Using Latent Semantic Analysis to Estimate Similarity

Sabrina Simmons (s.g.simmons@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry CV4 7AL, UK

Zachary Estes (z.estes@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry CV4 7AL, UK

Abstract

In three studies we investigated whether LSA cosine values estimate human similarity ratings of word pairs. In study 1 we found that LSA can distinguish between highly similar and dissimilar matches to a target word, but that it does not reliably distinguish between highly similar and less similar matches. In study 2 we showed that, using an expanded item set, the correlation between LSA ratings and human similarity ratings is both quite low and inconsistent. Study 3 demonstrates that, while people distinguish between taxonomic / thematic word pairs, LSA cosines do not. Although people rate taxonomically related items to be more similar than thematically related items, LSA cosine values are equivalent across stimuli types. Our results indicate that LSA cosines provide inadequate estimates of similarity ratings.

Latent Semantic Analysis (LSA) is a statistical model of language learning and representation that uses vector averages in semantic space to assess the similarity between words or texts (Landauer & Dumais, 1997). LSA has been used to model a number of cognitive phenomena and correlates well with many human behaviors relating to language use. Owing to the model's success, LSA ratings are now being used in place of human ratings as measures of semantic similarity. The purpose of this paper is to investigate whether LSA cosine values are adequate substitutes for human semantic similarity ratings.

Overview of LSA

The basic theory behind LSA is that the "psychological similarity between any two words is reflected in the way they co-occur in small sub-samples of language" (Landauer & Dumais, 1997 p. 215). The model begins with a matrix taking words as rows and contexts as columns, although, theoretically, many other types of information could also be used. The contexts may be anything, for example, newspaper articles, textbooks, or student essays and the words are simply those that appear in the training set. Importantly, the contexts with which the model is provided will determine what types of words it has experience with, so the training set should be relevant to the task the model is to perform. The first step is to associate each word with the contexts in which it is likely to appear. In addition to recording the frequency that a given word occurs in particular texts, the model weights entries to reflect the diagnosticity of a word for a given context. For example, a word that appears in a large number of very different contexts is not as diagnostic as a word that occurs less often and only in a small set of similar contexts.

The next steps in the process are essential to LSA's ability to uncover higher order associations between words. Through a combination of singular value decomposition (SVD) and dimension reduction, the representations of words that occur in the same or similar contexts become themselves, more similar (Kwantes, 2005). In this case, a word's representation is a vector in semantic space that summarizes information about the contexts in which that word is found. In this way the similarity between two words can be determined by the cosine between their vectors (although other methods are sometimes used, see Rehder, Schreiner, Wolfe, Laham, Landauer, & Kintsch, 1998). Thus, through a process that contains no information about semantic features, the definition of words, word order, or parts of speech, LSA is able to capture subtle relationships between words that might never have occurred together (Landauer & Dumais, 1997).

LSA in Application

LSA is able to model several human cognitive abilities and has a number of potential practical applications. To give a few examples, LSA can imitate the vocabulary growth rate of a school age child (Landauer & Dumais, 1997), is able to recognize synonyms about as accurately as prospective college students who speak English as a second language (Landauer & Dumais 1994), can pass a college level multiple choice test in psychology (Landauer, Foltz, & Laham, 1998), can successfully simulate semantic priming (Landauer & Dumais, 1997), assesses essay quality in a manner consistent with human graders (Landauer, et. al. 1998), and is able to determine what text a student should use in order to optimize learning (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch & Landauer, 1998).

Because of its success at modeling human performance on such a range of semantic tasks, LSA is sometimes used as a tool for stimuli norming and construction. For example, Howard and Kahana (2002) used LSA ratings as a measure of semantic similarity to predict a number of behaviors relating to free recall. Gagne, Spalding, and Ji (2005) have used LSA scores to control for semantic similarity between prime-target pairs in a relational priming study. Finally, in a very interesting application, LSA was used as a measure of the similarity of text samples in order to predict different health outcomes (Campbell & Pennebaker, 2003).

Using LSA in stimuli construction is an attractive idea. Collecting LSA scores does not require the time and money needed to acquire similarity ratings from human participants. Also, although they depend on the training corpus and number of dimensions, LSA ratings are relatively objective compared to human similarity judgments, which are influenced by a number of factors including task (Medin, Goldstone, & Gentner, 1993), context (Estes & Hasson, 2004), and expertise (Hardiman, Dufresne, & Mestre, 1989), to name just a few. But, if LSA cosines are to be used as estimates of similarity, it is essential to compare LSA scores directly with human similarity ratings.

In the following studies we provide a detailed evaluation of how well LSA scores predict human ratings of semantic similarity (cf. Steyvers, Shiffrin & Nelson, 2004). We address this question in a number of ways. Moving from the more general to the more specific, we first determine whether LSA can consistently distinguish a highly similar match to a given word from a dissimilar match to the same word (Study 1). Next, using an expanded item set, we investigate the ability of LSA scores to predict human similarity ratings (Study 2). Based on the outcomes of these studies, we examine whether LSA scores are differentially sensitive to taxonomic versus thematic relations between words (Study 3). For each study, cosine values were obtained from the publicly available internet version of LSA (http://lsa.colorado.edu), using the "General Reading Up to First Year College" corpus with 300 dimensions. Between 200 and 500 dimensions are usually recommended (see Landauer & Dumais, 1997), and Styvers et al. (2004, Figure 18.2) found that around 300 dimensions was optimal.

Study 1

As an initial step, we wanted to establish the extent to which LSA scores demonstrate a similarity "preference" for certain types of word pairs in a manner consistent with human similarity ratings. For example, given that people generally agree that *cat* is more similar to *dog* than it is to *geranium* will LSA give a higher cosine value to *cat / dog* than to *cat / geranium*? If so, then LSA scores might suffice for the construction of lists where a given target word has a similar and dissimilar match.

Method

We selected stimuli from three previously published experiments related to semantic similarity. Stimuli were 38 similar and 38 dissimilar word pairs from Gentner & Markman (1994), and 37 similar and 37 dissimilar word pairs from McRae & Boisvert (1998, Experiment 1) of the type described above, where a given target word has both similar and dissimilar matches. For example, the target *cart* was matched with *wagon* (highly similar) and *lemon* (dissimilar). The third stimuli set (McRae & Boisvert, 1998; Experiment 3) included a "less similar" match to the target word, in addition to similar and dissimilar matches. Continuing the above example, *cart* was matched with *wagon* (highly similar), *jet* (less similar) and *lemon* (dissimilar).

Results and Discussion

For each target we took the difference between the cosine of the similar match and the cosine of the dissimilar match. For the items from McRae & Boisvert (1998, Experiment 3) the difference was taken between both the highly similar and less similar, and between the less similar and dissimilar word pairs. We then calculated the percent agreement between the matches that LSA preferred (i.e. the match with the higher cosine value) and the matches preferred by human raters.

In the majority of cases LSA derived a higher cosine value for the similar word pair than for the dissimilar word pair. For two of the stimuli sets (Gentner & Markman, 1994; McRae & Boisvert, 1998; Experiment 1) 92% of the targets received a higher cosine when paired with the similar match than when paired with the dissimilar match. For the McRae and Boisvert (1998; Experiment 3) items, LSA correctly distinguished between highly similar and less similar word pairs 65% of the time and between the less similar and dissimilar items 88% of the time. The results suggest that LSA can consistently distinguish between very similar and very dissimilar matches, but that it performs less well at the higher end of the similarity scale.

Study 2

Based on the results of study 1, LSA cosine values seem sensitive to differences in similarity between items, at least when these differences are quite large. However, the data of McRae and Boisvert (1998) suggested that LSA was more predictive at the lower end of the similarity range. To investigate this further, we used an expanded item set and collected participants' ratings of the similarity of a variety of word pairs. The purpose was to sample word pairs across a wide range of LSA scores, so as to examine whether their predictive validity varies across the similarity range.

Method

Participants Forty-six psychology undergraduates from the University of Georgia were awarded partial course credit for their participation.

Materials We constructed 90 words pairs that spanned a range of LSA scores. Word pairs were generated by the authors and placed into one of three categories based on LSA rating until there were 30 word pairs in each of the following three intervals of cosine values: 0-.33, .34-.66, and .67-1.0. The final list of words included nouns (*penny*), adjectives (*impolite*), adverbs (*properly*), and verbs (smash). We were intentionally lenient in creating the word pairs. If LSA ratings are accurate only for certain items, then they cannot be taken *prima facie* to reflect human similarity.

Procedure Similarity ratings were collected using Direct RT software. Each participant rated all 90word pairs. On each trial a word pair was displayed in the center of the screen. After 1800 ms the sentence "On a scale from 1 (not at all similar) to 7 (very similar) how similar are X and Y?" appeared under the word pair and remained until the participant input his or her response.

Results and Discussion

The data were analyzed using linear regression with LSA score as the predictor and similarity ratings as the dependent variable. Separate analyses were conducted for the data set as a whole, and for the three LSA intervals. For the overall analysis, the regression was non-significant (F (1, 88) = .47, p = .5). Across all items, LSA did not predict human similarity ratings. When analyzed separately, the fit was significant for the lower and upper intervals (F (1, 28) = 8.14, p < .05; F (1, 28) = 10.87, p < .05), but not for the middle interval (F (1, 28) = .3, p = .59). Although LSA cosines and human similarity ratings are moderately correlated at the lower and upper intervals, the direction of this relationship changed from positive at the lower end, to negative at the upper end of the scale (Table 1; Figure 1).

Table 1: Correlation between LSA cosines and Similarity Ratings across cosine Intervals

	Total	Lower	Middle	Upper
r =	.07	.47**	.10	-.38*

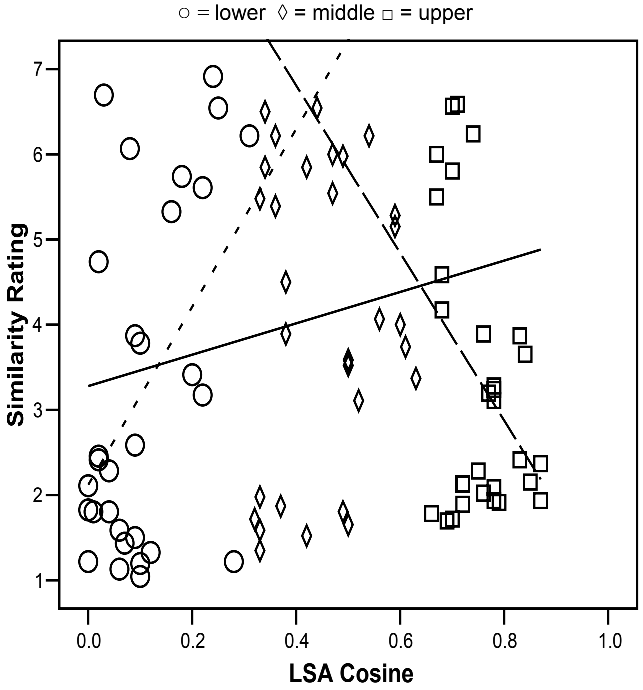
*p < .05, ** p < .001

This reversal of direction in the relationship likely reflects the type of stimuli that have a cosine greater than .60. Several antonymic pairs, such as *north / south* and *black / white* fell into this range. When all antonyms were excluded (19 items), the overall model fit was significant (F (1, 69) = 10.02, p < .05) as was the fit at the upper interval of LSA values (F (1, 14) = 6.33, p < .05). The correlation between LSA and human ratings at the upper interval increased and was positive (r= .56, p < .05). However, the correlation between cosine values and human ratings in the overall analysis remained low (r = .36, p < .05).

What can explain the failure of LSA cosine values to account for a substantial amount of variance in human ratings? One possibility is that LSA ratings do not distinguish between types of similarity in the same way that people do. People tend to use features when judging similarity. LSA has no direct access to features, but infers similarity on the basis of co-occurrence and contextual similarity (Landauer & Dumais, 1997). Assessing contextual similarity allows the model to capture sophisticated semantic relationships that might correspond to comparisons based on features. However, people do not base similarity judgments on co-occurrence, and LSA's use of this information could influence its similarity estimates of words that occur in the same scenarios, but are themselves dissimilar. A straightforward test of this is to compare LSA cosines with human ratings of taxonomically and thematically related word pairs.

The distinction between taxonomic and thematic relations is psychologically significant. Taxonomically related items tend to share attributes and have similar representational structures (Gentner & Gunn, 2001). Consider the taxonomically related concept pair, *cardinal* and *crow.* These concepts share a number of attributes, for instance, both are birds, have feathers, fly, etc. Furthermore, when these concepts differ, they do so along common dimensions, such as color, size, and beak shape. Thematically related items generally have dissimilar representational structures. Instead, they play complimentary roles in a given scenario (Lin & Murphy, 2001). For example, the thematically related concepts *hammer* and *nail* do not have any obvious common attributes. When these concepts differ, it is a difference of kind rather than value, i.e. one is a tool and one is not. Generally, thematically related items are judged to be less similar than taxonomic items (Wisniewski & Bassok, 1999). It is possible that LSA cosines are insensitive to this distinction, since they are products of both co-occurrence and contextual similarity.

Figure 1: Regression lines across the 3 cosine intervals



Study 3

The purpose of study 3 was to examine the possibility that the correlation between LSA scores and human similarity ratings is influenced by the type of relationship between the words (i.e. taxonomic or thematic). Given the psychological importance of the distinction between taxonomic and thematic relatedness, a failure of LSA to capture this difference would decrease the strength of the relationship between LSA scores and human ratings. The stimuli sets used in the previous studies were not intended to capture a specific sense of similarity. Although the majority of similar items from Study 1 appear to be taxonomically related, some items could also be interpreted thematically. For example, of the similar items in Gentner & Markman (1994) the pair *bowl / mug* can be interpreted taxonomically as *types of dishes*, or thematically as *items that one uses during breakfast*. On the same list, the item pair, *police car / ambulance* can be construed as *two types of emergency vehicles* (taxonomic) or as *two things likely to appear at the scene of an accident* (thematic). Likewise, the stimuli used in Study 2 comprised both taxonomically related (*cod / tuna*) and thematically related (*camel / desert*) word pairs.

Method

Participants Twenty-nine undergraduates from the University of Georgia were awarded partial course credit for their participation.

Materials Stimuli were 66 base words, each of which combined with two other words to create 66 taxonomically related and 66 thematically related pairs. For example, *dust* was matched with *soot* (taxonomic pair) and *sweep* (thematic pair). Twenty-eight of these triads were taken from Lin and Murphy (2001), 9 from Smiley and Brown (1979), and 31 were created by the authors. Taxonomic items belonged to the same category or shared salient features (*cane / pole*) and thematic items were "meaningfully and coherently related to the target" (*cane / limp*) (Lin & Murphy, 2001 p. 7). Four stimuli lists were constructed so that no given target word appeared twice in the same half of the list, and to control for the type of match (taxonomic or thematic) that appeared first.

Procedure Similarity ratings were collected using Direct RT software. Each participant rated all 132 word pairs from one of the four lists. On each trial a word pair was displayed in the center of the screen. After 1800 ms the sentence "On a scale from 1 (not at all similar) to 7 (very similar) how similar are X and Y?" appeared under the word pair and remained until the participant input his or her response.

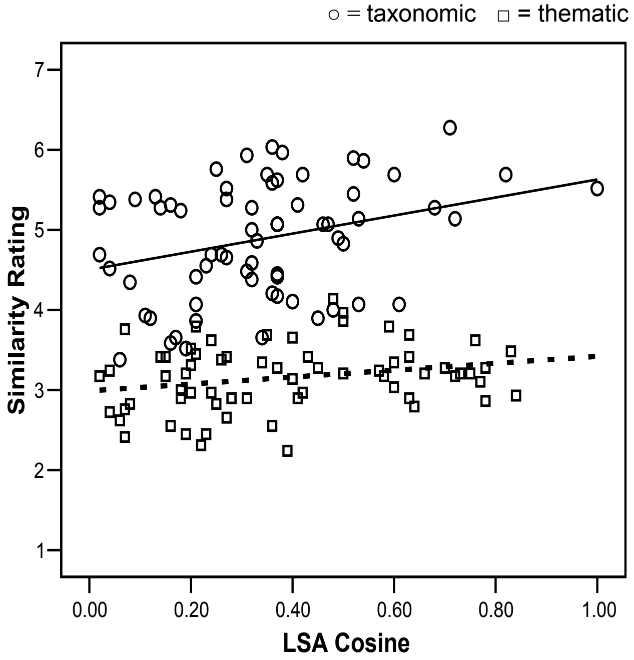
Results and Discussion

Responses were averaged across participants and the data were analyzed using linear regression with LSA score as the predictor and similarity ratings as the dependent variable. Separate analyses were conducted for taxonomic pairs and thematic pairs. For both taxonomic and thematic word pairs, LSA scores were significant predictors of human similarity ratings ($F(1, 65)_{\text{taxonomic}} = 6.98$, $p < .05$); $F(1, 65)_{\text{thematic}} = 4.20$, $p < .001$; Figure 2). The overall correlation between LSA scores and human similarity ratings was non-significant ($r = .06$, $p = .50$), but LSA cosine values did correlate with human ratings for taxonomic ($r = .31$, $p < .001$) and thematic ($r = .25$, $p < .05$) items separately. While the correlations are significant, very little of the variance in similarity ratings is explained by LSA cosine values.

This study was motivated by a hypothesis that LSA cosines do not distinguish between taxonomic and thematic similarity in the same way that people do. In fact, the cosine values did not differ between pair types ($M_{\text{taxonomic}} = .34$, SE = .02, $M_{\text{thematic}} = .38$, SE = .03; t (67) = -1.30, p = .2). Human similarity ratings, on the other hand, were reliably higher for taxonomic word pairs (M = 4.88, SE = .09) compared to thematic word pairs (M = 3.15, SE = .05; t (67) = 17.87, p<.001).

Although human ratings show a differential sensitivity to taxonomic versus thematic similarity, LSA does not. Given the large difference between human ratings across these stimuli types the failure to find a difference in LSA scores is surprising. Equally surprising is the failure of LSA cosines to correlate with human similarity ratings overall.

Figure 2: Regression lines for the two word pair types



General Discussion

In three studies we investigated the relationship between LSA and human similarity judgments. We found this relationship to be surprisingly inconsistent given LSA's ability to model other semantic behaviors. Our results suggest two explanations. Firstly, study 2 revealed that LSA cosines can be quite high for antonymous word pairs. The finding is consistent with the idea that "when applied in detail to individual cases of word pair relations or sentential meaning construal [LSA] often goes awry when compared to our intuitions" (Landauer, Foltz, & Laham, 1998 p. 25). For example, the model's tendency to overestimate the similarity of antonyms presents a serious problem in that, for any given word pair, a high LSA cosine value cannot be taken to indicate semantic similarity. It is only averaged across a large number of items that we can expect LSA scores to yield sensible results (Landauer, Foltz, & Laham, 1998). Although removing the antonymous items in Study 2 resolved the inconsistency in the direction of the relationship between LSA and human ratings, it did not appreciably increase the strength of this relationship. Secondly, the results of Study 3 suggest that an additional attenuating factor is LSA's insensitivity to different types of similarity. While the distinction between taxonomic and thematic similarity is important to humans, this difference is not reflected in LSA cosines.

We wish to emphasize that this is not a criticism of LSA. Proponents of the model have been quite up front about its unreliability with small samples of items (see Landauer et. al, 1998). LSA remains a good model of a number of linguistic phenomena. Our only claim is that, at this time, it is not well suited to this particular application.

We must acknowledge two limitations to the present studies. The potential range of LSA scores is -1 to +1, but the considerable difficulty in generating item pairs with negative cosine values led us to focus on the positive half of the scale. It is possible that a failure to sample from the negative range of LSA values could have influenced our results. More importantly, although we used the same corpus and number of dimensions across studies, the fit between LSA scores and human similarity judgments potentially could be improved by manipulating these factors. However, this type of intervention seems to preclude LSA's merits as an objective substitute for human similarity ratings. Future research could improve the fit between LSA cosines and human similarity ratings by determining the ideal number of dimensions to retain for this sort of assessment, and perhaps by developing a specialized corpus for collecting similarity ratings. Until that time, we must continue to rely on humans to rate the semantic similarity of word pairs.

References

Campbell, R. S. & Pennebaker, J. W. (2003). The secret life of pronouns. *Psychological Science, 14*, 60-65.

Estes, Z. & Hasson, U. (2004). The importance of being nonalignable: A critical test of the structural alignment theory of similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1082-1092.

Gagne, C. L., Spalding, T. L., Ji, H. (2005). Re-examining evidence for the use of independent relational representations during conceptual combination. *Journal of Memory and Language, 53*, 445-455.

Gentner, D, & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science, 5,* 152-158.

Gentner, D. & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory and Cognition, 9*, 565-577.

Hardiman, P.T., Dufresne, R., & Mestre, J.P. (1989). The relation between problem categorization and problem solving among experts and novices**.** *Memory & Cognition, 17*, 627-638.

Howard, M., W. & Kahana, M., J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*, 85-98.

Kwantes, P. J. (2005). Using context to build semantics. *Psychological Bulletin and Review, 12*, 703-710.

Landauer, T. K., & Dumais, S. T. (1994). Latent semantic analysis and the measurement of knowledge. In R. M. Kaplan & J. C. Burstein (Eds.), *Educational testing service conference on natural language processing techniques and technology in assessment and education*.
Princeton, NJ: Educational Testing Service.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the aquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.

Lin, E. L. & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General, 130*, 3-28.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory ,and Cognition*, *24*, 558-572.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review***,** *100***,** 254-278.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes,* 25, 337-354.

Smiley, S. S. & Brown, A. L (1979). Conceptual preference for thematic or taxonomic relations.: A nonmonotonic age trend from preschool to old age. *Journal of Experimental Child Psychology, 28*, 249-257.

Styvers, M, Shiffrin, R. M., & Nelson, D.L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. Washington, DC: American Psychological Association.

Wisniewski, E. J. & Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration*. Cognitive Psychology, 39*, 208-238.

Wolfe, M. B. W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W. & Landauer, T. K (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes*, *25*, 309-336.