Behavior Research Methods
2009, 41 (3), 647-656
doi:10.3758/BRM.41.3.647

More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis

Gabriel Recchia and Michael N. Jones
Indiana University, Bloomington, Indiana

Computational models of lexical semantics, such as latent semantic analysis, can automatically generate semantic similarity measures between words from statistical redundancies in text. These measures are useful for experimental stimulus selection and for evaluating a model's cognitive plausibility as a mechanism that people might use to organize meaning in memory. Although humans are exposed to enormous quantities of speech, practical constraints limit the amount of data that many current computational models can learn from. We follow up on previous work evaluating a simple metric of pointwise mutual information. Controlling for confounds in previous work, we demonstrate that this metric benefits from training on extremely large amounts of data and correlates more closely with human semantic similarity ratings than do publicly available implementations of several more complex models. We also present a simple tool for building simple and scalable models from large corpora quickly and efficiently.

Explaining how semantic representations of words are derived from experience is a central task for high-dimensional semantic space models such as the hyperspace analogue to language (HAL) framework of Burgess and Lund (2000), latent semantic analysis (LSA; Landauer & Dumais, 1997), and other lexical co-occurrence models of semantic memory. Although the models differ considerably in the algorithms used, they are all fundamentally based on the principle that a word's meaning can be induced by observing its statistical usage across a large sample of language. Evaluation typically proceeds by first training a model on a corpus of text, after which the model can generate similarity ratings between word pairs for comparison with human judgments. For example, if humans rate the words "cat" and "feline" as closer in meaning than the words "cat" and "cupboard," it would be desirable for a semantic space model to do so as well. Because the similarity ratings of semantic space models are used for experimental stimulus selection and for model evaluation, close correspondence with human data is critically important.

Starting from the premise that "simple associations in context are aggregated into conceptual representations" (Burgess & Lund, 2000), semantic space models typically apply techniques that go beyond direct lexical co-occurrence to induce more abstract semantic representations. Perhaps the best-known example is LSA's use of singular value decomposition (SVD), a technique from linear algebra. As described by Landauer, Foltz, and Laham (1998), LSA operates by first constructing a term–document matrix in which the value of each cell (*i, j*) represents the number of occurrences of word *i* in document *j*. Each term is then weighted with a log–entropy transform applied to the value of each cell in order to reduce the influence of very frequent words. Next, the matrix is factored, using SVD, allowing the construction of a new matrix of lower rank which can be thought of as a low-dimensional approximation of the original matrix. Finally, two rows of the final matrix can be correlated to obtain the semantic similarity between the rows' corresponding terms. Because the optimal choice of dimensionality for an LSA space varies depending on the choice of task and training corpus, it is generally chosen by recomputing the SVD over a wide range of possible choices and selecting the one for which the final matrix performs the best on the task at hand (Quesada, 2006). Similar probabilistic methods for inferring latent semantic components are arguably even more sophisticated and computationally intensive (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Hoffman, 1999).

The computational expense of the dimension reduction step raises several challenging issues for LSA and related semantic inference models. The first is lack of scalability. Because standard algorithms for computing SVDs require the entire term–document matrix to be held in memory, training LSA on corpora of many tens of millions of tokens is infeasible, even with high-end supercomputing resources. The problem is exacerbated by the fact that as the size of the corpus increases, the number of rows and columns in the matrix both increase significantly, the number of columns growing linearly with the number of docu-

G. Recchia, grecchia@indiana.edu

ments and the number of rows growing in approximate proportion to the square root of the number of tokens. A promising class of algorithms known as "out-of-core" methods for SVD calculation (Lin, 2000) have recently been shown to calculate large SVDs without having to store the entire matrix in memory (Martin, Martin, Berry, & Browne, 2007). However, no publicly available implementations of LSA currently make use of these methods. Even if out-of-core SVD algorithms can be shown to solve LSA's memory bottleneck, the high computational demands of computing extremely large SVDs may continue to be prohibitive for quite some time.

The standard LSA space used by most researchers (and available at lsa.colorado.edu) is trained on the Touchstone Applied Science Associates (TASA) corpus of textbook readings from a variety of grade levels (Zeno, Ivens, Millard, & Duvvuri, 1995). This corpus comprises some 11 million tokens,[1] a reasonable (perhaps even generous) approximation of the number of tokens most adults schooled in the United States have been exposed to through reading, given estimates that most children have read around only 3.8 million words by late grade school (Landauer & Dumais, 1997).

However, these estimates do not take into account the amount of language input that humans are exposed to through speech; indeed, 11 million tokens is approximately one third the number that children are estimated to have heard by the time they are 3 years old (Risley & Hart, 2006). Risley and Hart estimated that most American children hear between 10 and 33 million words in their first 3 years of life, not counting the 4 to 12 million that they produce during this time. By the age of 18—a lower bound for the age of study participants from whom semantic similarity judgments are collected—it is safe to assume that most would have experienced many times this amount. Although Landauer and Dumais (1997) argued convincingly that older children acquire most of their new word meanings from reading, this does not mean that these meanings cannot change over time as a result of additional exposure. If co-occurrence information plays a significant role in shaping humans' lexical semantic representations over time, one would expect our representations of word meaning to be shaped by co-occurrences in speech as well as in print. Given that semantic similarity judgments are collected from adult study participants, being able to scale semantic space models to large data sets is extremely important. We found training LSA or the topics model (Griffiths et al., 2007) on corpora significantly larger than TASA to be infeasible, given our resources and the massive computational demands of the learning algorithms.

Another challenge for LSA is a lack of incrementality: an inability to update semantic representations incrementally in response to a continual accumulation of language input. Humans update their semantic representations in response to new data continuously over time. In contrast, after an LSA space has been built, additional documents cannot be incorporated into the space without recomputing the low-rank approximation of the original matrix. This is problematic, because the original matrix is no longer presumed to be stored in memory after the SVD step has been computed (Landauer & Dumais, 1997). This lack of incrementality decreases the cognitive plausibility of LSA as a model of semantic organization (Lemaire & Denhière, 2004; Perfetti, 1998), and the appeal of static systems when learning from dynamically changing textbases.

Finally, there is the issue of complexity. Computational models of word segmentation, semantics, and syntax "often assume a computational complexity and linguistic knowledge likely to be beyond the abilities of developing young children" (Onnis & Christiansen, 2008). Given two models that correlate with human data equally well, the principle of scientific parsimony would tend to favor the one that makes fewer assumptions about the capabilities of the developing brain. SVD is arguably a complex process without an obvious explanation of how it might be implemented in humans. From the beginning, the SVD realization of LSA has been regarded more as a convenient expedient for dimensionality reduction than a claim about the specific analysis that humans apply to the data. Landauer and Dumais (1997, p. 218) stated that SVD should be considered as only one of a class of mathematical techniques worth exploration, and that additional features could be added to the dimensionality reduction step to "make it more closely resemble what we know or think we know about the basic processes of perception, learning, and memory."

These considerations motivated us to take a different approach. Rather than adding additional features to dimensional reduction algorithms to make them more scalable and psychologically plausible, our approach was to throw much more data at the problem, even if doing so required a less sophisticated learning algorithm. Recent discoveries in the field of computational linguistics have revealed that much more progress has been made toward systems for automatic parsing and sense disambiguation by training existing simple systems on more language data as it became available, rather than by investing in the development of superior algorithms. We take this lesson from computational linguistics as a useful indicator that computational models of semantics may reap greater benefits from more data rather than from developing more clever learning algorithms, and our goal in this article is to explore that hypothesis.

Pointwise mutual information (PMI) is one such measure that meets the previously discussed desiderata of scalability, incrementality, and simplicity, and which has been shown to yield high performance on forced-choice tests of semantic similarity (Bullinaria & Levy, 2006; Terra & Clarke, 2003; Turney, 2001). However, PMI has not yet been shown to outperform LSA on a wide variety of evaluation tasks. Budiu, Royer, and Pirolli (2007) found that PMI outperformed LSA on a variety of semantic tasks when trained on a larger corpus, but their work confounded corpus size and corpus quality. Bullinaria and Levy, using PMI to compare vectors of co-occurrence counts (i.e., the rows of a term $\times$ term matrix), found that PMI outper-

formed LSA on the test of English as a foreign language (TOEFL) forced-choice synonymy test when both methods were trained on a small corpus derived from Grolier's *Academic American Encyclopedia*. However, they did not correlate the performance of PMI and LSA with human judgments of semantic similarity.

First, we set out to determine whether PMI could provide a better match than LSA to human judgments of similarity when both measures were trained on the same type of text and only corpus size was varied. Finding this to be the case, we compared the performance of PMI trained on a large text corpus with several other measures of semantic relatedness, its high performance motivating us to release an easy-to-use resource for obtaining similarity judgments from arbitrary corpora. We end with a brief discussion of the theoretical and practical ramifications of these results.

Previous Research

PMI was first introduced in the context of word associations by Church and Hanks (1990). PMI is a very simple information-theoretic measure that, when computed between two words x and y, "compares the probability of observing x and y *together* (the joint probability) with the probabilities of observing x and y *independently* (chance)" (Church & Hanks, 1990, p. 23). It is defined as

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}. \quad (1)$$

In practice, $P(x)$ can be approximated as the number of times that x appears in the corpus, $P(y)$ as the number of times y appears in the corpus, and $P(x, y)$ as the number of times the two words co-occur in a context. Because a logarithm is a monotonically increasing function, relative ordinal rankings between PMI estimates are maintained if the log is dropped. Because we are interested in rank correlations to human similarity judgments, we use the term PMI in the remainder of this article to refer simply to the number of co-occurrences of x and y divided by the product of their individual frequencies.

Turney (2001) found that a version of PMI that used search engine hit counts to estimate lexical co-occurrences outperformed LSA on two forced choice vocabulary tests (TOEFL and ESL, described later), suggesting that PMI might be capable of approximating human similarity judgments more closely than LSA. However, LSA and PMI were trained on vastly different genres and sizes of text, leaving open the question of whether LSA would have performed better if trained on the same kind of text as PMI. Terra and Clarke (2003) also found high performance of PMI on TOEFL. They found that in general, performance seemed to increase with corpus size up to a point of around 850 GB, although the picture was muddied by the discrete nature of the evaluation metric. Furthermore, ESL and TOEFL constitute forced choice synonymy tests; correlations with human judgments of word similarity were not attempted in either study.

In an excellent comparison of algorithms and metrics, Bullinaria and Levy (2006) systematically investigated the effects of corpus size and quality on PMI. Their evaluation tasks consisted of TOEFL, a distance comparison (another forced-choice task in which the model must select the most semantically similar word out of a collection of alternatives), and two categorization tasks, one semantic and one syntactic. Rather than using the PMI score itself as the similarity value, Bullinaria and Levy constructed vectors for each word w in which each element corresponded to the PMI between w and another term from the training data. One can think of this as building up a term $\times$ term matrix; similarity scores between two words are estimated by calculating the distance between the corresponding two rows of the matrix. Using this method, the distance metric used to assess the similarity between vectors significantly affects the results. Bullinaria and Levy exhaustively tested a wide variety of distance metrics, along with other parameters, such as the number of vector components and the size of the context window. Although TOEFL was the only task on which Bullinaria and Levy directly compared LSA with PMI, they found that when these two methods were equated for corpus size and quality, comparing vectors of PMI scores using the cosine distance metric outperformed LSA even on a small corpus. They also found that additional training data greatly enhanced the performance of PMI on each of their evaluation tasks.

Finally, Budi, Royer, and Pirolli (2007) compared the performance of PMI with that of LSA and GLSA (Matveeva, Levow, Farahat, & Royer, 2005) on a variety of tasks across two corpora: the aforementioned TASA corpus of general textbook readings ranging from first grade through college, and the much larger Stanford corpus, consisting of the first 6.7 million pages of the WebBase project (Cho et al., 2006), a public-domain collection of Web page snapshots archived from 2004 to 2008. They found that a version of PMI trained on Stanford (PMI-Stanford) performed better than a version of LSA trained on TASA on several semantic similarity tasks. They also found that PMI-Stanford performed better on these tasks than a version of PMI trained on TASA did. They concluded that training on more data was what caused PMI's performance to improve.

However, the study confounded corpus size and corpus type, since TASA and the Stanford corpus represent two very different genres of text. TASA consists of a collection of readings "carefully put together to reflect college students' vocabulary, whereas Stanford is generated by a web crawl" (Budiu et al., 2007). It may be that PMI-Stanford produced a closer correspondence with human judgments, not because the Stanford corpus is larger, but rather because it is a better representation of the sort of language to which humans are exposed in everyday life. If LSA had been trained not on TASA but rather on a subset of the Stanford corpus, it might have performed better. Therefore, the experiment was inconclusive on the question of whether PMI would outperform LSA if both measures were trained on the same type of text. Likewise, if a version of PMI were trained on a representative subset of the Stanford corpus, that version might have done no better than PMI-Stanford did. In Experiment 1, we com-

pared the performance of LSA and PMI trained on a representative TASA-sized subset of the Wikipedia corpus, as well as with the performance of PMI trained on the full Wikipedia corpus (an infeasible task for LSA), in order to clarify whether corpus size or type of text was the causal factor behind PMI's success.

EXPERIMENT 1

Method

Evaluation Materials. We evaluated PMI and LSA on two forced choice synonymy tests and four lists of word pairs. The forced choice tests were the English as a second language (ESL) data set (Turney, 2001) and the TOEFL synonymy assessment (Landauer & Dumais, 1997). We used the lists of word pairs and accompanying human judgments of similarity compiled by Finkelstein et al. (2002), Miller and Charles (1991), Resnik (1995), and Rubenstein and Goodenough (1965). Both synonymy tests and all four word lists were also used as evaluation metrics by Budiu et al. (2007), and our abbreviations (ESL, TOEFL, MC, R, RG, and WS353, respectively) mimic theirs for consistency.

The ESL data set consists of a collection of synonymy questions for nonnative speakers of English. It was first used as a performance benchmark by Turney (2001) and later by Rohde, Gonnerman, and Plaut (2006, cited by Budiu et al., 2007). The data set consists of 50 questions, each consisting of a cue word (e.g., "envious") and a list of four possible choices (e.g., "jealous," "enthusiastic," "hurt," "relieved"). The object is to select from the list the word closest in meaning to the cue.

Consisting of 80 retired items from the Educational Testing Service's test of English as a foreign language, the TOEFL data set was used as a test bed for LSA by Landauer and Dumais (1997) and has later been employed extensively as a performance benchmark for many corpus-based approaches to capturing semantic similarity (Bullinaria & Levy, 2006; Matveeva et al., 2005; Pado & Lapata, 2007; Rapp, 2003). As with ESL, each item consists of a cue word paired with four possible synonyms.

Human judgments of word similarity were obtained from four collections of human similarity ratings. Rubenstein and Goodenough (1965) constructed a set of 65 noun pairs that varied widely in semantic similarity, ranging from near-synonyms (e.g., "car; automobile") to highly unrelated (e.g., "noon; string"), collecting judgments of semantic similarity from 51 participants. Miller and Charles (1991) selected a 30-pair subset of this data set and collected similarity ratings from 38 participants. In a replication study, Resnik (1995) published similarity judgments for 28 of the 30 Miller and Charles pairs, finding a correlation of .96 between his 10 participants' mean ratings and the mean ratings published by Miller and Charles. The Miller and Charles word pairs are also included in the WordSimilarity-353 test collection of Finkelstein et al. (2002), but the entire data set is over 10 times larger, consisting of a total of 353 word pairs. For each word pair, Finkelstein et al. instructed at least 13 participants to assess the relatedness on a scale ranging from 0 to 10. It is the only one of the four collections of similarity ratings to include proper nouns, adjectives, verbs, and gerunds, in addition to common nouns.

Training Materials. We trained PMI and LSA on a corpus derived from Wikipedia. The version of Wikipedia we refer to as the "Wikipedia corpus" consists of 2.5 GB of text from English Wikipedia articles downloaded in 2006, segmented into sentences and articles but otherwise stripped of formatting, markup, and nonalphabetic characters by Willits, D'Mello, Duran, and Olney (2007). We also trained each model on a small representative subset of the full Wikipedia corpus. For consistency, we decided to include 37,600 documents, the same number that TASA contains (Kanerva, Kristoferson, & Holst, 2000). Because Wikipedia's articles are on average larger than TASA documents, we originally planned to define a Wikipedia "document" as a single sentence. However, Wikipedia's average sentence length was only 17.7 words, which meant that the resulting subset of 37,600 Wikipedia sentences contained far fewer tokens than TASA did and, because of inadequate vocabulary coverage, yielded extremely poor performance for both models. Therefore, we redefined a document as any contiguous sequence of 10 sentences that were all part of the same article. Representing a wide variety of topics and articles, 37,600 documents were sampled randomly without replacement from the full corpus. This yielded a subset of Wikipedia with roughly the same document count and mean document length as TASA (document counts, 37,600 for each; mean document lengths, 166 and 162, respectively). The final subset contained 6,102,845 tokens and 251,703 types, in contrast to 417,775,181 tokens and 3,404,652 types in the full corpus.

Design and Procedure. We trained PMI on the Wikipedia subset and the full Wikipedia corpus. As previously discussed, LSA could be trained on the subset only. So that our work could be easily replicated, we implemented LSA according to Quesada's (2006) guidelines for constructing LSA spaces. The singular value decomposition was computed using Rohde's (2005) SVDLIBC implementation. To reduce memory demands, we restricted the terms in our term–document matrix to those terms which occurred in at least two documents in the Wikipedia subset. We judged that this should not impair LSA's performance, since terms occurring in only one document do not provide latent co-occurrence information useful for judging the similarity between any two other terms. In addition, terms occurring in any of our evaluation tasks were not excluded, even if they occurred in only one document. Like Budiu et al. (2007), we computed SVDs ranging from 100 to 500 dimensions in increments of 10 and selected the dimensionality for which LSA performed optimally for each evaluation task. Plotting LSA's performance against dimensionality on each task revealed U-shaped curves similar to those described in Landauer and Dumais (1997); it did not seem that smaller increments would have significantly increased performance.

PMI estimates for each pair of words x and y were calculated by dividing the number of times that x and y co-occurred within a single document with the product of the respective frequencies of x and y in the entire corpus. To assist with this task, we developed a software tool for efficiently calculating PMI scores from large corpora, described in more detail later in the article. As with the Wikipedia subset, documents were defined as contiguous sequences of 10 sentences that were part of the same article.

Results

For each model version and each list of word pairs (MC, R, RG, WS353), a Spearman rank correlation was calculated between the similarity judgments of the model and the normative human similarity judgments for that list. For

Table 1
Comparisons Between Human Judgments of Semantic Similarity and Model Estimates

	Trained on Wikipedia Subset		Trained on Full Wikipedia
Task	PMI	LSA	PMI
ESL	.35	.36	.62
TOEFL	.41	.44	.64
MC	.47	.62	.78
R	.46	.60	.86
RG	.46	.46	.76
WS353	.54	.57	.73

Note—For synonymy tests (ESL, TOEFL), values represent the percentage of correct responses. For all other tasks, values represent Spearman rank correlations between human judgments of semantic similarity and those of the model. Abbreviations for tasks are defined in the main text of the article.

the two synonymy tests (ESL, TOEFL), performance was assessed by the percentage of correct responses.[2] The results are displayed in Table 1. For all four sets of similarity judgments, LSA trained on the Wikipedia subset (LSA-Subset) produced judgments equally or more highly correlated with human data than did the version of PMI trained on the Wikipedia subset (PMI-Subset). However, for all sets of similarity judgments, the version of PMI trained on the full version of Wikipedia (PMI-Wiki) produced higher correlations with human data than did LSA-Subset. Similarly, LSA-Subset scored higher on both synonymy tests than PMI-Subset did, but scored more poorly than PMI-Subset did.

Discussion

We began this experiment with the question: Does PMI produce semantic ratings that more closely resemble judgments from humans, when provided with a quantity of data more representative of the amount that humans experience? Although previous work had suggested that the answer might be yes, no study had yet systematically investigated PMI's performance on a battery of semantic tasks while controlling for the type as well as the quantity of data. Are the additional data enough to approximate human semantic judgments more accurately than LSA, even though PMI does not take higher order co-occurrence information into account? Again, studies had shown that simple models exposed to huge amounts of data could outperform a TASA-trained LSA, but it remained unclear whether this was caused by differences in the amount of data, differences in the type of data, or to other uncontrolled factors such as document size. However, the present work suggests that PMI benefits from additional data, and benefits enough to outperform at least one more sophisticated but less scalable model. Given that PMI is an extremely simple, scalable, incremental model of semantic similarity, these results argue for the cognitive plausibility of semantic models that do nothing more sophisticated than increase the similarity of co-occurring terms, provided they have some mechanism to diminish the influence of highly frequent terms.

This work also has practical import, given PMI's ease of computation and low memory demands. Automatic approximations of semantic similarity are extremely useful for experimental stimulus development for any study in which semantic similarity might confound the factor of interest. For example, in investigations of orthographic or phonological priming, researchers must ensure that prime–target pairs in experimental and control groups do not systematically vary in semantic similarity. Additionally, the needs of researchers sometimes call for domain-specific similarity judgments. Examples include the use of an LSA training corpus, consisting of a small set of encyclopedia articles about the heart, to compute the coherence of texts on this topic (Foltz, Kintsch, & Landauer, 1998); and the work of Kaur and Hornof (2005), demonstrating that training semantic models on domain-specific corpora improved the models' correlations to human data in an information foraging task. Given the potential benefits to researchers in cognitive psychology, there is a need for an easy-to-use software tool that allows nonprogrammers to train reasonably well-performing models of semantic similarity on arbitrary corpora, using only a personal computer. To be useful in a practical sense, however, the tool would have to be shown to produce results competitive with other publicly available tools.

In Experiment 2, we compared PMI-Wiki with 19 other publicly available measures of semantic relatedness. Because these measures vary in the type and amount of data they are trained on, we make no claims about the superiority or inferiority of PMI to other metrics on the basis of the results of Experiment 2. However, if PMI correlates with human similarity judgments at a level comparable with other public tools, it would suggest that a tool allowing nonprogrammers to train PMI on any corpus could be a great benefit to researchers.

EXPERIMENT 2

Method

Materials. Experiment 2 employed the same six evaluation tasks and the same version of PMI-Wiki as described in Experiment 1. The metrics of semantic similarity against which we compared PMI-Wiki were those available at the Rensselaer measures of semantic relatedness (MSR) Web site (http://cwl-projects.cogsci.rpi.edu/msr/), a server which aggregates several publicly available measures of semantic relatedness into a single Web interface (Veksler, Grintsvayg, Lindsey, & Gray, 2007).

Design and Procedure. We used the MSR Web interface to gather similarity judgments on the word pairs in our evaluation tasks for all available models. The measures listed on the MSR Web site as NSS-GReuters and GLSA had to be omitted from the analysis because the Web interface reported errors in retrieving a large number of similarity judgments from these models on all tasks. Additionally, one question from TOEFL had to be omitted, since the Web interface reported an error in retrieving similarity judgments from almost every model, most likely because that question contained an extremely rare word likely to be missing from most models' lexicons.

Results

PMI-Wiki was compared against a total of 19 measures on the MSR Web site. The same criteria described in Experiment 1 were used for evaluation on the two synonymy tests and the four similarity judgment tasks. Table 2 displays the performance of PMI-Wiki with that of the eight measures that averaged the highest performance across all tasks. These were the WordNet::Similarity vector measure (Pedersen, Patwardhan, & Michelizzi, 2004), versions of PMI implemented by Veksler et al. (2007) that calculate co-occurrences via a Google search of Wikipedia as well as directly from the Factiva business news corpus (Dow Jones & Co., 2008), normalized search similarity (Cilibrasi & Vitányi, 2007; cited by Veksler, Gray, Gamard, Grintsvayg, & Lindsey, 2008) trained on Factiva and TASA, spreading activation (Anderson & Pirolli, 1984; cited by Farahat, Pirolli, & Markova, 2004)[3] trained on Factiva and TASA, and LSA trained on TASA (Landauer & Dumais, 1997). In Table 2, these measures are abbreviated WN, PMI.W, PMI.F, NSS.F, NSS.T, SA.N, SA.W, and LSA.T, respectively. Values in bold represent the highest and second highest values for each task. For both synonymy tests and three of the four sets of similarity judgments, PMI-Wiki was the second highest performing

Table 2
Comparisons Between Human Semantic Judgments and Measures of Semantic Relatedness

	Measure of Semantic Relatedness								
Task	PMI-Wiki	WN	PMI.W	PMI.F	NSS.F	NSS.T	SA.N	SA.W	LSA.T
ESL	**.62**	**.70**	.50	.42	.44	.56	.39	.51	.44
TOEFL	**.64**	**.87**	.42	.51	.59	.50	.61	.59	.55
MC	**.78**	**.88**	.50	.46	.62	.53	.49	.39	.69
R	**.86**	**.90**	.54	.41	.56	.54	.49	.52	.74
RG	**.76**	**.77**	.41	.51	.61	.56	.45	.45	.61
WS353	**.73**	.46	.29	.58	**.60**	.59	.40	.38	**.60**

Note—For synonymy tests (ESL, TOEFL), values represent the percentage of correct responses. For all other tasks, values represent Spearman rank correlations between human judgments of semantic similarity and those of the corresponding measure of semantic relatedness. Values in bold represent the highest and second-highest values in each row. Abbreviations for tasks and for measures of semantic relatedness are defined in the main text of the article.

measure, outperformed solely by the WordNet::Similarity vector measure (WN). On the remaining word similarity data set, PMI-Wiki was the highest performing measure.

Discussion

Given its simplicity, PMI-Wiki makes a strong showing when compared with numerous publicly available measures of semantic relatedness. WN also emerged as an extremely strong measure, outperforming all other measures on nearly all tasks. However, WN is also the only model based on hand-coded intelligence and is limited to the words in the WordNet lexicon, since it relies heavily on WordNet's dictionary glosses (Pedersen et al., 2004). This constitutes an important difference with scalable, incremental algorithms, which construct representations via unsupervised processing of large textbases. It is notable that WS353, the only task on which PMI-Wiki vastly outperformed WN, was also the only task to include adjectives and verbs, which may differ from nouns in terms of the characteristics of their WordNet dictionary glosses. Although this is not the only difference between WS353 and the other tasks, the inclusion of adjectives and verbs is a plausible explanation for WN's decreased performance on the WS353 data set. Thus, WN's impressive performance suggests that it is a viable choice for researchers looking for general approximations of semantic similarity between English nouns; for tasks that would benefit from training on domain-specific corpora, or that involve words not included in the WordNet hypernym hierarchy, PMI trained on a suitably large corpus may be a better choice.

Given the success of PMI-Wiki, it may seem puzzling that PMI.W (the version of PMI that estimated co-occurrences using a Google search of Wikipedia) performed relatively poorly. One potential reason is that Google reports approximate hit counts that vary with the order of the query terms and other factors, and thus do not reflect exact co-occurrences between search terms. For example, at the time of writing, the Google query *two AND three site:wikipedia.org* reports 368,000 hits, *three AND two site:wikipedia.org* reports 342,000, *two three site:wikipedia.org* reports 354,000, and *three two site:wikipedia.org* reports 346,000. The Web pages that Google indexes, which in this case correspond roughly to full Wikipedia articles, are also likely to be much larger than the 10-sentence documents of PMI-Wiki; this could also have hurt the performance of PMI.W, given that Turney (2001) and Bullinaria and Levy (2006) reported greater success with relatively small windows of text than with large documents (see also Hare, Jones, Thomson, Kelly, & McRae, in press). The version of PMI trained on Factiva, PMI.F, also performed competitively but not at the level of PMI-Wiki, probably because the restricted semantic domain of the Factiva corpus (business news) was not a good match for many of the evaluation tasks.

PMI-Wiki's impressive performance against most of the measures on the MSR Web site suggests that PMI is a viable choice for researchers looking to train a measure of semantic similarity on domain-specific corpora. However, we must emphasize that the goal of Experiment 2 was not to make a theoretical claim about PMI as opposed to other measures, since this study does not provide sufficient evidence to conclude that PMI is a superior measure overall. First, the measures are all trained on different corpora. In addition, such a comparison would be unfair to models such as LSA that are sensitive to parameterization; informal tests suggested that the MSR interface queries the LSA server with the default number of factors, rather than optimizing against some set of external human similarity judgments.

LMOSS: A Tool for Training Lightweight Metrics of Semantic Similarity on Large Corpora

Earlier, we pointed out the need for an easy-to-use resource for obtaining similarity judgments from arbitrary corpora. The promising performance of PMI-Wiki encouraged us to release a freely available version of Lightweight Metrics of Semantic Similarity (LMOSS, available at www.indiana.edu/~clcl/LMOSS/). LMOSS is a GUI interface to our implementation of PMI, allowing researchers to train PMI on their own corpora and retrieve similarity judgments quickly and efficiently. To our knowledge, this is the only free, publicly available tool that allows nonprogrammers to train a lexical co-occurrence metric on their own large corpora. The Web site offers a precompiled binary of LMOSS that has been tested on Windows XP and Vista systems. LMOSS was created in C# and makes use of Microsoft's .NET platform. Theoretically, it should be able to be compiled on other platforms via

the Mono 2.0 open source .NET development framework (Novell, 2008), but we have not yet attempted to do so.

LMOSS can be trained easily and efficiently on any text corpus that the user defines. This is particularly useful for calibrating stimuli when one is concerned about the decontextualized similarity metrics retrieved from systems trained on corpora such as TASA. The representation for a word learned by LSA from TASA is a melding of what would be multiple contextual senses in WordNet. For example, the vector for *heart* contains information about love and emotion, but also about arteries and pumping blood. Although techniques do exist for evaluating the contextual sense of a word in LSA from the composite representation (Kintsch, 2001), in many domains it is much better to train on a corpus related to the sense of the word relevant to an experiment. This was the original motivation for the "smallheart" version of LSA on the Colorado Web interface (it contains representations trained only on articles relating to the biological sense of the word *heart*), and more recent research has demonstrated that very large gains in performance can be seen in text classification systems trained on text corpora specific to the subject domain (Stone, Dennis, & Kwantes, 2008).

Speed and ease of use were primary considerations in the construction of LMOSS. On a 32-bit dual-core Dell Inspiron 1420 with 3 GB of memory and a processor speed of 1.83 GHz, we found that training PMI on the Wikipedia corpus (specifying 1,024 different words that the model would be tested on for purposes of evaluation) took less than 20 min. LMOSS also allows the resulting PMI model to be saved to a binary file, so that even models trained on extremely large corpora can be reloaded in seconds. Partly inspired by the LSA Web interface at lsa.colorado.edu, LMOSS offers a graphical user interface allowing similarity judgments to be evaluated in four different ways. *Matrix comparison* returns similarity judgments between every possible pairing of words on a list provided by the user; *one-to-many comparison* returns judgments between a single word and every word on a separate list; *pairwise comparison* returns judgments between arbitrary pairs of words; and *forced-choice comparison* evaluates PMI on a forced-choice task provided by the user (such as ESL or TOEFL). To assist users who have access to large corpora but who may be unfamiliar with basic scripting techniques, we provided an option to have LMOSS preprocess the training corpus text by stripping punctuation and nonalphanumeric characters and treating all words as lowercase.

Consonant with findings that counting co-occurrences within small windows of text produces better results than does counting co-occurrences within larger contexts (Bullinaria & Levy, 2006; Turney, 2001), LMOSS allows PMI to be trained using window sizes of a user-specified length. Counting co-occurrences within documents of arbitrary size, rather than windows of n words, can be accomplished by setting the window size to a number of words larger than any document in the corpus. LMOSS considers each line of the training input to constitute a separate document, and does not allow text windows to simultaneously include text from two adjacent documents.

Finally, the failure to incorporate word order information is a complaint commonly leveled at lexical co-occurrence models of semantic similarity. Information about word order can be combined with nonpositional co-occurrence information, via simple addition, to produce better correspondences to human semantic similarity judgments than would result from either information source alone (e.g., Jones, Kintsch, & Mewhort, 2006). Thus, LMOSS also includes an implementation of PMI/Order, an experimental metric we developed that takes positional information about surrounding words into account. To calculate the similarity of two terms x and y using PMI/Order, LMOSS first calculates PMI(x, y). Then it adds to this value [(# of times w appears m words after x)(# of times w appears m words after y)]/[(freq w $*$ freq x)(freq w $*$ freq y)($|m|$)] for each word w in the lexicon and for each value of m (excluding zero) from $-n$ to $+n$, where n is half the order window size set by the user. The $|m|$ in the denominator serves to weight words that appear closer to x and y as more important, making PMI/Order more sensitive to local context; for example, if the phrase "furry cat" were to appear much more often than would be expected if the distributions of "furry" and "cat" were independent, and if the same were true for the phrase "furry dog," *cat* and *dog* would get a large similarity boost. The system does this not only for the words appearing immediately before *cat* and *dog* ($m = -1$), but for each nonzero value of m from n to $-n$, with each position given a weight of $1/|m|$. We found slight improvements in correlations to human semantic similarity judgments when incorporating positional information about surrounding words in this way (an average improvement of .025 on correlations to human similarity judgments, and .024 on average forced-choice test accuracies).

Currently, PMI and PMI/Order are the only metrics available in LMOSS. In future versions, we plan to provide additional options allowing users to build up vectors of PMI scores in the fashion of Bullinaria and Levy (2006), allowing users to select the distance metric, vector length, and other parameters most appropriate for their particular task. In the meantime, however, our results in Experiment 2 show that even LMOSS's simplistic PMI estimate—provided that it is trained on a suitably large and high-quality corpus—rivals other publicly available measures of semantic similarity, serving as a valuable tool for researchers interested in obtaining corpus-specific approximations of human semantic similarity judgments. The system also makes it convenient for users to train their own corpora, selected specifically by genre or content domain, to create semantic metrics more sensitive to the contextual usage of words in their experiments.

GENERAL DISCUSSION

In Experiment 1, we found that PMI, a scalable, incremental, and simple measure of semantic similarity, greatly benefited from training on additional data, so much so that it outperformed a version of LSA trained on less of the same type of data over a variety of ex-

periments and tasks. This was found to be the case, even though LSA was trained on a quantity of data comparable to the TASA corpus (often used as a standard for LSA comparisons) and which was close to the limit that we could train LSA upon given our computational resources. Previous work had found that a version of PMI trained on a large amount of data produced higher correlations with human semantic similarity judgments than did a TASA-trained version of LSA, but this work had not controlled for factors such as type of data and document size.

Although ours is the first study to systematically investigate the effect of corpus size on PMI using a wide variety of semantic benchmarks, it is well known in computational linguistics that increasing training corpus size greatly improves performance in a variety of tasks that involve learning from unstructured natural language (Banko & Brill, 2001). The situation in the semantic modeling literature is similar to how Banko and Brill described the computational linguistics literature of 8 years ago:

> The empirical NLP community has put substantial effort into evaluating performance of a large number of machine learning methods over fixed, and relatively small, data sets. Yet since we now have access to significantly more data, one has to wonder what conclusions that have been drawn on small data sets may carry over when these learning methods are trained using much larger corpora. (p. 26)

These lessons from the field of computational linguistics, along with the present work, call into question whether present models of human semantic learning are overly complex as theories, and whether humans might apply a much simpler heuristic to create meaning from experience. Currently popular semantic inference models may tend to err on the side of complexity, because the field has previously been restricted to data not on a scale comparable to human experience; this forces these models to make do with far less data than is available to humans. For example, contemporary models of human semantic cognition (e.g., Rogers & McClelland, 2004) may inadvertently hardwire too much complexity into their processing architecture, because the models are trained on small data samples that are less complex than the large samples that humans experience. To make the model behave in a complex fashion—like a human but on less data—it becomes necessary to build complexity into the system. However, the truth may be that the requisite complexity for this behavior is already present in the structure of language if a realistic sample is taken, and humans may only require much simpler learning mechanisms than we needed to build into the model.

Notably, PMI achieves high performance despite its extreme simplicity—it is a straightforward tabulation of co-occurrence counts, with a normalization term to penalize words with high co-occurrence counts merely because they are highly frequent in the language as a whole. This suggests a possible connection between PMI and scalable vector accumulation techniques like BEAGLE (Jones & Mewhort, 2007) and random indexing (Kanerva et al., 2000; Sahlgren, 2006), which also increase the similarity of words that commonly co-occur while simultaneously correcting for global frequency; and which do so in a scalable, incremental, and simple manner. Because PMI is a general measure that can be implemented or approximated in many different ways, it is worth investigating models that share PMI's basic properties and that actually specify the details of how semantic representations are constructed.

In Experiment 2, we found that a version of PMI trained on Wikipedia outperformed several publicly available measures of semantic relatedness. This is interesting from a practical standpoint, since PMI similarity judgments are fast and easy to calculate, even on huge data sets. Because similarity measurements have practical applications for calibrating stimuli in a variety of experimental situations, we also released a free tool that lets researchers calculate PMI similarity estimates for themselves without high-performance cluster computing systems; one could even use it to create new large-scale data sets of PMI similarity estimates for use by others. Such resources should prove useful both for stimulus development and for improving our understanding of the mechanisms that humans use to organize meaning in memory.

One final note deserves comment. In making the criticism that an individual's ambient speech environment contains far more tokens than does the text corpora on which semantic models are commonly trained, we are aware that more tokens does not necessarily mean that humans experience more *information* than is present in current text corpora. For this claim to be true, it would require an implicit assumption that text and speech have the same statistical structure; and this is known to be false. For example, Hayes (1988) conducted a corpus comparison of the differences between written language and transcribed conversations. He found that the lexical diversity of conversation is much less rich than is the diversity in text. The greater volumes of data that we imply humans learn from may indeed be greater volumes of *redundant* data, and text corpora such as TASA may well contain the same or more bits of information, even though they contain far fewer tokens. Hence, it is possible that we provided the right number of tokens to PMI by using Wikipedia, but far more information than humans typically experience. Realistic scale corpora of ambient speech environments such as Infoture's (2008) Lena corpus will soon be available on which to train semantic models, and may help constrain plausible cognitive mechanisms for how humans learn semantic information from distributional experience.

AUTHOR NOTE

This work was presented at the 38th Meeting of the Society for Computers in Psychology, Chicago, IL. G.R.'s contribution received the John Castellan Award for best student paper. Correspondence concerning this article should be addressed to G. Recchia, Cognitive Science Program, Indiana University, 819 Eigenmann, 1910 E. 10th St., Bloomington, IN 47406-7512 (e-mail: grecchia@indiana.edu).

REFERENCES

ANDERSON, J. R., & PIROLLI, P. L. (1984). Spread of activation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 791-798.

BANKO, M., & BRILL, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Conference of the Association for Computational Linguistics* (pp. 26-33). Stroudsburg, PA: Association for Computational Linguistics.

BUDIU, R., ROYER, C., & PIROLLI, P. L. (2007). Modeling information scent: A comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In *Proceedings of the 8th Annual Conference of the Recherche d'Information Assistée par Ordinateur (RIAO)*. Pittsburgh, PA: Centre des Hautes Études Internationales d'Informatique Documentaire.

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

BURGESS, C., & LUND, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117-156). Mahwah, NJ: Erlbaum.

CHO, J., GARCIA-MOLINA, H., HAVELIWALA, T., LAM, W., PAEPCKE, A., RAGHAVAN, S., & WESLEY, G. (2006). Stanford WebBase components and applications. *ACM Transactions on Internet Technology*, **6**, 153-186.

CHURCH, K. W., & HANKS, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**, 22-29.

CILIBRASI, R., & VITÁNYI, P. M. B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge & Data Engineering*, **19**, 370-383.

DEANE, P., SHEEHAN, K. M., SABATINI, J., FUTAGI, Y., & KOSTIN, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading*, **10**, 257-275.

DOW JONES & CO. (2008). Available from Dow Jones Factiva Web site: http://factiva.com.

FARAHAT, A., PIROLLI, P., & MARKOVA, P. (2004). *Incremental methods for computing word pair similarity* (TR-04-6). Palo Alto, CA: Palo Alto Research Center, Inc.

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G., & RUPPIN, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, **20**, 116-131.

FOLTZ, P. W., KINTSCH, W., & LANDAUER, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, **25**, 285-307.

GRIFFITHS, T. L., STEYVERS, M., & TENENBAUM, J. B. (2007). Topics in semantic representation. *Psychological Review*, **114**, 211-244.

HARE, M., JONES, M. N., THOMSON, C., KELLY, S., & MCRAE, K. (in press). Activating event knowledge. *Cognition.*

HAYES, D. P. (1988). Speaking and writing: Distinct patterns of word choice. *Journal of Memory & Language*, **27**, 572-585.

HOFFMAN, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual Special Interest Group on Information Retrieval (SIGIR) Conference* (pp. 50-57). New York: ACM Press.

INFOTURE, INC. (2008, September). *Transcriptional analyses of the Infoture natural language corpus* (Report ITR-06-2). Retrieved December 15, 2008, from www.infoture.org/TechReport.aspx/Transcription/ITR-06-2/ITR-06-2_Transcription.pdf.

JONES, M. N., KINTSCH, W., & MEWHORT, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory & Language*, **55**, 534-552.

JONES, M. N., & MEWHORT, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, **114**, 1-37.

KANERVA, P., KRISTOFERSON, J., & HOLST, A. (2000). Random indexing of text samples for latent semantic analysis. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (pp. 103-106). Austin, TX: Cognitive Science Society. (Also available at www.rni.org/kanerva/cogsci2k-abstract.ps.)

KAUR, I., & HORNOF, A. J. (2005). A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. In G. C. van der Veer & C. Gale (Eds.), *Proceedings of the 2005 Conference on Human Factors in Computing Systems (CHI)* (pp. 51-60). New York: ACM Press.

KINTSCH, W. (2001). Predication. *Cognitive Science*, **25**, 173-202.

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

LANDAUER, T. K., FOLTZ, P., & LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259-284.

LEMAIRE, B., & DENHIÈRE, G. (2004). Incremental construction of an associative network from a corpus. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 825-830). Austin, TX: Cognitive Science Society.

LIN, M.-H. (2000). *Out-of-core singular value decomposition* (Report TR-83). New York: Stony Brook University, Experimental Computer Systems Laboratory.

MARTIN, D. I., MARTIN, J. C., BERRY, M. W., & BROWNE, M. (2007). Out-of-core SVD performance for document indexing. *Applied Numerical Mathematics*, **57**, 1230-1239.

MATVEEVA, I., LEVOW, G., FARAHAT, A., & ROYER, C. (2005, September). *Terms representation with generalized latent semantic analysis*. Presentation at the International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria.

MILLER, G. A., & CHARLES, W. G. (1991). Contextual correlates of semantic similarity. *Language & Cognitive Processes*, **6**, 1-28.

NOVELL, INC. (2008). Mono 2.0 [Computer software]. Retrieved August 1, 2008, from www.mono-project.com.

ONNIS, L., & CHRISTIANSEN, M. H. (2008). Lexical categories at the edge of the word. *Cognitive Science*, **32**, 184-221.

PADO, S., & LAPATA, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**, 161-199.

PEDERSEN, T., PATWARDHAN, S., & MICHELIZZI, J. (2004). WordNet::Similarity: Measuring the relatedness of concepts. In D. L. McGuinness & G. Ferguson (Eds.), *Proceedings of the 19th National Conference on Artificial Intelligence (Intelligent Systems Demonstrations)* (pp. 1024-1025). Cambridge, MA: MIT Press.

PERFETTI, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes*, **25**, 363-377.

QUESADA, J. (2006). Creating your own LSA space. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

RAPP, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th Machine Translation Summit* (pp. 315-322). New Orleans.

RESNIK, P. (1995). Using information content to evaluate semantic similarity. In C. S. Mellish (Ed.), *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 448-453). San Francisco: Morgan Kaufmann.

RISLEY, T. R., & HART, B. (2006). Promoting early language development. In N. F. Watt, C. Ayoub, R. H. Bradley, J. E. Puma, & W. A. LeBoeuf (Eds.), *The crisis in youth mental health: Critical issues and effective programs: Vol. 4. Early intervention programs and policies* (pp. 83-88). Westport, CT: Praeger.

ROGERS, T. T., & MCCLELLAND, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

ROHDE, D. (2005). SVDLIBC [Computer software]. Retrieved from http://tedlab.mit.edu/~dr/svdlibc/.

ROHDE, D., GONNERMAN, L., & PLAUT, D. (2006). *An improved model of semantic similarity based on lexical co-occurence*. Manuscript submitted for publication.

RUBENSTEIN, H., & GOODENOUGH, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, **8**, 627-633.

SAHLGREN, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Unpublished doctoral dissertation, Stockholm University.

STONE, B. P., DENNIS, S. J., & KWANTES, P. J. (2008). A systematic comparison of semantic models on human similarity rating data: The effectiveness of subspacing. In B. C. Love, K. McRae, & V. M. Slout-

sky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1813-1818). Austin, TX: Cognitive Science Society.

TERRA, E., & CLARKE, C. L. A. (2003). Frequency estimates for statistical word similarity measures. In *Proceedings of the 2003 Conference of the North American Chapter of HLT-NAACL* (pp. 165-172). Edmonton, AL, Canada: HLT-NAACL.

TURNEY, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In L. De Raedt & P. Flach (Eds.), *Proceedings of the 12th European Conference on Machine Learning* (pp. 491-502). Berlin: Springer.

TURNEY, P., & LITTMAN, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, **21**, 315-346.

VEKSLER, V. D., GRAY, W. D., GAMARD, S., GRINTSVAYG, A., & LINDSEY, R. (2008). *Measures of semantic relatedness*. Retrieved from Rensselaer MSR Web site: http://cwl-projects.cogsci.rpi.edu/msr/msr-about.html.

VEKSLER, V. D., GRINTSVAYG, A., LINDSEY, R., & GRAY, W. D. (2007). A proxy for all your semantic needs. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (p. 1878). Austin, TX: Cognitive Science Society.

WILLITS, J. A., D'MELLO, S. K., DURAN, N. D., & OLNEY, A. (2007). Distributional statistics and thematic role relationships. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 707-712). Austin, TX: Cognitive Science Society.

ZENO, S., IVENS, S., MILLARD, R., & DUVVURI, R. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone.

NOTES

1. The number of "tokens" refers to the number of words in a corpus, whereas the number of "types" refers to the number of *unique* words in a corpus. The total number of tokens in TASA is sometimes cited as 10 million (Kanerva, Kristoferson, & Holst, 2000; Turney & Littman, 2003), and sometimes as 17 million (Budiu, Royer, & Pirolli, 2007; Deane, Sheehan, Sabatini, Futagi, & Kostin, 2006). Although the version of the corpus that Zeno et al. (1995) used to compile their word frequency guide contained 17 million tokens, Landauer et al. (1998) clarified that the machine-readable version they used to construct their LSA space contained 11 million tokens.

2. A full point was awarded only if the model unambiguously picked out the correct answer as most similar to the cue. If the model judged n of the four answers as equally good (with the correct answer being among these n), $1/n$ points were awarded (the expected value if the model were to guess randomly among what it judged as the best answers).

3. The formula used is that of Anderson and Pirolli (1984), but Farahat et al. (2004) were the first to use it directly as a metric of word pair association.

(Manuscript received November 21, 2008;
accepted for publication December 11, 2008.)