Correlation Coefficients and Semantic Textual Similarity

Vitalii Zhelezniak, Aleksandar Savkov, April Shen & Nils Y. Hammerla
Babylon Health
{firstname.lastname}@babylonhealth.com

Abstract

A large body of research into semantic textual similarity has focused on constructing state-of-the-art embeddings using sophisticated modelling, careful choice of learning signals and many clever tricks. By contrast, little attention has been devoted to similarity measures between these embeddings, with cosine similarity being used unquestionably in the majority of cases. In this work, we illustrate that for all common word vectors, cosine similarity is essentially equivalent to the Pearson correlation coefficient, which provides some justification for its use. We thoroughly characterise cases where Pearson correlation (and thus cosine similarity) is unfit as similarity measure. Importantly, we show that Pearson correlation is appropriate for some word vectors but not others. When it is not appropriate, we illustrate how common nonparametric rank correlation coefficients can be used instead to significantly improve performance. We support our analysis with a series of evaluations on word-level and sentence-level semantic textual similarity benchmarks. On the latter, we show that even the simplest averaged word vectors compared by rank correlation easily rival the strongest deep representations compared by cosine similarity.

1 Introduction

Textual embeddings are immensely popular because they help us reason about the abstract and fuzzy notion of semantic similarity in purely geometric terms. Distributed representations of words in particular (Bengio et al., 2003; Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017; Joulin et al., 2017) have had a massive impact on machine learning (ML), natural language processing (NLP), and information retrieval (IR).

Recently, much effort has also been directed towards learning representations for larger pieces of text, with methods ranging from clever compositions of word embeddings (Mitchell and Lapata, 2008; De Boom et al., 2016; Arora et al., 2017; Wieting et al., 2016; Wieting and Gimpel, 2018; Zhelezniak et al., 2019) to sophisticated neural architectures (Le and Mikolov, 2014; Kiros et al., 2015; Hill et al., 2016; Conneau et al., 2017; Gan et al., 2017; Tang et al., 2017; Zhelezniak et al., 2018; Subramanian et al., 2018; Pagliardini et al., 2018; Cer et al., 2018).

Comparatively, there is little research into similarity measures for textual embeddings. Despite some investigations into alternatives (Camacho-Collados et al., 2015; De Boom et al., 2015; Santus et al., 2018; Zhelezniak et al., 2019), cosine similarity has persistently remained the default and unquestioned choice across the field. This is partly because cosine similarity is very convenient and easy to understand. Sometimes, however, we have to resist what is convenient and instead use what is appropriate. The core idea behind our work is to treat each word or sentence embedding as a sample of (e.g. 300) observations from some scalar random variable. Hence, no matter how mysterious word vectors appear to be, just like any samples, they become subject to the full power of traditional statistical analysis. We first show that in practice, the widely used cosine similarity is nothing but the Pearson correlation coefficient computed from the paired sample. However, Pearson's r is extremely sensitive to even slight departures from normality, where a single outlier can conceal the underlying association. For example, we find that Pearson's r (and thus cosine similarity) is acceptable for word2vec and fastText but not for GloVe embeddings. Perhaps surprisingly, when we average word vectors to represent sentences, cosine similarity remains acceptable for word2vec, but not for fastText any longer. We show that this seemingly counterintuitive behaviour can be predicted

Proceedings of NAACL-HLT 2019, pages 951–962
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

by elementary univariate statistics, something that is already well known to researchers and practitioners alike. Furthermore, when there are clear indications against cosine similarity, we propose to repurpose rank-based correlation coefficients, such as Spearman's ρ and Kendall's τ, as *similarity measures* between textual embeddings. We support this proposition by a series of experiments on word- and sentence-level semantic textual similarity (STS) tasks. Our results confirm that rank-based correlation coefficients are much more effective when the majority of vectors break the assumptions of normality. Moreover, we show how even the simplest sentence embeddings (such as averaged word vectors) compared by rank correlation easily rival recent deep representations compared by cosine similarity.

2 Related Work

At the heart of our work is a simple statistical analysis of pre-trained word embeddings and exploration of various correlation coefficients as proxies for semantic textual similarity. Hence, any research that combines word embeddings with tools from probability and statistics is relevant. Of course, word embeddings themselves are typically obtained as the learned parameters of statistical machine learning models. These models can be trained on large corpora of text to predict a word from its context or vice versa (Mikolov et al., 2013a). Alternatively, there are also supervised approaches (Wieting et al., 2015, 2016; Wieting and Gimpel, 2017, 2018).

A different line of research tries to move away from learning word embeddings as point estimates and instead model words as parametric densities (Vilnis and McCallum, 2014; Barkan, 2017; Athiwaratkun and Wilson, 2017). These approaches are quite appealing because they incorporate semantic uncertainty directly into the representations. Of course, such representations need to be learned explicitly. In some cases one could estimate the densities even for off-the-shelf embeddings, but this still requires access to the training data and the usefulness of such post-factum densities is limited (Vilnis and McCallum, 2014). In other words, these approaches are not very helpful to practitioners who are accustomed to using high-quality pre-trained word embeddings directly.

Arguably, statistical analysis of pre-trained word embeddings is not as principled as applying a probabilistic treatment end-to-end. Any such analysis, however, is very valuable as it provides insights and justifications for methods that are already in widespread use. For example, removing the common mean vector and a few top principal components makes embeddings even stronger and is now a common practice (Mu and Viswanath, 2018; Arora et al., 2016, 2017; Ethayarajh, 2018). These works view word embeddings as observations from some D-dimensional distribution; such treatment is naturally suitable for studying the overall geometry of the embedding space. We, on the other hand, are interested in studying the similarities between individual word vectors and require a completely different perspective. To this end, we see each word embedding itself as a sample of D observations from a scalar random variable. It is precisely this shift in perspective that allows us to reason about semantic similarity in terms of correlations between random variables and make the connection to the widely used cosine similarity.

Finally, we propose using rank-based correlation coefficients when cosine similarity is not appropriate. Recently, Santus et al. (2018) introduced a rank-based similarity measure for word embeddings, called APSynP, and demonstrated its efficacy on outlier detection tasks. However, the results on the word-level similarity benchmarks were mixed, which, interestingly enough, could have been predicted in advance by our analysis.

3 Correlation Coefficients and Semantic Similarity

Suppose we have a vocabulary of N words $\mathcal{V} = \{w_1, w_2, \ldots, w_N\}$ and the word embeddings matrix $\mathbf{W} \in \mathbb{R}^{N \times D}$, where each row $\mathbf{w}^{(i)}$ for $i = 1, \ldots, N$ is a D-dimensional word vector. Popular pre-trained embeddings in practice typically have dimension $D = 300$, while the vocabulary size N can range from thousands to millions of words.

We now consider the following: what kinds of statistical analyses can we apply to $\mathbf{W}$ in order to model semantic similarity between words? One option is to view all word embeddings $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots \mathbf{w}^{(N)}$ as a sample of N observations from some D-variate distribution $P(E_1, \ldots E_D)$. For example, we can fit a Gaussian and study how all 300 dimensions correlate with each other. Perhaps we can fit a mixture model and see how the embeddings cluster. We

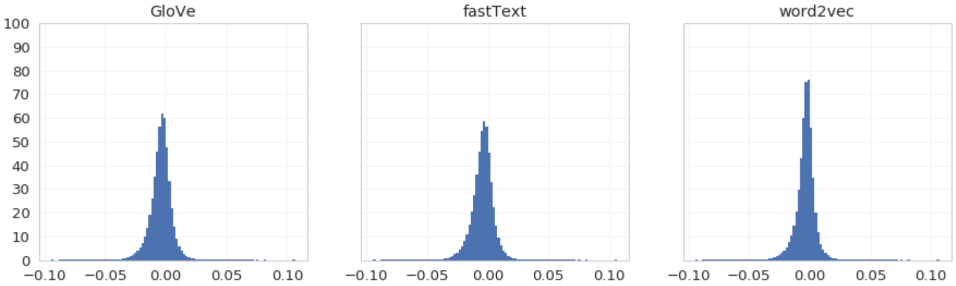

Figure 1: Normalised histograms of the mean distribution for three commonly used word embedding models: GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), and word2vec (Mikolov et al., 2013b,c).

could also normalise them and study their distribution on the unit sphere. It is clear by now that $P(E_1, \dots, E_D)$ is suitable for describing the overall geometry of the embedding space but is not very useful for our goals.

If we are to reason about similarities between individual word vectors, we should instead be looking at the transpose of $\mathbf{W}$. Putting it differently, we see $\mathbf{W}^T$ as a sample of D observations from an N-variate distribution $P(W_1, W_2, \dots, W_N)$, where W_i is a scalar random variable corresponding to the word w_i. This distribution is exactly what we need because the associations between W_i captured by P will become a proxy for semantic similarity. Often we are only interested in pairwise similarities between two given words w_i and w_j; thus the main object of our study is the bivariate marginal $P(W_i, W_j)$. To lighten up the notation slightly, we denote the two words as w_x and w_y, and the corresponding random variables as X and Y. We also refer to $P(X, Y)$ as the joint and $P(X)$, $P(Y)$ as the marginals. In practice, of course, the actual $P(X, Y)$ is unknown but we can make inferences about it based on our sample $(\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), (x_2, y_2), \dots (x_D, y_D)\}$.

First, we might want to study the degree of linear association between X and Y, so we compute the sample Pearson correlation coefficient

$$\hat{r} = \frac{\sum_{i=1}^{D}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{D}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{D}(y_i - \bar{y})^2}}, \quad (1)$$

where $\bar{x}$ and $\bar{y}$ are the sample means

$$\bar{x} = \sum_{i=1}^{D} x_i, \qquad \bar{y} = \sum_{i=1}^{D} y_i. \quad (2)$$

Let's view $\mathbf{x}$ and $\mathbf{y}$ as word embeddings momentarily and compute cosine similarity between them

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{D} x_i y_i}{\sqrt{\sum_{i=1}^{D} x_i^2}\sqrt{\sum_{i=1}^{D} y_i^2}}. \quad (3)$$

We see now that Equation (1) and Equation (3) look very similar; when the sample means $\bar{x}$, $\bar{y}$ are zero, cosine similarity and Pearson's $\hat{r}$ are equal. The real question here is whether or not they coincide in practice. Putting it differently, if we take any single word vector $\mathbf{w}$ and compute the mean (across the D dimensions), is this mean close to zero? It turns out that it is, and we can show this by plotting the distribution of the means across the whole vocabulary for various popular word embeddings (see Figure 1). We find that the means are indeed highly concentrated around zero; quantitatively, only 0.03% of them are above 0.05 in magnitude. It follows that in practice when we compute cosine similarity between word vectors, we are actually computing Pearson correlation between them.

However, is this always the right thing to do? When the joint $P(X, Y)$ is bivariate normal, Pearson correlation indeed provides a complete summary of association between X and Y, simply because the covariance is given by $\operatorname{cov}(X, Y) = r_{XY}\sigma_X\sigma_Y$. However, Pearson correlation is extremely sensitive to even the slightest departures from normality – a single outlier can easily conceal the underlying association (Pernet et al., 2013). When the normality of $P(X, Y)$ is in doubt, it is preferable to use robust correlation coefficients such as Spearman's $\hat{\rho}$ or Kendall's $\hat{\tau}$.

Spearman's $\hat{\rho}$ is just a Pearson's $\hat{r}$ between ranked variables

$$\hat{\rho} = \frac{\sum_{i=1}^{D}(r[x_i] - \overline{r[x]})(r[y_i] - \overline{r[y]})}{\sqrt{\sum_{i=1}^{D}(r[x_i] - \overline{r[x]})^2}\sqrt{\sum_{i=1}^{D}(r[y_i] - \overline{r[y]})^2}}, \quad (4)$$

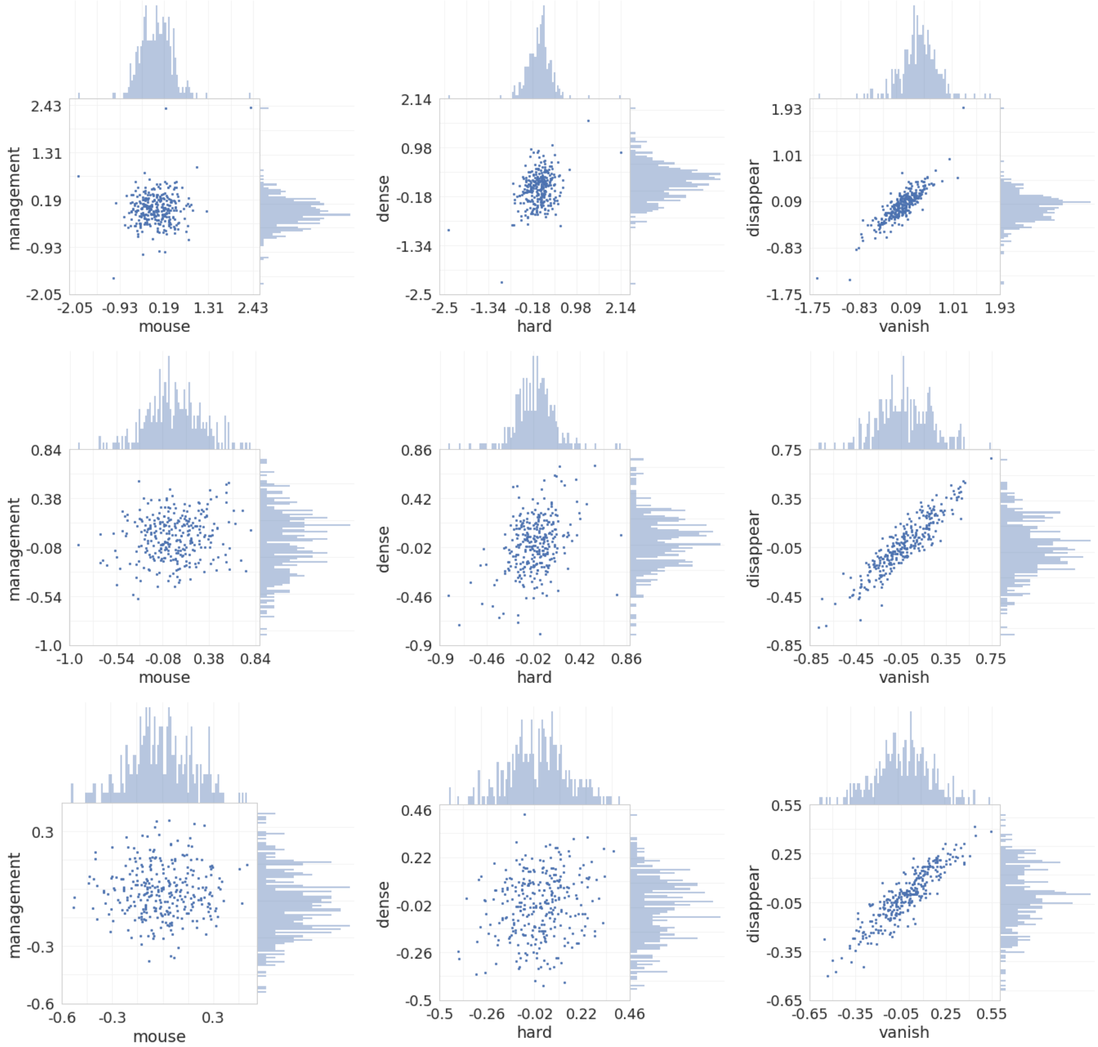

Figure 2: Scatter plots of paired word vectors, along with histograms (100 bins) of individual word vectors. Rows from top to bottom correspond to one of three common models: GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), and word2vec (Mikolov et al., 2013b,c). Columns from left to right correspond to increasing degrees of semantic similarity between the words, and accordingly increasingly pronounced linear correlation between the word vectors. Both the scatter plots and the histograms exhibit the presence of heavy outliers for GloVe vectors, which damage the efficacy of Pearson correlation in reliably capturing statistical associations. The outliers are relatively less pronounced for fastText vectors and much less pronounced for word2vec vectors.

where $r[x_i]$ denotes the integer rank of x_i in a vector $\mathbf{x}$ (similarly $r[y_i]$), while $\overline{r[x]}$ and $\overline{r[y]}$ denote the means of the ranks. Kendall's $\hat{\tau}$ is given by

$$\hat{\tau} = \frac{2}{D(D-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j)\operatorname{sgn}(y_i - y_j) \quad (5)$$

and can be interpreted as a normalised difference between the number of concordant pairs and the number of discordant pairs. These rank correlation coefficients are more robust to outliers than Pearson's $\hat{r}$ because they limit the effect of outliers to their ranks: no matter how far the outlier is, its rank cannot exceed D or fall below 1 in our case. There are also straightforward extensions to account for the ties in the ranks.

The main point here is the following. It is tempting to chose cosine similarity as the default and apply it everywhere regardless of the embedding type. Sometimes, however, we should resist using what is convenient and instead use what is appropriate. For example, if the samples corresponding to the marginals $P(X)$ and $P(Y)$ already look non-normal, then we conclude the joint $P(X, Y)$ cannot be a bivariate normal and the appropriateness of cosine similarity should be seriously questioned. In some of these cases, using a rank-based coefficient as a similarity measure be-

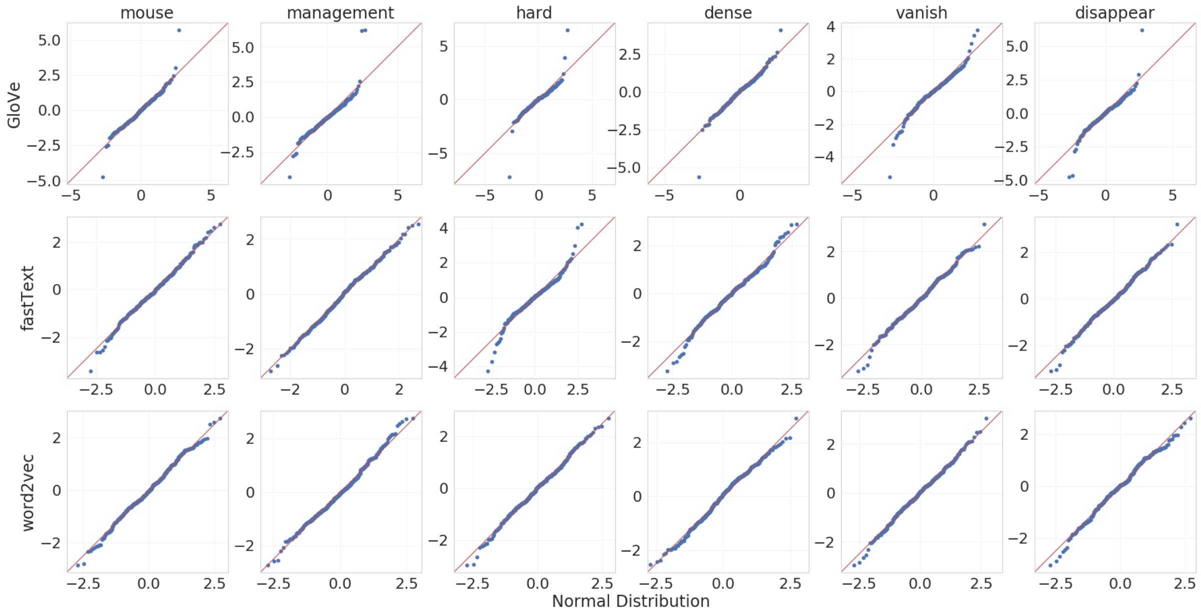

Figure 3: Q-Q plots comparing the theoretical quantiles of a standard normal distribution (horizontal axis) against the sample quantiles of standardised (Mean 0, SD 1) word vectors from three commonly used models: GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), and word2vec (Mikolov et al., 2013b,c). Perfect fit to the 45-degree reference line would indicate perfect normality. Note the pronounced discrepancy between the normal distribution and GloVe vectors due to the presence of heavy outliers. The discrepancy is relatively less pronounced for fastText vectors and much less pronounced for word2vec vectors. Figure 2 provides an alternative visualisation of the same phenomena.

tween word embeddings would be a much better alternative. It will capture the association better, which could in turn lead to large improvements in performance on the downstream tasks. In general, of course, even normal marginals do not imply a normal joint and care should be exercised either way; however we found the normality of marginals to be a good indication for cosine similarity within the scope of the present work. In the next section we illustrate how the ideas discussed here can be applied in practice.

4 Statistical Analysis of Word Embeddings: A Practical Example

No matter how mysterious word vectors appear to be, just like any samples, they are subject to the full power of traditional statistical analysis. As a concrete example, let's say we decided to use GloVe vectors (Pennington et al., 2014). We treat each vector $\mathbf{w}_i$ as if it was a sample of 300 observations from some scalar random variable W_i. We take a few hundred of these vectors, run a normality test such as Shapiro-Wilk (Shapiro and Wilk, 1965) and find that the majority of them look non-normal ($p < 0.05$). As there is a considerable evidence against normality, we flag these vectors as 'suspicious' and look at them closer. We pick a few vectors and examine their histograms and Q-Q plots, seen in Figure 2 and Figure 3 respectively; the latter in particular is a statistical tool used to compare empirical and theoretical data distributions, and is explained further in the caption of Figure 3. In both cases we observe that while the bulk of the distribution looks bell-shaped, we always get a couple of very prominent outliers.

Next, we can also visualise our word vectors in a way more directly relevant to the task at hand. We take some pairs of words that are very similar (e.g. 'vanish' and 'disappear'), moderately similar ('hard' and 'dense'), and completely dissimilar ('mouse' and 'management') and make the scatter plots for the corresponding pairs of word vectors. These are also presented in Figure 2. We see that for similar pairs the relationship is almost linear; it becomes less linear as the similarity decreases, until we see a spherical blob (no relationship) for the most dissimilar pair. However, we again face the presence of bivariate outliers that are too far away from the main bulk of points.

Given this evidence, which course of action shall we take? Based on the presence of heavy outliers, we reject the normality of GloVe vectors and rule out the use of Pearson's r and cosine similarity. Instead we can use rank correlation coefficients, such as Spearman's ρ or Kendall's τ, as they offer more robustness to outliers. Note that in this particular case, it may also be acceptable to winsorize (clip) the vectors and only then proceed with the standard Pearson's r. We evaluate the proposed solution on word-level similarity tasks and observe good improvement in performance over cosine similarity, as seen in Table 1.

Of course this exploration is in no way specific to GloVe vectors. Note that from Figure 2 and Figure 3, we also see that word2vec vectors in particular tend to be much more normally distributed, meaning that we don't find strong evidence against using Pearson correlation; this is again backed up by Table 1.

This example helps illustrate that proper statistical analysis applied to existing textual embeddings is extremely powerful and comparatively less time-consuming than inventing new approaches. Of course, this analysis can be made as fine-grained as desired. Quite coarsely, we could have rejected the use of cosine similarity right after the Shapiro-Wilk test; on the other hand, we could have used even more different tests and visualisations. The decision here rests with the practitioner and depends on the task and the domain.

5 Experiments

To empirically validate the utility of the statistical framework presented in Section 3, we run a set of evaluations on word- and sentence-level STS tasks. In all experiments we rely on the following publicly available word embeddings: GloVe (Pennington et al., 2014) trained on Common Crawl (840B tokens), fastText (Bojanowski et al., 2017) trained on Common Crawl (600B tokens), and word2vec (Mikolov et al., 2013b,c) trained on Google News. All the source code for our experiments is available on GitHub[1]; in the case of the sentence-level tasks we rely also on the SentEval toolkit (Conneau and Kiela, 2018).

	task	**N**	**V**	**COS**	**PRS**	**SPR**	**KEN**
GloVe	YP-130	.01	=	57.1	57.0	60.2	59.9
	MTURK-287	.13	=	69.3	69.3	70.8	70.9
	SIMLEX-999	.04	R	40.8	40.9	46.0	46.0
	MC-30	.10	=	78.6	79.2	77.0	77.4
	SIMVERB-3500	.04	R	28.3	28.3	34.3	34.3
	RG-65	.14	=	76.2	75.9	71.0	71.1
	WS-353-SIM	.06	=	80.3	80.2	80.1	80.1
	VERB-143	.00	=	34.1	33.9	37.8	37.4
	RW-STANFORD	.16	R	46.2	46.2	52.8	52.9
fastText	YP-130	.73	=	62.5	62.6	65.3	65.0
	MTURK-287	.88	=	72.6	72.7	73.4	73.3
	SIMLEX-999	.76	=	50.3	50.2	50.4	50.2
	MC-30	.90	=	85.2	85.2	84.6	84.5
	SIMVERB-3500	.68	=	42.6	42.6	42.6	42.5
	RG-65	.90	N	85.9	85.8	83.9	84.1
	WS-353-SIM	.84	N	84.0	83.8	82.4	82.2
	VERB-143	.21	=	44.7	44.9	43.8	44.3
	RW-STANFORD	.80	=	59.5	59.4	59.0	58.9
word2vec	YP-130	.95	=	55.9	56.1	55.0	54.7
	MTURK-287	.94	=	68.4	68.3	67.1	67.2
	SIMLEX-999	.94	=	44.2	44.2	43.9	44.0
	MC-30	.92	=	78.8	77.9	76.9	76.9
	SIMVERB-3500	.96	=	36.4	36.4	36.0	36.0
	RG-65	.94	=	75.0	74.3	73.9	74.2
	WS-353-SIM	.92	N	77.2	76.9	75.8	75.8
	VERB-143	.98	=	49.7	50.1	48.9	49.0
	RW-STANFORD	.95	N	53.4	53.5	52.5	52.5

Table 1: Spearman's ρ on word similarity tasks for combinations of word vectors and the following similarity metrics: cosine similarity (COS), Pearson's r (PRS), Spearman's ρ (SPR), and Kendall τ (KEN). **N** indicates the proportion of sentence vectors in a task for which the null hypothesis of normality in a Shapiro-Wilk test was *not* rejected at $\alpha = 0.05$. The **V** column indicates the type of the best performing method: a rank-based correlation coefficient (R), a non-rank-based correlation or measure (N), or a tie (=). The winners in **V** were determined by comparing the top rank-based method for that vector/task combination with the top non-rank-based method. Winners were assigned only when the difference was statistically significant as determined by 95% BCa confidence intervals.

First we consider a group of word-level similarity datasets that are commonly used as benchmarks in previous research: *WS-353-SIM* (Finkelstein et al., 2001), *YP-130* (Yang and Powers, 2005), *SIMLEX-999* (Hill et al., 2015), *SimVerb-3500* (Gerz et al., 2016), *RW-STANFORD* (Luong et al., 2013), *Verb-143* (Baker et al., 2014), *MTurk-287* (Radinsky et al., 2011), *MC-30* (Miller and Charles, 1991). These datasets contain pairs of words and a human-annotated similarity score for each pair. The success metric for the experiments is the Spearman correlation between the human-

[1] https://github.com/Babylonpartners/corrsim

annotated similarity scores and the scores generated by the algorithm. To avoid any confusion whatsoever, note that here Spearman correlation serves as an evaluation criterion; this is completely unrelated to using Spearman correlation as a similarity measure between word embeddings as proposed in Section 3. Bias-corrected and accelerated bootstrap (Efron, 1987) 95% confidence intervals were used to determine statistical significance. We report the results for different combinations of word vectors and similarity measures in Table 1. The main takeaways from these experiments are the following:

- There is no significant difference between the results obtained with cosine similarity and Pearson correlation. This is because empirically, the means across dimensions of these word vectors are approximately zero, in which case cosine similarity and Pearson correlation are approximately the same.

- Rank correlation coefficients tend to perform on par or better than cosine and Pearson on tasks and word vectors where there is a high proportion of non-normally distributed word vectors (over 90%). This makes sense because it is precisely in the non-normal cases where Pearson correlation fails.

- When word vectors seem mostly normal, our analysis does not tell us definitively whether cosine similarity or rank correlation should perform better, and indeed we see that cosine and Pearson perform on par or better than Spearman and Kendall.

In the second set of experiments, we use the datasets from the sentence-level Semantic Textual Similarity shared task series 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017). The success metric for these experiments is the Pearson correlation between the human-annotated sentence similarity scores and the scores generated by the algorithm. Again, this use of Pearson correlation as an evaluation criterion is completely unrelated to its use as a similarity measure between sentence embeddings. Note that the dataset for the STS13 SMT subtask is no longer publicly available, so the mean Pearson correlations reported in our experiments involving this task have been re-calculated accordingly.

For these experiments we use averaged word vectors as a sentence representation for various types of word vector, with similarity computed by the different correlation coefficients as well as cosine similarity and APSynP (Santus et al., 2018). We report these results in Table 2, and the full significance analysis for each subtask in Table 4. We also compare the top performing combination of averaged word vectors and correlation coefficient against several popular approaches from the literature that use cosine similarity: BoW with ELMo embeddings (Peters et al., 2018), Skip-Thought (Kiros et al., 2015), InferSent (Conneau et al., 2017), Universal Sentence Encoder with DAN and Transformer (Cer et al., 2018), and STN multitask embeddings (Subramanian et al., 2018). These results are presented in Table 3. Our observations for the sentence-level experiments are as follows:

	task	N	COS	PRS	SPR	KEN	APS
GloVe	STS12	.01	52.1	52.0	53.4	52.6	**53.8**
	STS13	.00	49.6	49.6	56.2	**56.7**	55.9
	STS14	.00	54.6	54.5	**63.2**	63.0	63.0
	STS15	.00	56.1	56.0	64.5	**65.3**	64.2
	STS16	.00	51.4	51.4	62.1	**63.7**	60.8
fastText	STS12	.01	58.3	58.3	**60.2**	59.0	58.4
	STS13	.01	57.9	58.0	65.1	**65.3**	61.8
	STS14	.00	64.9	65.0	**70.1**	69.6	68.5
	STS15	.00	67.6	67.6	74.4	**74.6**	72.7
	STS16	.00	64.3	64.3	73.0	**73.5**	70.7
word2vec	STS12	.95	51.6	51.6	51.7	**53.1**	45.3
	STS13	.94	58.2	**58.3**	57.9	58.2	57.2
	STS14	.96	**65.6**	**65.6**	65.5	**65.6**	64.1
	STS15	.96	67.5	67.5	67.3	**68.3**	66.5
	STS16	.96	64.7	64.7	64.6	**65.6**	63.9

Table 2: Mean Pearson correlation on STS tasks for methods using combinations of word vectors and similarity metrics. All methods use averaged word vectors to represent sentences. The similarity measures are: cosine similarity (COS), Pearson's r (PRS), Spearman's ρ (SPR), Kendall τ (KEN) and APSynP (APS). **N** indicates the proportion of sentence vectors in a task for which the null hypothesis of normality in a Shapiro-Wilk test was *not* rejected at $\alpha = 0.05$

- The conclusions from the word-level tasks continue to hold and are even more pronounced: in particular, cosine and Pearson are essentially equivalent, and the increase in performance of rank-based correlation coefficients over cosine similarity on non-normal sentence vectors is quite dramatic.

Approach	STS 12	13	14	15	16
ELMo (BoW)	55	53	63	68	60
Skip-Thought	41	29	40	46	52
InferSent	**61**	56	68	71	71
USE (DAN)	59	59	68	72	70
USE (Transformer)	**61**	64	**71**	74	**74**
STN (multitask)	60.6	54.7†	65.8	74.2	66.4
fastText - COS	58.3	57.9	64.9	67.6	64.3
fastText - SPR	60.2	65.1	70.1	74.4	73.0
fastText - KEN	59.0	**65.3**	69.6	**74.6**	73.5

Table 3: Mean Pearson correlation on STS tasks for a variety of methods in the literature compared to averaged fastText vectors with different similarity metrics: cosine similarity (COS), Spearman's ρ (SPR), and Kendall τ (KEN). Values in bold indicate best results per task. Previous results are taken from Perone et al. (2018) (only two significant figures provided) and Subramanian et al. (2018). † indicates the only STS13 result (to our knowledge) that includes the SMT subtask.

- Averaged word vectors compared with rank correlation easily rival modern deep representations compared with cosine similarity.

Finally, the fraction of non-normal word vectors used in sentence-level tasks is consistent with the results reported for the word-level tasks in Table 1. However, we observe the following curious phenomenon for fastText. While there is no evidence against normality for the majority of fastText vectors, perhaps surprisingly, when we average them to represent sentences, such sentence embeddings are almost entirely non-normal (Table 2). Empirically we observe that many high-frequency words or stopwords have prominently non-normal fastText vectors. Although stopwords constitute only a small fraction of the entire vocabulary, they are very likely to occur in any given sentence, thus rendering most sentence embeddings non-normal as well. While it's tempting to invoke the Central Limit Theorem (at least for longer sentences), under our formalism, averaging word vectors corresponds to averaging scalar random variables used to represent words, which are neither independent nor identically distributed. In other words, there are no easy guarantees of normality for such sentence vectors.

6 Discussion

In this work, we investigate statistical correlation coefficients as measures for semantic textual similarity and make the following contributions:

- We show that in practice, for commonly used word vectors, cosine similarity is equivalent to the Pearson correlation coefficient, motivating an alternative statistical view of word vectors as opposed to the geometric view, which is more prevalent in the literature.
- We illustrate via a concrete example the power and benefits of using elementary statistics to analyse word vectors.
- We characterise when Pearson correlation is applied inappropriately and show that these conditions hold for some word vectors but not others, providing a basis for deciding whether or not cosine similarity is a reasonable choice for measuring semantic similarity.
- We demonstrate that when Pearson correlation is not appropriate, non-parametric rank correlation coefficients, which are known to be more robust to various departures from normality, can be used as similarity measures to significantly improve performance on word- and sentence-level STS tasks.
- Finally, we show in particular that sentence representations consisting of averaged word vectors, when compared by rank correlation, can easily rival much more complicated representations compared by cosine similarity.

We hope that these contributions will inspire others to carefully investigate and understand alternative measures of similarity. This is particularly important in the realm of sentence representations, where there are many more complex ways of constructing sentence representations from word embeddings besides the simple averaging procedure tested here. It is worth exploring whether a more subtle application of rank correlation could help push these more complex sentence representations to even better performance on STS tasks.

A final and fascinating direction of future work is to explain the non-normality of certain types of word vectors (and in particular the presence of outliers) by analysing their training procedures. Preliminary investigations suggest that unsupervised

		GloVe			**fastText**			**word2vec**		
		SPR	**COS**	**Δ95% BCa CI**	**SPR**	**COS**	**Δ95% BCa CI**	**SPR**	**COS**	**Δ95% BCa CI**
STS12	MSRpar	35.90	**42.55**	[-10.74, -2.52]	**39.66**	**40.39**	[-3.22, 1.80]	38.79	**39.72**	[-1.77, -0.16]
	MSRvid	**68.80**	66.21	[1.31, 4.09]	**81.02**	73.77	[6.16, 8.53]	**77.88**	**78.11**	[-0.52, 0.06]
	SMTeuroparl	**48.73**	**48.36**	[-5.26, 6.48]	50.29	**53.03**	[-5.41, -0.17]	**16.96**	16.06	[0.21, 1.34]
	surprise.OnWN	**66.66**	57.03	[6.89, 12.76]	**73.15**	68.92	[2.19, 6.56]	**70.75**	**71.06**	[-0.73, 0.09]
	surprise.SMTnews	**47.12**	**46.27**	[-4.27, 5.50]	**56.67**	**55.20**	[-2.50, 5.50]	**53.93**	**52.91**	[-0.13, 2.09]
STS13	FNWN	**43.21**	**38.21**	[-0.54, 10.24]	**49.40**	39.83	[2.74, 16.46]	**40.73**	**41.22**	[-2.07, 1.07]
	headlines	**67.59**	63.39	[2.58, 5.89]	**71.53**	**70.83**	[-0.17, 1.58]	**65.48**	**65.22**	[-0.12, 0.66]
	OnWN	**57.66**	47.20	[8.10, 13.02]	**74.33**	63.03	[9.27, 13.50]	67.49	**68.29**	[-1.29, -0.33]
STS14	deft-forum	**39.03**	30.02	[5.24, 13.52]	**46.20**	40.19	[2.88, 10.00]	**42.95**	**42.66**	[-0.43, 1.03]
	deft-news	**68.99**	**64.95**	[-0.39, 8.72]	**73.08**	**71.15**	[-0.36, 4.39]	**67.33**	**67.28**	[-0.70, 0.91]
	headlines	**61.87**	58.67	[1.15, 5.48]	**66.33**	**66.03**	[-0.68, 1.28]	**62.09**	**61.88**	[-0.22, 0.66]
	images	**70.36**	62.38	[6.30, 10.00]	**80.51**	71.45	[7.44, 10.96]	76.98	**77.46**	[-0.89, -0.09]
	OnWN	**67.45**	57.71	[7.89, 11.97]	**79.37**	70.47	[7.42, 10.50]	74.69	**75.12**	[-0.81, -0.08]
	tweet-news	**71.23**	53.87	[13.98, 21.67]	**74.89**	70.18	[2.60, 7.21]	68.78	**69.26**	[-0.92, -0.01]
STS15	answers-forums	**50.25**	36.66	[10.18, 17.55]	**68.28**	56.91	[7.99, 15.23]	**53.74**	**53.95**	[-1.28, 0.86]
	answers-students	**69.99**	63.62	[4.25, 9.59]	**73.95**	71.81	[0.69, 3.56]	**72.45**	**72.78**	[-0.70, 0.04]
	belief	**58.77**	44.78	[10.11, 19.05]	**73.71**	60.62	[9.64, 19.50]	**61.73**	**61.89**	[-0.84, 0.46]
	headlines	**69.61**	66.21	[1.65, 5.29]	**72.93**	**72.53**	[-0.40, 1.20]	**68.58**	**68.72**	[-0.48, 0.23]
	images	**73.85**	69.09	[3.45, 6.29]	**83.18**	76.12	[5.76, 8.58]	**80.04**	**80.22**	[-0.55, 0.18]
STS16	answer-answer	**43.99**	40.12	[0.90, 7.36]	**54.51**	45.13	[5.14, 15.93]	**43.41**	**43.14**	[-1.03, 1.43]
	headlines	**67.05**	61.38	[2.43, 9.44]	**71.00**	**70.37**	[-0.93, 2.13]	**66.55**	**66.64**	[-0.66, 0.51]
	plagiarism	**72.25**	54.61	[12.69, 23.74]	**84.45**	74.49	[6.38, 14.81]	75.21	**76.46**	[-2.31, -0.37]
	postediting	**69.03**	53.88	[12.01, 19.06]	**82.73**	68.76	[7.55, 22.96]	**73.87**	**73.35**	[-0.08, 1.21]
	question-question	**58.32**	47.21	[7.02, 18.18]	**72.29**	62.62	[6.35, 13.64]	**63.94**	**63.74**	[-1.03, 1.38]

Table 4: Pearson correlations between human sentence similarity score and a generated score. Generated scores were produced via measuring Spearman correlation (SPR), as explained in Section 3, and cosine similarity (COS) between averaged word vectors. Values in bold represent the best result for a subtask given a set of word vectors, based on a 95% BCa confidence interval (Efron, 1987) on the differences between the two correlations. In cases of no significant difference, both values are in bold.

objectives based on the distributional hypothesis are probably not to blame, as word vectors trained without relying on the distributional hypothesis, such as those of Wieting et al. (2015), still exhibit non-normality to some degree. The actual causes remain to be determined. We believe that understanding the reasons for these empirically-observed characteristics of textual embeddings would be a significant step forwards in our overall understanding of these crucial building blocks for data-driven natural language processing.

Acknowledgements

We would like to thank Dan Busbridge for his useful comments and suggestions.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation*

(SemEval 2014), pages 81–91. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43. Association for Computational Linguistics.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. *International Conference on Learning Representations*.

Ben Athiwaratkun and Andrew Wilson. 2017. Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1645–1656. Association for Computational Linguistics.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 278–289. Association for Computational Linguistics.

Oren Barkan. 2017. Bayesian neural word embedding. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 3135–3143. AAAI Press.

Yoshua Bengio, Rjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Nasari: a novel approach to a semantically-aware representation of items. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 567–577. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. Association for Computational Linguistics.

C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B. Dhoedt. 2015. Learning semantic similarity for very short texts. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1229–1234.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recogn. Lett.*, 80(C):150–156.

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Kawin Ethayarajh. 2018. Unsupervised random walk sentence embeddings: A strong but simple baseline.

In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 406–414, New York, NY, USA. ACM.

Zhe Gan, Yunchen Pu, Ricardo Henao, Chunyuan Li, Xiaodong He, and Lawrence Carin. 2017. Learning generic sentence representations using convolutional neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2390–2400. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning*, pages 1188–1196.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Cyril R. Pernet, Rand Wilcox, and Guillaume A. Rousselet. 2013. Robust correlation analyses: False positive and power validation using a new open source matlab toolbox. *Frontiers in Psychology*, 3.

Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA. ACM.

Enrico Santus, Hongmin Wang, Emmanuele Chersoni, and Yue Zhang. 2018. A rank-based similarity metric for word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 552–557. Association for Computational Linguistics.

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Shuai Tang, Hailin Jin, Chen Fang, Zhaowen Wang, and Virginia R. de Sa. 2017. Exploring asymmetric encoder-decoder structure for context-based sentence representation learning. *CoRR*, abs/1710.10380.

Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. *CoRR*, abs/1412.6623.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards Universal Paraphrastic Sentence Embeddings. In *International Conference on Learning Representations*.

John Wieting and Kevin Gimpel. 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2088. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462. Association for Computational Linguistics.

Dongqiang Yang and David M. W. Powers. 2005. Measuring semantic similarity in the taxonomy of wordnet. In *Proceedings of the Twenty-eighth Australasian Conference on Computer Science - Volume 38*, ACSC '05, pages 315–322, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Vitalii Zhelezniak, Dan Busbridge, April Shen, Samuel L. Smith, and Nils Y. Hammerla. 2018. Decoding Decoders: Finding Optimal Representation Spaces for Unsupervised Similarity Tasks. *CoRR*, abs/1805.03435.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, Francesco Moramarco, Jack Flann, and Nils Y. Hammerla. 2019. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. In *International Conference on Learning Representations*.