Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words

Aida Nematzadeh, **Stephan C. Meylan**, and **Thomas L. Griffiths**
University of California, Berkeley
{nematzadeh, smeylan, tom_griffiths}@berkeley.edu

Abstract

Vector-space models of semantics represent words as continuously-valued vectors and measure similarity based on the distance or angle between those vectors. Such representations have become increasingly popular due to the recent development of methods that allow them to be efficiently estimated from very large amounts of data. However, the idea of relating similarity to distance in a spatial representation has been criticized by cognitive scientists, as human similarity judgments have many properties that are inconsistent with the geometric constraints that a distance metric must obey. We show that two popular vector-space models, Word2Vec and GloVe, are unable to capture certain critical aspects of human word association data as a consequence of these constraints. However, a probabilistic topic model estimated from a relatively small curated corpus qualitatively reproduces the asymmetric patterns seen in the human data. We also demonstrate that a simple co-occurrence frequency performs similarly to reduced-dimensionality vector-space models on medium-size corpora, at least for relatively frequent words.

Keywords: word representations; vector-space models; word associations

Introduction

Finding good representations of the meaning of words is a fundamental problem in cognitive science and related disciplines. Vector-space models of semantics represent words as points in an N-dimensional Euclidean space where words with similar meanings are expected to be close together. These models have been successful in both modeling human semantic processing (*e.g.*, Landauer and Dumais, 1997) and natural language processing applications (for a review, see Turney and Pantel, 2010). However, relating the similarity between words to their distance in a vector space means that these representations are subject to certain geometric constraints. Previous research has criticized this property of spatial representations because aspects of human semantic processing do not conform to these same constraints (*e.g.*, Tversky, 1977). For example, people's interpretation of semantic similarity does not always obey the triangle inequality, *i.e.*, the words w_1 and w_3 are not necessarily similar when both pairs of (w_1, w_2) and (w_2, w_3) are similar. While "asteroid" is very similar to "belt" and "belt" is very similar to "buckle", "asteroid" and "buckle" are not similar (Griffiths et al., 2007).

Recent work has resulted in significant advances in vector-space models of semantics, making it possible to train models on extremely large datasets (Mikolov et al., 2013a; Pennington et al., 2014). The resulting vector-space models—Word2Vec and GloVe—achieve state-of-the-art results for a wide range of tasks requiring machine representations of word meanings. However, the similarity between words in these models is typically measured using the cosine of the angle between word vectors (*e.g.*, Mikolov et al., 2013b; Pennington et al., 2014).

In this paper, we examine whether these constraints imply that Word2Vec and GloVe representations suffer from the same difficulty as previous vector-space models in capturing human similarity judgments. To this end, we evaluate these representations on a set of tasks adopted from Griffiths et al. (2007) in which the authors showed that the representations learned by another well-known vector-space model, Latent Semantic Analysis (Landauer and Dumais, 1997), were inconsistent with patterns of semantic similarity demonstrated in human word association data. We show that Word2Vec and GloVe suffer from similar problems. Recent probabilistic interpretations of Word2Vec (Levy and Goldberg, 2014; Arora et al., 2015) provide a way to construct a conditional probability from vector-space representations, although we show that this does not result in a significant improvement in performance over cosine similarity.

A probabilistic topic model performs less well than these vector-space models in predicting overall associations, but provides a better fit to human data on tasks where vector-spaced models are subject to geometric constraints. However, two advantages of the recent models are that they can produce word representations for very large vocabularies (millions of types) and can be trained on very large corpora (hundreds of billions of tokens). We investigate whether the performance of co-occurrence frequency—easily obtainable from large corpora—is comparable to the recent models. We find that vectors of simple co-occurrence frequency provide comparable performance to the above models, suggesting that dimensionality reduction may not be necessary feature for machine representations of words.

Vector-Space Models

We first provide high-level descriptions of two recent vector-space models that have received significant attention in the machine learning, natural language processing, information retrieval, and cognitive science communities.

Word2Vec

Word2Vec (Mikolov et al., 2013b) is a shallow neural network model with a single hidden layer that learns similar vector representations for words with similar distributional properties. They present two variants: *continuous bag of words* or *CBOW*, in which a word token is predicted from its unordered context, and *skipgram*, in which a given word token is used to predict words in its context. Both variants perform well predicting associations, analogies, and can be used to identify idiomatic multi-word phrases. We focus here on the skipgram formulation given its higher obtained performance in a variety of natural language processing tasks.

The objective of a Word2Vec model is to maximize the average log probability of each word's context following

$$J = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t), \tag{1}$$

where T is the number of training words and c is the number of context words. $p(w_{t+j}|w_t)$ is given by the softmax function,

$$p(w_o|w_i) = \frac{\exp(v'^{\top}_{w_o} v_{w_i})}{\sum_{w=1}^{W}\exp(v'^{\top}_{w} v_{w_i})}, \tag{2}$$

where W is the number of unique words (type) in the corpus $w_1 \ldots w_T$, and v_w and v'_w are the input and output vector representations of word w.

Computing the normalizing term in the softmax is prohibitively expensive for large datasets in that the cost of the computation is proportional to W (which may be in millions), thus an approximation is obtained through hierarchical softmax (Morin and Bengio, 2005), Noise Contrastive Estimation (Gutmann and Hyvärinen, 2012), or a related novel technique they introduce, negative sampling. In negative sampling, the model updates the representations of a small number of words such that the network predicts an observed "positive" word pair (e.g., *chicken salad*), and does not predict any of a number of "negative" pairs that are unlikely to be observed in the text (e.g. *chicken battleship* or *chicken advantageously*). The negative pairs are drawn from an explicitly specified noise distribution, typically a unigram model. Because a small number of negative samples are used—usually fewer than 20—a relatively small number of weights need to be adjusted each time the model updates the representation of a word. Mikolov et al. find additional performance gains by sampling less from high frequency words.

Performance of Word2Vec model thus depends on the number of hidden units (typically 50-600), the size of the context window, the degree to which frequent words are undersampled, and the choice of approximation to the full softmax; if negative sampling is used then the number of negative samples can have a significant effect on performance.

GloVe

GloVe (Pennington et al., 2014) is a weighted bilinear regression model that uses global co-occurrence statistics to derive a real-valued vector representation of each word. Like Word2Vec, GloVe learns similar vector representations for words that appear in similar contexts, however the latter model differs significantly in that it fits co-occurrence frequencies from an entire corpus rather than iterating through local context windows. GloVe exhibits particularly strong performance in analogy tasks, but also performs well on similarity tasks and named entity recognition (NER).

In GloVe, the best word representations W and $\widetilde{W}$ are found by minimizing a least squares objective:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w}_j + b_i + \widetilde{b}_j - \log X_{ij})^2 \tag{3}$$

where V is the vocabulary, i and j pick out words in the vocabulary, f(X_{ij}) is a weighting term (explicated below), w_i is the representation of the ith word, $\widetilde{w}$ is the representation of the jth word, b_i and $\widetilde{b}_j$ are bias terms, and $logX_{ij}$ is the co-occurrence count of words i and j. If X is symmetric, W and $\widetilde{W}$ are equivalent (differing only according to their random initialization). GloVe additionally introduces a weighting into the cost function of the model to avoid $\log 0$ errors and to dampen the effect of high frequency co-occurrences:

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} \text{ if } x < x_{max} \\ 1 \text{ otherwise} \end{cases} \tag{4}$$

where x is the co-occurrence count, and α allows for an exponential weighting of for counts between 0 and the threshold X_{max}. The performance of a GloVe model thus depends on the dimensionality of the word vector (typically 50 - 300), X_{max}, α , and the size of the window used to compute co-occurrence statistics around each word.

Co-occurrence Frequency

We also consider a baseline model that simply uses normalized co-occurrence frequencies of words to measure their similarity. In other words, given sufficient data, is a term-by-term matrix sufficient to predict human association norms? We note that this baseline is used by previous work to model human semantic and syntactic processing, as well as in infomration retrieval (*e.g.*, Burgess and Lund, 1997; Azzopardi, 2005).

Shortcomings of Spatial Models

Similarity between two words in a vector-space model is usually computed using the cosine of the angle or the Euclidean distance between the vectors representing the words. While intuitive, this approach has at least one significant shortcoming: cosine and Euclidean distance cannot capture the observed asymmetries in human similarity judgments because they are inherently symmetric measures. Tversky (1977) famously argued that spatial representations cannot capture human similarity judgments because the latter often violate the metric axioms. For example, elicited word (or phrase) similarity is asymmetric: when queried, most participants considered "North Korea" to be very similar to "China," while the reverse relationship was rated as significantly less strong ("China" is not very similar to "North Korea").

Griffiths et al. (2007) extended this argument to spatial representations of the semantic relationships between words, showing that similar violations of the metric axioms can be demonstrated for vector-space representations. We now revisit these analyses, examining the extent to which they are problematic for new vector-space models.

One of the properties of metric spaces is that the distance between each 3-word tuple must satisfy the *triangle inequality*: given three points x, y, and z, $d(x, z) \leq d(x, y)+d(y, z)$, where $d()$ is a distance function. This inequality constrains the possible distance values among the vector representations for each of the three words: if distances between the words in two of the pairs are very small, the distance between the words in the third pair is also expected to be small.

After demonstrating that cosine—as a monotonic function of the angle between two vectors—satisfies an analogue of the triangle inequality Griffiths et al. (2007) studied to what extent this is true among the cue–target pairs in the Nelson norms. For the words w_1, w_2, and w_3, they plot the distribution of $p(w_3|w_1)$ when both $p(w_2|w_1)$ and $p(w_3|w_2)$ are greater than a given threshold τ. They observe that even for large values of τ, there are a lot of very small values of $p(w_3|w_1)$. Consistent with the intuition that human similarity judgments are not always transitive, they find many cases where two of the pairs (w_2–w_1 and w_3–w_2) in a tuple (w_1, w_2, and w_3) are highly similar, but the words in the third pair (w_1 and w_3) are not.

As a result, by using cosine (or any distance measure more generally) on vector-space representations, we cannot replicate the asymmetric patterns of similarities observed in human judgments. To enable word representations derived with vector space models to account for a greater range of phenomena, we propose an elaborated, non-metric similarity measure for vector-space representations. Following recent work that provides a probabilistic interpretation of Word2Vec (Levy and Goldberg, 2014; Arora et al., 2015) we calculate the conditional probability for a given pair of words w_1 and w_2 using a softmax function:

$$p(w_2|w_1) = \frac{\exp(\mathbf{w_2}.\mathbf{w_1})}{\sum_{\mathbf{w_j}} \exp(\mathbf{w_j}.\mathbf{w_1})} \tag{5}$$

where $\mathbf{w_j}$ is the vector representation of w_j and $w_2.w_1$ is the dot product of the two vectors. Using this probabilistic measure, we can now examine how well Word2Vec and GloVe representations perform on tasks that do not satisfy the geometric constraints, *i.e.*, triangle inequality and asymmetries in similarity judgments.

Evaluating Vector-Space Representations

In this section, we describe the evaluation data and explain the tasks that we use to examine how well vector-space representations predict human word associations.

Data: Nelson Association Norms

Following Griffiths et al. (2007), we use the association norms from Nelson et al. (1998) as our gold-standard evaluation data. Nelson et al. (1998) performed an extensive free association experiment where they asked 6000 participants to record the first word they can think of given a cue word. The experiment resulted in a set of 5018 cues, the target words produced in response of each cue (associates), and the probability of producing each target word for a given cue. Approximately 45% of the target words are present as cues in the dataset. The Nelson norms are well-suited for the evaluation of semantic similarity because unlike most gold-standard similarity lexicons (*e.g.*, Hill et al., 2015), word associations obtained in this way potentially encode asymmetric relations: the Nelson association norms encode for many words both how likely people are to produce w_1 when cued with w_2, as well as w_2 when cued with w_1.

Evaluation Tasks

We evaluate the word representations found by these models on four tasks to assess whether they capture empirical phenomena of interest in the Nelson norms. The first two, *coefficient of correlation* and *median rank of associates*, test whether these representations capture the strength of associations between each cue–target pair. The remaining two, the *triangle inequality* and *ratio of asymmetries* specifically test whether these representations can account for human behavior on tasks with asymmetric associations.

Coefficient of correlation. Computing the correlation between two list of scores is a standard way for measuring their similarity (Budanitsky and Hirst, 2006). We created a gold-standard list of similarity scores that, for each cue–target pair in the norms, includes $p(\text{target}|\text{cue})$. We then retrieved a list of similarities for the same cue–target pairs from the representations under study, measuring similarity as either $\text{cosine}(\mathbf{w}_{\text{target}}, \mathbf{w}_{\text{cue}})$ or $p(\mathbf{w}_{\text{target}}|\mathbf{w}_{\text{cue}})$, where $\mathbf{w}_x$ is the vector representation of x. To assess the extent to which these representations can predict human similarity judgments of semantic associations, we calculated the Spearman's rank correlation coefficient (ρ_{assoc}) between these two lists.

Median rank of associates. We also assess the quality of the representations by checking whether they produce similar rankings of target words (associates) for each cue in the Nelson norms. For each cue, we rank all its associates based on their conditional probabilities (given the cue) from the Nelson norms, and also get a similar ranking for each cue in the model. For the first associate of each cue, *i.e.*, the one with the highest probability per the Nelson norms ranking, we check its rank in the model list. We take the median rank of the first associate across all the cues from the Nelson norms, and repeat this process for second and third associates.

Triangle inequality. We extend the analysis in Griffiths et al. (2007) to the evaluate whether word representations satisfy the triangle inequality. For every w_1, w_2, and w_3 such that similarity of w_1–w_2 and w_2–w_3 are greater than a threshold τ, we plot the distribution of similarity values of w_1–w_3. For the Nelson norms, similarity of words in a pair is their conditional probability; for other models similarity is given by the cosine or conditional probability. We select thresholds (τ) such that for each threshold, the number of pairs selected for each model is similar to that of the Nelson norms; The thresholds for the norms are taken from Griffiths et al. (2007).

Asymmetry ratio. Griffiths et al. (2007) show that the similarity of more than 85% of cue-target pairs in Nelson norms are asymmetric by the criterion of at least an order of magnitude difference between $p(w2|w1)$ and $p(w1|w2)$. However, distance measures are inherently symmetric and for any distance function $d()$, we have $d(w_1, w_2) = d(w_2, w_1)$. To measure the performance of vector-space representations in predicting the asymmetries, for each cue–target pair in the Nelson norms, we calculate the ratio of asymmetry as follows:

$$\mathrm{asym}(w_1, w_2) = \frac{p(w_2|w_1)}{p(w_1|w_2)} \tag{6}$$

We then calculate the Spearman's rank correlation coefficient between the asymmetry scores of these similarities and those from the Nelson norms.

Corpora and Model Training

To support comparison with Griffiths et al. (2007) we trained GloVe, Word2Vec skipgram, and collected co-occurence frequencies on TASA, the Touchstone Applied Sciences Corpus (Landauer and Dumais, 1997). This corpus consists of approximated 8M tokens taken from reading materials appropriate for a high school English students. In addition to TASA, we trained Word2Vec skipgram and GloVe, and collected co-occurrence frequencies on English Wikipedia (3.91B tokens). This corpus is too large for training a Latent Dirichlet Allocation (LDA) topic model using Gibbs Sampling. While we tried to replicate the LDA results for TASA with more scalable variational methods (Hoffman et al., 2010), the resulting topics produced associations that were significantly worse than those obtained through Gibbs sampling or either of the vector space models.

Preprocessing was matched to the extent possible across model inputs. All words were translated to their nearest lowercase ASCII equivalent. For both TASA and Wikipedia we discarded function words using the Python `stopwords` package. For TASA we removed the same set of low-information words and enforced the same frequency cutoff as Griffiths et al. (2007). For Wikipedia, we removed words that appeared on too many pages or too few, and retained only the top 100k most frequent remaining words.

To evaluate the performance of the Word2Vec skipgram model we trained 20 models across a range of hyperparameter settings, varying the size of the embedding vector (50, 100, 200, 300 or 400 hidden units), the choice of optimization method (hierarchical softmax or negative sampling), and for models with negative sampling the number of negative examples (5, 10, 15). Words with unigram probability higher than .001 are downsampled following Mikolov et al. (2013b).

Because of an implementation error, we were unable to explore a large parameter space with GloVe, and report only the results with the default parameters ($X_{max} = 10$, $\alpha = .75$, 50-dimensional vectors, and a 7-word symmetric window on either side of the target word). This leaves open the possibility that GloVe may exhibit even higher performance on TASA and Wikipedia with appropriate parameter settings. We also compute association using the LDA results (sampled document-topic and topic-word assignments) from Griffiths et al. (2007).

Finally, we used large-scale pre-trained models distributed by the authors of Word2Vec and GloVe. These largest-available models often exhibit best-in-class performance because they reflect extensive parameter search, proprietary corpora, and distributed implementations that can handle more training data than publicly-distributed single-machine implementations. For Word2Vec we used a pre-trained 300-dimensional model obtained by using the continuous bag of words architecture (CBOW) on a corpus of 100 billion words from Google News. For GloVe we used a 300-dimensional model trained by Pennington et al. (2014) using a 2014 export of Wikipedia and the Gigaword 5 corpus, consisting of approximately 6 billion tokens in total.[1]

Results

Overall associations. We first look at the coefficient of correlation that shows how the various models perform in predicting the overall associations. We find that using conditional probability in place of cosine results in slightly better performance in predicting the semantic associations when the models are trained on medium or large corpora (see cosine ("cos.") and conditional probability ("cond. pr.") columns in Table 1). We also observe that given small and medium corpora (first and second row of Table 1), the Word2Vec skip-gram has the highest correlation with human word associations; but, given the largest corpora, GloVe performs slightly better than the Word2Vec model. Interestingly, given the small and medium corpora, simple co-occurrence frequencies perform similarly to or better than the Word2Vec CBOW and GloVe representations. Looking at the second measure of associations, the median rank of the associates (Table 2), we observe that the LDAmodel and co-occurrence frequencies perform similar to Word2Vec on TASA and Wikipedia; both models exhibit better performance than GloVe. The representations of the pre-trained GLoVe model (on the largest corpus) have the lowest (best) median ranks.

Geometric constraints. The results for the triangle inequality analysis using the conditional probability measure are shown in Figure 1 (cosine results are omitted as they cannot produce the pattern). We observe the expected pattern for the Nelson norms, the LDA model, and co-occurrence frequency (see Figure 1 a-c): even for large values of τ, there are a lot of pairs that have probabilities close to zero. However, as shown in Figure 1 d-f, we do not see pairs with very small values of similarity when examining large thresholds for any of the recent vector-space representations. Our results reveal that even with a probabilistic measure, Word2Vec representations cannot predict the triangle inequality: for very high thresholds on the similarity of w_1–w_2 and w_2–w_3, there are

[1] The pre-trained Word2Vec model is available at `https://code.google.com/archive/p/word2vec/`; The pre-trained GloVe model is available at `http://nlp.stanford.edu/projects/glove/`

Table 1: The Spearman's rank correlation coefficient (ρ_{assoc}) between gold-standard association scores from Nelson norms and different models of word representations. "cos." and "cond. pr." refer to cosine and conditional probability, respectively. [*] Data unavailable or infeasible to compute given current resources.

	Word2Vec CBOW		Word2Vec skip-gram		GloVe			
	cos.	cond. pr.	cos.	cond. pr.	cos.	cond. pr.	LDA	Co-occurrence
Small (TASA)	.22	.21	**.25**	**.25**	.21	.20	.20	.21
Medium (Wikipedia)	.22	.22	.23	**.24**	.16	.19	[*]	.20
Largest available	.25	.26	[*]	[*]	.24	**.27**	[*]	[*]

Table 2: The median rank of first, second, and third associates (1^{st}/2^{nd}/3^{rd}) for different models of word representation using conditional probabilities. The number of possible targets is 3951 for all corpora.[*] Data unavailable or infeasible to compute given current resources.

	Word2Vec CBOW	Word2Vec skip-gram	GloVe	LDA	Co-occurrence
Small (TASA)	48/112/160	26/72/106	56/138/215	23/69/103.5	21/58/122
Medium (Wikipedia)	23/48/75	21/46/74	52/92/129	[*]	23/48/70
Largest available	13/29/47	[*]	**11/25/40.5**	[*]	[*]

no w_1–w_3 pairs with low similarity. These results suggest that using a probabilistic measure do not address the limitations of the vector-space models with respect to the triangle inequality.

Finally, we examine whether the representations capture the observed asymmetry in human similarity judgements as calculated in Eqn. (6). Note that we can only use conditional probabilities in this analysis because the cosine measure is symmetric. This probabilistic measure of similarities in both Word2Vec and GloVe to some extent predicts the asymmetric patterns of similarity observed in the Nelson norms (Table 3). We observe that the performance of the LDA model is comparable to the GloVe representations trained the largest corpora. The GloVe models performs significantly better than the Word2Vec models, which we believe is a result of its objective function—it uses the ratio of conditional probabilities of word pairs in training.

Table 3: The Spearman's rank correlation coefficient (ρ_{asym}) between asymmetry scores of Nelson norms and representations from the models. In our data, there are 7096 cue–target pairs for which target–cue also exits. [*] Data unavailable or infeasible to compute given current resources.

	Word2Vec CBOW	Word2Vec Skipgram	GloVe	LDA
Small	.18	.01	.32	**.49**
Medium	.20	.19	.43	[*]
Largest avail.	.20	[*]	.48	[*]

Discussion

The selection of models, corpora, and tasks presented above suggests that LDA and co-occurrence frequencies have certain advantages when compared with the vector-space representations produced by Word2Vec and GloVe. We expound on a few key points below to contextualize our results and set the stage for future research.

Most of the targets and queues analyzed here are of relatively high frequency rank. In future work we would like to investigate exactly how robust each of these models are to sparsity to test the hypothesis that reduced-dimensionality models are better at generalizing, such that they better predict associations for low frequency words.

The two vector-space models investigated here were both developed with the explicit objective of capturing meaningful linguistic difference in the linear substructure of the model (*e.g.*, the vector produced by *king* - *man* + *woman* is closest to *queen*). As such, these models show strong performance on analogy tasks, while LDA typically fairs poorly. One question is thus whether a single representation could predict word associations, while preserving linear substructure.

Conclusion

We show that representations from two new vector-space models, Word2Vec and GloVe, suffer from the same geometric constraints as predecessors, and are consequently unable to predict some of the characteristics of human similarity judgments, such as asymmetric similarity relations between two words or triangle inequality. Besides performing well in the above task, word representations derived from LDA topic modeling show remarkable predictive power with respect to human judgments given that they are learned from a dataset two orders of magnitude smaller than comparably performing vector-space models.

References

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Randwalk: A latent variable model approach to word embeddings. *arXiv preprint arXiv:1502.03520*, 2015.

M. M. Azzopardi, L.;Girolami. Probabilistic hyperspace analogue to language. In *Proceedings of the 28th Annual ACM Conference on Research and Development in Infomration Retrieval (SIGIR 2005)*, pages 575–576, 2005.

A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

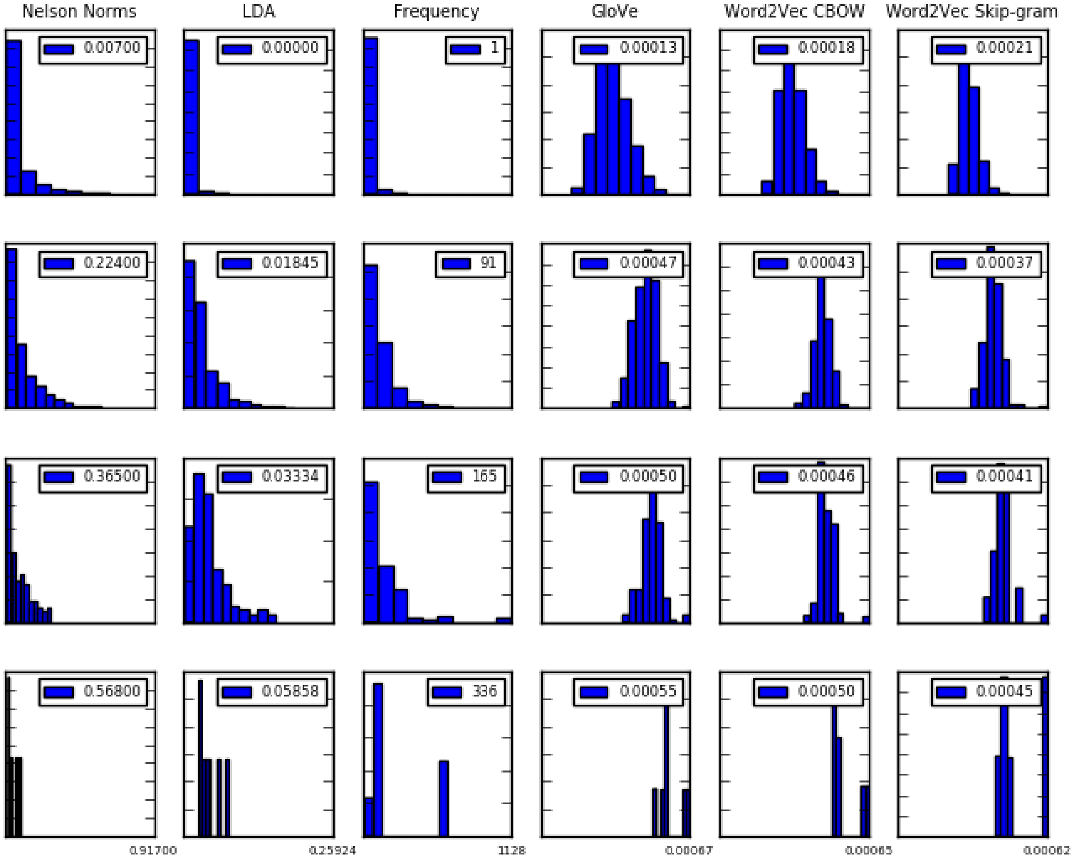

Figure 1: The triangle inequality histograms on TASA: conditional probability for the third pair of words in a tuple (w_1–w_3) when the first two pairs (w_1–w_2 and w_2–w_3) are above the given threshold.

C. Burgess and K. Lund. Modelling parsing constraints with high-dimensional context space. *Language and cognitive processes*, 12(2-3):177–210, 1997.

T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psych. Rev.*, 114(2):211, 2007.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

F. Hill, R. Reichart, and A. Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2015.

M. Hoffman, D. M. Blei, and F. R. Bach. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864, 2010.

T. K. Landauer and S. T. Dumais. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, pages 211–240, 1997.

O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185. 2014.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119. 2013b.

F. Morin and Y. Bengio. Hierarchical probabilistic neural network language model. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 246–252. Society for Artificial Intelligence and Statistics, 2005.

D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The university of south florida free association, rhyme, and word fragment norms. 1998.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

A. Tversky. Features of similarity. *Psychological Review*, 84 (4):327, 1977.