Neutralizing Gender Bias in Word Embeddings 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), Daejeon, Korea
{tmdwo0910,gtshs2,adkto8093,khm0308,es345,icmoon}@kaist.ac.kr

Abstract

Recent research demonstrates that word embeddings, trained on the human-generated corpus, have strong gender biases in embedding spaces, and these biases can result in the discriminative results from the various downstream tasks. Whereas the previous methods project word embeddings into a linear subspace for debiasing, we introduce a *Latent Disentanglement* method with a siamese auto-encoder structure with an adapted gradient reversal layer. Our structure enables the separation of the semantic latent information and gender latent information of given word into the disjoint latent dimensions. Afterwards, we introduce a *Counterfactual Generation* to convert the gender information of words, so the original and the modified embeddings can produce a gender-neutralized word embedding after geometric alignment regularization, without loss of semantic information. From the various quantitative and qualitative debiasing experiments, our method shows to be better than existing debiasing methods in debiasing word embeddings. In addition, Our method shows the ability to preserve semantic information during debiasing by minimizing the semantic information losses for extrinsic NLP downstream tasks.

1 Introduction

Recent researches have disclosed that word embeddings contain unexpected bias in their geometry on the embedding space (Bolukbasi et al., 2016; Zhao et al., 2019). The bias reflects unwanted stereotypes such as the correlation between gender[1] and occupation words. Bolukbasi et al. (2016) enumerated that the automatically generated analogies of (she, he) in the Word2Vec (Mikolov et al., 2013b) show the gender biases in significant level. An example of the analogies is the relatively closer distance of *she* to *nurse*; and *he* to *doctor*. Garg et al. (2018) demonstrated that the embeddings, from Word2Vec (Mikolov et al., 2013a) to 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 results in the downstream tasks (Caliskan et al., 2017a; Kiritchenko and Mohammad, 2018; Bhaskaran and Bhallamudi, 2019) and gender discrimination in NLP systems.

From the various gender debiasing methods for pre-trained word embeddings, the widely recognized method is a post-processing method, which projects word embeddings to the space that is or-

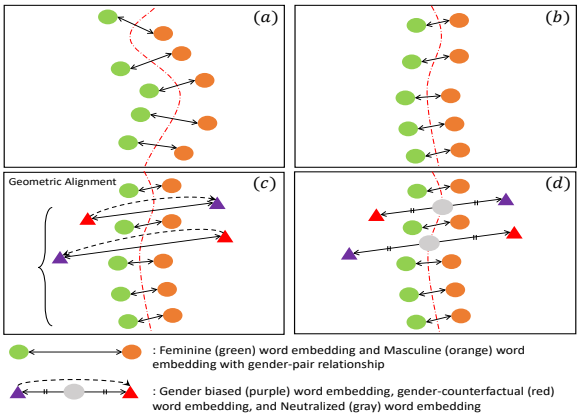


Figure 1: The process view of our method. We can improve the embedding space from (a) to (b) with a better-aligned structure between gender word pairs by the proposed latent disentanglement. Afterwards, (c) we generate the gender-counterfactual embedding of the gender-biased word while keeping a geometrically aligned relationship with the gender word pairs to guarantee that the pair of word embeddings only differs from gender information, not hurting semantic information. (d) We obtain the gender-neutralized word embedding by interpolating the embedding from the pair of original-counterfactual word embeddings.

[1]While we acknowledge a potential and expanded definition on gender as stated in Larson (2017), we only cover the gender bias between the male and female in this paper.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3126–3140

thogonal to the gender direction vector defined by a set of gender word pairs. However, if the gender direction vector includes a component of semantic information[2], 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 into 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 into two sub-spaces, we use a gradient reversal layer by prohibiting the inference on the gender latent information from the semantic information. Then, we generate a counterfactual word embedding by converting the encoded gender latent into the opposite gender. Afterwards, the original and the counterfactual word embeddings are geometrically interpreted to neutralize the gender information of given word embeddings, see Figure 1 for the illustration on our debiasing method.

Our contributions are summarized as follows:

- We propose a method for disentangling the latent information of the word embedding by utilizing the siamese auto-encoder structure with an adapted gradient reversal layer.
- We propose a new gender debiasing method, which transforms the original word embedding into gender-neutral embedding, with the gender-counterfactual word embedding.
- We propose a generalized alignment with a kernel function that enforces the embedding shift, during the debiasing process, in a direction that does not damage the semantics of word embedding.

We evaluated the proposed method and other baseline methods with several quantitative and qualitative debiasing experiments, and we found that the proposed method shows significant improvements from the existing methods. Additionally, the results from several NLP downstream tasks show that our proposed method minimizes performance degradation than the existing methods.

[2]Throughout this paper, we define the semantics of words to be the meanings and functionality of words other than the gender information by following Shoemark et al. (2019).

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 in 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 other embedding dimensions are inferred to be gender-neutral. However, learning new word embeddings for large-scale corpus can be difficult and expensive.

The second mechanism post-processes trained word embeddings to debias them after the training. An example of such post-processings is a linear projection of gender-neutral words toward 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 gender pronouns, such as *he* and *she*. In addition to the linear projections, Dev and Phillips (2019) utilizes other alternatives, such as flipping and subtraction, to reduce the gender bias more effectively. Beyond simple projection methods, Kaneko and Bollegala (2019) proposed a neural network based encoder-decoder framework to add a regularization on preserving the gender-related information in feminine and masculine words.

3 Methodology

Our model introduces 1) the siamese network structure (Bromley et al., 1994; Weston et al., 2012) with an adapted gradient reversal layer for latent disentanglement and 2) the counterfactual data augmentation with geometric regularization for gender debiasing. We process the gender word pairs through the siamese network with auxiliary classifiers to reflect the inference of gender latent dimensions. Afterwards, we debias the gender-neutral words by locating it to be at the middle between a reconstructed pair of original gender latent variable and counterfactually generated gender latent variable.

Same as previous researches (Kaneko and Bollegala, 2019), we divide a whole set of vocabulary V into three mutually exclusive categories : *feminine*

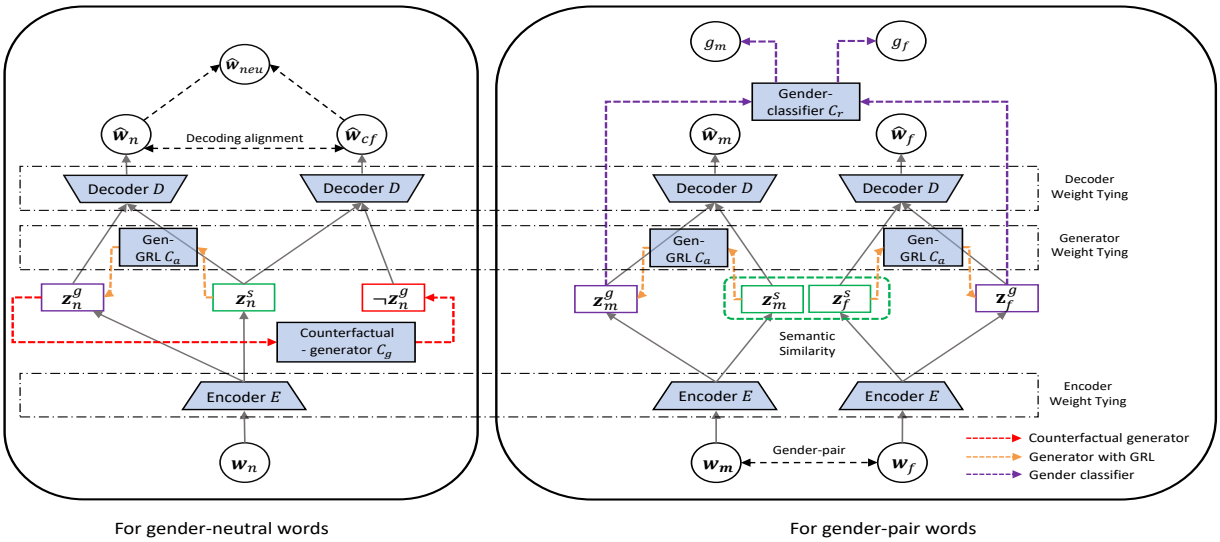


Figure 2: The framework overview of our proposed model. We characterize specialized regularization and network parameters with colored dotted lines and boxes with blue color, respectively.

word set V_f; *masculine* word set V_m; and *gender neutral* word set V_n, such that $V = V_f \cup V_m \cup V_n$. In most cases, words in V_f and V_m exist in pairs, so we denote Ω as the set of feminine and masculine word pairs, such that $(w_f, w_m) \in \Omega$.

3.1 Overall Model Structure

Figure 2 illustrates the overall structure of our proposed method for pre-trained word embeddings, which we named *Counterfactual*-Debiasing, or *CF*-Debias. 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 a loss for disentanglement and L_{cf} to be a loss for counterfactual generation. λ can be seen as a balancing hyper-parameter between two-loss terms.

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

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 to 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 the pre-defined value for the gender latent dimension.[3]

3.2 Siamese Auto-Encoder for Latent Disentanglement

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.

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

For the disentanglement, our fundamental assumption is maintaining the identical semantic information in $\boldsymbol{z}^s$ for the gender word pairs, $(w_f, w_m) \in \Omega$. Under this assumption, we introduce a latent disentangling method by utilizing the siamese auto-encoder with gender word pairs. The data structure of the gender word pairs provide an opportunity to adapt the siamese auto-encoder structure because the gender word pairs almost always have two words in pair[4].

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 of a gender word pair should be the same regardless of the gender.

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

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

[4]This structure can be expanded as our gender coverage changes.

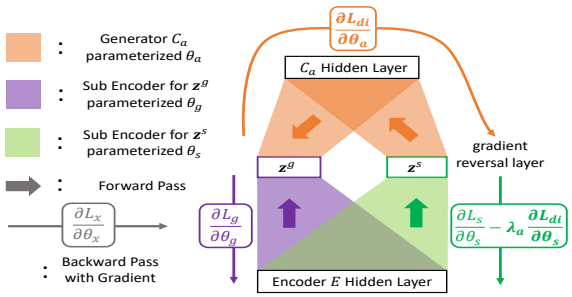


Figure 3: Gradient reversal layer utilized for the latent disentanglement. We follow similar description in Ganin et al. (2016)

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 masculine words, labeled as $g_m = 1$, and to produce zero in feminine words, $g_f = 0$, respectively. After training, the output of C_r can be an indicator of the gender information for each word.[5]

$$L_{ge} = -\sum_{w_m \in V_m} g_m \log C_r(\boldsymbol{z}^g_m) - \sum_{w_f \in V_f} (1 - g_f) \log(1 - C_r(\boldsymbol{z}^g_f)) \quad (4)$$

Disentanglement of Semantic and Gender Latent The above two regularization terms 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. We modify the flipping gradient idea of (Ganin et al., 2016) to the latent disentanglement between the semantic and the gender latent dimensions. The sufficient 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).

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

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 reversed gradient, see Figure 3.

$$\frac{\partial L_{ld}}{\partial \theta_s} = \frac{\partial L_s}{\partial \theta_s} - \lambda_a \frac{\partial L_{di}}{\partial \theta_s} \quad (6)$$

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

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

3.3 Gender-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).

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

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

To switch $\boldsymbol{z}^g_n$, 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 $\boldsymbol{z}^g_n$ by C_r, see Eq. (9). For instance, if C_r returns 0.8 for the original gender latent, $\boldsymbol{z}^g_n$, then we regularize the virtually generated gender latent, $\neg \boldsymbol{z}^g_n$, to lead C_r to return 0.2.

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

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

[5] 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 Appendix D.

that is close to the original gender latent variable for regularizing the counterfactural generation.

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

Though we keep the semantic latent variable, $\boldsymbol{z}^s$, and switch the gender latent variable, $\boldsymbol{z}^g$, to generate the gender-counterfactual word embedding, their concatenation during decoding can be vulnerable to the semantic information changes because of variances in the individual latent variables. Consequently, we constrain that the reconstructed word embedding with the counterfactual gender latent, $\hat{\boldsymbol{w}}_{cf}$, differs only in the gender information from $\hat{\boldsymbol{w}}_n$, which is the reconstructed word embedding with the original gender latent.

Linear Alignment For this purpose, we introduce the linear alignment, which regularizes $\hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf}$ by measuring the alignment to the gender direction vector $\boldsymbol{v}_g$ in Eq. (11), which is an averaged gender difference vector from the gender word pairs.

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

This regularization suggests that we constrain the embedding shift of the gender-neutral word to be the direction of $\boldsymbol{v}_g$. This alignment can be accomplished by maximizing the absolute inner product between $\hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf}$ and $\boldsymbol{v}_g$ as given in Eq. (12). We introduce CF-Debias-LA, which adds the below linear alignment regularization, $\lambda_{la} L_{la}$, to L_{cf}.

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

Kernelized Alignment While the linear alignment computes the gender direction vector $\boldsymbol{v}_g$ as a simple average, the gender information of word embedding can have a nonlinear structure. Therefore, we introduce the *kernelized alignment*, which enables the nonlinear alignment between 1) $\hat{\boldsymbol{w}}_m^i - \hat{\boldsymbol{w}}_f^i$ of each gender word pair (w_f^i, w_m^i) and 2) $\hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf}$ of gender-neutral words w_n.

We hypothesize a nonlinear mapping function f, which projects a word embedding $\boldsymbol{w}_i \in \mathbb{R}^d$ into a newly introduced feature space, $f(\boldsymbol{w}_i) \in \mathbb{R}^m$. We can utilize the kernel trick (Schölkopf et al., 1998) for computing pairwise operation on the nonlinear space introduced by f . Let $\mathrm{k}(\boldsymbol{w}, \boldsymbol{w}') = f(\boldsymbol{w}) \cdot f(\boldsymbol{w}')$ be a kernel representing an inner-product of two vectors in the feature space. Also, we set ϕ_k to be k-th eigenvector for the projected outputs of the given embeddings $\{f(\boldsymbol{w}_i)\}_{i=1}^N$. By following Appendix A, PC_k is the k-th principal component of new word embedding $\boldsymbol{w}'$ on the introduced feature space: $PC_k = f(\boldsymbol{w}') \cdot \phi_k$. Then, we find the k-th principal component for embedding $\boldsymbol{w}'$ as given in Eq. (15), when a_k^i is i-th component of k-th eigenvector of $\boldsymbol{K}$, which is a $N \times N$ kernel matrix of given data.

$$\begin{aligned} PC_k = f(\boldsymbol{w}') \cdot \phi_k &= \Sigma_{i=1}^N a_k^i f(\boldsymbol{w}_i) \cdot f(\boldsymbol{w}') \\ &= \Sigma_{i=1}^N a_k^i \, \mathrm{k}(\boldsymbol{w}_i, \boldsymbol{w}') \end{aligned} \quad (13)$$

Substituting the inner product in Eq. (12) with Eq. (14), we design the nonlinear alignment between the gender difference vector, $\hat{\boldsymbol{w}}_m - \hat{\boldsymbol{w}}_f$, and the gender neutral vector, $\hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf}$, by maximizing the Top-$K$ kernel principal components as Eq. (14). We introduce CF-Debias-KA, which adds the kernelized alignment regularization, $\lambda_{ka} L_{ka}$, to L_{cf}. We use Radial Basis Function kernel for our experiment.

$$\begin{aligned} L_{ka} = -\sum_{k=1}^{K} \sum_{w_n \in V_n} \sum_{(w_f^i, w_m^i) \in \Omega} \\ a_k^i \, \mathrm{k}\left(\hat{\boldsymbol{w}}_m^i - \hat{\boldsymbol{w}}_f^i, \hat{\boldsymbol{w}}_n - \hat{\boldsymbol{w}}_{cf}\right) \end{aligned} \quad (14)$$

3.4 Post-Processing by 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 preserves the 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 $w_m \in V_m$. For each $w \in V_f \cup V_m$, we can safely preserve gender information of given word by using reconstructed embedding such that $\boldsymbol{w} := \hat{\boldsymbol{w}}$.

4 Experiments

4.1 Datasets and Experimental Settings

We used the set of gender word pairs created by Zhao et al. (2018) as V_f and V_m, respectively. All models utilize GloVe on 2017 January dump of English Wikipedia with 300-dimension embeddings for 322,636 unique words. Additionally, to investigate the debiasing effect on languages other than

	English (GloVe)						Spanish (Fasttext)						Korean (Fasttext)					
	Sembias			Sembias subset			Sembias			Sembias subset			Sembias			Sembias subset		
Embeddings	Def ↑	Stereo↓	None ↓	Def ↑	Stereo ↓	None ↓	Def ↑	Stereo↓	None ↓	Def ↑	Stereo ↓	None ↓	Def ↑	Stereo ↓	None ↓	Def ↑	Stereo ↓	None ↓
Original	80.22	10.91	8.86	57.5	20.0	22.5	70.98†	17.38†	11.63†	84.61†	11.86†	3.52†	80.38†	7.48†	12.14†	76.26	8.87	14.88
Hard-Debias	87.95*	8.41	3.64*	50.0	32.5	17.5	41.76	27.55	30.68	21.12	38.54	40.33	41.39	15.31	43.30	**89.23***	2.62*	**8.15***
GN-Debias	97.73$^{\dagger *}$	1.36$^{\dagger *}$	0.91$^{\dagger *}$	75.0†	15.0	10.0	—	—	—	—	—	—	—	—	—	—	—	—
ATT-Debias	80.22	10.68	9.09	60.0	17.5	22.5	75.23*†	13.02*†	11.74†	83.44†	9.80$^{\dagger *}$	6.76†	82.98$^{\dagger *}$	7.70†	**9.33**$^{\dagger *}$	79.59*	8.87	11.55*
CPT-Debias	73.63	5.68	20.68	45.0	12.5	42.5	69.62†	18.26†	12.11†	84.62†	11.86†	3.52†	61.31†	10.57†	28.12†	38.52	15.76	45.72
AE-Debias	84.09	7.95	7.95	65.0†	15.0	20.0	73.19*†	15.56*†	11.26†	86.38$^{\dagger *}$	10.10$^{\dagger *}$	3.52†	57.66†	11.91†	30.44†	55.72	10.76	33.53
AE-GN-Debias	98.18$^{\dagger *}$	1.14$^{\dagger *}$	0.68$^{\dagger *}$	80.0$^{\dagger *}$	12.5†	7.5	—	—	—	—	—	—	—	—	—	—	—	—
GP-Debias	84.09	8.18	7.73	65.0†	15.0	20.0	72.93$^{\dagger *}$	15.87$^{\dagger *}$	**11.19**$^{\dagger *}$	86.37$^{\dagger *}$	10.09$^{\dagger *}$	3.52†	55.85†	15.62	28.53†	68.00	16.19	15.81
GP-GN-Debias	98.41$^{\dagger *}$	1.14$^{\dagger *}$	0.45$^{\dagger *}$	82.5$^{\dagger *}$	12.5†	5.0*	—	—	—	—	—	—	—	—	—	—	—	—
CF-Debias	98.18$^{\dagger *}$	0.68$^{\dagger *}$	1.13$^{\dagger *}$	80.0$^{\dagger *}$	7.5†	12.5	78.93$^{\dagger *}$	**3.83**$^{\dagger *}$	17.23†	96.15$^{\dagger *}$	**0.0**$^{\dagger *}$	3.85†	83.02$^{\dagger *}$	2.44$^{\dagger *}$	14.53†	80.98*	**0.0**$^{\dagger *}$	19.02
CF-Debias-LA	**100.00**$^{\dagger *}$	**0.00**$^{\dagger *}$	**0.00**$^{\dagger *}$	**100.0**$^{\dagger *}$	**0.0**$^{\dagger *}$	**0.0**$^{\dagger *}$	69.33†	9.05$^{\dagger *}$	21.61†	**100.0**$^{\dagger *}$	**0.0**$^{\dagger *}$	**0.0**$^{\dagger *}$	**85.07**$^{\dagger *}$	2.37$^{\dagger *}$	12.5†	88.04*	**0.0**$^{\dagger *}$	11.95*
CF-Debias-KA	92.04$^{\dagger *}$	3.41$^{\dagger *}$	4.55*	62.5	17.5	20.0	**80.35***†	6.73*†	12.91†	**100.0**$^{\dagger *}$	**0.0**$^{\dagger *}$	**0.0**$^{\dagger *}$	84.28$^{\dagger *}$	**2.09**$^{\dagger *}$	13.62†	82.27*	2.38*	15.35

Table 1: Percentage of predictions of each category on sembias analogy task, for each language. † and ∗ denote the statistically significant differences for Hard-Debias and Original embedding, respectively. The best model is indicated as boldface. We denote "—" for the skipped cases, whose methods are closely tied to GloVe embedding.

English; we conducted one of the debiasing experiments for Spanish, which is the Subject-Verb-Object language as English; and Korean, one of the Subject-Object-Verb language. We used Fasttext (Bojanowski et al., 2016) for experiments of Spanish and Korean. Accordingly, we excluded the baselines, whose methods are closely tied to GloVe, for the experiments of other languages. We specify the dimensions of $\boldsymbol{z}$, l, as 300, which is divided into 295 semantic latent dimensions and 5 gender latent dimensions. Also, we utilize the sequential hyperparameter schedule, which updates the weight for L_{ld} more at the initial step and gradually increases updating the weight for the L_{cf}, by changing λ in Eq. (1) from 1 to 0. Further information on experimental settings can be found in Appendix G.

4.2 Baselines

We compare our proposed model with below baseline models, and we utilize the authors' implementations.[6] Hard-Debias (Bolukbasi et al., 2016) utilizes linear projection technique for gender-debiasing. GN-Debias (Zhao et al., 2018) trains the word embedding from scratch by preserving the gender information into the specific dimension and regularizing the other dimensions to be gender-neutral. CPT-Debias (Karve et al., 2019) introduces a debiasing mechanism by utilizing the conceptor matrix. ATT-Debias (Dev and Phillips, 2019) defines gender subspace with common names and proposes the subtraction and the linear projection methods based on gender subspace.[7] AE-Debias and AE-GN-Debias (Kaneko and Bollegala, 2019) utilize the autoencoder structure for debiasing, and utilize the original word embedding and GN-Debias, respectively. Besides, GP-Debias and GP-GN-Debias adopt additional losses to neutralize gender bias and preserve gender information for gender-definition words.

4.3 Quantitative Evaluation for Debiasing

4.3.1 Sembias Analogy Test

We perform the *Sembias* gender analogy test (Zhao et al., 2018; Jurgens et al., 2012) to evaluate the degree of gender bias in embeddings. The *Sembias* dataset in English contains 440 instances, and each instance consists of four-word pairs : 1) a gender-definition word pair (Def), 2) a gender-stereotype word pair (Stereo), and 3,4) two none-type word pairs (None). We test models by calculating the linear alignment between each word pair difference vector, $\overrightarrow{a} - \overrightarrow{b}$; and $\overrightarrow{he} - \overrightarrow{she}$, which we refer to as *Gender Direction*. This test regards an embedding model to be better debiased if the alignment is larger for the word pair of Def compared to the word pairs of None and Stereo. By following the past practices, we test models with 40 instances from a subset of *Sembias*, whose gender word pairs are not used for training. To investigate the result of *Sembias* analogy test in Spanish and Korean, we translated the words in *Sembias* into the other languages with human corrections.

Table 1 shows the percentages of the largest alignment with *Gender Direction* for all instances. For English, CF-Debias-LA selects all the pairs of Def, which shows the sufficient maintenance of the gender information for those words. Also, CF-Debias-LA selects neither stereotype nor none-type words, so the difference vectors of Stereo and None always have less alignment to *Gender Direction* than the difference vectors of Def. We further refer to the experimental settings of Spanish and Korean in Appendix J.

[6] We provided link of the authors' implementations in Appendix H.

[7] We use the subtraction method as an ATT-Debias.

	career vs family		math vs art		science vs art		intellect vs appear		strong vs weak	
Embeddings	p-value	d	p-value	d	p-value	d	p-value	d	p-value	d
Original	0.000	1.605	0.276	0.494	0.014	1.260	0.009	0.706	0.067	0.640
Hard-Debias	0.100	0.842	0.090	-1.043	0.003	-0.747	0.693	-0.121	0.255	0.400
GN-Debias	0.000	1.635	0.726	-0.169	0.081	1.007	0.037	0.595	0.083	0.620
ATT-Debias	0.612	0.255	0.007	-0.519	0.000	0.843	0.129	0.440	0.211	0.455
CPT-Debias	0.004	1.334	0.058	1.029	0.000	1.417	0.001	0.906	0.654	-0.172
AE-Debias	0.000	1.569	0.019	0.967	0.024	1.267	0.007	0.729	0.027	0.763
AE-GN-Debias	0.001	1.581	0.716	0.317	0.139	0.639	0.006	0.770	0.028	0.585
GP-Debias	0.000	1.567	0.019	0.966	0.027	1.253	0.006	0.733	0.028	0.758
GP-GN-Debias	0.000	1.599	**0.932**	0.109	0.251	0.591	0.004	0.791	0.098	0.610
CF-Debias	0.210	0.653	0.759	0.261	0.725	−0.363	0.256	-0.328	0.305	0.371
CF-Debias-LA	**0.874**	**-0.089**	0.669	-0.125	0.360	0.480	0.678	-0.124	**0.970**	**0.013**
CF-Debias-KA	0.196	0.673	0.887	**0.083**	**0.919**	**-0.235**	**0.893**	**-0.039**	0.373	0.338

Table 2: WEAT hypothesis test results for five gender-stereotypical word categories. The best and second-best models are indicated as boldface and underline, respectively. The absolute value of the effect size denotes the degree of bias. A value of d closer to 0 means that there is no gender bias.

4.3.2 WEAT

We apply the Word Embedding Association Test (WEAT) (Caliskan et al., 2017b) for debiasing test. WEAT uses permutation test to compute the effect size (d) and p-value in Table 2, as a measurement of the bias in word embeddings. The effect size computes differential association of the sets of stereotypical target words, i.e. *career vs family*, and the gender word pair sets from Chaloner and Maldonado (2019a). A higher value of effect size indicates a higher gender bias between the two sets of target words. The p-value is used to check the significant level of bias. We provide the detailed description of WEAT in Appendix C. The variations of our method show the best performances for whole categories except *math vs art*, see Table 2.

Embeddings	no gender bias	semantic validity
Original	0.447±0.179	**0.875**±0.132
Hard-Debias	0.491±0.142	0.652±0.123
ATT-Debias	0.610±0.136	0.761±0.131
CPT-Debias	0.552±0.128	0.827±0.138
GP-GN-Debias	0.328±0.241	0.421±0.149
CF-Debias-LA	**0.644**±0.124	0.683±0.152
CF-Debias-KA	0.615±0.107	0.744±0.142

Table 3: Human-based evaluation for the gender bias and semantics of generated analogy, with standard deviation. The best model is indicated as boldface.

4.3.3 Analogy Test with Human based Validation

We conducted a human experiment on the analogy generated by the debiased embeddings to evaluate the debiasing efficacy of each embedding. each embeddings generate a word based on the question "a is to b as c is to what?", when words a, b are selected from the gender word pairs of *Sembias* dataset; and c is given as a gender stereotypical word, i.e. homemaker, housekeeper, from Bolukbasi et al. (2016). The answer word from each question is generated by $argmax_{d \in V}(\overrightarrow{d} \cdot (\overrightarrow{c} - \overrightarrow{a} + \overrightarrow{b}))$. 18 Human subjects were asked to evaluate the generated analogies from two perspectives; 1) existence of gender bias in the analogy, 2) semantic validity of the analogy.[8] Table 3 shows that our method indicates the least gender bias while competitively maintaining the semantic validity.

4.4 Debiasing Qualitative Analysis

To demonstrate the indirect gender bias in the word embedding, we perform two qualitative analyses from Gonen and Goldberg (2019). We take the top 500 male-biased words and the top 500 female-biased words, which becomes a word collection of the top 500 and the bottom 500 inner product between the word embeddings and $\overrightarrow{he} - \overrightarrow{she}$. From the debiasing perspective, these 1,000 word vectors should not be clustered distinctly. Therefore, we create two clusters with K-means and check the heterogeneity of the clusters through the cluster majority classification. The left side on Figure 4 shows that CF-Debias-KA generates a gender-invariant embedding for gender-biased wordsets by showing the lowest cluster classification accuracy.

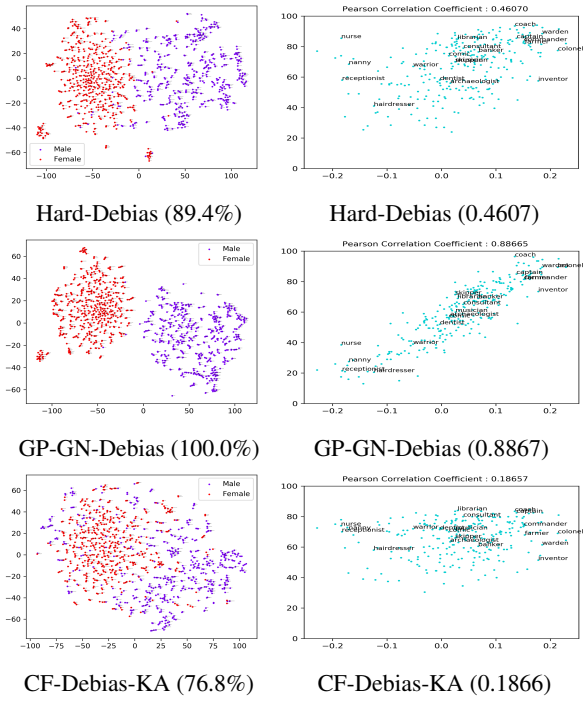
Gonen and Goldberg (2019) demonstrates that the original bias[9] has a high correlation with

[8] We enumerate the embeddings utilized in an experiment and detailed description of the human experiment in Appendix I.

[9] the dot-product between the original word embedding from GloVe and $\overrightarrow{he} - \overrightarrow{she}$

	POS Tagging		POS Chunking		Named Entity Recognition	
Embeddings	Δ F1	Δ Recall	Δ F1	Δ Recall	Δ F1	Δ Recall
Hard-Debias	-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-Debias	-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-Debias	-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-Debias	-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-Debias	-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-Debias	-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-Debias	-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-Debias	-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-Debias	-0.327±0.248	-0.621±0.564	**0.000**±0.000	**−0.001**±0.001	**0.000**±0.000	**−0.001**±0.001
CF-Debias-LA	-0.287±0.118	-0.506±0.260	-0.002±0.001	-0.006±0.004	-0.002±0.001	-0.007±0.005
CF-Debias-KA	**−0.123**±0.135	**−0.186**±0.208	**0.000**±0.000	**−0.001**±0.001	**0.000**±0.000	**−0.001**±0.001

Table 4: Performance degradation percentage with standard deviation for downstream tasks of POS Tagging, POS Chunking, and NER. The best model is indicated as boldface.



Hard-Debias (89.4%) Hard-Debias (0.4607)

GP-GN-Debias (100.0%) GP-GN-Debias (0.8867)

CF-Debias-KA (76.8%) CF-Debias-KA (0.1866)

Figure 4: The t-SNE views for 500 male, female-biased word embeddings from original embedding, with the cluster-based classification accuracy in parentheses. (left) The percentage of male neighbors for each profession as a function of original bias, with the Pearson correlation coefficient in parentheses. (right)

the male/female ratio of the gender-biased words among the nearest neighbors of the word embedding. The right side of Figure 4[10] shows each profession word at (the dot-product, the male/female ratio). CF-Debias-KA shows the minimal Pearson correlation coefficient between the two axes.

[10]Full plots of other baselines for two qualitative analyses are available in Appendix E and F, respectively.

4.5 Downstream Task of Debiased Word Embeddings

We compared multiple downstream task performances of the original and the debiased word embeddings, to check the ability to preserve semantic information in debiasing procedures. Following CoNLL 2003 shared task (Sang and Erik, 2002), we selected Part-Of-Speech tagging, Part-Of-Speech chunking, and Named Entity Resolution as our tasks. Table 4 shows that there are constant performance degradation effects for all debiasing methods from the original embedding. However, our methods minimized the degradation of performances across the baseline models. Especially, CF-Debias-KA shows the minimal performance degradations by utilizing the nonlinear alignment regularization.

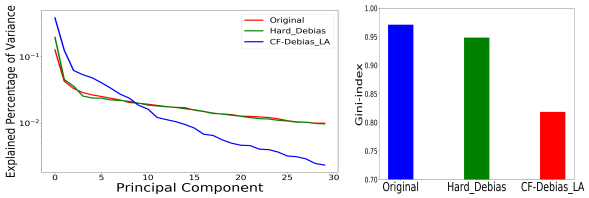


Figure 5: The proportion (Left) and Gini-index (Right) from the variance vector for top 30 PCs of difference vectors for gender word pairs

4.6 Analyses on Alignment Regularization

If the difference vectors of gender word pairs are not linearly aligned, the gender direction vector $\boldsymbol{v}_g$ in Eq. (11) cannot be a pure direction of the gender information. Hence, we compared the variances explained by the top 30 principal components (PC) of difference vectors for gender word pairs, as a measurement for the linear alignment. The left plot

in Figure 5 shows the proportion of variances from each PC. Our method shows the largest concentration of the variances on a few components, other than Hard-Debias and Original embedding. The right plot in Figure 5 shows Gini-index (Gini, 1912) for the variance proportion vector from PCs. Our method shows minimal Gini-index, which indicates the monopolized proportion of variances.

Also, Figure 6 shows two example plots of a selected gender word pairs in the original embedding space (Upper) and the CF-Debias-LA embedding space (Lower), by Locally Linear Embedding (LLE), (Roweis and Saul, 2000). The lower plot in Figure 6 shows the consistency of the gender direction, and the plot visually describes the neutralization of *housekeeper, statistician* by utilizing the counterfactually augmented word embeddings.

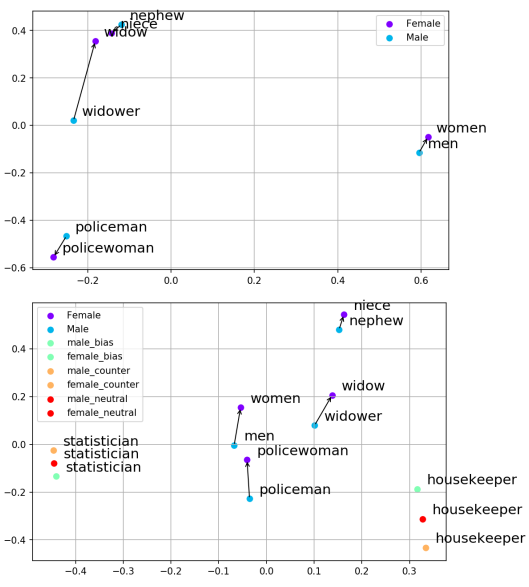


Figure 6: LLE projection view of selected gender word pairs and biased word for original embedding space (left) and debiased embedding space (right)

5 Conclusions

This work contributes to natural language processing society in two folds. For gender debiasing application, our model produces the debiased embeddings that has the most neutral gender latent information as well as the efficiently maintained semantics for the various NLP downstream tasks. For methodological modeling, CF-Debias suggests a new method of disentangling the latent information of word embeddings with the gradient reversal layer and creating the counterfactual embeddings by exploiting the geometry of the embedding space. It should be noted that these types of latent modeling can be applied to diverse natural language tasks to control expressions on emotions, prejudices, ideologies, etc.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1C1B6008652)

References

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

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. Säckinger, 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. 2019a. 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.

Kaytlin Chaloner and Alfredo Maldonado. 2019b. 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.

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.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

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

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

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.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. 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.

A The Derivation of Principal Component on Kernelized Alignment

Let's assume that we want to align a word embedding $\boldsymbol{w}'$ to the set of the word embeddings $\{\boldsymbol{w}_i\}_{i=1}^N$. Then, we introduce nonlinear mapping function f, which projects a word embedding $\boldsymbol{w}_i \in \mathbb{R}^d$ into a newly introduced feature space, $f(\boldsymbol{w}_i) \in \mathbb{R}^m$. If we assume that the mapped outputs from the word embeddings $\{f(\boldsymbol{w}_i)\}_{i=1}^N$ are zero-centered, the covariance matrix can be estimated as follows:

$$\Sigma_f = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{w}_i) f(\boldsymbol{w}_i)^T$$

Same as the main paper, we set ϕ_k and λ_k to be k-th eigenvector and eigenvalue for the projected outputs of the given embeddings $\{f(\boldsymbol{w}_i)\}_{i=1}^N$, respectively. Then, we can get following equation, which describes the eigen-decomposition of the covariance matrix.

$$\begin{aligned}\Sigma_f \phi_k &= \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{w}_i) f(\boldsymbol{w}_i)^T \phi_k \\ &= \frac{1}{N}\sum_{i=1}^{N} (f(\boldsymbol{w}_i)\cdot\phi_k) f(\boldsymbol{w}_i) = \lambda_k \phi_k\end{aligned}$$

From above function, ϕ_k can be represented as a linearly-weighted combination of the N mapped outputs of word embeddings as follows:

$$\phi_k = \frac{1}{N\lambda_k}\sum_{i=1}^{N} (f(\boldsymbol{w}_i)\cdot\phi_k) f(\boldsymbol{w}_i)$$

Then, we multiply $f(\boldsymbol{w}_j)$ for $j = 1, ..., N$ to both sides of the equation.

$$\begin{aligned}f(\boldsymbol{w}_j)\cdot\phi_k &= \frac{1}{N\lambda_k} f(\boldsymbol{w}_j)\sum_{i=1}^{N} (f(\boldsymbol{w}_i)\cdot\phi_k) f(\boldsymbol{w}_i) \\ &= \sum_{i=1}^{N} \frac{1}{N\lambda_k} (f(\boldsymbol{w}_i)\cdot\phi_k)(f(\boldsymbol{w}_i)\cdot f(\boldsymbol{w}_j))\end{aligned}$$

We can replace an inner-product of the two mapped outputs, $(f(\boldsymbol{w}_i)\cdot f(\boldsymbol{w}_j))$, into kernel $\mathrm{k}(\boldsymbol{w}_i, \boldsymbol{w}_j)$, which represents an inner product of two vectors in the projected space, for the case when computing mapped results of given data is complex or impossible.

$$f(\boldsymbol{w}_j)\cdot\phi_k = \sum_{i=1}^{N} \frac{1}{N\lambda_k} (f(\boldsymbol{w}_i)\cdot\phi_k)\, \mathrm{k}(\boldsymbol{w}_i, \boldsymbol{w}_j)$$

By letting $a_k^i = \frac{1}{N\lambda_k}(f(\boldsymbol{w}_i)\cdot\phi_k)$, we get

$$f(\boldsymbol{w}_j)\cdot\phi_k = \lambda_k N a_k^j = \sum_{i=1}^{N} a_k^j\, \mathrm{k}(\boldsymbol{w}_i, \boldsymbol{w}_j)$$

The above equation can be represented as the j-th component of the k-th eigenvector-decomposition problem of $\boldsymbol{K}$, which is a matrix of $N \times N$ kernel elements $\mathrm{k}(\boldsymbol{w}_i, \boldsymbol{w}_j)$ for $i, j = 1, ..., N$. See the below equation, which is k-th eigenvector-decomposition problem of $\boldsymbol{K}$, when $\boldsymbol{a}_k = [a_k^1, ..., a_k^N]^T$.

$$\lambda_k N \boldsymbol{a}_k = \boldsymbol{K}\, \boldsymbol{a}_k$$

This implication means that a_k^j is j-th component of k-th eigenvector of $\boldsymbol{K}$ and we can compute a_k^j by solving eigen-decomposition problem of $\boldsymbol{K}$.

Substituting $f(\boldsymbol{w}_j)$ on above equation into $f(\boldsymbol{w}')$, which is mapped result of the target word embedding $\boldsymbol{w}'$, we get PC_k, k-th principal component of new word embedding $\boldsymbol{w}'$ on the projected space as follows:

$$\begin{aligned}PC_k = f(\boldsymbol{w}')\cdot\phi_k &= \Sigma_{i=1}^N a_k^i f(\boldsymbol{w}_i)\cdot f(\boldsymbol{w}') \\ &= \Sigma_{i=1}^N a_k^i\, \boldsymbol{K}(\boldsymbol{w}_i, \boldsymbol{w}') \quad (15)\end{aligned}$$

It should be noted that above derivation is based on Schölkopf et al. (1998). The proposed Kernelized alignment can be seen as an example which applies an nonlinear alignment to the word embeddings, by utilizing the kernel trick provided from Schölkopf et al. (1998).

B Notation table

Notation	Description
$\boldsymbol{w}_f$	The embedding of *feminine* word
$\boldsymbol{w}_m$	The embedding of *masculine* word
$\boldsymbol{w}_n$	The embedding of *gender neutral* word
V_f	The *feminine* word set
V_m	The *masculine* word set
V_n	The *gender neutral* word set
$\boldsymbol{z}_f^s$	The semantic latent variable of w_f
$\boldsymbol{z}_m^s$	The semantic latent variable of w_m
$\boldsymbol{z}_n^s$	The semantic latent variable of w_n
$\boldsymbol{z}_f^g$	The gender latent variable of w_f
$\boldsymbol{z}_m^g$	The gender latent variable of w_m
$\boldsymbol{z}_n^g$	The gender latent variable of w_n
$\neg\boldsymbol{z}_n^g$	The counterfactual-gender latent variable
$\hat{\boldsymbol{w}}_f$	The reconstructed word embedding of $\boldsymbol{w}_f$
$\hat{\boldsymbol{w}}_m$	The reconstructed word embedding of $\boldsymbol{w}_m$
$\hat{\boldsymbol{w}}_n$	The reconstructed word embedding of $\boldsymbol{w}_n$
$\hat{\boldsymbol{w}}_{cf}$	The counterfactually reconstructed word embedding
$\hat{\boldsymbol{w}}_{neu}$	The gender neutralized word embedding
g_f	The output of gender classifier for z_f^g
g_m	The output of gender classifier for z_m^g
$\boldsymbol{v}_g$	The gender direction vector
Ω	The gender word pairs set
E	The encoder of our method
D	The decoder of our method
C_r	The auxilary gender classifier
C_a	The gender latent generator

Table 5: The description of the notations in this paper.

C WEAT Hypothesis test

WEAT hypothesis (Caliskan et al., 2017b) test quantifies the bias with effect size and p-value. We can compute the effect size of the two target words set against two attribute words set. To quantify the gender bias, we use (Chaloner and Maldonado, 2019b) subset of masculine (A_1) and feminine words(A_2) as an attribute words, and use career (T_1) and family (T_2) related words for target words set. We compare the effect size and p-value for different experiment environment by changing the attribute words, as shown in Table 2 in the paper.

We can compute the association measure s, between target word t and the attribute word set as follows:

$$s(t) = \frac{1}{|A_1|}\sum_{a_1 \in A_1} \cos(t, a_1) - \frac{1}{|A_2|}\sum_{a_2 \in A_2} \cos(t, a_2)$$

We compute the effect size, the degree of bias, based on the difference between mean of association value as follows:

$$\frac{\text{Mean}_{t_1 \in T_1} s(t_1) - \text{Mean}_{t_2 \in T_2} s(t_2)}{\text{std}_{t \in T_1 \cup T_2} s(t)}$$

To check the significant level of bias, we need to compute the test statistics, $s(T_1, T_2)$, and one-sided p-value. We compute the p-value based on $\{T_1^{(i)}, T_2^{(i)}\}$, the all partition of $T_1 \cup T_2$ as follows:

$$s(T_1, T_2) = \sum_{t_1 \in T_1} s(t_1) - \sum_{t_2 \in T_2} s(t_2)$$

$$\text{p-value} = P\{|s(T_1^{(i)}, T_2^{(i)})| > |s(T_1, T_2)|\}$$

If the word embedding has a conventional gender bias, effective size can have a positive value, and negative value, otherwise. To measure the gender bias properly, we need to consider both of conventional gender bias, and anti-conventional gender bias. We compute the p-value based on the absolution value of test statistics to measure gender bias properly.

D Performance Test Result for Gender Classifier C_r

To test gender indicating the ability of the gender classifier $C_r : \boldsymbol{z}^g \rightarrow [0, 1]$, we tested indicating accuracy of the gender-definition words, i.e., he, she, etc.; and gender-stereotypical words, i.e., doctor, nurse, etc. We utilized 53 gender word pairs as test word pairs from entire gender word pairs, utilizing the remaining words for training. We selected well known gender-biased occupation words for examples of gender-stereotypical words, 10 for each gender case as follows:

$[doctor, programmer, boss, maestro, warrior, john, politician, statistician, athlete, nurse, homemaker, cook, cosmetics, dancer, mary, violinist, housekeeper, secretary]$.

The test accuracy for gender-definition words are 0.8490, 0.8867 for masculine and feminine words, respectively. For gender-stereotypical words, C_r indicates correct gender biases for all male-biased words except the word *athlete* and all female-biased words. Figure 7 shows the visual separation of gender latent variables for masculine words and feminine words.

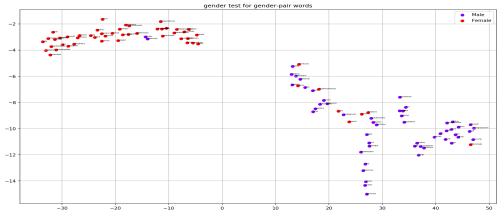
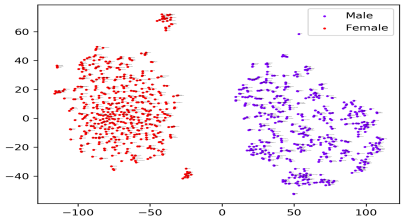
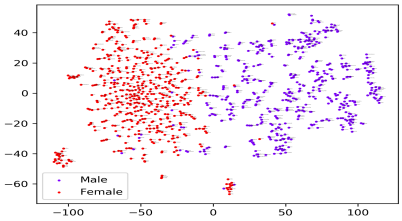


Figure 7: The t-SNE projection view of gender latent variables of the test gender word pairs

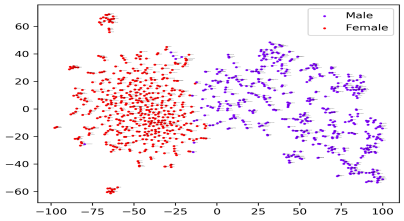
E Full Plots for the Clustering Analysis



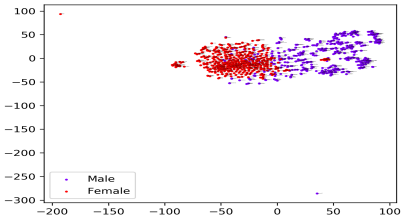
(a) Original (99.9%)



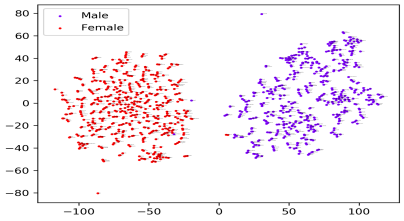
(b) Hard-Debias (89.4%)



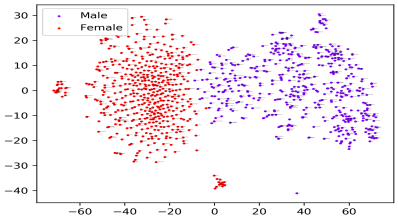
(c) GN-Debias (97.3%)



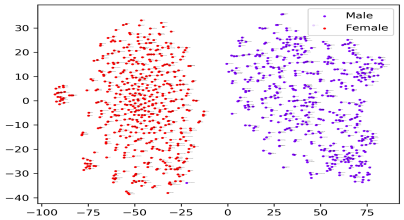
(d) ATT-Debias (99.1%)



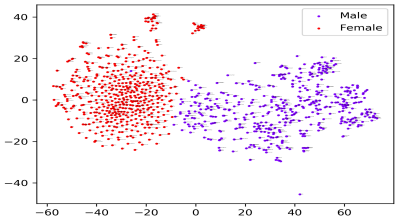
(e) CPT-Debias (100.0%)



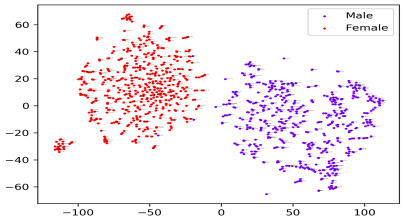
(f) AE-Debias (92.6%)



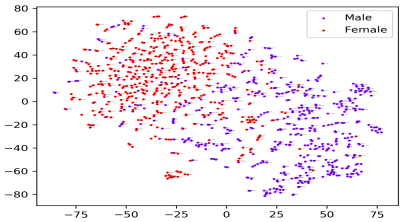
(g) AE-GN-Debias (100.0%)



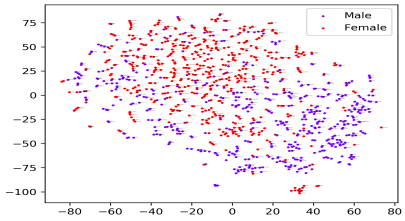
(h) GP-Debias (92.5%)



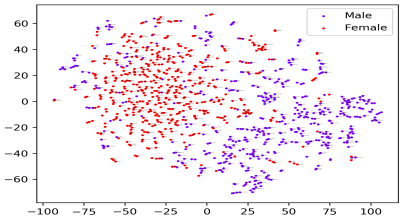
(i) GP-GN-Debias (100.0%)



(j) CF-Debias (78.1%)



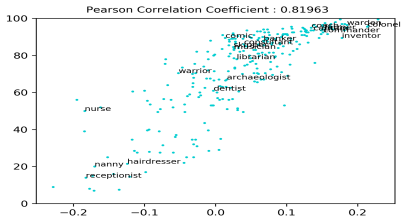
(k) CF-Debias-LA (63.1%)



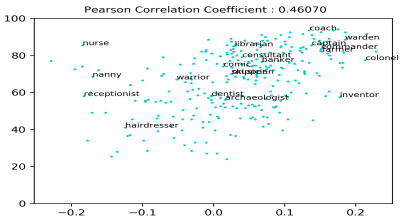
(l) CF-Debias-KA (76.8%)

Figure 8: 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.

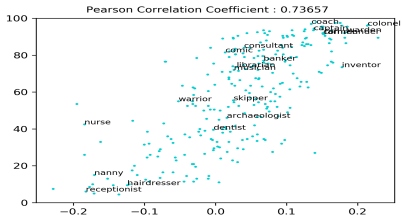
F Full Plots for Correlation Analysis between Original Bias and Nearest Neighbors



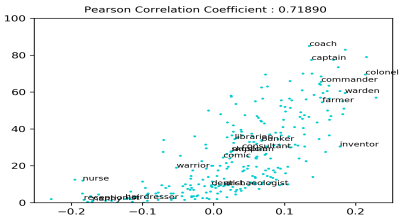
(a) Original (0.8196)



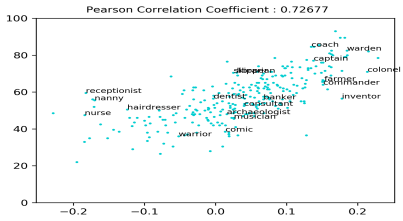
(b) Hard-Debias (0.4607)



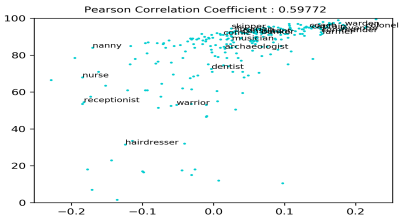
(c) GN-Debias (0.7366)



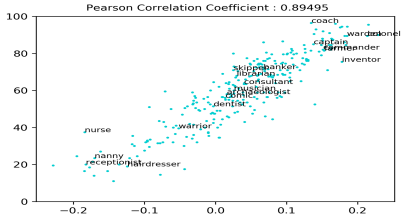
(d) ATT-Debias (0.7189)



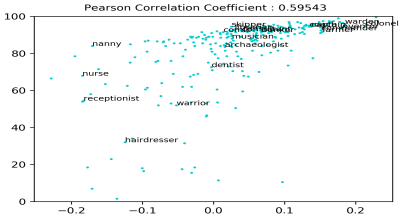
(e) CPT-Debias (0.7268)



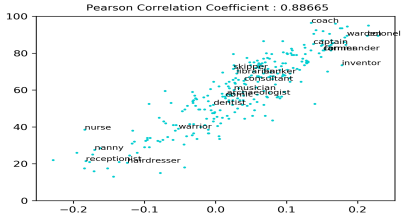
(f) AE-Debias (0.5977)



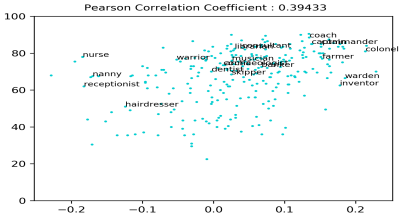
(g) AE-GN-Debias (0.8950)



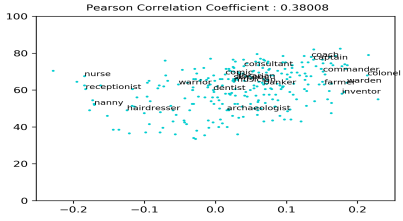
(h) GP-Debias (0.5954)



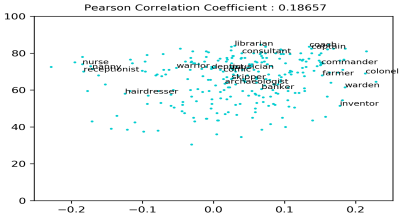
(i) GP-GN-Debias (0.8867)



(j) CF-Debias (0.3943)



(k) CF-Debias-LA (0.3801)



(l) CF-Debias-KA (0.1865)

Figure 9: The percentage of male neighbors for each profession as a function of original bias for whole embeddings, we show only a limited number of professions on the plot to make it readable. The pearson correlation coefficient is added in parenthesis.

G Experimental Setup for Our Method

We implement the encoder E and the decoder D with one hidden layer and hyperbolic tangent function as an activation function. The generators C_a and C_g are implemented as feed-forward neural network with one hidden layer, followed by the hyperbolic tangent function as an activation function. The gender classifier C_r is similarly implemented as the feed-forward neural network with one hidden layer, followed by sigmoid activation function for the output layer. The whole training was performed using the Adam optimizer with learning rate 10^{-5}. We trained our model using a single Titan-RTX GPU. Each run takes approximately 2 hours including the time for saving the post-processed word embeddings.

As described in Appendix D, to test classification accuracy of the gender classifier $C_r : \boldsymbol{z}^g \rightarrow [0, 1]$ for gender-definition words and gender stereotypical words, we only used 143 gender word pairs from entire gender word pairs on the training procedure. The remaining 53 gender word pairs were utilized for gender classification test in Appendix D.

H The Link of Implementation for Each Baseline

Hard-GloVe : https://github.com/tolga-b/debiaswe.
GN-GloVe : https://github.com/uclanlp/gn_glove.
CPT-GloVe : https://github.com/jsedoc/ConceptorDebias.
ATT-GloVe : https://github.com/sunipa/Attenuating-Bias-in-Word-Vec.
AE-GloVe, AE-GN, GP-GloVe and GP-GN : https://github.com/kanekomasahiro/gp_debias.

I The Experimental Setting of Human Experiment

We conducted an human validation test on the linear analogies generated by the debiased embeddings to evaluate debiasing efficacy of each embedding. For the question "a is to b as c is to ?", words a, b were selected from gender word pairs of $Sembias$ dataset and c was sampled from gender stereotypical words, i.e. homemaker, given by Bolukbasi et al. (2016).

The question word is chosen from $argmax_{d \in V}(\overrightarrow{d} \cdot (\overrightarrow{c} - \overrightarrow{a} + \overrightarrow{b}))$. In order to enable human subjects to efficiently compare generated words of each debiased word embedding, We compared only 5 baseline methods; Original GloVe embedding, Hard-Debias, ATT-Debias, CPT-Debias, GP-GN-Debias with our methods; CF-Debias-LA and CF-Debias-KA. As stated in section 4.4 of main paper, 18 Human subjects were asked to evaluate the 84 generated analogies from two perspectives; 1) the existence of gender bias on generated analogy, 2) the semantic validity of analogy. The semantic validity in our experiment equals to the question, "Is it possible to infer semantic relationship from generated analogy?". The representative examples of the analogy questions are given as follows: "man is to $woman$ as $boss$ is to ?" , "$female$ is to $male$ as $weak$ is to ?".

J The Experimental Settings for Other Languages; Spanish and Korean

We used Fasttext (Bojanowski et al., 2016) pre-trained on $CommonCrawl$ and $Wikipedia$, with 300 dimensional embeddings for 2,000,000 unique words for the experiments of Spanish. Also, we used Fasttext (Bojanowski et al., 2016) pre-trained on $Wikipedia$, with 300 dimensional embeddings for 879,125 unique words for the experiments of Korean. For the gender word pairs required for gender debiasing, the query words used in the English version were translated into Spanish and Korean. In this procedure, some words, which are not present in the given corpus, were excluded.