Revisiting GloVe, Word2Vec and BERT: On the Homogeneity of Word Vectors

Ruiyu Wang *
University of Toronto
rwang@cs.toronto.edu

Abstract

Popular embeddings such as GloVe , Word2Vec , and BERT are often considered distinct to each other. Through a series of experiments, we examine their linear separability and the potential for approximation between different embeddings. Our findings reveal that word embeddings exhibit a remarkable level of homogeneity, as utilizing methods like Support Vector Machines and a 2-layer Multi-Layer Perceptrons (MLP), we are able to produce effective approximations to embeddings, which is evidenced by high cosine similarity scores and strong alignment in multiple tests. Furthermore, our exploration of MLP application on contextualized embeddings highlights the similarity between traditional and contextualized representations, which infers the homogeneity between them. These results underscore the need for a re-evaluation of the fundamental nature of word embeddings.

1 Introduction

Word embeddings play a pivotal role in the landscape of Natural Language Processing (NLP) research. Traditional word embedding techniques, exemplified by GloVe [Pennington et al., 2014] and Word2Vec [Mikolov et al., 2013a], alongside their respective training methodologies, continue to demonstrate efficacy across diverse domains [Schlechtweg et al., 2019, Tahmasebi et al., 2021]. Concurrently, contextualized embeddings based on BERT [Devlin et al., 2018] have consistently delivered state-of-the-art performance across a wide spectrum of NLP tasks. Even amidst the emergence of Large Language Models (LLM), where the principles of scalability govern various tasks, the efficacy of producing efficient representations remains central to an effective LLM-based pipeline.

Previous research endeavors have predominantly focused on comparing and applying various word embeddings to novel use cases [Rezaeinia et al., 2019, Naili et al., 2017, Wu et al., 2018], despite that some studies suggest the training methodologies of these embeddings may not be entirely distinct [Shi and Liu, 2014, Levy et al., 2015]. This observation becomes even more pronounced with the advent of BERT , which has revolutionized the NLP landscape. Apart from the remarkable performance of BERT embeddings, it is essential to acknowledge that they cannot be conclusively deemed perfect, particularly concerning their vector distribution [Liang et al., 2021, Ethayarajh, 2019]. Table 1 underscores that BERT embeddings may not conform precisely to the envisioned contextualization. Specifically, with an increase in fixed context, words at the center tend to exhibit a higher likelihood of sharing similar representations. Moreover, instead of disseminating long-distance information across the context, this context-awareness appears to converge quickly within a relatively confined context window. As a result, a more thorough examination of the representations and their origins is imperative to foster a deeper understanding of the intricacies involved.

*Work done under the supervision of Prof. Gerald Penn.

Preprint. Under review.

Context Window Size	1	3	5	7	9
Average Pairwise Cosine Similarity	0.512	0.737	0.820	0.877	0.908

Table 1: Average pairwise cosine similarity for BERT representations within specific context window sizes. The cosine similarity between the central word and other words under the same context is computed. While words appear distant when individually examined (i.e., context size = 1), they quickly cluster together with fixed contexts. Training data sourced from the British National Corpus.

In this study, we performed tests[2] to quantify the disparities and facilitate approximation between embeddings. Specifically, we draw linear transformations between embeddings, using the Support Vector Machines (SVM) to classify the differences between them, and approximate one embedding by utilizing a multi-layer perceptron (MLP) on the other. Our primary contributions are as follows:

- We introduce a novel approach to assess the classifiability between embeddings. Through the concatenation of diverse word vectors into a single sequence, we demonstrate that an SVM can effectively discern whether they originate from the same word or not.
- We highlight that word embeddings such as `GloVe` and `Word2Vec` should not be regarded as disparate entities but rather as distinct perspectives on a universal representation. Notably, their discrepancies can be readily classified by an SVM, and an MLP can adeptly capture and transfer all pertinent information between their respective embedding spaces.
- We extend the test paradigm to contextualized vectors. Particularly, We devise an efficient simulation of BERT by employing a centered vector complemented by fixed context vectors. Our findings elucidate that, following the efficient approximation through context, BERT 1. exhibits no distinction from a context-rich iteration of the universal representation, and 2. can be fully replaced by the surrounding context vectors.

2 Related Works

The classification of various word embeddings often hinges upon their distinct training objectives and the methodologies employed to construct the vector space. Based on their training procedures, these embeddings can be categorized into three distinct types:

Counted Based Representations Count-based representations are derived from word co-occurrence frequencies [Hamilton et al., 2018]. In this approach, each cell in the matrix M_{ij} corresponds to the number of instances where the words i and j co-occur within a specified context window. Subsequently, various transformations are applied to the raw co-occurrence matrix to enhance its performance and reduce dimensionality. Two notable methods that adopt this training procedure are Positive Pointwise Mutual Information (PPMI) and `GloVe` [Pennington et al., 2014]. PPMI computes the mutual informativeness between words and sets a cut-off threshold at zero. Conversely, `GloVe` learns word vectors such that the dot product of two vectors equals the log probability of their co-occurrence.

Prediction Based Representations Prediction-based representations entail predicting either the word itself or its surrounding context. One approach involves predicting the word representations based on the surrounding context, known as the Continuous Bag-of-Words (CBOW) model [Zhang et al., 2010]. Conversely, predicting the surrounding context given the word in the middle is termed Skip-Gram [Mikolov et al., 2013b]. The `Word2Vec` model is based on Skip-Gram with Negative Sampling (SGNS), which not only aims to predict the correct surrounding words but also learns to discriminate against incorrect ones, known as negative samples.

BERT BERT [Devlin et al., 2018] is a language representation model that has significantly advanced various natural language understanding tasks. Central to BERT's training objective are two key components: Masked Language Model (MLM) and Next Sentence Prediction (NSP). The MLM objective involves randomly masking a certain percentage of input tokens in a sequence and tasking the model with predicting the masked tokens based on the context provided by the surrounding

[2]The GitHub link will be provided for further reference.

tokens. This approach encourages the model to understand the bidirectional context of words within a sentence, enabling it to capture deeper semantic relationships. Concurrently, the NSP objective focuses on training the model to determine whether two sentences in a pair are contiguous or not. By predicting whether the second sentence in a pair follows the first in the original text, BERT learns to grasp the coherence and flow of discourse, enhancing its ability to comprehend and generate meaningful textual representations. Together, these training objectives empower BERT to acquire a robust understanding of language structure and semantics, thereby enabling it to excel in a wide range of natural language processing tasks.

3 Experiments

3.1 Procrustes

The Procrustes is a least-square problem that finds the optimal transformation from matrix A to B. Solving this problem enables us to achieve an optimal alignment between the embedding spaces and assess the quality of the alignment.

Given a vocabulary V with size n and an embedding space E, we construct a matrix $M_{n\times d}$ such that the line i of the matrix M represents the i-th word in the vocabulary. For two matrices M_a and M_b that are constructed using the above procedures from embeddings E_a and E_b, we seek an optimal transformation W^* such that the sum of the squared euclidean distance between BW^* and A is minized. In particular, we want:

$$W^* = \arg\min_w \sum_i ||B_i W - A_i||^2$$

The Procrustes method and its variant, Orthogonal Procrustes [Schönemann, 1966], are commonly utilized for aligning embedding spaces if the embeddings can be effectively transferred via a single orthogonal linear transformation. In our study, we evaluate the performance of Procrustes alignment between `GloVe` vectors and `Word2Vec` vectors by computing the average cosine similarity between the vectors in the transformed embedding space (BW^*) and the target embedding space (A).

3.2 Support Vector Machine

The Support Vector Machine (SVM) is a widely employed method in various classification tasks. However, directly classifying the origin of vectors from different embeddings poses a challenge, primarily because word vectors are typically normalized during their production [Levy et al., 2015]. To address this challenge, we propose a novel approach.

Given arbitrary word vectors $w_a \in E_a$ and $w_b \in E_b$, we define a mixed representation $r = w_a \oplus w_b$ where $\oplus$ denotes vector concatenation. We then label the representations by if the two component word vectors come from the same word. By creating such a dataset and training Support Vector Machines with various kernel setups, we aim to elucidate the disparities between different embeddings and assess whether an SVM can effectively discern and account for these differences. Specifically, the dataset is constructed using `GloVe` and `Word2Vec` vectors, and we ensure a balanced distribution between true and false labels.

3.3 Neural Network

Neural networks represent the cornerstone of contemporary deep learning methodologies. Despite the emergence of new architectures showcasing state-of-the-art performances, multi-layer perceptrons (MLPs) remain extensively utilized as either heads or adaptors on backbones.

In our experiment, we leverage neural networks to approximate one embedding space to another. Through training a neural network on a dataset comprising word vectors from one embedding space alongside their corresponding vectors from another space, our objective is to acquire a mapping function capable of transforming embeddings from one space to another. Given a word vector $w_a \in E_a$, the objective is to minize $MSE(MLP(w_a) - w_b)$, where MSE denotes the mean squared error and $w_b \in E_b$.

Given the trained MLP model, we denote the vector mapping as $w_m = MLP(w_a)$. We evaluate the quality of the mapping by 1. the cosine similarity between w_m and w_b, 2. the psychometric tests, such as the analogy tests and the consistency with human annotators, and 3. the downstream task results. For 1 and 2, we apply the same evaluations to the residual vectors $w_r = w_a - w_b + w_m$ to investigate the importance of the residuals that has been ruled out during the learning.

Apart from `GloVe` and `Word2Vec` , we extend the paradigm to contextualized vectors. In partcular, we evaluate the MLP on contextualized embeddings. Given a n-word sequence $w_1, w_2, w_3, ..., w_n$, with w_i to be the word in the center, we define the model input to be $b_1 \oplus b_2 \oplus ... \oplus g_i \oplus ... \oplus b_n$, where $b_n = \texttt{BERT}(w_n)$, $g_i = \texttt{GloVe}(w_i)$, and the output to be $b_i = \texttt{BERT}(w_i)$, i.e. given a `GloVe` vector surrounding by `BERT` vector contexts, approximate the produced `BERT` vector. To assess the quality of approximation, we analyze the cosine similarities and conduct a comparative study of MLP weights. Initially, we employ a standard model input configuration. Subsequently, we deploy an ablation study, in which the central GloVe vector is masked out.

3.4 Implementation

We utilize the `GoogleNews-vectors-negative300` version as the Word2Vec (`Word2Vec`) embedding. For `GloVe` , we employ the 300-dimensional version trained on the 6B training corpus. The tests are conducted using the overlapping vocabulary between `GloVe` and `Word2Vec` embeddings, comprising a total of 105,990 vectors. Context-dependent embeddings are obtained by extracting n-word sentences from the British National Corpus[3], where we select $n = 7$ resulting in 192,257 instances.

We conduct tests using both the Procrustes and the Orthogonal Procrustes. The primary distinction between them lies in the fact that the former transformation does not have to be orthogonal. The Support Vector Machine is evaluated using various kernels: `linear`, `polynomial`, `rbf`, and `sigmoid`. Specifically, the `polynomial` kernel is tested with degrees ranging from 2 to 6.

The Multi-Layer Perceptron utilized in our experiments comprises two hidden layers with a size of 10,000 and ReLU. It undergoes training for 100 epochs employing the Adam optimizer.

All psychometric tests are conducted based on the benchmark provided by Jastrzębski et al. (2017) [Jastrzebski et al., 2017]. The analogy test is performed using the `Google Analogy` dataset [4]. Human similarity estimation is based on three datasets: WS353 [Agirre et al., 2009], SIMLEX999 [Hill et al., 2014], and MEN [Bruni et al., 2014]. For the downstream tasks, we evaluate semantic composition using the Big BiRD dataset [Asaadi et al., 2019], and perform sentiment analysis on the Movie Review (MR) dataset [Pang and Lee, 2005], utilizing a Convolutional Neural Network (CNN) approach as described by [Kim, 2014].

It is noteworthy that all tests and evaluations are performed on the **seen** data instead of the test data. This choice is motivated by two factors: Firstly, unlike other machine learning tasks, word embeddings are finite. Secondly, the primary objective of this study is not to learn the representation pattern of each embedding but rather to investigate whether a clear connection can be established among different datasets.

4 Results

4.1 Procrustes

In this experiment, the `GloVe` embeddings serve as the target space, and the `Word2Vec` space is transformed to align with it. As outlined in Section 3.4, both Procrustes analysis and Orthogonal Procrustes are employed, and the average cosine similarity between the mapped vectors and the target vectors is computed. The Procrustes method yields a cosine similarity of 0.416, while the Orthogonal Procrustes method results in a slightly improved value of 0.460. This suggests that a linear transformation may not be entirely suitable for capturing the necessary information accurately.

[3] http://www.natcorp.ox.ac.uk/

[4] http://download.tensorflow.org/data/questions-words.txt

Kernel Selection	Seen	Unseen
Linear	0.563	0.452
Polynomial (2)	**0.970**	**0.819**
Polynomial (3)	0.953	0.556
Polynomial (4)	0.955	0.322
Polynomial (5)	0.962	0.209
Polynomial (6)	0.967	0.179
RBF	0.966	0.714
Sigmoid	0.395	0.425

Table 2: The SVM classification results for different kernel setups on seen and unseen data. The number in parentheses for the polynomial kernels indicates the degree. The polynomial kernel with degree 2 demonstrates the best performance in classifying both the seen and unseen data.

Approximation Direction	Cosine similarity	L1-distance	L2-distance
`Word2Vec` to `GloVe`	0.950	28.688	2.074
`GloVe` to `Word2Vec`	0.896	16.871	1.220
`GloVe` to `BERT`	0.906	133.511	6.048
masked `GloVe` to `BERT`	0.895	139.966	6.344

Table 3: Cosine similarity, L1 distance, and L2 distance between the approximation vectors and the target vectors in the specified approximation direction. Masked-`GloVe` represents the special comparative analysis mentioned in Section 3.3. All approximations achieve a cosine similarity of approximately 0.9 with the target embeddings, consistent with their close performance as presented in Table 4. The significant variance among the L1 and L2 distances is attributed to the differing dimensions of the vectors.

4.2 SVM

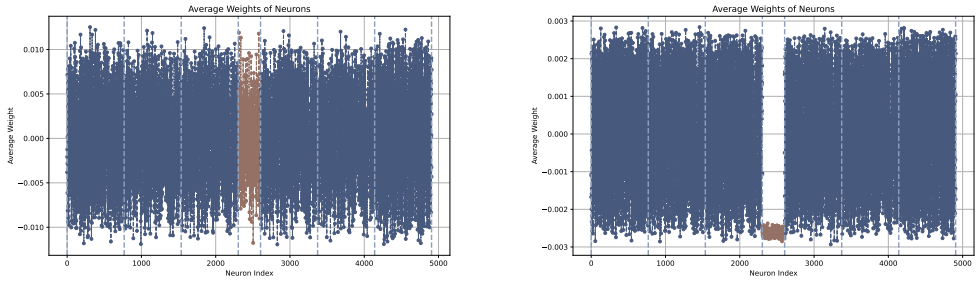
In this experiment, the `GloVe` and `Word2Vec` embeddings are concatenated together with a label indicating whether the embeddings originate from the same word. Table 2 presents the results of different kernel setups for classifying the word origins, both on seen and unseen data. Notably, the model utilizing the second-degree polynomial kernel demonstrates the best performance on both the seen and unseen data.

Overall, the majority of the models exhibit exceptional performance in classifying the vectors based on their origins. Specifically, the 2nd-degree polynomial kernel, the 6th-degree polynomial kernel, and the RBF kernel achieve the highest performance on the seen data. However, the 6th-degree polynomial kernel exhibits a considerable drop in performance on the unseen data, indicating potential overfitting. The other two kernels demonstrate relatively minor decreases in performance on unseen data. Nevertheless, the consistent ability to address and classify the differences between `GloVe` and `Word2Vec` embeddings confirms the feasibility of the task designed.

4.3 MLP

On Context-Independent Vectors The MLP is utilized to approximate a target embedding from a template embedding. The effectiveness of this approximation is assessed through various metrics including their distances and cosine similarities to the target vectors, results from psychometric tests, and performance in downstream tasks. The outcomes, as presented in Tables 3, 4, and 5, demonstrate that the MLP-transformed word embeddings are comparable to the target embeddings across all evaluated dimensions from vector distances to downstream task performance.

On BERT The MLP is employed to approximate the central BERT vector using a combination of a central `GloVe` embedding and surrounding BERT context vectors. Two experiments were conducted; in the second, the central GloVe embeddings were replaced with zero-masked vectors. As shown in Table 3, both configurations yielded high cosine similarity values, indicating a high quality of approximation. The in-depth analysis of the MLP weights in these two experiments was performed and shown in Figure 1. The glove-central approximation task, as illustrated by Figure 1a, shows a normal MLP neuron weight distribution, where no interpretable trends can be captured; however, in



(a) BERT surrounding + `GloVe` central

(b) BERT surrounding + masked central

Figure 1: Neuron weights of the first layer in contextualized BERT setups. Figure 1a represents the approximation experiment to a BERT vector from a centered `GloVe` vector with BERT context surroundings, and Figure 1b refers to an identical setup but with the centre `GloVe` masked by zeroes. When the `GloVe` vector is masked, the MLP disregards this vector yet maintains approximation quality solely through the use of surrounding context vectors, as demonstrated in Table 3.

Vector Type	MEN	WS353	SIMLEX999	Analogy
`Word2Vec` -like embeddings				
`Word2Vec`	0.764	0.667	0.441	0.165
mapped-`Word2Vec`	0.757	0.677	0.434	0.155
`Word2Vec` -residual	0.731	0.619	0.418	0.148
`GloVe` -like embeddings				
`GloVe`	0.698	0.543	0.371	0.955
mapped-`GloVe`	0.696	0.540	0.368	0.939
`GloVe` -residual	0.696	0.543	0.366	0.958

Table 4: Psychometric test results of `GloVe` , `Word2Vec` , the approximated `GloVe` (mapped-`GloVe`), the residual vector from the approximated `GloVe` (`Word2Vec` -residual), the approximated `Word2Vec` (mapped-`Word2Vec`), and the residual vector from the approximated `Word2Vec` (`GloVe` -residual), evaluated on MEN [Bruni et al., 2014], WS353 [Agirre et al., 2009], SIMLEX999 [Hill et al., 2014], and the Google Analogy dataset. Through the use of an MLP, vectors can be closely approximated to the target embeddings, while the residual retains the performance of the original embeddings.

the zero-masked setup, the MLP neurons turns abnormal. As recorded by Figure 1b, the latter setup nearly disregards the central vector and constructs the approximation solely based on the surrounding BERT vectors.

4.4 Discussion

Embeddings are not Linearly-Separable, but not overly Complex The experimental findings detailed in Section 4.1 indicate that word embeddings are not linearly separable. Procrustes analysis fails to achieve satisfactory classification results, suggesting the inadequacy of linear transformations for distinguishing between embedding spaces. However, it is noteworthy that the classification task is not excessively complex. The SVM experiments presented in Table 2 demonstrate that the differences between embeddings are discernible, and the relatively low degree of the kernel utilized in the SVM classification further supports the notion that the classification task does not necessitate a highly complex solution.

Current Word Vectors Exhibit Homogeneity The findings suggest that current word vectors exhibit homogeneity, characterized by the absence of high-dimensional differences that impede classification tasks. This observation aligns with the results discussed earlier. Moreover, our experiments with MLPs in Section 4.3 demonstrate that one embedding can be accurately approximated from another, as evidenced by the cosine similarity scores exceeding 0.9 (see Table 3). This strong alignment between embeddings, coupled with the performance of MLP-mapped vectors on other psychometric tests and downstream tasks (see Table 4, 5), suggests that the mapping process does capture meaningful information inherent in the embeddings, but not an accurate but meaningless ap-

Tasks	GloVe	mapped-GloVe	Word2Vec	mapped-Word2Vec
Big-BiRD methods				
head only	0.334	0.332	0.347	0.337
modifier only	0.423	0.423	0.430	0.428
addition	0.533	0.533	0.590	0.593
multiplication	0.234	0.238	0.364	0.363
tensor product	0.377	0.378	0.374	0.382
dilation	0.492	0.490	0.493	0.486
Sentiment Analysis				
MR	0.793	0.778	0.811	0.796

Table 5: Results of word embedding tests on downstream tasks. The first section details performance on the Big BiRD dataset [Asaadi et al., 2019], which assesses bigram relatedness through semantic composition, with results reported via Pearson correlation to human annotations. The second section presents outcomes of sentiment analysis on the movie review dataset [Pang and Lee, 2005] for each vector type. Notably, the approximated vectors (mapped-GloVe and mapped-Word2Vec) demonstrate a very close relationship with their respective target vectors (GloVe and Word2Vec) in these tasks.

proximation. Consequently, we can infer that the information contained in current word embeddings can be effectively conveyed to other embeddings by an MLP, or their differences are accountable and classifiable by a specifically-designed SVM. Despite representing different training objectives, these embeddings do not exhibit significant, classifiable disparities among each other as expected.

Understanding BERT's Functionality BERT has been heralded as a transformative development in traditional NLP frameworks due to its exceptional performance across a range of tasks at its inception. Its success is often attributed to the attention mechanism and its training objectives, which are focused on understanding context. However, our experiments suggest that BERT's behavior is not markedly different from that of other embedding models. By equipping GloVe vectors with contextual data, we found that a double-layer Multilayer Perceptron (MLP) can approximate GloVe templates to BERT's context-sensitive embeddings with a high degree of accuracy. This indicates that when provided with identical contextual information, GloVe embeddings do not significantly deviate from BERT embeddings to an extent that cannot be reconciled by an MLP. Consequently, BERT essentially offers a similar representation to that of GloVe and Word2Vec , with the primary difference being BERT's incorporation of context into its representations, as opposed to the static, singular word representations provided by the latter models.

Furthermore, Tables 1 and 3, along with Figure 1b, present additional complexities in our comprehension of BERT's functionality on contextualization. Table 1 indicates that BERT-generated vector representations exhibit minimal variation within specific context windows, showing a strong dependency on those contexts. Tables 3 and Figure 1b not only reinforce this context dependency but also reveal that the contextual information is encoded so robustly that the central word can be omitted without significantly affecting the semantic meaning reconstruction. The mechanisms by which BERT addresses long-term dependency challenges remain unclear, as does the nature of information distribution within these context windows that render BERT embeddings more deterministic. Given the evidence at hand, BERT appears to function similarly to an n-gram-based system that redistributes and equalizes information of initial embeddings across a given context.

5 Conclusion

In conclusion, our investigation delved into the homogeneity of word embeddings, encompassing GloVe , Word2Vec , and BERT . Through a series of experiments, we illustrated that while word embeddings aren't entirely linearly separable, they exhibit homogeneity and can be effectively approximated from one another using methods such as a 2nd-degree polynomial SVM and a 2-layer MLP. These approaches demonstrated that word embeddings can be closely approximated, with high cosine similarity scores and close test results indicating strong alignment between embeddings.

Furthermore, our exploration of MLP application on contextualized GloVe to BERT highlighted its similarity to traditional embeddings when contexts aren't considered. Our findings prompt a reevaluation of the essence of word embeddings, as popular embeddings at this stage appear

homogeneous, and contextualized ones like BERT. This suggests avenues for further inquiry into the nature and utility of word embeddings in natural language processing research.

References

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013a.

Dominik Schlechtweg, Anna Hätty, Marco del Tredici, and Sabine Schulte im Walde. A wind of change: Detecting and evaluating lexical semantic change across times and domains, 2019.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1), 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139–147, 2019.

Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112:340–349, 2017.

Lingfei Wu, Ian EH Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J Witbrock. Word mover's embedding: From word2vec to document embedding. *arXiv preprint arXiv:1811.01713*, 2018.

Tianze Shi and Zhiyuan Liu. Linking glove with word2vec. *arXiv preprint arXiv:1411.5595*, 2014.

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. doi: 10.1162/tacl_a_00134. URL `https://aclanthology.org/Q15-1016`.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. Learning to remove: Towards isotropic pre-trained bert embedding. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, pages 448–459. Springer, 2021.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change, 2018.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1:43–52, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013b.

Peter Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31 (1):1–10, 1966. URL `https://EconPapers.repec.org/RePEc:spr:psycho:v:31:y:1966:i:1:p:1-10`.

Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks, 2017.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In Mari Ostendorf, Michael Collins, Shri Narayanan, Douglas W. Oard, and Lucy Vanderwende, editors, *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-1003.

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation, 2014.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, jan 2014. ISSN 1076-9757.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1050. URL https://aclanthology.org/N19-1050.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124, 2005.

Yoon Kim. Convolutional neural networks for sentence classification. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://aclanthology.org/D14-1181.