Information Processing and Management 60 (2023) 103272

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Examining the effect of whitening on static and contextualized word embeddings

Shota Sasaki [a,b,*], Benjamin Heinzerling [a,b], Jun Suzuki [b,a], Kentaro Inui [b,a]

[a] RIKEN, Sendai, 980-8579, Miyagi, Japan
[b] Tohoku University, Sendai, 980-8579, Miyagi, Japan

ARTICLE INFO

Keywords:
Static word embeddings
Contextualized word embeddings
Whitening
Frequency bias

ABSTRACT

Static word embeddings (SWE) and contextualized word embeddings (CWE) are the foundation of modern natural language processing. However, these embeddings suffer from spatial bias in the form of anisotropy, which has been demonstrated to reduce their performance. A method to alleviate the anisotropy is the "whitening" transformation. Whitening is a standard method in signal processing and other areas, however, its effect on SWE and CWE is not well understood. In this study, we conduct an experiment to elucidate the effect of whitening on SWE and CWE. The results indicate that whitening predominantly removes the word frequency bias in SWE, and biases other than the word frequency bias in CWE.

1. Introduction

Static word embeddings (SWE) (Mikolov, Sutskever, et al., 2013; Pennington et al., 2014) and contextualized word embeddings (CWE) (Devlin et al., 2019; Peters et al., 2018; Raffel et al., 2020, i.a.) are the foundation of modern natural language processing systems. However, while the aim of creating such embeddings is to provide accurate representations of word, phrase, and sentence meaning, they also reflect and sometimes amplify biases inherent in the training data, such as gender bias (Zhao et al., 2019), social bias (Kaneko & Bollegala, 2022), and word frequency bias (Gong et al., 2018). For SWE, prior research has demonstrated that the embedding space exhibits a spatial frequency bias; namely, frequent words tend to concentrate along a particular direction (Mu & Viswanath, 2018). Generally, this *anisotropy*, i.e., the non-uniform angular distribution of word vectors, is undesirable because it leads to inefficient use of the embedding space. Furthermore, frequency-based anisotropy causes frequent words to be represented by similar vectors simply by virtue of their high frequency, although their meaning may not be similar.

Aiming to reduce the negative impact of anisotropy, several *isotropization* methods have been proposed. These methods make embeddings more *isotropic*, i.e., transform embedding vectors so that they have a more uniform angular distribution. Isotropization methods can be divided into supervised debiasing methods and unsupervised post-processing methods. The primary goal of supervised debiasing methods is to remove biases with respect to specific categories, such as gender, nationality, and word frequency. If the bias manifests itself as an uneven distribution of word embeddings, then debiasing results in a more isotropic embedding space. A representative example of such an approach is the adversarial removal of protected social variables proposed by Zhang et al. (2018). In contrast to supervised debiasing methods, unsupervised post-processing methods aim to improve word embeddings without relying on word meaning or associated categories.

The focus of this study is the arguably simplest and most common unsupervised isotropization method, namely *whitening*. Informally, whitening is a linear operation that transforms a set of spatially correlated (and therefore anisotropic) vectors into a set

* Corresponding author at: RIKEN, Sendai, 980-8579, Miyagi, Japan.
E-mail addresses: shota.sasaki.yv@riken.jp (S. Sasaki), benjamin.heinzerling@riken.jp (B. Heinzerling), jun.suzuki@tohoku.ac.jp (J. Suzuki), kentaro.inui@tohoku.ac.jp (K. Inui).

https://doi.org/10.1016/j.ipm.2023.103272
Received 20 May 2022; Received in revised form 30 November 2022; Accepted 8 January 2023
Available online 24 January 2023

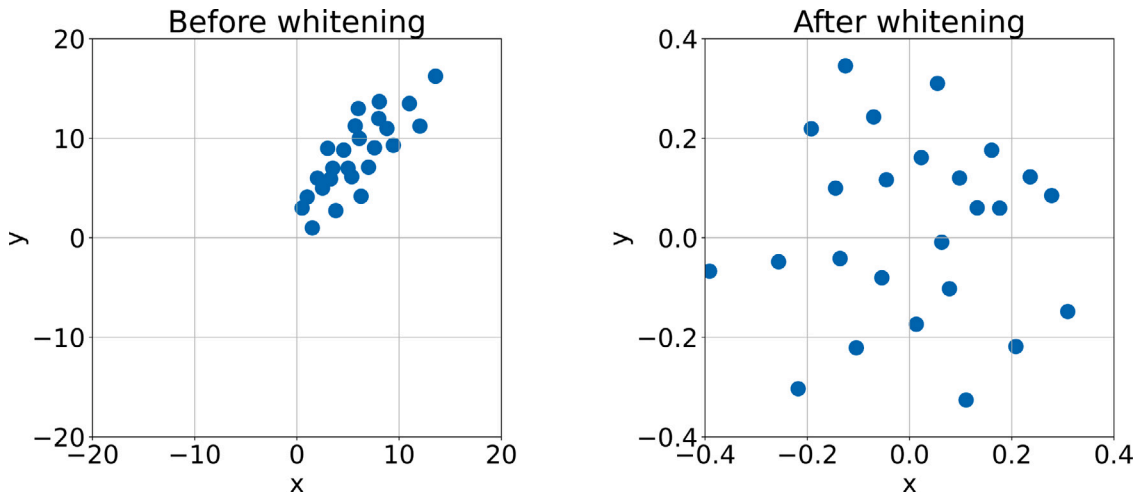


Fig. 1. Examples of plots before and after applying whitening to a set of two-dimensional vectors. Before applying whitening, x and y are correlated, whereas after applying whitening, they are no longer correlated.

of uncorrelated (isotropic) vectors (see Section 2.2 for a formal definition). Although whitening is a standard data transformation technique, applications to NLP and CWE, in particular, have appeared only recently (Huang et al., 2021; Su et al., 2021). These applications have demonstrated that whitening performs better than other isotropization methods for CWE. However, a major disadvantage of whitening and other unsupervised post-processing methods is that their impact on various forms of bias and other semantic properties of embeddings is not yet understood, although understanding bias in SWE and CWE is a prerequisite for their ethical use in real-world applications. In this paper, we present an initial analysis of the semantic impact of unsupervised post-processing. In particular, we analyze changes in the frequency bias when applying the whitening transformation to SWE and CWE.

Our preliminary analysis indicates that the effect of whitening partially includes the effect of frequency debiasing. Our research question is thus whether the effect of whitening consists of frequency debiasing only. To increase the granularity of the effect of whitening, we employ a method whose effect is frequency debiasing only; specifically, we propose a reconstruction-based frequency debiasing (RFD), which focuses only on removing frequency bias in embeddings. We then compare the behavior of whitening with that of RFD. Our experimental results indicate that whitening removes word frequency bias in SWE as well as biases other than word frequency bias in CWE.

2. Background

2.1. Anisotropy in static and contextualized word embeddings

Mu and Viswanath (2018) reported an anisotropy problem that leads to reduce expressiveness of SWE. Ethayarajh (2019) reported that anisotropy also existed in CWE. Accordingly, there has been much discussion about the causes of anisotropy in embeddings. Mu and Viswanath (2018) reported that word frequency information is embedded in the first and second principal components of SWE, and is the cause of anisotropy in SWE. In CWE, frequency bias is the most common issue. Li et al. (2020) empirically demonstrated that there is a frequency bias in the vectors of the word embedding layer in BERT (Devlin et al., 2019). Specifically, they demonstrated that vectors of frequent words are embedded closer to the origin, while infrequent words are embedded farther from the origin. Moreover, they showed that vectors of high-frequency words are densely embedded, while vectors of low-frequency words sparsely dispersed. Liang et al. (2021) also reported that there is a correlation between the logarithm of word count and the norm/average cosine similarity of word vectors. In addition to frequency bias, outlier dimensions in CWE have also recently received attention. Luo et al. (2021) and Kovaleva et al. (2021) identified dimensions in the embeddings of BERT and RoBERTa (Liu et al., 2019) that were significantly higher than other dimensions, suggesting that they were the cause of anisotropy in the embeddings.

2.2. Isotropization via the whitening transformation

Whitening is a linear transformation that transforms a set of vectors into a new set where the covariance matrix is the identity matrix. The fact that the covariance matrix is an identity matrix signifies that the transformation makes each dimension uncorrelated (*uncorrelation*) and sets the variance to 1 (*variance flattening*). Through the transformation, the resulting whitened embeddings become more isotropic (Fig. 1). Let $X \in \mathbb{R}^{N\times d}$ be a set of d-dimensional N vectors such as word embeddings. The transformed embeddings $X' \in \mathbb{R}^{N\times d}$ are defined as follows:

$$X' = (X - m)U\sqrt{S^{-1}}, \tag{1}$$

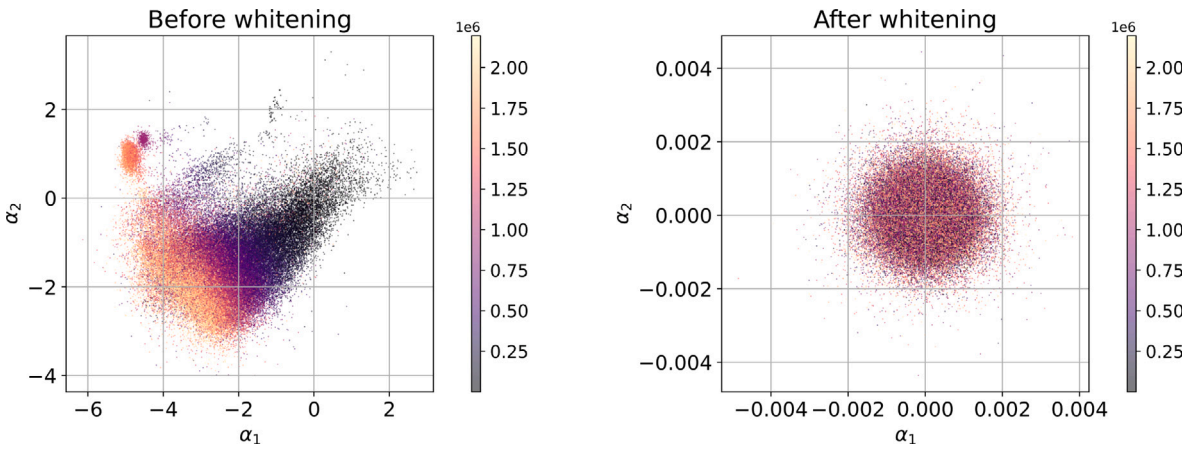


Fig. 2. Plots of the first and second principal components (α_1 and α_2) before and after applying whitening to GloVe embeddings. Colors correspond to word frequency ranks. Black represents frequent words, while yellow represents infrequent words. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $m \in \mathbb{R}^d$ is the mean vector of the vectors in X, and U and S are given by singular value decomposition (SVD) of the covariance matrix Σ of X. $\Sigma \in \mathbb{R}^{d \times d}$ is calculated as follows:

$$\Sigma = (X - m)^T (X - m). \tag{2}$$

U and S satisfy the following equation by the definition of SVD:

$$\Sigma = U S U^T, \tag{3}$$

where S is a diagonal matrix having the eigenvalues of Σ and U is the corresponding orthogonal matrix of eigenvectors.

Whitening is commonly used in machine learning to reduce bias in the training data, and it has been applied to a set of feature vectors as a feature preprocessing method (Coates et al., 2011; Ranzato et al., 2010). It has been reported that reducing bias helps deep learning models learn high-quality representation and speeds up model convergence. Since whitening is a general-purpose algorithm that can be applied to a set of vectors, it can also be applied to sentence vectors obtained by CWE. Huang et al. (2021) applied whitening to CWE to address the anisotropy problem and demonstrated that it improved the performance of the CWE. Whitening is mathematically well defined; however, analysis has not yet been performed to clarify what information in SWE and CWE is processed and how the information is transformed by whitening. In this study, we aim to clarify the mechanism of whitening (i.e., uncorrelation and variance flatting) in SWE and CWE.

2.3. Other isotropization methods

In addition to whitening, several other methods to address anisotropy have been proposed. One class of isotropization methods removes the principal components of embeddings. (Mu & Viswanath, 2018) suggested that there is a word frequency bias in SWE, and reported that the bias negatively affects task performance. Specifically, they observed a frequency bias in the first and second principal components of GloVe (Pennington et al., 2014) and Word2Vec (Mikolov, Sutskever, et al., 2013) embeddings, and proposed a method to remove the top-D principal components, denoted the RPC method. They found that the RPC method improves the performance of tasks and reduces the anisotropy of SWEs. Rajaee and Pilehvar (2021) proposed a cluster-based version of the RPC method, while (Liang et al., 2021) proposed a weighted version.

Several mathematically-motivated methods have also been proposed to handle anisotropic embeddings. Li et al. (2020) proposed `BERT-flow`, which learns a flow function that projects embeddings obtained from BERT to a standard Gaussian latent space. Their theoretical motivation was that embeddings with a standard Gaussian distribution are a sufficient condition for isotropy and an efficiently embedded space, which they described as lacking holes.

3. Preliminaries

Following the work of Mu and Viswanath (2018), we performed an analysis to explore the effect of whitening in preliminary experiments.Figs. 2 and 3 present plots of the first and second principal components of the embeddings before and after applying whitening to GloVe and BERT embeddings, respectively. Before applying whitening, we observed a frequency bias, i.e., a correlation between the principal components of word embeddings and their word frequencies in both GloVe and BERT embeddings. In particular, GloVe embeddings had a strong frequency bias, which is consistent with the reports from Mu and Viswanath (2018). However, after applying whitening, we found that there was no bias in the first and second principal components of the embeddings, which indicates that whitening has the effect of frequency debiasing.

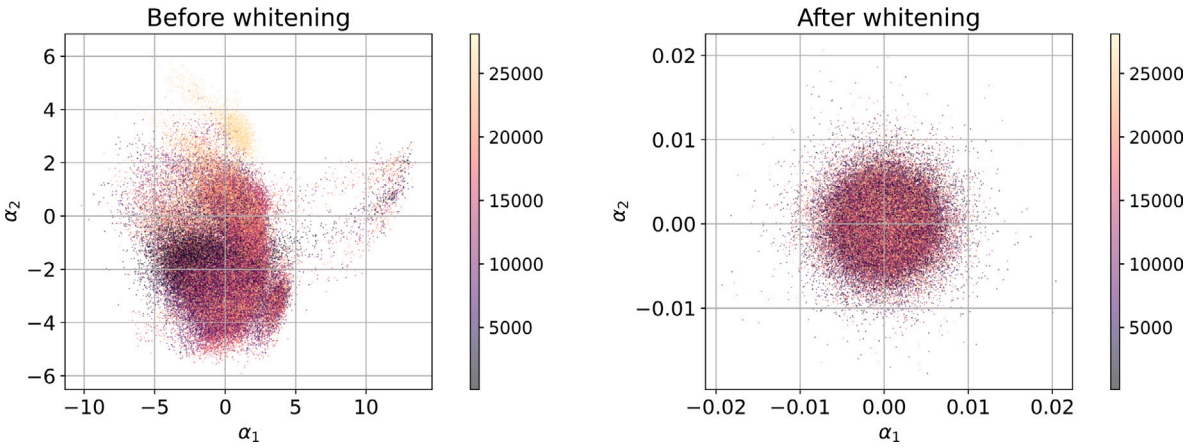


Fig. 3. Plots of the first and second principal components (α_1 and α_2) before and after applying whitening to BERT embeddings. Colors correspond to word frequency ranks. Black represents frequent words, while yellow represents infrequent words. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Based on this analysis, we aim to clarify the effect of whitening. Our research question is whether whitening is equivalent to frequency debiasing or whether it has effects other than frequency debiasing. To address this question, we conduct an experiment in which we apply both whitening and a frequency debiasing method, which is introduced in Section 4, to the model at the same time. When both are applied to the model, if the effects of whitening and frequency debiasing are independent, then respective gains can be expected, while if there is an overlap between the two effects, then the improvements can be limited.

4. Frequency debiasing method

In this section, we introduce a frequency debiasing method that focuses only on the effect of frequency debiasing without affecting the original quality of the embeddings. Gong et al. (2018) proposed a method of frequency debiasing for word embeddings using adversarial training. Given a certain task, such as text classification, their method optimizes each vector in the word embeddings layer with the debiasing loss simultaneously as learning. With these settings, the resulting word embeddings are task-specific representations. We aim to obtain general representations rather than task-specific representations; thus, we propose a reconstruction-based frequency debiasing (RFD) inspired by Gong et al. (2018). Like (Gong et al., 2018), we assume that word embeddings are trained to fool a discriminator attempting to identify whether the words are rare or popular. Instead of a task-specific loss, we introduce a reconstruction loss that aims to preserve the pretrained representations such as embeddings from GloVe or BERT.

First, we present a reconstruction loss for SWE such as GloVe. Let $\mathcal{W}$ be a set of words in a vocabulary, $\boldsymbol{e}(w)$ be the pretrained fixed embeddings of word w and $\boldsymbol{v}(w;\theta^{\text{emb}})$ be the learnable embeddings of word w, where $\theta^{\text{emb}} \in \mathbb{R}^{d\times V}$ is the parameter matrix of the word embeddings. Here d is the number of dimensions of the embeddings and V $(= |\mathcal{W}|)$ is the vocabulary size. The reconstruction loss for SWE is defined as follows:

$$L_{R_{\text{swe}}}(\mathcal{W};\theta^{\text{emb}}) = \sum_{w\in\mathcal{W}} \left\| \boldsymbol{e}(w) - \boldsymbol{v}(w;\theta^{\text{emb}}) \right\|_2^2. \tag{4}$$

Regarding the discriminator part for SWE, we follow the settings used by Gong et al. (2018)'s settings. First, we divide the vocabulary $\mathcal{W}$ into two parts: $\mathcal{W}_{\text{pop}}$ and $\mathcal{W}_{\text{rare}}$. Words in $\mathcal{W}_{\text{pop}}$ are the top-t% frequent words, while $\mathcal{W}_{\text{rare}} = \mathcal{W} \setminus \mathcal{W}_{\text{pop}}$. Let f_{θ^D} represent a discriminator with parameters θ^D that takes word embeddings as input and returns a probability score indicating whether the word is rare or not. The loss of the discriminator for SWE is defined as follows:

$$L_{D_{\text{swe}}}(\mathcal{W};\theta^D,\theta^{\text{emb}}) = \frac{1}{|\mathcal{W}_{\text{pop}}|}\sum_{w\in\mathcal{W}_{\text{pop}}} \log f_{\theta^D}(\boldsymbol{v}(w;\theta^{\text{emb}})) + \frac{1}{|\mathcal{W}_{\text{rare}}|}\sum_{w\in\mathcal{W}_{\text{rare}}} \log(1 - f_{\theta^D}(\boldsymbol{v}(w;\theta^{\text{emb}}))). \tag{5}$$

Lastly, we optimize θ^{emb} and θ^D using the adversarial training procedure with the min–max objective as follows:

$$\arg\min_{\theta^{\text{emb}}} \arg\max_{\theta^D} L_{R_{\text{swe}}}(\mathcal{W};\theta^{\text{emb}}) - \lambda L_{D_{\text{swe}}}(\mathcal{W};\theta^D,\theta^{\text{emb}}), \tag{6}$$

where λ is a hyperparameter used as a weight coefficient. Following Gong et al. (2018), we alternate between optimizing the argmin objective for θ^{emb} and optimizing the argmax objective for θ^D.

Table 1
The model parameters of the embeddings we used in the experiments.

	V	d
GloVe840B	2,196,016	300
GloVe6B	40,000	300
Gnews	3,000,000	300
BERT-base	28,996	768
DistilBERT-base	28,996	768
RoBERTa-base	50,265	768

For CWE, we prepare a training corpus C and optimize the embeddings obtained by encoding sentences from C. Let s represent a sentence from corpus C, $\mathcal{W}_s$ represent a set of words[1] in s, and $\mathcal{L}$ be a target set of layers in a CWE model such as BERT. Let $\boldsymbol{e}^l(w,s)$ be the pretrained lth layer embeddings of word w obtained by encoding sentence s. Embeddings $\boldsymbol{v}^l(w,s;\theta^{\text{emb}})$ is similar to $\boldsymbol{e}^l(w,s)$ but embeddings from a new CWE model with learnable parameters θ^{emb}. The reconstruction loss and discriminator loss for CWE are defined as follows:

$$L_{R_{\text{cwe}}}(C;\theta^{\text{emb}}) = \sum_{s\in C}\sum_{w\in\mathcal{W}_s}\sum_{l\in\mathcal{L}}\left\|\boldsymbol{e}^l(w,s) - \boldsymbol{v}^l(w,s;\theta^{\text{emb}})\right\|_2^2, \tag{7}$$

$$L_{D_{\text{cwe}}}(C;\theta^D,\theta^{\text{emb}}) = \sum_{s\in C}\sum_{w\in\mathcal{W}_s}\sum_{l\in\mathcal{L}} L'_{D_{\text{cwe}}}(w,l;\theta^D,\theta^{\text{emb}}), \tag{8}$$

$$L'_{D_{\text{cwe}}}(w,l;\theta^D,\theta^{\text{emb}}) = \frac{1}{|\mathcal{W}_{s,\text{pop}}|}\sum_{w\in\mathcal{W}_{s,\text{pop}}}\log f_{\theta^D}(\boldsymbol{v}^l(w,s;\theta^{\text{emb}})) + \frac{1}{|\mathcal{W}_{s,\text{rare}}|}\sum_{w\in\mathcal{W}_{s,\text{rare}}}\log(1 - f_{\theta^D}(\boldsymbol{v}^l(w,s;\theta^{\text{emb}}))), \tag{9}$$

where $\mathcal{W}_{s,\text{pop}} = \mathcal{W}_s \cap \mathcal{W}_{\text{pop}}$ and $\mathcal{W}_{s,\text{rare}} = \mathcal{W}_s \setminus \mathcal{W}_{s,\text{pop}}$. The objective for CWE is defined as follows:

$$\arg\min_{\theta^{\text{emb}}}\arg\max_{\theta^D} L_{R_{\text{cwe}}}(C;\theta^{\text{emb}}) - \lambda L_{D_{\text{cwe}}}(C;\theta^D,\theta^{\text{emb}}). \tag{10}$$

5. Experiments

We conduct experiments with several models to investigate the effect of whitening. When applied simultaneously, if the performance of whitening and frequency debiasing is equivalent to that of a single applied model, this indicates that whitening has the same effect as frequency debiasing.

5.1. Task

Dataset. To evaluate the quality of the embeddings, we adopt a semantic textual similarity (STS) task. Specifically, we use the STS Benchmark dataset (Cer et al., 2017) created to provide a standard setup of STS for fair comparison.[2] The dataset consists of sentence pairs with manually annotated scores as sentence similarities. The scores range from 0 to 5.

Evaluation. Following previous studies (Huang et al., 2021; Reimers & Gurevych, 2019), we compute the Spearman rank correlation between the annotated ground truth scores and the similarities predicted by models. We calculate the cosine similarities of the sentence vectors as the similarities between sentences.

5.2. Settings

We use two types of GloVe embeddings[3] (`GloVe840B` and `GloVe6B`) and Google news embeddings[4] (`GNews`) as the SWE. `GloVe840B` is trained on the Common Crawl dataset containing 840 billion tokens, while `GloVe6B` is trained on the Wikipedia and Gigaward dataset containing 6 billion tokens. `GNews` is trained by using a CBOW algorithm (Mikolov, Chen, et al., 2013) and the training corpus is Google news dataset containing 100 billion tokens. We use a BERT-base (Devlin et al., 2019), DistilBERT-base (Sanh et al., 2019) and RoBERTa-base (Liu et al., 2019) model from Huggingface Transformer Library (Wolf et al., 2020) as the CWE. The model parameters are summarized in Table 1.

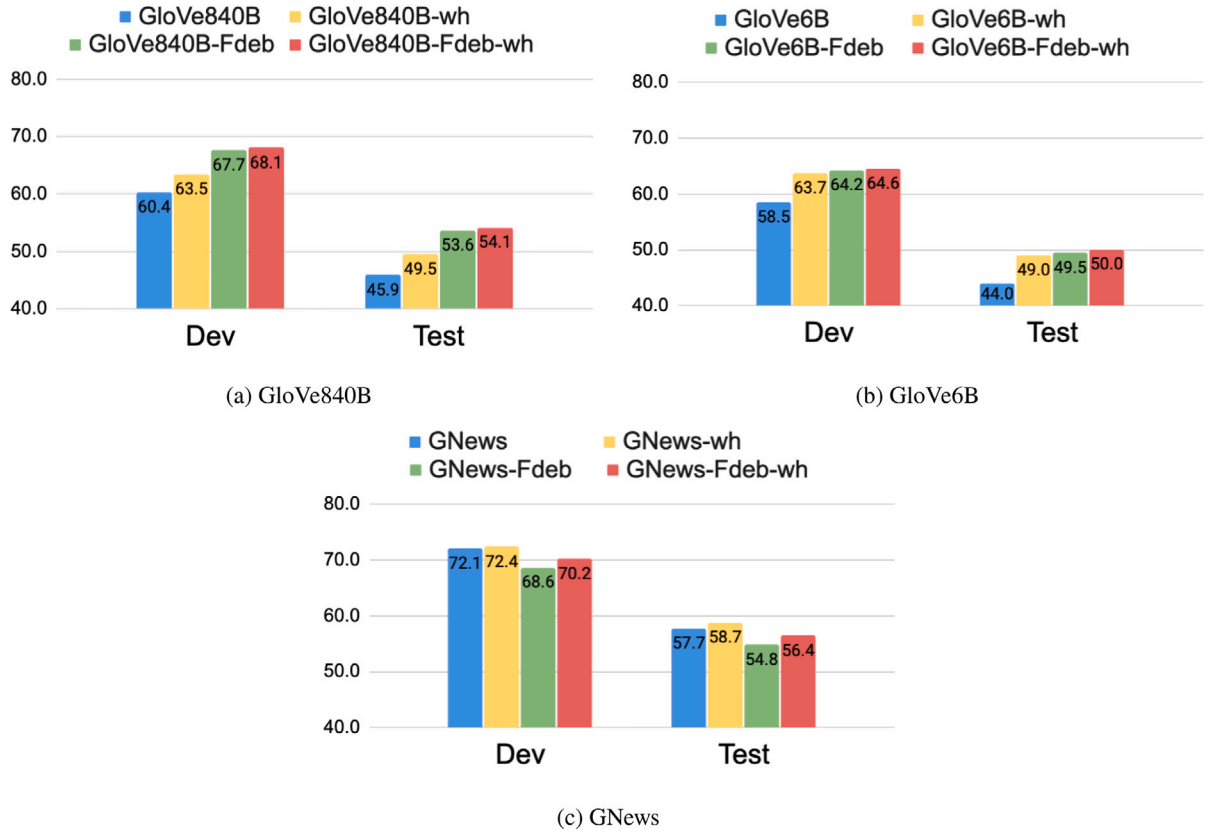
We compare four distinct settings of SWE and CWE:

[1] Note that we use "word" to refer to "token" for the sake of clarity even though words are separated into tokens in CWE models; for example, "interferometer" is separated into "inter", "##fer" and "##ometer".

[2] https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

[3] https://nlp.stanford.edu/projects/glove/

[4] https://code.google.com/archive/p/word2vec/



(a) GloVe840B (b) GloVe6B

(c) GNews

Fig. 4. The results of the SWE experiments on STS benchmark development (`Dev`) and test (`Test`) set. The performance is evaluated by the Spearman rank correlation between the annotated ground truth scores and the similarities predicted by the models.

1. Vanilla models: models that use raw embeddings without post-processing.
2. Whitening models (`wh`): models with whitening applied.
3. Frequency debiasing models (`Fdeb`): models with frequency debiasing applied.
4. Frequency debiasing and whitening models (`Fdeb-wh`): models with frequency debiasing applied followed by whitening.

For frequency debiasing, we use RFD, which is introduced in Section 4. For the SWE models, the sentence vectors are calculated by averaging the vectors of all words. For the CWE models, following Huang et al. (2021), we average the vectors of words at the first and last hidden layers. There are various options as the target layer list $\mathcal{L}$ in Eqs. (7) and (8). For example, we can set the target layers to be all layers. In this study, we set $\mathcal{L} = [1, 12]$ for BERT and RoBERTa and $\mathcal{L} = [1, 6]$ for DistilBERT, which are the simplest choices among the target layer lists that include the first and last layer. We use $t = 10$ as the threshold for a set of frequent words. We simply fix the batch size to 128 because we confirmed the performance is almost same when changing the batch size in preliminary experiments. We search λ in Eqs. (6) and (10) from $[0.02, 0.1]$,[5] the learning rate from $[1e-3, 5e-3, 1e-2]$, and the number of epochs from $[1, 3, 5, 10, 20, 30, 40, 50]$ in the development set. We use the sentences from the STS datasets as the training corpus $\mathcal{C}$ for RFD for the CWE models.

5.3. Results

Figs. 4 and 5 show the performance of the models on the STS Benchmarks development and test sets. The results are as follows:

(i) In both the SWE and CWE experiments, whitening improved the performance. In particular, the performance of the CWE models increased significantly. For example, the performance of `RoBERTa` improved by 10.7 points on the test set.
(ii) We observed performance improvement by the effect of frequency debiasing in all the models except for `GNews`. Notably, the performance of `GloVe840B-Fdeb` was 7.7 points higher than that of `GloVe840B` on the test set.

[5] We search λ from $[0.02, 0.1]$ because the default and recommended setting in Gong's implementation on the Github (https://github.com/ChengyueGongR/Frequency-Agnostic) is $\lambda = 0.02$, while (Gong et al., 2018) used $\lambda = 0.1$ in their experiments.

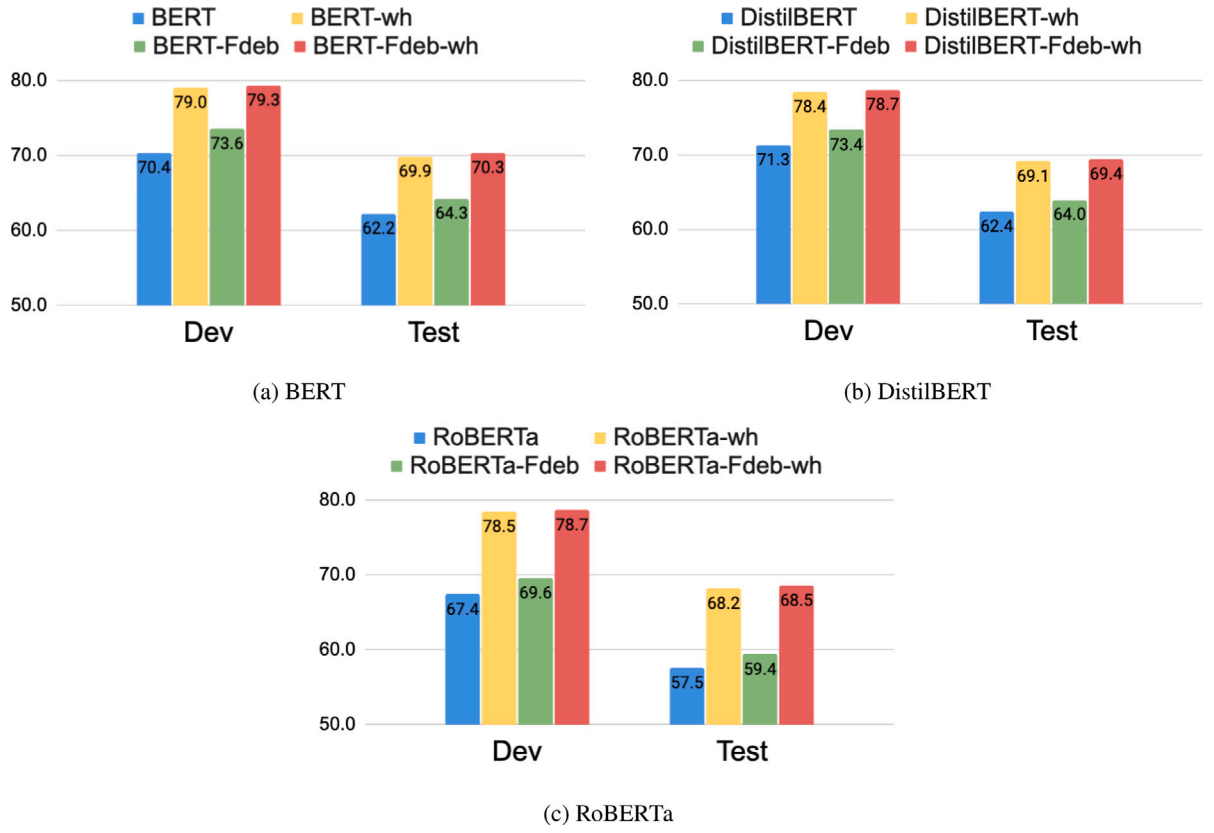


Fig. 5. The results of the CWE experiments on STS benchmark development (`Dev`) and test (`Test`) set. The performance is evaluated by the Spearman rank correlation between the annotated ground truth scores and the similarities predicted by the models.

(iii) The `Fdeb-wh` models for SWE such as `GloVe-Fdeb-wh` had no significant improvement over the `Fdeb` models such as `GloVe-Fdeb`.
(iv) Unlike the SWE results, the `Fdeb-wh` models for CWE had significant improvement over the `Fdeb` models. For example, the performance of `RoBERTa-Fdeb-wh` was 9.1 points higher than that of `RoBERTa-Fdeb` on the test set.

The observation of result (i) is consistent with the results from a previous study (Huang et al., 2021). Regarding result (ii), the gain from frequency debiasing for GloVe was higher than that for the CWE models. We assumed that this was because GloVe had a stronger frequency bias, as suggested by the plot of the PCA coefficients in Section 3. To verify this assumption, we present an analysis in Section 6. Of all the models, only `GNews` had no improvement by applying frequency debiasing. We suspect the training algorithms of SWE embeddings affect the performance of frequency debiasing.

Regarding results (iii) and (iv), if the effects of whitening and frequency debiasing are independent, then respective gains can be expected when both are applied to the model. However, we observed that no significant difference between `Fdeb` and `Fdeb-wh` for SWE (result (iii)). This observation reveals that the effect of whitening on SWE is almost the same as that of frequency debiasing, or that there is a large overlap between the two. In contrast, result (iv) indicates that on CWE, whitening has effects other than frequency debiasing, such as the correction of defects inherent in CWE that do not exist in SWE. We speculate that one of these effects may be the correction of the outlier problems reported by Kovaleva et al. (2021), Luo et al. (2021).

6. Analysis

In Section 5, we assume that whitening has effects on CWE other than frequency debiasing only. In this section, we describe a more in-depth analysis that we conduct to support this assumption. Specifically, we investigate how the quality of embeddings in each model varies with the frequency of words in the sentence. Given a sentence pair s_1 and s_2 in the STS dataset, we first combine these two sentences to create a paired sentence p. Then we compute its average word frequency ranks defined as follows:

$$R_{\text{sent}}(p) = \frac{1}{N} \sum_{w} r(w), \tag{11}$$

where N is the number of words in p, w represents a word in p, and $r(\cdot)$ is a function that takes a word w and returns the frequency rank of w in a predefined vocabulary. Based on R_{sent}, we sort the sentence pairs in the STS evaluation data, and exclude sentence

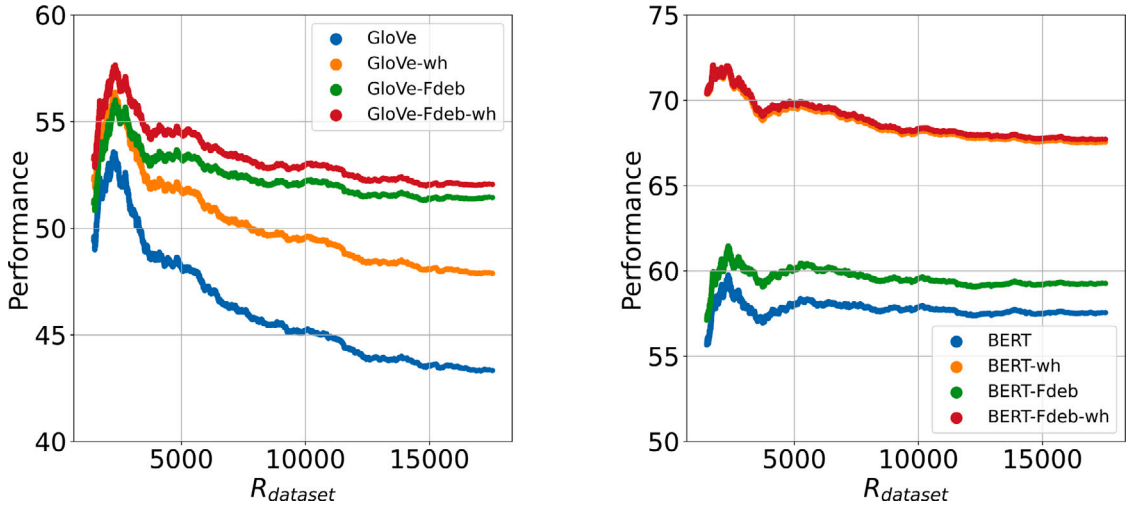


Fig. 6. Performance/R_{dataset} curves when using models based on GloVe (left) and BERT (right). The x-axis represents R_{dataset}, while the y-axis represents the STS performance evaluated by the Spearman rank correlation between the annotated ground truth scores and the similarities predicted by the models.

pairs with the highest R_{sent} one by one, and then we evaluate the models on each subset of the dataset. By this procedure, we reduce an average R_{sent} across the dataset, which is defined as follows:

$$R_{\text{dataset}} = \frac{1}{P}\sum_{p} R_{\text{sent}}(p), \tag{12}$$

where P is the number of pairs in the dataset. A Higher R_{dataset} signifies that the dataset contains more rare words, while a lower R_{dataset} signifies that the dataset contains more frequent words.

Fig. 6 presents graphs of the STS performance when varying R_{dataset} for the STS-14 dataset. We observed that for `GloVe840B`, the larger the $R_{dataset}$, the lower the performance. This indicates that the embeddings of rare words in GloVe had a negative impact on STS performance. In contrast, we did not observe this tendency for BERT. This suggests that the embeddings of rare words in BERT had a small effect on the STS performance. This supports the fact that the representation learning of BERT, which separates infrequent words into subwords, was effective. As mentioned in Section 5, we observed significant improvements when whitening was applied to CWE, whereas the impact of frequency debiasing was limited. Thus, we conclude that whitening has other effects other than frequency debiasing.

7. Conclusion

In this study, we investigated the effect of whitening on SWE and CWE. In a preliminary experiment, we confirmed the existence of a frequency bias in SWE and CWE by visualizing the PCA coefficients of the embeddings and proposed that the application of whitening can remove the bias in the embeddings. In our main experiment, we empirically examined whether whitening had effects other than the effect of frequency debiasing. The results indicated that there was a large overlap between the effects of whitening and the effect of frequency debiasing on SWE. However, on CWE, whitening had effects other than frequency debiasing only. In our analysis, we investigated how the quality of embeddings in each model varies with the frequency of words in the sentence. We found that whitening and our frequency debiasing method, namely `RFD`, improved the quality of GloVe embeddings for low-frequency words. However, we observed no such trend for BERT embeddings. This result could be caused by the training procedures of the embeddings, i.e., GloVe embeddings are optimized based on word frequency information and BERT separates words into subword tokens.

The main contributions of this work are the discovery of differences in the effects of whitening on CWE and SWE, and the suggestion that whitening has effects on CWE other than the removal of frequency bias, which is the most studied cause of anisotropy in word embeddings. It remains for future work to identify the remaining effects of whitening on CWE. We speculate that one of the effects may be the correction of the outlier problems reported by Kovaleva et al. (2021), Luo et al. (2021).[6]

[6] Our codes are available at https://github.com/losyer/whitening_effect

CRediT authorship contribution statement

Shota Sasaki: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Benjamin Heinzerling:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Jun Suzuki:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition. **Kentaro Inui:** Conceptualization, Resources, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR20D2, Japan, and by JSPS KAKENHI Grant Number 21K17814. The work of Jun Suzuki was supported by JST Moonshot R&D Grant Number JPMJMS2011 (fundamental research).

References

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th international workshop on semantic evaluation* (pp. 1–14). Association for Computational Linguistics.

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of machine learning research*: *vol. 15*, *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 215–223). PMLR.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)* (pp. 4171–4186). Association for Computational Linguistics.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 55–65). Association for Computational Linguistics.

Gong, C., He, D., Tan, X., Qin, T., Wang, L., & Liu, T.-Y. (2018). FRAGE: Frequency-agnostic word representation. In *Advances in neural information processing systems, vol. 31* (p. 12). Curran Associates, Inc.

Huang, J., Tang, D., Zhong, W., Lu, S., Shou, L., Gong, M., Jiang, D., & Duan, N. (2021). Whitening BERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics* (pp. 238–244). Association for Computational Linguistics.

Kaneko, M., & Bollegala, D. (2022). Unmasking the mask – evaluating social biases in masked language models. In *Proceedings of the 36th AAAI conference on artificial intelligence* (p. 13).

Kovaleva, O., Kulshreshtha, S., Rogers, A., & Rumshisky, A. (2021). BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics* (pp. 3392–3405). Association for Computational Linguistics.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 9119–9130). Association for Computational Linguistics.

Liang, Y., Cao, R., Zheng, J., Ren, J., & Gao, L. (2021). 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* (pp. 448–459). Springer-Verlag.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Luo, Z., Kulmizev, A., & Mao, X. (2021). Positional artefacts propagate through masked language model embeddings. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 5312–5327). Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of advances in neural information processing systems* (pp. 3111–3119).

Mu, J., & Viswanath, P. (2018). All-but-the-top: Simple and effective postprocessing for word representations. In *Proceedings of international conference on learning representations* (p. 25).

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1532–1543).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (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)* (pp. 2227–2237). Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Rajaee, S., & Pilehvar, M. T. (2021). A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 2: short papers)* (pp. 575–584). Association for Computational Linguistics.

Ranzato, M., Krizhevsky, A., & Hinton, G. (2010). Factored 3-way restricted Boltzmann machines for modeling natural images. In Y. W. Teh, & M. Titterington (Eds.), *Proceedings of machine learning research*: *vol. 9*, *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 621–628). PMLR.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3982–3992). Association for Computational Linguistics.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Su, J., Cao, J., Liu, W., & Ou, Y. (2021). Whitening sentence representations for better semantics and faster retrieval, CoRR.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45). Association for Computational Linguistics.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 335–340).

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K. W. (2019). Gender bias in contextualized word embeddings. In *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)* (pp. 629–634). Association for Computational Linguistics.