Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Journal Track

Context Vectors Are Reflections of Word Vectors in Half the Dimensions (Extended Abstract)*

Zhenisbek Assylbekov and **Rustem Takhanov**
School of Sciences and Humanities, Nazarbayev University
{zhassylbekov, rustem.takhanov}@nu.edu.kz

Abstract

This paper takes a step towards the theoretical analysis of the relationship between word embeddings and context embeddings in models such as word2vec. We start from basic probabilistic assumptions on the nature of word vectors, context vectors, and text generation. These assumptions are supported either empirically or theoretically by the existing literature. Next, we show that under these assumptions the widely-used word-word PMI matrix is approximately a random symmetric Gaussian ensemble. This, in turn, implies that context vectors are reflections of word vectors in approximately half the dimensions. As a direct application of our result, we suggest a theoretically grounded way of tying weights in the SGNS model.[1]

1 Introduction and Main Result

Today word embeddings play an important role in many natural language processing tasks, from predictive language models and machine translation to image annotation and question answering, where they are usually plugged into a larger model. An understanding of their properties is of interest as it may allow the development of embeddings that are better both in interpretability and quality of models built upon them. This paper takes a step in this direction.

Notation: We let $\mathbb{R}$ denote the real numbers. Bold-faced lowercase letters ($\mathbf{x}$) denote vectors in Euclidean space, bold-faced uppercase letters ($\mathbf{X}$) denote matrices, plain-faced lowercase letters (x) denote scalars, plain-faced uppercase letters (X) denote scalar random variables, 'i.i.d.' stands for 'independent and identically distributed'. We use the sign $\sim$ to abbreviate the phrase 'distributed as', and the sign $\propto$ to abbreviate 'proportional to'.

Assuming that words have already been converted into indices, let $\{1, \dots, n\}$ be a finite vocabulary of words. Following the setup of the widely used WORD2VEC model [Mikolov *et al.*, 2013], we will use *two* vectors per each word i:

- $\mathbf{w}_i$ is an embedding of the word i when i is a center word,
- $\mathbf{c}_i$ is an embedding of the word i when i is a context word.

We make the following key assumptions in our work.

Assumption 1. *A priori word vectors $\mathbf{w}_1, \dots, \mathbf{w}_n \in \mathbb{R}^d$ are i.i.d. draws from isotropic multivariate Gaussian distribution:*

$$\mathbf{w}_i \overset{iid}{\sim} \mathcal{N}\left(\mathbf{0},\, \tfrac{1}{d}\mathbf{I}\right), \tag{1}$$

where $\mathbf{I}$ is the $d \times d$ identity matrix.

This is motivated by the work of Arora *et al.* [2016], where the ensemble of word vectors consists of i.i.d. draws generated by $\mathbf{v} = s \cdot \hat{\mathbf{v}}$, with $\hat{\mathbf{v}}$ being from the spherical Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and s being a scalar random variable with bounded expectation and range. In their work, the norm $\|\mathbf{v}_i\|$ of the word vector for a word i is related to its unigram probability $p(i)$, and to allow a sufficient dynamic range for these probabilities they needed the multiplier s. In our work, unigram probabilities are not mapped to vector lengths, and this is why we do not need such multiplier. Direct relationship between word probabilities and word vector norms is also implied by the model of Hashimoto *et al.* [2016].

Assumption 2. *Context vectors $\mathbf{c}_1, \dots, \mathbf{c}_n$ are related to word vectors according to*

$$\mathbf{c}_i = \mathbf{Q}\mathbf{w}_i, \quad i = 1, \dots, n, \tag{2}$$

for some orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$.

This is mainly guided by the work of Press and Wolf [2017], who showed that context vectors in the SGNS model of Mikolov *et al.* [2013] are distributed similarly to word vectors in the sense that pairwise cosine distances between word (input) embeddings strongly correlate with the corresponding pairwise cosine distances between context (output) embeddings (see their Table 4). This is why we choose the transform from word vectors to context vectors to be orthogonal as it preserves inner products and consequently Euclidean norms. Notice, that $\mathbf{c}_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \tfrac{1}{d}\mathbf{I}\right)$.

Assumption 3. *Given a word j, the probability of any word i being in its context[2] is given by*

$$p(i \mid j) \propto p_i \cdot e^{\mathbf{w}_j^\top \mathbf{c}_i} \tag{3}$$

*This paper is an extended abstract of an article in the Journal of Artificial Intelligence Research [Assylbekov and Takhanov, 2019].

[1] Our modification of the SGNS is available at https://github.com/zh3nis/word2vec_wt

[2] Context is a fixed-size symmetric window around the given word.

where $p_i = p(i)$ is the unigram probability for the word i, which is inverse proportional to its smoothed frequency rank r_i, i.e.

$$p_i \propto \frac{1}{r_i^{1-\alpha}}, \qquad \alpha \in (0, 1]. \tag{4}$$

This is similar to the log-linear model of Arora *et al.* [2016], but differs in the following aspects: $\mathbf{c}_i$ is not assumed to do a random walk over the unit sphere with bounded displacement; we use the factor p_i to directly capture word frequencies and do not model them via vector norms. Equation (3) can be interpreted as follows: probability that the word i occurs in the context of the word j is the probability that the word i occurs anywhere in a large corpus, corrected for the relationship between words i and j. This approach was already considered by Melamud *et al.* [2017] but in their work i is the entire left context of the word j, and $\mathbf{c}_i$ is a vector representation of this entire context. Also, like Arora *et al.* [2016] but unlike Melamud *et al.* [2017], we use the model (3) for a theoretical analysis rather than for fitting to data. Smoothing of the unigram probabilities (i.e. raising them to power $1-\alpha$) is motivated by the works of Mikolov *et al.* [2013], Levy *et al.* [2015], Pennington *et al.* [2014], where $\alpha = 0.25$ is a typical choice. We notice here that $\alpha = 0^+$ gives us Zipf's law [Zipf, 1935], whereas $\alpha = 1$ gives us uniform distribution of word frequencies which is not valid empirically but on the other hand can be used to explain additivity of word vectors [Gittens *et al.*, 2017].

The relationship between word (input) and context (output) vectors was addressed in several previous works. E.g., in recurrent neural network language modeling (RNNLM), tying input and output embeddings is a useful regularization technique introduced earlier [Bengio *et al.*, 2001] and studied in more details recently [Press and Wolf, 2017; Inan *et al.*, 2017]. This technique improves language modeling quality (measured as perplexity of a held-out text) while decreasing the total number of trainable parameters almost two-fold since most of the parameters in RNNLM are due to embedding matrices. The direct application of this regularization technique to SGNS worsens the quality of word vectors as was shown empirically by Press and Wolf [2017] and by Gulordava *et al.* [2018]. This worsening was predicted earlier by Goldberg and Levy [2014] using a simple linguistic observation that words usually do not appear in the contexts of themselves. This basically means that $\mathbf{Q} \neq \mathbf{I}$ in (2). At the same time, there is empirical evidence that the relationship between input and output embeddings *is* linear [Mimno and Thompson, 2017; Gulordava *et al.*, 2018]. In this paper, we provide a theoretical justification for this and reveal the exact form of the transform $\mathbf{Q}$. Our main contribution is the following

Theorem 1. *Under Assumptions 1, 2, and 3 above, the context vector $\mathbf{c}_i$ for a word i is a reflection of the word vector $\mathbf{w}_i$ in approximately half of the dimensions.*

In general, our word and context vectors live in a d-dimensional vector space over real numbers ($\mathbb{R}^d$). By Theorem 1 we can settle them in a $d/2$-dimensional vector space over complex numbers ($\mathbb{C}^{d/2}$) in such way that the context vector $\widetilde{\mathbf{c}}_i \in \mathbb{C}^{d/2}$ for a word i is the *complex conjugate* of the word vector $\widetilde{\mathbf{w}}_i \in \mathbb{C}^{d/2}$. This is in line with the results of Allen *et al.* [2019], however they use a completely different set of basic assumptions and their primary goal is to encode statistical properties of words directly into word vectors.

2 Proof Sketch of Theorem 1

The proof is divided into three steps: first we show that the partition function in (3) concentrates around 1, and thus $\propto$ can be replaced by $\approx$; using this fact we show that $\mathbf{Q}$ is (approximately) an involutary matrix, i.e. similar to $\mathrm{diag}(\mathbf{q})$, $\mathbf{q} \in \{+1, -1\}^d$; and finally we show that the word-word pointwise-mutual information matrix (PMI) is approximately a symmetric Gaussian random matrix with weakly dependent entries. Based on the theory of random matrices, the latter fact immediately implies a symmetric around 0 distribution of the PMI eigenvalues, and thus the statement of the Theorem 1.

The complete proof can be found in the full version [Assylbekov and Takhanov, 2019].

3 Empirical Verification

In this section we empirically verify two predictions of our theory: a symmetric distribution of a PMI spectrum and the involutarity of a matrix transforming input embeddings into output ones.

3.1 Symmetry of a PMI Spectrum

To verify that the real-world PMI matrices have indeed a symmetric (around 0) distribution of their eigenvalues, we consider a widely-used dataset `text8`[3] that consists of 17M tokens and from which we extract PMI matrices using the HYPERWORDS tool of Levy *et al.* [2015]. We use the default settings for all hyperparameters, except the word frequency threshold and context window size. We were ignoring words that appeared less than 100 times, resulting in a vocabulary of 11,815 words. We additionally experiment with the context window size 5, which by default is set to 2, and which we believe could affect the results. By default, HYPERWORDS zeroes out an entry $\mathrm{PMI}_{i,j}$ if the words i and j do not co-occur in the corpus.[4] The eigenvalues of the PMI matrices are calculated using the TENSORFLOW library. The histograms of eigenvalues are provided in Figure 1. As we can see, the distributions are not perfectly symmetric with a little right skewness, but in general they seem to be symmetric. Notice, that this is in stark contrast with the equation (2.5) from Arora *et al.* [2016], which claims that the PMI matrix should be approximately positive semi-definite, i.e. that it should have mostly positive eigenvalues. Also, notice that the shapes of distributions are far from resembling the Wigner semicircle law $x \mapsto \frac{1}{2\pi}\sqrt{4-x^2}$, which is the limiting distribution for the eigenvalues of many random symmetric matrices with i.i.d. entries [Wigner, 1955; Wigner, 1958]. This means that the entries of a typical PMI

[3] http://mattmahoney.net/dc/textdata.html.

[4] This means that apriori each i and j are assumed to be independent, and we follow this convention. Thus, our PMI matrices do not have any $-\infty$'s, instead, they have lots of 0's, i.e. they are sparse.

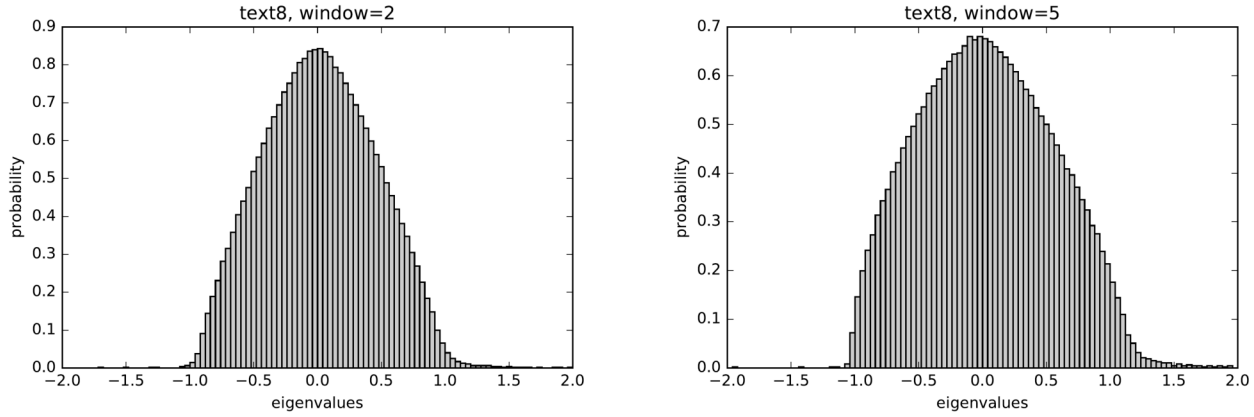

Figure 1: Empirical distribution of eigenvalues of PMI matrices.

matrix *are* dependent, otherwise we would observe approximately semicircle distributions for its eigenvalues. Interestingly, there is a striking similarity between the shapes of distributions in Figure 1 and of spectral densities of the scale-free random graphs with strong clustering [Farkas *et al.*, 2001] which arise in physics and network science. Notice that the connection between human language structure and scale-free random graphs with strong clustering was observed previously by Cancho and Solé [2001], and we believe it is worth investigating this connection deeper.

3.2 Involutarity of Q for SGNS Embeddings

Our theory suggests that the matrix $\mathbf{Q}$ in Assumption 2 should be approximately involutary. To test this claim we train off-the-shelf skip-gram embeddings $\{\mathbf{w}_i\}$ and $\{\mathbf{c}_i\}$ on `text8` dataset using the reference WORD2VEC implementation from the TENSORFLOW codebase. Let $\mathbf{W} \in \mathbb{R}^{n\times d}$ be a word embedding matrix in which i-th row is $\mathbf{w}_i^\top$, and $\mathbf{C} \in \mathbb{R}^{n\times d}$ be a context embedding matrix in which i-th row is $\mathbf{c}_i^\top$. According to (2),

$$\mathbf{C}^\top = \mathbf{Q}\mathbf{W}^\top \quad \Leftrightarrow \quad \mathbf{C} = \mathbf{W}\mathbf{Q}^\top.$$

Thus, $\hat{\mathbf{Q}} := \mathbf{W}^\dagger\mathbf{C}$ should give an approximately involutary matrix, where $\mathbf{W}^\dagger$ is the pseudo-inverse of $\mathbf{W}$. This means that $\hat{\mathbf{Q}}^2$ should be approximately an identity matrix. The distribution of diagonal and off-diagonal elements of $\hat{\mathbf{Q}}^2$ is given in Fig. 2. We see that the diagonal elements are concentrated around their mean 0.68, while the off-diagonal elements are concentrated around 0, i.e. $\hat{\mathbf{Q}}^2 \approx 0.68 \cdot \mathbf{I}$. However, our theory predicts $\hat{\mathbf{Q}}^2 \approx \mathbf{I}$. We attribute this mismatch to the underlying gap between Assumption 2 and the empirical observations: in SGNS the transform between word and context vectors is *not exactly* orthogonal. We stress here that our assumptions are *motivated by* but are not exactly consistent with the skip-gram embeddings. Despite this, our theory is quite applicable to the SGNS model as is shown in Section 4.

4 Weight Tying in the Skip-gram Model

We would like to apply our results to tie embeddings in the skip-gram model of Mikolov *et al.* [2013] in a theoretically grounded way. One may argue that our key Assumption 3 differs from the softmax-prediction of the skip-gram model. Although this is true, in fact the softmax normalization is never used in practice when training skip-gram. Instead it is common to replace the softmax cross-entropy by the negative sampling objective (Eq. (4) in Mikolov *et al.* [2013]), and its optimization is almost equivalent to finding a low-rank approximation of the shifted word-word PMI matrix in the form $\mathbf{w}_i^\top \mathbf{c}_j \approx \mathrm{PMI}_{ij} - \log k$ [Levy and Goldberg, 2014b]. Since our Assumptions lead to the same conclusion up to a constant shift [Assylbekov and Takhanov, 2019, Eq. (26)], we believe that Theorem 1 can be directly applied to tie word ($\mathbf{w}_i$) and context ($\mathbf{c}_i$) embeddings in the SGNS model. For this purpose we form a vector $\mathbf{q} = [+1, \dots, +1, -1 \dots, -1] \in \mathbb{R}^d$ consisting of equal number of $+1$'s and -1's, and then put

$$\mathbf{c}_i = \mathbf{q} \odot \mathbf{w}_i \tag{5}$$

for all words i in the vocabulary. This is equivalent to (2) when the matrix $\mathbf{Q}$ has a diagonal form with the first $d/2$ diagonal entries being $+1$ and the rest $d/2$ entries being -1. Such modification of the SGNS is refered to as 'SGNS + WT'. We also experiment with random flipping of signs: for this purpose we form $\mathbf{q} \in \mathbb{R}^d$ as a random vector consisting of d i.i.d. draws from the Rademacher distribution[5] and then put $\mathbf{c}_i$ as in (5). Such variant of weight tying is referred to as 'SGNS + WTR'.

The word embeddings $\mathbf{w}_i$ are initialized randomly, and then trained on `text8` and `enwik9` using the reference WORD2VEC implementation from the TENSORFLOW codebase with all hyperparameters set to their default values[6] except that we choose the learning rate to decay 20% faster in the weight-tied model. This additional tuning of the learning rate decay is not surprising: the model with tied embeddings has two times fewer parameters compared to the model

[5]Rademacher distribution is a discrete probability distribution where a random variate X has a 50% chance of being $+1$ and a 50% chance of being -1.

[6]Embedding size $d = 200$, 15 epochs to train, initial learning rate 0.2 on `text8` and 0.15 on `enwik9`, 100 negative samples per training example, batch size 16, windows size 5.

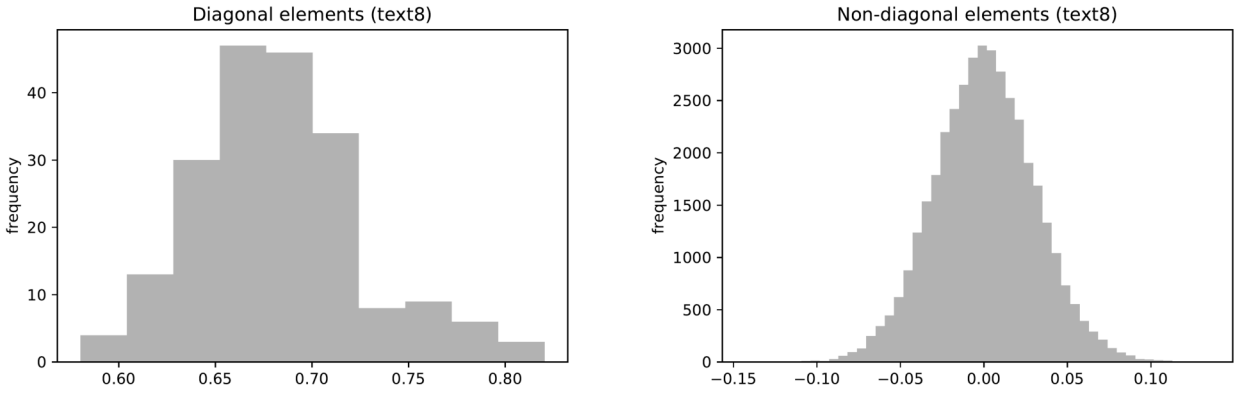

Figure 2: Diagonal and off-diagonal elements of $(\mathbf{W}^\dagger \mathbf{C})^2$.

Data	Model	Size	Finkelstein *et al.* WordSim	Bruni *et al.* MEN	Radinsky *et al.* M. Turk	Luong *et al.* Rare Words	Google	MSR
text8	SGNS	28M	.681	.241	.631	.072	.307	.286
	SGNS+WT	14M	.638	.216	.642	.058	.309	.319
	SGNS+WTR	14M	.637	.215	.624	.057	.314	.319
enwik9	SGNS	87M	.671	.268	.662	.213	.558	.410
	SGNS+WT	44M	.640	.237	.615	.188	.515	.425
	SGNS+WTR	44M	.633	.236	.639	.175	.516	.429

Table 1: Evaluation of word embeddings on the analogy tasks (Google and MSR) and on the similarity tasks (the rest). For word similarities evaluation metric is the Spearman's correlation with the human ratings, while for word analogies it is the percentage of correct answers. Model sizes are in number of trainable parameters.

with untied weights, and this leads to a significant change of the optimization landscape, which in turn results in the need to tune the most sensitive hyperparameter — the learning rate (or its decay schedule). As is standard nowadays the trained embeddings are evaluated on several word similarity and word analogy tasks. We used the HYPERWORDS tool of Levy *et al.* [2015] and we refer the reader to their paper for the methodology of evaluation. We only mention here a few key points:

- Our goal is not to beat state of the art, but to empirically validate the statement of Theorem 1. This is why we were evaluating only word (input) embeddings for both SGNS and SGNS+WT. I.e., we were not adding context vectors to word vectors in the similarity tasks, as it is usually done nowadays.
- Word similarity datasets contain word pairs together with human-assigned similarity scores. The word vectors are evaluated by ranking the pairs according to their cosine similarities and measuring the correlation (Spearman's ρ) with the human ratings.
- For answering analogy questions (a is to b as c is to ?) we use the 3COSMUL of Levy and Goldberg [2014a] and the evaluation metric for the analogy questions is the percentage of correct answers.

The results of evaluation are provided in Table 1. First of all, notice that random flipping of signs (SGNS+WTR) gives practically the same results as non-random flipping (SGNS+WT). Next, SGNS+WT produces embeddings comparable in quality with those produced by the baseline SGNS model despite having 50% fewer parameters. This also empirically validates the statement of our Theorem 1. We notice that similar results can be obtained by letting the linear transform $\mathbf{Q}$ be a trainable matrix as shown by Gulordava *et al.* [2018]. The main difference of our approach is that we know exactly the form of $\mathbf{Q}$, and thus we do not need to learn it.

5 Conclusion

There is a remarkable relationship between human language and other branches of science, and we can get interesting and practical results by studying such relationships deeper. For example, the modern theory of random matrices is replete with theoretical results that can be immediately applied to models of natural language once such models are cast into the appropriate probabilistic setting, as is done in this paper.

Acknowledgements

This work is supported by the Nazarbayev University Collaborative Research Program 091019CRP2109.

References

[Allen *et al.*, 2019] Carl Allen, Ivana Balazevic, and Timothy Hospedales. What the vec? towards probabilistically grounded embeddings. In *Advances in Neural Information Processing Systems*, pages 7465–7475, 2019.

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

[Assylbekov and Takhanov, 2019] Zhenisbek Assylbekov and Rustem Takhanov. Context vectors are reflections of word vectors in half the dimensions. *Journal of Artificial Intelligence Research*, 66:225–242, 2019.

[Bengio *et al.*, 2001] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. 2001.

[Bruni *et al.*, 2012] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. Distributional semantics in technicolor. In *Proceedings of ACL*, pages 136–145. Association for Computational Linguistics, 2012.

[Cancho and Solé, 2001] R Ferrer I Cancho and Ricard V Solé. The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*, 268(1482):2261, 2001.

[Farkas *et al.*, 2001] Illés J Farkas, Imre Derényi, Albert-László Barabási, and Tamas Vicsek. Spectra of "real-world" graphs: Beyond the semicircle law. *Physical Review E*, 64(2):026704, 2001.

[Finkelstein *et al.*, 2002] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131, 2002.

[Gittens *et al.*, 2017] Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. Skip-gram-zipf+ uniform= vector additivity. In *Proceedings of ACL*, volume 1, pages 69–76, 2017.

[Goldberg and Levy, 2014] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[Gulordava *et al.*, 2018] Kristina Gulordava, Laura Aina, and Gemma Boleda. How to represent a word and predict it, too: Improving tied architectures for language modelling. In *Proceedings of EMNLP*, pages 2936–2941, 2018.

[Hashimoto *et al.*, 2016] Tatsunori B Hashimoto, David Alvarez-Melis, and Tommi S Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286, 2016.

[Inan *et al.*, 2017] Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. In *Proceedings of ICLR*, 2017.

[Levy and Goldberg, 2014a] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of CoNLL*, pages 171–180, 2014.

[Levy and Goldberg, 2014b] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Proceedings of NeurIPS*, pages 2177–2185, 2014.

[Levy *et al.*, 2015] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[Luong *et al.*, 2013] Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, pages 104–113, 2013.

[Melamud *et al.*, 2017] Oren Melamud, Ido Dagan, and Jacob Goldberger. A simple language model based on pmi matrix approximations. In *Proceedings of EMNLP*, pages 1860–1865, 2017.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*, pages 3111–3119, 2013.

[Mimno and Thompson, 2017] David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In *Proceedings of EMNLP*, pages 2873–2878, 2017.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, 2014.

[Press and Wolf, 2017] Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of EACL*, volume 2, pages 157–163, 2017.

[Radinsky *et al.*, 2011] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM, 2011.

[Wigner, 1955] Eugene P Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, pages 548–564, 1955.

[Wigner, 1958] Eugene P Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, pages 325–327, 1958.

[Zipf, 1935] GK Zipf. *The psycho-biology of language: an introduction to dynamic philology.* Houghton Mifflin, 1935.