Neutralizing Gender Bias in Word Embedding with Latent Disentanglement and Counterfactual Generation

Seungjae Shin, Kyungwoo Song, JoonHo Jang ,Hyemi Kim, Weonyoung Joo, Il-Chul Moon
Korea Advanced Institute of Science and Technology (KAIST), Korea
{tmdwo0910,gtshs2,adkto8093,khm0308,es345,icmoon}@kaist.ac.kr

Abstract

Recent researches demonstrate that word embeddings, trained on the human-generated corpus, have strong gender biases in embedding spaces, and these biases can result in the prejudiced results from the downstream tasks, i.e. sentiment analysis. Whereas the previous debiasing models project word embeddings into a linear subspace, we introduce a *Latent Disentangling* model with a siamese auto-encoder structure and a gradient reversal layer. Our siamese auto-encoder utilizes gender word pairs to disentangle semantics and gender information of given word, and the associated gradient reversal layer provides the negative gradient to distinguish the semantics from the gender. Afterwards, we introduce a *Counterfactual Generation* model to modify the gender information of words, so the original and the modified embeddings can produce a gender-neutralized word embedding after geometric alignment without loss of semantic information. Experimental results quantitatively and qualitatively indicate that the introduced method is better in debiasing word embeddings, and in minimizing the semantic information losses for NLP downstream tasks.

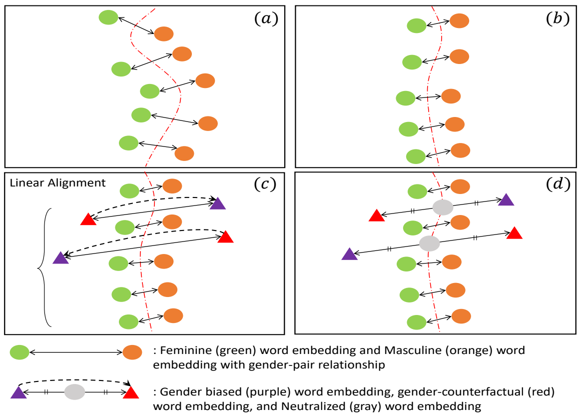


Figure 1: The process view of our debiasing method. We ideally draw gender decision boundary by linking neutral points from multiple gender word pairs. We can improve the embedding space from (a) to (b) with better linear aligned structure between gender word pairs by the proposed latent disentanglement. Afterwards, (c) We generate gender-counterfactual word embedding for the gender-biased word embedding, while keeping linear relationship with gender word pairs to guarantee that the pair of word embeddings only differs from gender information, not hurting semantic information. (d) We get gender-neutralized word embedding by interpolating it from the pair of word embeddings.

1 Introduction

Recent researches have disclosed that the word embeddings contain unexpected biases in their geometry on the embedding space (Bolukbasi et al., 2016; Garg et al., 2018; Ethayarajh et al., 2019; Zhao et al., 2019; Agarwal et al., 2019). The biases reflect unwanted stereotypes such as the correlation between the gender and the occupation from texts. Bolukbasi et al. (2016) enumerated automatically generated analogies of the pair (she, he) in the Word2Vec embedding (Mikolov et al., 2013a,b). An example of the analogies include *she* is relatively closer to *nurse*; and *he* is located near *doctor*. Garg et al. (2018) demonstrated that the embeddings, from Google News Word2Vec (Mikolov et al., 2013a) and Glove (Pennington et al., 2014), have strong associations between value-neutral words and population-segment words, i.e. a strong association between *housekeeper* and *Hispanic*. This unwanted bias can cause biased result in the downstream tasks (Caliskan et al., 2017a; Kiritchenko and Mohammad, 2018; Bhaskaran and Bhallamudi, 2019).

To mitigate gender biases in word embeddings, researchers proposed various debiasing methods for pre-trained word embeddings (Bolukbasi et al., 2016; Dev and Phillips, 2019; Kaneko and Bollegala, 2019). The widely recognized method is a post-processing method projecting word embeddings to the space that is orthogonal to the gender direction vector defined by a set of gender word pairs. Here, the challenge is extracting the proper gender direction vector component. If the gender direction vector includes a component of semantic[1] information, the semantic information will be lost through the post-processing projections.

To balance between the gender debiasing and the semantic information preserving, we propose an encoder-decoder framework that disentangles a latent space of a given word embedding to be two encoded latent spaces: the first part is the gender latent space, and the second part is the semantic latent space that is independent to the gender information. To disentangle the latent space, we use a gradient reversal layer approach by prohibiting the generation for the gender latent information from the semantic latent information. The adaptation of the gradient reversal layer in the latent disentanglement is one of our methodological contributions in this paper. From this disentanglement, we have a learned encoder to identify the gender and the semantic latent for gender-neutral words. Then, we use the counterfactual approach to neutralize the gender information in gender-neutral words. For instance, we generate a counterfactual word embedding by turning the encoded gender latent into the opposite gender. Afterward, the original and the counterfactual word embeddings are geometrically interpreted to neutralize the gender information, and to preserve the semantic information, see Figure 1 for process view of our debiasing.

We evaluate the debiased word embeddings from the proposed method and other baselines for the debiasing tasks. Quantitatively, we compared the methods by Sembias and WEAT (Word Embedding Association Test), and we found the proposed method shows improvements. Also, we perform qualitative evaluations through clustering analysis for most biased words and nearest neighbor analysis. The visual inspection of t-SNE (Maaten and Hinton, 2008) from the debiased embedding space supports the ability of the proposed method to mitigate indirect gender bias. Finally, the results from several NLP downstream tasks show that our proposed method minimizes the performance degradation from debiasing less than the existing methods.

[1]Throughout this paper, we define the semantics of words to be the meanings of words other than the gender information.

2 Gender Debiasing Mechanisms for Word Embeddings

We can divide existing gender debiasing mechanisms for word embeddings into two categories. The first mechanism is neutralizing the gender aspect of word embeddings on the training procedure. Zhao et al. (2018) proposed the learning scheme to generate a gender-neutral version of Glove, called GN-Glove, which forces preserving the gender information in pre-specified embedding dimensions while ensuring that other embedding dimensions are inferred to be gender-neutral. However, learning new word embeddings for large-scale corpus can be ineffective for time constraints.

Because of this limitation, the second mechanism post-processes trained word embeddings for debiasing. Simple post-processing can be a linear projection of gender-neutral words to a subspace, which is orthogonal to the gender direction vector defined by a set of gender-definition words (Bolukbasi et al., 2016). Another way of constructing the gender direction vector is using common names, e.g. *john*, *mary*, etc (Dev and Phillips, 2019), while the previous approach used pronouns, such as *he* and *she*. In addition to simple linear projections, Dev and Phillips (2019) utilize other alternatives, such as flipping and subtraction, to reduce gender bias more effectively. Beyond simple projection methods, Kaneko and Bollegala (2019) proposed a neutral-network based encoder-decoder framework to remove gender biases from gender-stereotyped words and to preserve gender-related information in feminine and masculine words.

3 Methodology

This paper improves the debiasing mechanism through the better latent disentanglement. Our model introduces 1) the siamese network structure (Bromley et al., 1994; Koch, 2015; Weston et al., 2012) for latent disentangling and 2) the counterfactual data augmentation for gender debiasing. We process the gender word pairs through the siamese network with auxiliary classifiers to reflect the inference of gender-dependent latent dimensions. Afterward, we debias the gender-neutral words through the learned encoder-decoder networks, and this debiasing requires the gender-neutral word to be located at the middle between a reconstructed pair of semantic latent variable and counterfactually generated gender latent variables.

Same as previous researches (Kaneko and Bolle-

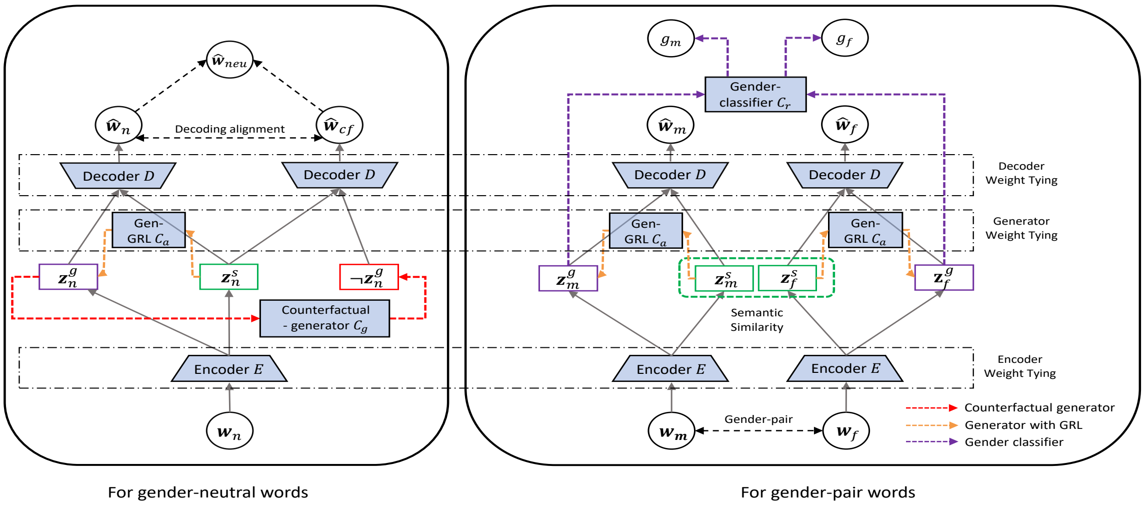


Figure 2: The framework overview of our proposed model. We characterize specialized regularization and network parameters with colored dotted lines and boxes with blue color, respectively.

gala, 2019), we divide a whole set of vocabulary V into three mutually exclusive categories : *feminine* word set V_f; *masculine* word set V_m; and *gender neutral* word set V_n, such that $V = V_f \cup V_m \cup V_n$. Our debiasing objective follows the below criteria:

- Word $w_f \in V_f$ and word $w_m \in V_m$ have the embedding information with the opposite direction with respect to the gender decision boundary in the embedding space, respectively.
- Word $w_n \in V_n$ has zero direction with respect to the gender decision boundary in the embedding space.
- Word $w \in V$ preserves the semantic information in embedding space.

In most cases, words in V_f and V_m exist in pairs, so we denote Ω as the feminine and masculine word pairs set, such that $(w_f, w_m) \in \Omega$.

3.1 Overall Model Structure

Figure 2 illustrates the overall structure of our proposed gender debiasing method for pre-trained word embeddings, which we named *Counterfactual*-Debiasing, or *CF*-Debiasing. Eq. (1) specifies the entire loss function of the whole network parameters in Figure 2. The entire loss function is divided into two types of losses: L_{ld} to be losses for disentanglement and L_{cf} to be losses for counterfactual generation. λ can be seen as a balancing hyper-parameter between two-loss terms.

$$L = \lambda L_{ld} + (1 - \lambda) L_{cf}, 0 \leq \lambda \leq 1 \qquad (1)$$

Here, we use pre-trained word embeddings $\{\boldsymbol{w}_i\}_{i=1}^{V} \in \mathbb{R}^d$ for the debiasing mechanism. In the encoder-decoder framework, we denote the latent variable of $\boldsymbol{w}_i$ to be $\boldsymbol{z}_i \in \mathbb{R}^l$, which is mapped into the latent space by the encoding function, $E : \boldsymbol{w}_i \rightarrow \boldsymbol{z}_i$; and the decoding function, $D : \boldsymbol{z}_i \rightarrow \hat{\boldsymbol{w}}_i$. After the disentanglement of the latent space, $\boldsymbol{z}_i$ is divided into two parts, such that $\boldsymbol{z}_i = [\boldsymbol{z}_i^s, \boldsymbol{z}_i^g]$: $\boldsymbol{z}_i^s \in \mathbb{R}^{l-k}$ is the semantic latent variable of $\boldsymbol{w}_i$; and $\boldsymbol{z}_i^g \in \mathbb{R}^k$ is the gender latent variable of $\boldsymbol{w}_i$, where k is pre-defined value for the gender latent dimension.[2]

3.2 Siamese Auto-Encoder for Latent Disentangling

This section provides the construction details of L_{ld}. Eq. (2) defines the objective function for latent disentanglement as a linearly-weighted sum of the losses to be introduced in this section.

$$L_{ld} = \lambda_{se} L_{se} + \lambda_{ge} L_{ge} + \lambda_{di} L_{di} + \lambda_{re} L_{re} \quad (2)$$

For the disentanglement, our fundamental assumption is maintaining the identical semantic information in $\boldsymbol{z}^s$ for the gender word pairs, $(w_f, w_m) \in \Omega$, by excluding the gender latent dimension, $\boldsymbol{z}^g$. Under this assumption, we introduce a latent disentangling method by utilizing the siamese auto-encoder with gender word pairs.

Siamese Auto-Encoder. The data structure of the gender word pairs provide an opportunity to adapt

[2] For the simplicity in notations, we skip the word-index i in the losses of our proposed method.

the siamese auto-encoder structure because the gender word pairs always have two words in pairs. Our siamese auto-encoder shares the network weights of the encoder, E, and the decoder, D, across the gender word pairs as well as the gender-neutral word. Afterward, this paper focuses on how to manipulate latent variables, $\boldsymbol{z}^s$ and $\boldsymbol{z}^g$, produced by the shared encoder and decoder structure.

Semantic Latent Formulation. First, we regularize a pair of semantic latent variables $(\boldsymbol{z}_f^s, \boldsymbol{z}_m^s)$, from a gender word pair, (w_f, w_m), to be same by minimizing the squared ℓ_2 distance as Eq. (3), since the semantic information should be the same regardless of the gender.

$$L_{se} = \sum_{(w_f, w_m)\in\Omega} \|\boldsymbol{z}_m^s - \boldsymbol{z}_f^s\|_2^2 \tag{3}$$

Gender Latent Formulation. To formulate the gender-dependent latent dimensions, we introduce an auxiliary gender classifier, $C_r : \boldsymbol{z}^g \rightarrow [0, 1]$, given in Eq. (4), and C_r is asked to produce one in highly masculine words, labeled as $g_m = 1$, and to produce zero in highly feminine words, $g_f = 0$, respectively. After training, the output of C_r can be an indicator of the gender information for each word.[3]

$$\begin{aligned} L_{ge} = &- \sum_{w_m \in V_m} g_m \log C_r(\boldsymbol{z}_m^g) \\ &- \sum_{w_f \in V_f} (1 - g_f) \log(1 - C_r(\boldsymbol{z}_f^g)) \end{aligned} \tag{4}$$

Disentanglement of Semantic and Gender Latent. The above two regularization terms extract the semantic and the gender latent dimensions, but these regularizations do not guarantee the independence between the semantic and the gender latent dimensions. To enforce the independence between two latent dimensions, we introduce a *Generator* with *Gradient Reversal Layer* (GRL), $C_a : \boldsymbol{z}^s \rightarrow \boldsymbol{z}^g$ (Ganin et al., 2016), which generates the gender latent dimension with the semantic latent dimension. (Ganin et al., 2016) use GRL to train a shared representation for two classifiers: the label and the domain classifiers. When adopting new domains, the feature should be less susceptible to domain changes. So the loss for label classifier is optimized as usual, and the loss for the domain

[3]We report the test performances of the gender classifier for gender-definition words, i.e., he, she, etc.; and gender-stereotypical words, i.e., doctor, nurse, etc., in Section C, Supplementary Material.

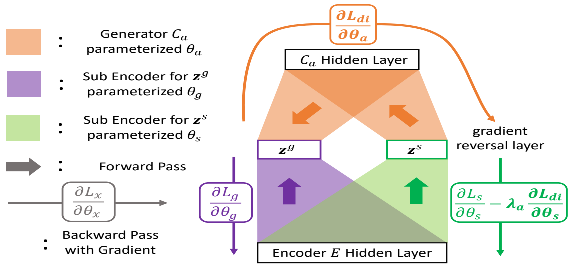


Figure 3: Gradient reversal layer utilized for the latent disentanglement. We follow similar description in Ganin et al. (2016)

classifier is adversarially optimized. This is the reason for flipping the gradient by GRL.

We modify this flipped gradient idea to the latent disentanglement between the semantic and the gender latent dimensions. Our idea is connecting $\boldsymbol{z}^g$ and $\boldsymbol{z}^s$ by generating $\boldsymbol{z}^g$ from $\boldsymbol{z}^s$ with a generator, C_a. The generation of $\boldsymbol{z}^g$ from $\boldsymbol{z}^s$ means that $\boldsymbol{z}^s$ has enough information on $\boldsymbol{z}^g$, so the generation should be prohibited to make $\boldsymbol{z}^g$ and $\boldsymbol{z}^s$ independent. Hence, our feedback of the gradient reversal layer is maximizing the loss of generating $\boldsymbol{z}^g$ from $\boldsymbol{z}^s$, which is represented as L_{di} in Eq. (5). Our design is distinct from the previous work in that our objective for adopting GRL is the latent disentanglement of two-dimension sets instead of the learning shared representation with a particular focus.

$$L_{di} = \sum_{w\in V} \|C_a(\boldsymbol{z}^s) - \boldsymbol{z}^g\|_2^2 \tag{5}$$

In the learning stage, the gradient of the encoder for $\boldsymbol{z}^s$, which is parameterized as θ_s, becomes the summation of 1) $\frac{\partial L_s}{\partial \theta_s}$, which is the gradient for the loss L_s, the latent disentanglement losses of the encoder for $\boldsymbol{z}^s$ excluding L_{di}; and 2) $-\lambda_a \frac{\partial L_{di}}{\partial \theta_s}$, which is the λ_a-weighted negative gradient of the loss L_{di} which is reversed after passing the GRL, because we intend to train the encoder for $\boldsymbol{z}^s$ by preventing the generation of $\boldsymbol{z}^g$. Eq. (5) specifies the loss function for the disentanglement by GRL, and Eq. (6) specifies the above motivation of the reversed gradient, see Figure 3 for details.

$$\frac{\partial L_{ld}}{\partial \theta_s} = \frac{\partial L_s}{\partial \theta_s} - \lambda_a \frac{\partial L_{di}}{\partial \theta_s} \tag{6}$$

Reconstruction. We add the reconstruction loss given in Eq. (7) for this encoder-decoder framework.

$$L_{re} = \sum_{w\in V} \|\boldsymbol{w} - \hat{\boldsymbol{w}}\|_2^2 \tag{7}$$

3.3 Regularization for Counterfactual Generation

This section provides the construction details of L_{cf}. Same as L_{ld}, We define the objective function for the counterfactual generation as the linearly-weighted sum of the losses, introduced in this section, as in Eq. (8).

$$L_{cf} = \lambda_{mo} L_{mo} + \lambda_{mi} L_{mi} + \lambda_{al} L_{al} \quad (8)$$

Unlike the gender word pairs, a word in the gender neutral word set $w_n \in V_n$ utilizes the counterfactual generator, $C_g : \boldsymbol{z}_n^g \to \neg \boldsymbol{z}_n^g$, which switches the original gender latent, $\boldsymbol{z}_n^g$, to the opposite gender, $\neg \boldsymbol{z}_n^g$. It should be noted that C_g is only activated for optimizing the losses in only L_{cf}, which assumes that other parameters learned for the latent disentanglement are freezed.

To switch $\boldsymbol{z}_n^g$, we utilize a prediction from the gender classifier, C_r, which is trained through the disentanglement loss. The modification loss, L_{mo}, originates from indicating the opposite gender with $\boldsymbol{z}_n^g$ by C_r, see Eq. (9). For instance, if C_r returns 0.8 for the original gender latent, $\boldsymbol{z}_n^g$, then we regularize the virtually generated gender latent, $\neg \boldsymbol{z}_n^g$, to lead C_r to return 0.2.

$$L_{mo} = \sum_{w_n \in V_n} \| C_r(\neg \boldsymbol{z}_n^g) - (1 - C_r(\boldsymbol{z}_n^g)) \|_2^2 \quad (9)$$

While Eq. (9) focuses on the gender latent switch, Eq. (10) emphasizes the minimal change of the gender latent, $\boldsymbol{z}_n^g$. The combination of these two losses guides to the efficiently switched gender latent variable from the original gender latent variable.

$$L_{mi} = \sum_{w_n \in V_n} \| \neg \boldsymbol{z}_n^g - \boldsymbol{z}_n^g \|_2^2 \quad (10)$$

Though we switch the gender latent information, we need to maintain the semantic latent information of the word embedding. To enable this maintenance, we constraint that $\hat{\boldsymbol{w}}_{cf}$, the reconstructed word embedding with the counterfactual gender latent, differs only in the gender information from $\hat{\boldsymbol{w}}_n$, the reconstructed word embedding with the original gender latent.

For this purpose, we introduce the gender direction vector $\boldsymbol{v}_g$ given in Eq. (11). The gender direction vector, $\boldsymbol{v}_g$, identifies a linear alignment between gender word pairs by utilizing the reconstructed word embeddings of the original and the counterfactual gender latents.

$$\boldsymbol{v}_g = \frac{1}{|\Omega|} \sum_{(w_f, w_m) \in \Omega} (\hat{\boldsymbol{w}}_m - \hat{\boldsymbol{w}}_f) \quad (11)$$

Afterwards, we regularize the difference of $\hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf}$ by measuring the alignment to $\boldsymbol{v}_g$ from the gender word pairs. This suggests that we constraint the embedding shift of the gender-neutral word to be the gender direction of $\boldsymbol{v}_g$, which argues that the counterfactual only influences on the gender information. This alignment can be accomplished by maximizing the absolute cosine similarity between the difference vector of $\hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf}$ and $\boldsymbol{v}_g$ as given in Eq. (12).

$$L_{al} = \sum_{w_n \in V_n} -|\cos(\boldsymbol{v}_g, \hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf})| \quad (12)$$

3.4 Post-Processing based on the Word's Category

After learning the network parameters, we post-process words by its categories of V_f, V_m, and V_n. We gender-neutralize the embedding vector of $w_n \in V_n$ by relocating the vector to the middle point of the reconstructed original-counterfactual pair embeddings, such that $\boldsymbol{w} := \frac{\hat{\boldsymbol{w}}_{cf} + \hat{\boldsymbol{w}}_n}{2} = \hat{\boldsymbol{w}}_{neu}$. We utilize a reconstructed word embedding which preserve gender information in embedding space, $\boldsymbol{w} := \hat{\boldsymbol{w}_f}$ for $w_f \in V_f$ and $\boldsymbol{w} := \hat{\boldsymbol{w}_m}$ for $w_m \in V_m$.

4 Experiments

4.1 Datasets and Experimental Settings

We used the feminine and the masculine word set created by Zhao et al. (2018) as V_f and V_m, respectively. All models utilize GloVe on the 2017 January dump of English Wikipedia with 300-dimensional word embeddings for 322,636 unique words, same as Zhao et al. (2018). Our experiments specify the latent dimension of $\boldsymbol{z}$, l, as 300, which is divided into 295 semantic latent dimensions and five gender latent dimensions. Same as Kaneko and Bollegala (2019), we pre-train the autoencoder for better-balanced training between the reconstruction loss and the disentanglement losses. Also, we utilize the time-based learning rate schedule, which updates the weight for disentangling more at the initial step and gradually increases updating the weight for the counterfactual generation, by changing λ in Eq. (1) from 1 to 0.

4.2 Baselines

We compare our proposed model with below baseline models, and we utilize each author's implementations.[4] Hard-GloVe (Bolukbasi et al., 2016) debiases word embedding by projecting the gender-neutral words to a subspace, which is orthogonal to the gender direction vector. GN-GloVe (Zhao et al., 2018) trains the gender-neutral embedding from scratch for a given corpus by preserving the gender information into the specific dimension and regularizing the other dimensions to be gender-neutral. CPT-GloVe (Karve et al., 2019) introduces a debiasing mechanism by utilizing the conceptor matrix. ATT-GloVe (Dev and Phillips, 2019) defines gender subspace with common names and proposes the subtraction and the linear projection methods based on gender subspace. We use the subtraction method as an ATT-GloVe, which shows better performance in our experiments. AE-GloVe and AE-GN (Kaneko and Bollegala, 2019) utilize the autoencoder for preserving the semantic information of word embedding. AE-GloVe and AE-GN utilize the GloVe and GN-GloVe, respectively. Besides, GP-GloVe and GP-GN adopt additional losses to preserve gender information for feminine and masculine words and remove gender biases for gender-neutral words.

4.3 Quantitative Evaluation for Debiasing Performance

4.3.1 Sembias Analogy Test

We perform the gender relational analogy test with the *Sembias* dataset (Zhao et al., 2018; Jurgens et al., 2012) to evaluate the degree of gender bias in word embeddings. The dataset contains 440 instances, and each instance consists of four pairs of words: 1) a gender-definition word pair (Definition), 2) a gender-stereotype word pair (Stereotype), and 3,4) two none-type word pairs (None). A tested model chooses a word pair (a, b) whose difference vector, $\overrightarrow{a} - \overrightarrow{b}$, has the highest cosine similarity with $\overrightarrow{he} - \overrightarrow{she}$ as a classification for the gender-definition word pair. By following the past practice (Zhao et al., 2018), we test models with 40 instances from *Sembias subset*, whose gender-definition word pairs are not used for training.

Table 1 shows the percentages of prediction for each category: Definition, Stereotype, and None. Our model clearly selects all the gender-definition word pairs, which demonstrates the maintenance of the gender latent information for those words. Also, our model selects neither gender-stereotype words nor none-type words, so the difference vector of $\overrightarrow{a} - \overrightarrow{b}$ has a minimal linear correlation with those words after applying our debiasing method.

	Sembias			Sembias subset		
Embeddings	Definition ↑	Stereotype ↓	None ↓	Definition ↑	Stereotype ↓	None ↓
GloVe	80.22	10.91	8.86	57.5	20.0	22.5
Hard-Glove	87.95*	8.41	3.64*	50.0	32.5	17.5
GN-GloVe	97.73$^{\dagger *}$	1.36$^{\dagger *}$	0.91$^{\dagger *}$	75.0†	15.0	10.0
ATT-GloVe	80.22	10.68	9.09	60.0	17.5	22.5
CPT-GloVe	73.63	5.68	20.68	45.0	12.5	42.5
AE-GloVe	84.09	7.95	7.95	65.0	15.0	20.0
AE-GN	98.18$^{\dagger *}$	1.14$^{\dagger *}$	0.68$^{\dagger *}$	80.0$^{\dagger *}$	12.5†	7.5
GP-GloVe	84.09	8.18	7.73	65.0$^{\dagger *}$	15.0	20.0
GP-GN	98.41$^{\dagger *}$	1.14$^{\dagger *}$	0.45$^{\dagger *}$	82.5$^{\dagger *}$	12.5†	5.0*
CF-GloVe	**100.00**$^{\dagger *}$	**0.00**$^{\dagger *}$	**0.00**$^{\dagger *}$	**100.0**$^{\dagger *}$	**0.0**$^{\dagger *}$	**0.0**$^{\dagger *}$

Table 1: Percentage of predictions for each category on gender relational analogy task. We can expect a high percentage for Definition and low percentages for Stereotype and None for well-debiased word embeddings. † and $*$ denote the statistically significant differences comparing with Hard-GloVe and Glove, respectively. The best performing model is indicated as boldface.

4.3.2 WEAT Hypothesis Test

To quantify the degree of gender bias, we apply the Word Embedding Association Test (WEAT) (Caliskan et al., 2017b). WEAT measures the effect size and the hypothesis statistics based on the gender-definition words and the well-known gender-stereotypical words set, such as *strength* and *weakness*. We follow the same experimental setting of WEAT in Chaloner and Maldonado (2019), and we provide details about WEAT hypothesis test in Section B, Supplementary Material.

Our proposed model, CF-GloVe, shows the best performances for B1, B3, and B5 categories, and CF-GloVe also exhibits the competitive performance for B2 and B4 categories, see Table 2. While all baseline models record worse performance than GloVe on at least one of the categories, our model always shows better performances than GloVe.

4.4 Debiasing Qualitative Analysis

To demonstrate the indirect gender bias hidden in the word embedding space, we perform two qualitative analysis tasks proposed by Gonen and Goldberg (2019).

4.4.1 Clustering Analysis for The Most Biased Words

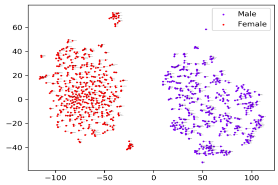
Same as Gonen and Goldberg (2019), we take the top 500 male-biased words and the top 500 female-

[4]We provided link of each author's implementation in Section G, Supplementary Material.

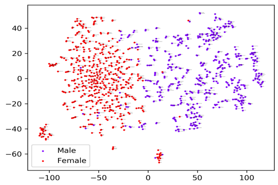
	B1 : career vs family		B2 : maths vs arts		B3 : science vs arts		B4 : intelligence vs appearance		B5 : strength vs weakness	
Embeddings	p-value	Effect size	p-value	Effect size	p-value	Effect size	p-value	Effect size	p-value	Effect size
GloVe	0.000	1.605	0.276	0.494	0.014	1.260	0.009	0.706	0.067	0.640
Hard-GloVe	0.100	0.842	0.090	-1.043	0.003	-0.747	**0.693**	**-0.121**	0.255	0.400
GN-GloVe	0.000	1.635	<u>0.726</u>	-0.169	0.081	1.007	0.037	0.595	0.083	0.620
ATT-GloVe	<u>0.612</u>	<u>0.255</u>	0.007	-0.519	0.000	0.843	0.129	0.440	0.211	0.455
CPT-GloVe	0.004	1.334	0.058	1.029	0.000	1.417	0.001	0.906	<u>0.654</u>	<u>-0.172</u>
AE-GloVe	0.000	1.569	0.019	0.967	0.024	1.267	0.007	0.729	0.027	0.763
AE-GN	0.001	1.581	0.716	0.317	0.139	0.639	0.006	0.770	0.028	0.585
GP-GloVe	0.000	1.567	0.019	0.966	0.027	1.253	0.006	0.733	0.028	0.758
GP-GN	0.000	1.599	**0.932**	**0.109**	<u>0.251</u>	<u>0.591</u>	0.004	0.791	0.098	0.610
CF-GloVe	**0.874**	**-0.089**	0.669	<u>-0.125</u>	**0.360**	**0.480**	<u>0.678</u>	<u>-0.124</u>	**0.970**	**0.013**

Table 2: WEAT hypothesis test results for five popular gender-biased word categories. The best performing model is indicated as boldface. The second-best model is indicated as underline. The absolute value of the effect size denotes the degree of bias, and the p-value denotes the statistical significance of the results.

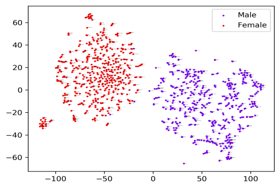
biased words, which is a word collection of the top 500 and the bottom 500 dot-product value between word embeddings and $\overrightarrow{he} - \overrightarrow{she}$ vector. Ideally, from the debiasing perspective, these 1,000 word vectors should not be clustered as two distinct groups. Therefore, we create two clusters with K-means, and we check the heterogeneity of the clusters through the classification by the cluster majority. Figure 4 shows that CF-GloVe generates gender-invariant representations for gender-biased word sets by showing the lowest cluster classification accuracy.



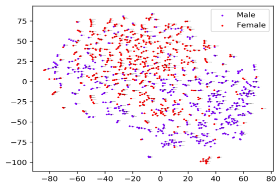
(a) GloVe (99.9%)



(b) Hard-GloVe (89.4%)



(c) GP-GN (100.0%)

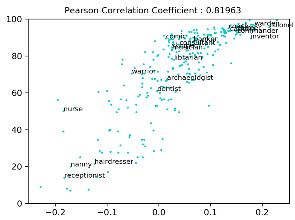


(d) CF-GloVe (63.1%)

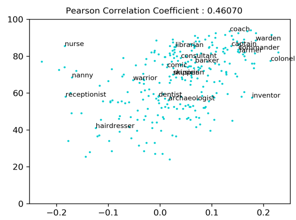
Figure 4: The t-SNE projection views for embeddings of 500 male-biased words and 500 female-biased words according to the original Glove, the cluster majority based classification accuracy is added in parenthesis.

4.4.2 Correlation Analysis between Original Bias and Nearest Neighbors

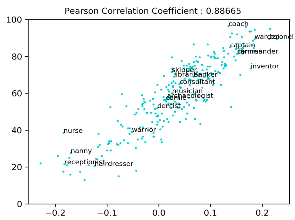
Gonen and Goldberg (2019) demonstrates that the original bias, the dot-product between the original word embedding from GloVe and $\overrightarrow{he} - \overrightarrow{she}$, has a high correlation with the male/female ratio of



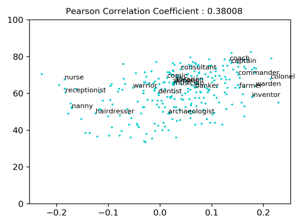
(a) GloVe (0.8196)



(b) Hard-GloVe (0.4607)



(c) GP-GN (0.8867)



(d) CF-GloVe (0.3801)

Figure 5: The percentage of male neighbors for each profession as a function of original bias for each embedding, we show only a limited number of professions on the plot to make it readable. The Pearson correlation coefficient is added in parenthesis.

gender-biased words among the nearest neighbors of the word embedding. Figure 5[5] shows each profession word at (the dot-product, the male/female ratio). If the word embeddings are truly debiased, the dot-product and the male/female ratio should not be correlated, and CF-GloVe shows the minimal Pearson correlation coefficient between the two axes.

4.5 Downstream Task of Debiased Word Embeddings

We compared the downstream task performance degradation by comparing the original and the debiased word embeddings. By following CoNLL 2003 shared task (Sang and Erik, 2002), we select

[5] Full plots of other baselines for two qualitative analyses are available in Section D and E, Supplementary Material.

	POS Tagging		POS Chunking		Named Entity Recognition	
Embeddings	Δ F1	Δ Recall	Δ F1	Δ Recall	Δ F1	Δ Recall
Hard-GloVe	-0.657±0.437	-1.220±0.938	-0.007±0.001	-0.025±0.003	-0.004±0.001	-0.015±0.005
GN-GloVe	-0.594±0.367	-1.115±0.821	-0.003±0.001	-0.010±0.003	**−0.002 ± 0.001**	-0.008±0.002
ATT-GloVe	-0.689±0.474	-1.279±1.000	-0.024±0.005	-0.091±0.019	-0.013±0.003	-0.046±0.011
CPT-GloVe	-0.501±0.277	-0.959±0.674	-0.004±0.001	-0.016±0.005	**−0.002 ± 0.000**	-0.008±0.001
AE-GloVe	-2.862±1.632	-8.647±5.072	-2.108±0.558	-7.753±1.996	-1.669±0.547	-5.895±1.893
AE-GN	-3.505±1.498	-10.766±4.525	-4.765±0.402	-16.760±1.299	-4.460±0.485	-5.097±1.524
GP-GloVe	-2.911±1.664	-8.810±5.156	-2.058±0.555	-7.573±1.988	-1.611±0.542	-5.696±1.877
GP-GN	-3.560±1.506	-10.943±4.557	-4.791±0.391	-16.843±1.262	-4.485±0.468	-5.176±1.471
CF-GloVe	**−0.287 ± 0.118**	**−0.506 ± 0.260**	**−0.002 ± 0.001**	**−0.006 ± 0.004**	**−0.002 ± 0.001**	**−0.007 ± 0.005**

Table 3: Performance degradation percentage with standard deviation for downstream tasks of POS Tagging, POS Chunking, and NER. The best performing model is indicated as boldface.

Part-Of-Speech tagging, Part-Of-Speech chunking, and Named Entity Resolution (NER) as our downstream tasks. Table 3 shows the percentage of performance degradations, and we observed that there are constant degradation effects for all debiasing methods. However, our method minimized the degradation of task performance across baseline models.

4.6 Qualitative Analysis on Linear Alignment of Gender Word Pairs

We tested the linear alignment of gender word pairs, which is a condition of using the gender direction vector $\boldsymbol{v}_g$ in Eq. (11). If the difference vectors of gender word pairs are not linearly aligned, the gender direction vector cannot be a pure representation of the gender bias.

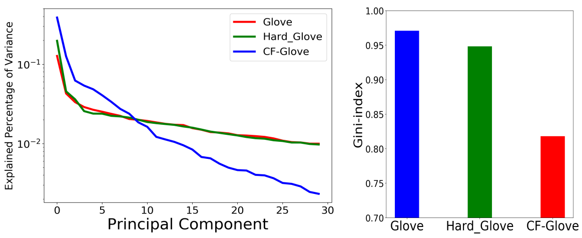


Figure 6: The proportion of variances from top 30 PCs (left) and Gini-index for the variance proportion vector for top 30 PCs(right)

We first compared the variances explained by the top 30 principal components (PC) of gender word pairs. The left plot in Figure 6 shows the proportion of variances from each PC. CF-GloVe shows the largest concentration of the variances on a few components. On the contrary, Hard-GloVe and GloVe show the long spread of the variance without the concentration like CF-GloVe. The right plot in Figure 6 shows Gini-index (Gini, 1912) for the variance proportion vector from PCs of each embedding. CF-GloVe shows minimal Gini-index, which quantitatively indicates the monopolized proportion of variances on a few components.

Also, Figure 7 shows two example plots of a selected gender word pairs in the original embedding space and the debiased embedding space, by Locally Linear Embedding (LLE) (Roweis and Saul, 2000) which preserves the local linear relationship of high-dimensional data in reducing dimensions. The right plot in Figure 7 shows the consistency of the gender direction vector, and the plot indicates the neutralization of *homemaker* by the counterfactually generated word embeddings.

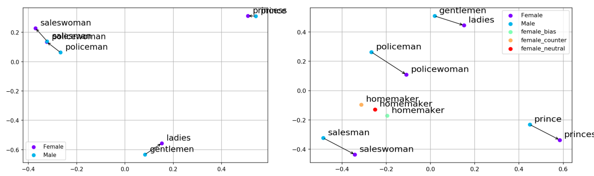


Figure 7: LLE projection view of selected gender word pairs and biased word for original embedding space (left) and debiased embedding space (right)

5 Conclusions

This work makes contributions in two layers. At the application layer, CF-GloVe produces the debiased word embeddings that has the most neutral gender latent information as well as the efficiently maintained semantic latent information for the downstream tasks. At the modeling layer, CF-GloVe suggests a new method of disentangling the latent information of word embeddings with the gradient reversal layer and creating the counterfactual latent variables by exploiting the geometry of the embedding space. It should be noted that these types of latent modeling methods can be applied to diverse natural language tasks to control expressions on emotions, prejudices, ideologies, etc.

References

Oshin Agarwal, Funda Durupınar, Norman I Badler, and Ani Nenkova. 2019. Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 205–211.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? occupational gender stereotypes in sentiment analysis. *arXiv preprint arXiv:1906.10256*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

J. Bromley, I. Guyon, Y. LeCun, E. Sckinger, and R. Shah. 1994. Signature verification using a "siamese" time delay neural network. *Neural Information Processing Systems*.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017a. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017b. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Corrado Gini. 1912. Variabilità e mutabilità (variability and mutability). *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1955) ed. Bologna*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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)*, pages 609–614.

David A Jurgens, Peter D Turney, Saif M Mohammad, and Keith J Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 356–364. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742*.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Gregory Koch. 2015. Siamese neural networks for one-shot image recognition.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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

Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

Tjong Kim Sang and F Erik. 2002. Introduction to the conll-2002 shared task: language-independent named entity recognition. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4. Association for Computational Linguistics.

Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. 2012. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 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)*, pages 629–634.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.