 2001). Therefore, LSA's misjudgments relative to human judgments could vary depending on the parameters.

In this study, we attempted to investigate the nature of misjudgment by LSA under its optimal parameter settings. Therefore, we first identified LSA's optimal parameter settings by employing as many as 441 unique parameter combinations. Then, under the selected optimal parameter settings, we identified misjudgments by LSA as misses or false positives. Finally, we measured the degree of misjudgment of the two types using normalized scores.

The remainder of this paper has the following structure: (1) introduction to LSA, (2) an experiment identifying optimal parameter settings, (3) identification of misjudgments as misses and false positives under the optimal parameter settings, and (4) discussion of the underlying causes of the misses and false positives.

LSA

LSA is based on a vector space model in which documents are first transformed into a word-by-document matrix. Rows of the matrix correspond to the unique words across documents, whereas columns correspond to individual documents. Cell values are the frequencies of words within each document. The cell values can be weighted in two respects: to what degree a word is important in representing a document's topic (local weighting), and to what degree a word is important in distinguishing one document from another according to their topics (global weighting). Using the weighted cell values, each document can be represented as a vector in a multidimensional space, where the dimensions correspond the unique words. Finally, the sematic similarity between two document vectors is typically measured using the cosine similarity score.

The core process that distinguishes LSA from the vector space model is singular value decomposition (SVD) implemented on the word-by-document matrix. SVD is a matrix factorization method that decomposes an original matrix (A) into three sub-matrices, USV^T, where U is a unitary w * r matrix (word-by-dimension matrix), S is an r * r diagonal matrix with non-negative real numbers on its diagonal (singular value matrix), and V^T is a unitary r * d matrix (dimension-by-document matrix). By multiplying these three sub-matrices, the original matrix can be retrieved, and this type of SVD is called full SVD.

In a modified version of the full SVD, called reduced SVD, small singular values located in the lower right corner of S are intentionally discarded, while preserving the first k largest singular values. The corresponding columns and rows of U and V^T, respectively, are discarded, too. The original USV^T, after discarding some values, then can be denoted as $U'S'(V^T)'$, where U' is a w * k matrix whose columns are the first k columns of U, S' is a k * k diagonal matrix whose diagonal elements are the k largest singular values of S, and $(V^T)'$ is a k * d matrix whose rows are the first k rows of V^T. By multiplying these three reduced sub-matrices, one can obtain the least squares approximation of the original matrix. Finally, documents can be represented as vectors on a k dimensional singular-value-space, which has k orthogonal axes. These dimensions are constructed so that the first axis explains the largest amount of variance of A, and the second axis explains the second largest amount of variance of A, and so on.

Furnas et al. (1988) was the first to apply the reduced SVD to the vector space model. This method was later called latent semantic analysis by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), who also demonstrated that LSA retrieves information better than traditional word-matching methods. Deerwester et al. argued that SVD uncovers latent semantic relations across documents that are buried in the corpus by removing noise (small singular values) in the original word-by-document matrix.

Identification of Misses and False Positives under LSA's Optimal Parameter Settings

Stimuli and Procedure

Target Text Documents We used the 1,225 document pairs from Pincombe (2004), which Lee et al. (2005) also adopted. These document pairs were generated by pairing 50 target news articles selected from Australian Broadcasting Corporation's news mail service. Each news article had a single paragraph containing 51 to 126 words (average: 82 words). They covered a variety of topics, such as terrorism and hunger in Africa. For each of the 1,225 document pairs, Pincombe collected about 10 human ratings by asking 83 university students to each rate the relatedness of a subset of the document pairs. Participants used a five-point scale, with one indicating "highly unrelated" and five indicating "highly related".

Background Documents Lee et al. (2005) used 314 news articles from the same Australian news corpus as background documents[2]. In this study, to explore the optimal parameter settings of LSA, we employed 4,172 additional news articles from the same news corpus (total of 4,486). These new background documents contained a single paragraph (average: 152 words). They also covered a variety of topics as the 50 target news articles did. In addition to the background document size used in Lee et al. (314) and the maximum size available in this study (4,486), we examined

[2] Background documents are included in the original corpus subject to LSA, along with the target documents. They are employed only for constructing the multidimensional space in which the target documents are represented. It is generally regarded that LSA's performance improves as the number of background documents increases (Bullinaria & Levy, 2006).

five intermediate background document sizes by randomly selecting the following numbers of documents from the new set of 4,172 articles: 314, 750, 1,000, 2,000, and 3,000 (see Table 1).

Table 1. Seven background document conditions.

Size	Source
314	The same 314 news articles as in Lee et al. (2005)
314	Randomly selected from the 4,172 articles
750	Randomly selected from the 4,172 articles
1,000	Randomly selected from the 4,172 articles
2,000	Randomly selected from the 4,172 articles
3,000	Randomly selected from the 4,172 articles
4,486	Combination of the 314 news articles from Lee et al. and the new 4,172 articles

Dimensionality Regarding the dimensionality of the reduced SVD, the maximum possible dimension for a given background document size corresponds to the total number of documents subjected to SVD (50 + the number of background documents). For example, in the 314-background document condition, the maximum dimension is 364 (= 50 + 314). In most of the background document conditions employed in this study, higher dimensions than 364 were possible. However, following some researchers' arguments for the importance of maintaining 300 dimensions (Landauer & Dumais, 1997), we selected the following seven dimensions for the reduced SVD (i.e., LSA): 50, 100, 150, 200, 250, 300, and 364.

Other LSA Parameters Stemming, normalization, and removal of stopwords and alphanumeric words are known to improve LSA's document representation (Pincombe, 2004; Stone, Dennis, & Kwantes, 2011). Therefore, they were applied to all LSA runs. Three local weighting schemes (tf, log, and alt-log) and three global weighting schemes (idf, entropy, and p-inverse) were selected based on their significant effects observed in a pilot study (not reported here).

LSA cosine scores were computed for every possible (441) combination of the above parameters: 7 background document sizes * 7 dimensions * 3 local weighting schemes * 3 global weighting schemes.

Identifying Misses and False Positives To classify LSA ratings as misses and false positives relative to human ratings, we first normalized the human ratings and LSA's cosine scores using z-score[3]. The degree of misjudgment was measured as the absolute difference between the two normalized scores for a given document pair. If a document pair's normalized cosine score was smaller than the normalized human rating by at least 1.0, then the LSA's cosine score was considered a miss. But if a document pair's normalized cosine score was greater than the normalized human rating by 1.0, then the LSA cosine score was considered a false positive.

Results and Discussion

Optimal Parameters of LSA To determine which parameter settings are optimal for LSA's document similarity representation, we examined the correlation between LSA cosine scores and human ratings. The correlation was affected more systematically and strongly by the interaction of background document size and dimensionality than the local-global weighting schemes. Therefore, for the sake of simplicity, we merged the correlations across the nine weighting schemes at a given background document size and dimensionality. As shown in Table 2, the correlation increased markedly as we added more background documents, consistent with previous research (Bullinaria & Levy, 2006). But this effect was more prominent at relatively high dimensions than at low dimensions.

Table 2. Correlations between human ratings and LSA cosine scores as a factor of the background document size and dimensionality. Correlations were merged across the nine local-global weighting schemes at a given background document size and dimensionality. Relatively high correlations ($r \geq .67$) are shaded.

	Dimension							
Background	50	100	150	200	250	300	364	Average
314	0.53	0.55	0.55	0.56	0.56	0.57	0.59	0.56
New 314	0.62	0.59	0.58	0.56	0.57	0.58	0.61	0.59
750	0.65	0.66	0.63	0.61	0.61	0.60	0.59	0.62
1000	0.62	0.66	0.66	0.64	0.62	0.61	0.60	0.63
2000	0.51	0.61	0.64	0.68	0.67	0.68	0.67	0.64
3000	0.53	0.61	0.64	0.64	0.66	0.67	0.69	0.63
4486	0.58	0.65	0.67	0.69	0.68	0.68	0.66	0.66
Average	0.58	0.62	0.62	0.63	0.62	0.63	0.63	0.62

To identify optimal parameter settings of LSA, we first selected the 10 combinations of background document size and dimension that yielded correlations of at least .67, averaged across all weighting schemes (see the shaded cells in Table 2). Then, for each of these 10 combinations, we chose the local-global weighting scheme that yielded the highest correlation with human ratings. Table 3 shows the specific parameter settings of these 10 selected combinations as optimal parameter settings. The table also shows the correlation, number of misses and false positives, and the average absolute z-score errors.

[3] We considered the approach of transforming scores into a 0-1 scale, as in Lee et al. (2005). However, this approach is overly sensitive to the minimum and maximum values. On the other hand, the z-score normalization yields more reliable results with respect to the frequency of misses and false positives.

Table 3. Ten optimal parameter settings of LSA selected for the identification of misses and false positives. The parameters, correlation with human ratings, number of misses and false positives, and the average of the absolute z-score errors are shown.

Background document size	Dimension	Local Weighting	Global Weighting	Correlation	Number of misses	Number of false positives	Average misjudgment (absolute z-score error) for misses	Average misjudgment (absolute z-score error) for false positives
2000	200	tf	p-inverse	0.70	112	86	1.57	1.43
2000	250	tf	p-inverse	0.68	121	67	1.60	1.56
2000	300	tf	idf	0.68	117	72	1.61	1.56
2000	364	alt-log	p-inverse	0.68	118	78	1.60	1.54
3000	300	tf	p-inverse	0.68	115	86	1.63	1.47
3000	364	alt-log	entropy	0.69	104	76	1.61	1.42
4486	150	alt-log	p-inverse	0.68	119	95	1.57	1.42
4486	200	tf	p-inverse	0.70	120	83	1.62	1.48
4486	250	tf	p-inverse	0.68	124	78	1.64	1.58
4486	300	tf	idf	0.68	119	69	1.61	1.49
			Average	0.69	117	79	1.57	1.43

Nature of Misjudgments by LSA To determine the nature of LSA's misjudgments under optimal parameters, we used the z-score errors obtained from the 10 parameter settings shown in Table 3. As shown at the bottom of the table, misses (M_{Miss} = 117) were much more common than false positives ($M_{False\ Positive}$ = 79), χ^2 (1, N = 1,959) = 73.324, p < .001, just as suggested by Lee et al. (2005). Also, as the error magnitude increases, the ratio of misses to false positives also increases, which is consistent across the 10 optimal parameter settings. Figure 1 shows the frequency of the two types of errors (misses vs. false positives) as a function of the absolute z-score error.

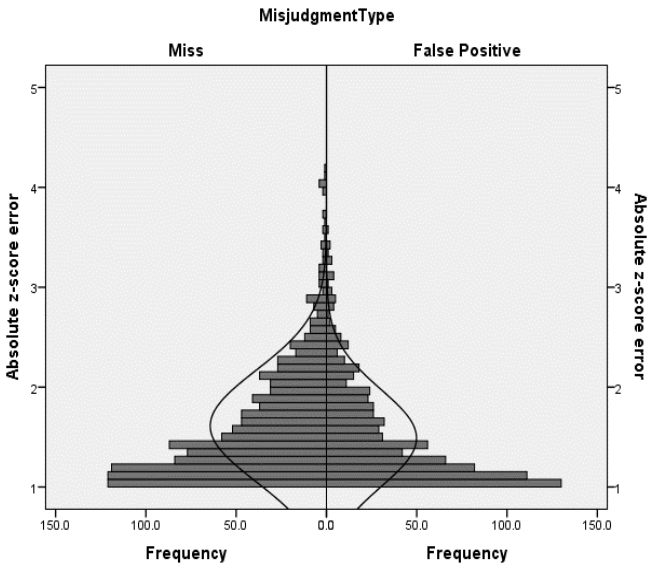

Figure 1. The frequency of the two types of misjudgment by LSA as a function of the absolute z-score error.

Effect of the Parameters on the Frequency of Misses and False Positives Although the distribution of the two types of errors by LSA at optimal parameters was the primary focus of this study, we also examined the ratio of misses to false positives across all the 441 parameter settings. The results showed that the ratios were systematically affected by the interaction between the background document size and dimensionality. That is, the ratio of misses to false positives increased as the dimensionality increased. However, the degree of increase is getting less prominent as the background document size increases. In other words, although there were more misses than false positives in general, the disproportion of misses over false positives is more prominent at high dimensions with small number of background documents.

Effect of the Number of Background Documents on Correlation between Humans and LSA One of the most striking findings above was the strong effect of background document size on LSA's document similarity representation. As shown in Table 2, employing more background documents (combined with an appropriate dimensionality) tends to significantly improve LSA's document similarity judgments. To illustrate the significant effect of background documents, we plotted the correlation between LSA and human ratings for three background document sizes (0, 314, and 4,486) and the nine weighting schemes as a function of dimensionality (Figure 2). The graph illustrates (a) the strong effect of the number of background documents, (b) important effect of dimensionality when the background document size is small (i.e., the left side of the graph), and (c) the relative unimportance of weighting schemes.

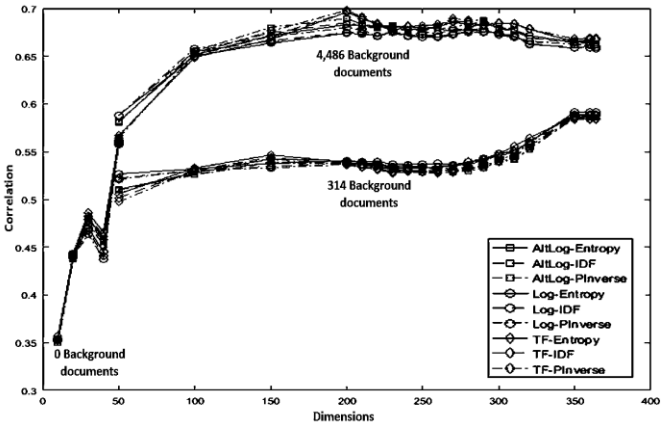

Figure 2. The correlation of ratings between humans and LSA for three background document conditions (0, 314 and 4,486) and nine weighting schemes as a function of dimensionality.

One may suspect that including even more background documents would further increase the correlation. However, to make a positive impact on LSA's performance, the background documents should not only be numerous but also relevant to the content of the target documents (Foltz, Britt, & Perfetti, 1995). For example, Stone, Dennis, and Kwantes (2011) tested the effect of various kinds of background documents on LSA's document similarity judgments, using the same 50 target news articles examined in this study. They tested 55,021 Canada Toronto Star newspaper articles (miscellaneous gossip paragraphs, from the year 2005) and 10,000 articles selected by the researchers from online encyclopedia, Wikipedia (http://www.wikipedia.org/). However, the highest correlation between humans and LSA obtained was about .10 with the gossip news articles as background and .40 with the Wikipedia background documents. These correlations were significantly lower than the highest correlation of .60 obtained in Lee et al. (2005) and .70 in the current study, despite utilizing only 314 and 4,486 background documents, respectively. Therefore, not only the size but also the relevance of background documents to the target documents seem to be critical for LSA's document similarity judgments.

If target documents came from a certain population (e.g., specific news corpus), we recommend using documents from the same population as background documents. In our case, employing background documents of 4,172 news articles that came from the same population as the target articles increased the correlation between humans and LSA from .60 to .70.

Conclusion

Lee et al. (2005) suggested that the primary weakness of computational models of document similarity is failing to pick up on some of the strong document similarities that humans easily detect. To test this hypothesis, we compared the document similarity ratings made by humans and LSA based on a range of parameter combinations. Then we identified the frequency and degree of large misses and large false positives under optimal parameters of LSA. The results confirmed that LSA makes more misses than false positives, especially among the most severe errors.

The results also suggest that if one attempts to further improve models of text document similarity by reducing its errors relative to humans, the misses rather than the false positive would be the primary focus of the revision. More specifically, one should look for ways to help models pick up on some of the strong semantic similarities that they currently miss.

Potential Causes of Misses and False Positives

An obvious follow-up question of this study is what causes LSA's greatest misses and false positives. Considering that LSA's basis, the vector space model, judges document similarity based on the overall word similarity between two documents, a potential cause of error is that LSA misses or falsely overestimates the semantic similarity of some word pairs from two documents. In fact, there are various cases where LSA cannot help but miss some of the word similarities, which in turn would cause one type of error, miss. For example, although "United States", "US", "U.S.", and "U.S.A" refer to the same country, they may not be recognized as the same entity in the word-by-document matrix for various reasons: because they are not a single word (United States), too short to be included (US or U.S. after the special character removal), or happen to match an excluded stop word (US and the pronoun us). However, humans would correctly recognize them and utilize these words for document similarity judgment.

Also, some words (especially proper nouns including human names) may occur in the target documents but not in the background documents, preventing LSA from utilizing those words in judging document similarity. However, those words could be critical for humans to judge the document similarity. Then, LSA may judge document pairs including those words to be less related than humans would do (i.e., leading to a miss).

The above-mentioned potential cause of misses (i.e., LSA misses document similarity because it misses word similarity in document pairs) could be further supported if LSA's document similarity scores do correspond to the overall word similarity between two documents. To confirm this, we calculated the correlation between the 1,225 document pairs' LSA cosine scores and the average LSA cosine scores of every possible word pair from each of the document pairs. We found a correlation of .73 from this analysis, indicating that LSA's document similarity is heavily relying on the overall word similarity in document pairs.

Similar to the potential cause of misses by LSA addressed in the above, a potential cause of false positives by LSA is that LSA mistakenly perceives semantic similarity between words that are in fact unrelated. Table 4 shows 10 word pairs that were judged to be highly related by humans and LSA, respectively in one of the document pairs used in this study. Although LSA does generally make reasonable judgments on

word relatedness, some word pairs judged to be highly related by LSA do not seem to have a meaningful relationship. For example, *design* and *document* were the most strongly related word pair to LSA, despite being seemingly unrelated. Thus, LSA will occasionally overestimate the relatedness of the document pairs that include this word pair.

Table 4. The 10 most related words pairs to humans and LSA from a pair of news articles.

Highly related word pairs by human	Ratings (1-5 scale)	Highly related word pairs by LSA	Ratings (z-score)
dollar-money	5.00	design-document	6.87
job-money	5.00	increase-rise	5.61
angrily-attack	4.90	paid-worker	3.97
plan-target	4.83	effect-target	3.38
increase-profit	4.80	group-work	3.37
money-profit	4.80	effect-increase	3.08
cost-lawsuit	4.78	disclosure-profit	2.89
job-meet	4.75	disclosure-financial	2.83
agreement-plan	4.73	commonwealth-deal	2.41
job-paid	4.73	australia-target	2.32

An alternative hypothesis regarding the misses and false positives of LSA is that, when judging document similarity, humans do not rely on the overall word similarity as much as LSA does. As Griffiths, Steyvers, and Tenenbaum (2007) suggested, humans may catch the gist of each document and compare the semantic representations of the gist rather than relying on the overall similarity of words in the documents. Then, two documents with a large overlap of words but with different topics would be regarded unrelated by humans although they could be highly related to LSA (resulting in false positives). To assess to what degree human document similarity judgments rely on the overall word similarity, one could examine the correlation between human document similarity ratings and the average of the human similarity ratings for all the possible word pairs in a given document pair. If humans do not rely on the overall word similarity as much as LSA does, then the correlation would not be as high as the corresponding correlation of LSA.

References

Bullinaria, J. A., & Levy, J. P. (2006). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*, 510-526.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391-407.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, and Computers, 23*, 229-236.

Foltz, P. W., Britt, M. A., & Perfetti, C. A. (1995). Measuring text influence on learning history. *Proceedings of the Fifth Annual Winter Text Conference*, Jackson, WY.

Furnas, G. W., Deerwester, S. C., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., & Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. *Proceedings of the Eleventh Annual International ACM* 81 *SIGIR Conference on Research and Development in Information Retrieval.* 465-480. Grenoble, France.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 211-244.

Jung, K. (2013). *Mismatches between humans and latent semantic analysis in document similarity judgments* (Doctoral dissertation).

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*, 211-240.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*, 43-58.

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). An empirical evaluation of models of text document similarity. *Proceedings of Twenty Seventh Annual Conference of the Cognitive Science Society* (pp. 1254-1259). Mahwah, NJ: Erlbaum.

Lintean, M., Moldovan, C., Rus, V., & McNamara D. S. (2010). The role of local and global weighting in assessing the semantic similarity of texts using Latent Semantic Analysis. *Proceedings of the Twenty Third International Florida Artificial Intelligence Research Society Conference*. Daytona Beach, FL.

Nakov, P., Popova, A., & Mateev, P. (2001). Weight functions impact on LSA performance. *Proceedings of the Recent Advances in Natural Language Processing* (pp. 187-193). Tzigov Chark, Bulgaria.

Pincombe, B. M. (2004). *Comparison of human and LSA judgments of pairwise document similarities for a news corpus* (Tech. Rep. DSTO-RR-0278). Adelaide, Australia: Australian Defense Science and Technology Organization, Intelligence, Surveillance and Reconnaissance Division.

Stone, B., Dennis, S., & Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis, *Topics in Cognitive Science, 3*, 92-122.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Yeh, E., Ramage, D., Manning, C. D., Agirre, E., & Soroa, A. (2009). WikiWalk: Random walks on Wikipedia for semantic relatedness. *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing* (pp. 41-49). Stroudsburg, PA, USA: Association for Computational Linguistics.