Word Embeddings via Causal Inference: Gender Bias Reducing and Semantic Information Preserving

Lei Ding[1], Dengdeng Yu[3], Jinhan Xie[1],Wenxing Guo[1], Shenggang Hu[2], Meichen Liu[1], Linglong Kong[1], Hongsheng Dai[2], Yanchun Bao[2], Bei Jiang[1]

[1]Department of Mathematical and Statistical Sciences, University of Alberta
{lding1,jinhan3,wenxing2,meichen,lkong,bei1}@ualberta.ca
[2]Department of Mathematical Sciences, University of Essex
{sh19509, hdaia, ybaoa}@essex.ac.uk
[3] Department of Mathematics, University of Texas at Arlington
{dengdeng.yu}@uta.edu

Abstract

With widening deployments of natural language processing (NLP) in daily life, inherited social biases from NLP models have become more severe and problematic. Previous studies have shown that word embeddings trained on human-generated corpora have strong gender biases that can produce discriminative results in downstream tasks. Previous debiasing methods focus mainly on modeling bias and only implicitly consider semantic information while completely overlooking the complex underlying causal structure among bias and semantic components. To address these issues, we propose a novel methodology that leverages a causal inference framework to effectively remove gender bias. The proposed method allows us to construct and analyze the complex causal mechanisms facilitating gender information flow while retaining oracle semantic information within word embeddings. Our comprehensive experiments show that the proposed method achieves state-of-the-art results in gender-debiasing tasks. In addition, our methods yield better performance in word similarity evaluation and various extrinsic downstream NLP tasks.

Introduction

Word embeddings are dense vector representations of words trained from human-generated corpora (Mikolov et al. 2013; Pennington, Socher, and Manning 2014). Word embeddings have become an essential part of natural language processing (NLP). However, it has been shown that stereotypical bias can be passed from human-generated corpora to word embeddings (Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018; Zhao et al. 2019).

With wide applications of NLP systems to real life, biased word embeddings have the potential to aggravate and possibly cause serious social problems. For example, translating 'He is a nurse' to Hungarian and back to English results in 'She is a nurse' (Douglas 2017). In word analogy tasks appears in Bolukbasi et al. (2016), wherein $\overrightarrow{she}$ is closer to $\overrightarrow{nurse}$ than $\overrightarrow{he}$ is to $\overrightarrow{doctor}$. Zhao et al. (2018) shows that biased embeddings can lead to gender-biased identification outcomes in co-reference resolution systems.

Current studies on word embedding bias reductions can be divided into two camps: word vector learning methods (Zhao et al. 2018) and post-processing algorithms (Bolukbasi et al. 2016; Kaneko and Bollegala 2019). Word vector learning methods are time-consuming and suffer from the high computational cost required to train word embeddings from scratch. To overcome these limitations, post-processing algorithms have emerged as popular alternatives. Yang and Feng (2020), for example, proposes a simple and efficient algorithm that projects embeddings into a space that is orthogonal to gender-specific words such as *mother* and *father* and is successful in reducing gender bias.

However, the critical issue of using gender-specific word vectors remains: information on gender and semantics entangled within these words. For example, the gendered word pair *bride* and *bridegroom* exhibits gender information as well as semantic information pertaining to weddings. Therefore, eliminating gender information through pairs of gendered words such as *policeman* and *policewoman* or *wizard* and *witch*, also eliminates intrinsic semantic information: this is clearly not ideal.

As a solution, we propose utilizing the differences between vectors corresponding to paired gender-specific words to better eliminate gender bias while retaining important semantic information. These differences are between embedded vectors for male- and female-gendered words, such as $\overrightarrow{father-mother}$ or $\overrightarrow{bridegroom-bride}$. As a motivating example, Table 1 demonstrates that this simple change from gender-specific word vectors to the differences between word-pair vectors indeed retains more semantic information than do state-of-the-art post-processing frameworks (Yang and Feng 2020).

In this paper, we propose novel causal frameworks for reducing bias in word embeddings while maximally preserving semantic and lexical information. Our contributions are summarized as follows.

- We develop two causal inference frameworks for reducing biases in word embeddings that improves upon existing state-of-the-art methods.
- We find an intuitive and effective way to better represent gender-related information that needs to be removed and

	Task 1 *Wedding*	Task 2 *Service*	Task 3 *Family*	Task 4 *Religion*
Oracle	11.22 (0.20)	9.96 (0.11)	13.51 (0.30)	20.27 (0.30)
DeSIP	7.01 (0.15)	6.67 (0.10)	10.69 (0.25)	13.59 (0.25)
HSR	4.34 (0.14)	5.61 (0.10)	8.90 (0.22)	9.85 (0.20)
Win-loss	100.00%	99.00%	100.00 %	100.00%

Table 1: Semantic information preservation experiment results. For each of the four pre-determined words *Wedding*, *Service*, *Family*, and *Religion*, we identify the top 200 most cosine-correlated words. For each of the 200 words, we fit a ridge regression against gender-specific words defined in Yang and Feng (2020) (HSR), and a linear regression against the differences between gender-specific word pairs from this paper (DeSIP). The fitted word vectors are used as reduced-bias word vectors. To quantify the semantic information preservation, the mean absolute dot product between the pre-determined words and their bias-reduced versions over the 200 most related words are presented, with standard errors in parentheses. Note that, the oracle preservation semantic information is achieved by using the original word vector instead of the fitted one. The last row shows the proportion of these 200 words for which DeSIP outperforms HSRwith respect to semantic information preservation.

use this approach to achieve oracle-like semantic and lexical information retention.

- We show that our methods outperform other *state-of-the-art* debiasing methods in various downstream NLP tasks.

The rest of this paper is organized as follows. We first present a thorough review of current studies on word embedding bias evaluation and debiasing algorithms. We then define two types of bias and propose frameworks for dealing with each. The comprehensive experimental results on a series of gender bias evaluation and semantic evaluation tasks demonstrate the effectiveness of our proposed methods.

Related Works

Quantifying Gender Bias

Numerous studies have demonstrated that word embeddings trained by human-generated corpora exhibit human stereotype bias. Caliskan, Bryson, and Narayanan (2017) develops the Word Embedding Association Test (WEAT) as an analogue to the Implicit Association Test used in psychology (Greenwald, McGhee, and Schwartz 1998) to detect implicit stereotypes. WEAT measures the association between a word and an attribute using cosine similarity; the test compares two sets of target words against a pair of attribute sets.

Bolukbasi et al. (2016) applies word analogy tests as a way to demonstrate bias. The task uses a word embedding to find an output to pair with a given input word, say, *doctor*, such that the (target, output) pair is in analogy to the gender pair (he, she). The word embedding passes the test if the output is stereotype-free, say, *physician* instead *nurse* for the input *doctor*. However, this task requires crowd-sourcing to set the benchmark and has been replaced by other evaluation methods in more recent works.

Another approach from Bolukbasi et al. (2016) for evaluating gender bias involves computing projections onto a gender direction, the difference between vector embeddings of a pair of gender-specific words (e.g., he and she, as the most widely accepted definition). This debiasing metric is used in many other studies (Manzini et al. 2019). Such a method has failed to become the gold standard because a "true" gender direction if it exists, is used in the evaluation.

Gonen and Goldberg (2019) later points out that direct projection does not eliminate gender bias from the geometry of the embedding and that biased words tend to cluster together even after debiasing. To account for this, the neighborhood bias metric was introduced to measure the bias of a word by counting the difference in the number of (socially) male- and female-biased neighbors among the word's K-nearest neighbors.

Prior Debiasing Methods

Bolukbasi et al. (2016) formulates the core idea underlying most later debiasing methods, which is to detect the subspace that contains the most information related to gender (or other stereotypes) (Manzini et al. 2019; Kaneko and Bollegala 2019). Other works have incorporated different considerations into their strategies, e.g., maximizing the distance between masculine and feminine words (Zhao et al. 2018), detecting gender direction using partial projection (Dev and Phillips 2019), or detecting and mitigating distortion in gender direction due to differences in word frequency (Wang et al. 2020). Shin et al. (2020) utilizes the idea of counterfactual generation in machine learning. This approach trains an autoencoder together with a classifier to disentangle latent gender information from semantic information and mitigating bias by generating a counterfactual copy of each gender-neutral words.

Yang and Feng (2020) approaches the problem using a causal framework in which it is assumed that latent gender information affects both gendered and gender-biased words. The model aims to recover gender-specific information in gender-biased words from the gendered words through a linear ridge regression.

However, Yang and Feng (2020) does not fully leverage the power of causal inference, as the authors' framework only models the subspace formed by gender information and overlooks the effect of gender on semantic information.

Methodology

Preliminary Definitions

We characterize two types of gender bias in the causal framework and propose algorithms for removing each type. Specifically, we use model intervention techniques to determine causal effects in a causal model. It is more manageable to apply the model intervention to proxy variables of the gender bias rather than the gender bias variables themselves (represented by the differences between gender-specific word pair vectors, such as $\overrightarrow{he}-\overrightarrow{she}$ or $\overrightarrow{male}-\overrightarrow{female}$), since the latter are generally regarded as inherited attributes for which interventions are often impossible in practice.

We consider five types of variables corresponding to five word-related matrices: an s_1-dimensional pure gender bias variable D with a corresponding matrix $\mathbf{D} \in \mathcal{R}^{N \times s_1}$ composed of pure gender bias vectors such as $\overrightarrow{he-she}$ and $\overrightarrow{male-female}$; an s_2-dimensional gender bias variable proxy P with a corresponding matrix $\mathbf{P} \in \mathcal{R}^{N \times s_2}$ composed of vectors that are directly influenced by D that should not affect the final prediction; an m-dimensional resolving, non-gender-specific word variable Z with a corresponding matrix $\mathbf{Z} \in \mathcal{R}^{N \times m}$ composed of vectors that are influenced by D in a manner that we accept as non-discriminatory; a d-dimensional, non-gender-specific word variable Y with a corresponding matrix $\mathbf{Y} \in \mathcal{R}^{N \times d}$ composed of word vectors potentially containing gender bias that needs to be removed, such as $\overrightarrow{nurse}$ and $\overrightarrow{engineer}$; and another p-dimensional, non-gender-specific word variable X with a corresponding matrix $\mathbf{X} \in \mathcal{R}^{N \times p}$ that may retain semantic information. Here N is the dimension of the word embedding vector, and s_1, s_2, m, d, and p are the sizes of the variables D, P, Z, Y and X, respectively.

It is clear that using the vectors in $\mathbf{D}$ can eliminate pure gender bias information contained in word embeddings. In this way, semantic information can be preserved. As shown in Figures 1 and 2, we generally allow influence along the pathway $D \to X \to Y$ in our framework. Motivated by Kilbertus et al. (2017) and these conventions, we introduce the following definitions.

Definition 1 *(Potential proxy bias.) A variable Y in a causal graph exhibits potential proxy bias if there exists a directed path from D to Y that is blocked by a proxy variable P and if Y itself is not a proxy.*

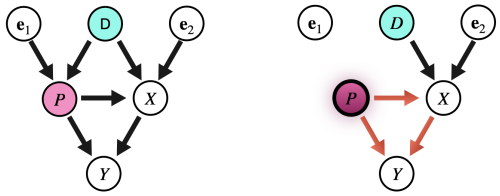
This definition indicates that potential proxy bias from P articulates a causal criterion that is in a sense dual to unresolved bias from Z.

Definition 2 *(Unresolved bias.) A variable Y in a causal graph exhibits unresolved bias if there exists a directed path from D to Y that is not blocked by a resolving variable Z and Y itself is non-resolving.*

This definition implies that all paths from a gender-bias variable D are problematic unless they are justified by a resolving variable Z.

Removing Potential Proxy Bias

We now develop a practical procedure for removing proxy bias in a linear structural equation model. For each $\mathbf{y} \in \mathcal{R}^N$, the column vector of $\mathbf{Y}$, it can be decomposed into two parts as $\mathbf{y} = \mathbf{y}_\Delta + \mathbf{y}_{\Delta^\perp}$, where $\mathbf{y}_\Delta$ and $\mathbf{y}_{\Delta^\perp}$ are the projections of $\mathbf{y}$ onto the mutually orthogonal spaces Δ and $\Delta^\perp$. In particular, let $\boldsymbol{\phi}_j \in \mathcal{R}^N$ denote the basis vectors for Δ and $\boldsymbol{\psi}_{j'} \in \mathcal{R}^N$ denote the basis vectors for $\Delta^\perp$. The whole space $\Omega = \Delta \cup \Delta^\perp$. We can write $\mathbf{y} = \sum_{j:\boldsymbol{\phi}_j \in \Delta} \xi_j \boldsymbol{\phi}_j + \sum_{j':\boldsymbol{\psi}_{j'} \in \Delta^\perp} \kappa_{j'} \boldsymbol{\psi}_{j'}$, where $\xi_j, \kappa_{j'} \in \mathcal{R}$. In this paper, we take $\Delta = \mathrm{Span}(\mathbf{D})$, namely, the linear space spanned by the column vectors of $\mathbf{D}$. Consequently, $\Delta^\perp$ contains the semantic information not described by $\mathbf{D}$. As bias reduction is primarily concerned with reducing bias along paths starting from D, we do not remove information from $\mathbf{y}_{\Delta^\perp}$.



(a) Proxy bias (b) Intervention on proxy bias

Figure 1: A causal graph for proxy bias removal.

We next propose an algorithm for debiasing non-gender-specific word vectors $\mathbf{y}$. As illustrated in Figure 1, the corresponding linear structural equations are

$$\begin{aligned} \mathbf{P} &= \mathbf{D}\boldsymbol{\alpha}_0 + \boldsymbol{e}_1 \\ \mathbf{X} &= \mathbf{D}\boldsymbol{\alpha}_1 + \mathbf{P}\boldsymbol{\alpha}_2 + \boldsymbol{e}_2 \\ \mathbf{Y} &= \mathbf{P}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\beta}_2, \end{aligned} \tag{1}$$

where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are unobserved errors and $\boldsymbol{\alpha}_0 \in \mathcal{R}^{s_1 \times s_2}$, $\boldsymbol{\alpha}_1 \in \mathcal{R}^{s_1 \times p}$, $\boldsymbol{\alpha}_2 \in \mathcal{R}^{s_2 \times p}$, $\boldsymbol{\beta}_1 \in \mathcal{R}^{s_2 \times d}$ and $\boldsymbol{\beta}_2 \in \mathcal{R}^{p \times d}$ are parameters. Here, we note that the proxy matrix $\mathbf{P}$ contains vectors of words that are direct descendants of $\mathbf{D}$ and should not affect the prediction of $\mathbf{Y}$. In this paper, we prespecify $\mathbf{P}$ using the gendered-word pairs listed in Zhao et al. (2018). We build predictors that remove proxy bias by intervening on P, that is, by setting $P = p'$, where p' is a random variable: this is similar to the approach in Kilbertus et al. (2017). In particular, we want to guarantee that P has no overall influence on the prediction of the non-gender-specific variable Y by adjusting the $P \to Y$ pathway to cancel the influence along $P \to X \to Y$. We do not generally prohibit the potential for the gender bias variable D to influence the non-gender-specific variable Y in this case: see Figure 1. The non-gender-specific word matrix $\widehat{\mathbf{Y}}$ with potential proxy bias removed is

$$\widehat{\mathbf{Y}} = (\mathbf{X} - \mathbf{P}\widehat{\boldsymbol{\alpha}}_2)\widehat{\boldsymbol{\beta}}_2, \tag{2}$$

where the parameters $\widehat{\boldsymbol{\alpha}}_2$ and $\widehat{\boldsymbol{\beta}}_2$ are estimated by partial least squares (PLS), a regression method that is a supervised dimension reduction technique and works particularly well when variable dimensionality is large and sample size is small (Vinzi et al. 2010). However, since the debiasing procedure above does not retain any information of $\mathbf{Y}_{\Delta^\perp}$ since $\widehat{\mathbf{Y}}$ is a descendant of $\mathbf{D}$, we must find a way to restore the information of $\mathbf{Y}_{\Delta^\perp}$.

In particular, we propose obtaining a least-squares estimate $\widehat{\mathbf{Y}}_\Delta$ of $\mathbf{Y}_\Delta$ through multivariate linear regression of $\mathbf{Y}$ on $\mathbf{D}$. We then use the residual $\widehat{\mathbf{Y}}_{\Delta^\perp}$ as an estimate of $\mathbf{Y}_{\Delta^\perp}$. Finally, we compute $\widehat{\mathbf{Y}}_{\text{P-DeSIP}} = \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}}_{\Delta^\perp}$ as the bias-reduced version of $\mathbf{Y}$. This post-processing algorithm is formally presented in Algorithm 1.

Algorithm 1: (P-DeSIP) Removing potential proxy bias.

Input: $\mathbf{D}$, $\mathbf{P}$, $\mathbf{X}$, and $\mathbf{Y}$.

1: Solve $\mathbf{X} = \mathbf{D}\boldsymbol{\alpha}_1 + \mathbf{P}\boldsymbol{\alpha}_2 + \boldsymbol{e}_2$ by PLS to get $(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2)$

2: Solve $\mathbf{Y} = \mathbf{P}\boldsymbol{\beta}_1 + \mathbf{X}\boldsymbol{\beta}_2$ by PLS to get $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$

3: Compute $\widehat{\mathbf{Y}} = (\mathbf{X} - \mathbf{P}\widehat{\boldsymbol{\alpha}}_2)\widehat{\boldsymbol{\beta}}_2$

4: Compute $\widehat{\mathbf{Y}}_{\Delta^\perp} = \mathbf{Y} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Y}$

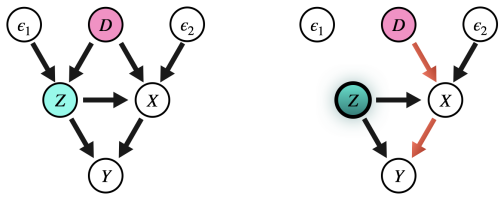
5: Compute $\widehat{\mathbf{Y}}_{\text{P-DeSIP}} = \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}}_{\Delta^\perp}$

Output: $\widehat{\mathbf{Y}}_{\text{P-DeSIP}}$ as debiased word matrix.

In practice, when the dimensionality of $\mathbf{X}$ is extremely high, the computational cost of this algorithm becomes a concern. With this in mind, we introduce a preliminary screening step to reduce ultrahigh dimensionality to a moderate level before conducting a refined analysis. Before conducting a simple screening procedure using correlation learning, each column of $\mathbf{X}$ and $\mathbf{Y}$ are standardized to a mean of zero and a standard deviation of one. Inspired by Fan and Lv (2008), we propose the following marginal screening utility to measure the dependence between $\mathbf{Y}$ and the columns $\mathbf{x}_k$ $(k = 1, \ldots, p)$ of $\mathbf{X}$: $\tau_k = \max_{j=1,\ldots,d} |\mathbf{x}_k^\top \mathbf{y}_j|/N$, where $\mathbf{y}_j$ $(j = 1, \ldots, d)$ denotes the j-th column of $\mathbf{Y}$. We propose ranking $\mathbf{x}_k$ by sorting τ_k from largest to smallest. We denote the reduced non-gender-specific word matrix by $\mathbf{X}_{\widehat{\mathcal{M}}}$, where $\widehat{\mathcal{M}} = \{k : \tau_k \geq \gamma_n\}$ and γ_n is a pre-specified threshold value.

Removing Unresolved Bias

We take a similar approach to remove unresolved bias when a proxy gender bias matrix $\mathbf{P}$ is not attainable. We consider the resolving non-gender-specific word matrix $\mathbf{Z} \in \mathcal{R}^{N\times m}$ that directly affects $\mathbf{X}$ instead of the proxy bias matrix $\mathbf{P}$: this is illustrated in Figure 2.



(a) Unresolved bias (b) Intervention on unresolved bias

Figure 2: A causal graph for unresolved bias removal.

Resolving variables are influenced by $\mathbf{D}$ in a manner that we accept as non-discriminatory: therefore, $\mathbf{Z}$ is chosen to directly affect $\mathbf{X}$ and have some correlation with $\mathbf{D}$. In particular, we choose $\mathbf{Z}$ containing the adjectives and nouns correlated to $\mathbf{D}$ based on mean cosine similarity, while $\mathbf{X}$ includes the words that are otherwise contained by $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{D}$. Since all adjectives in English have an adverb form, this ensures that the path from $\mathbf{Z}$ to $\mathbf{X}$ exists.

The causal dependencies in the corresponding linear structural equation model are equivalent to those in Figure 1 for potential proxy bias:

$$\begin{aligned} \mathbf{Z} &= \mathbf{D}\boldsymbol{\gamma}_0 + \boldsymbol{\epsilon}_1 \\ \mathbf{X} &= \mathbf{D}\boldsymbol{\gamma}_1 + \mathbf{Z}\boldsymbol{\gamma}_2 + \boldsymbol{\epsilon}_2 \\ \mathbf{Y} &= \mathbf{Z}\boldsymbol{\theta}_1 + \mathbf{X}\boldsymbol{\theta}_2, \end{aligned} \tag{3}$$

where $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are unobserved errors and $\boldsymbol{\gamma}_0 \in \mathcal{R}^{s_1\times m}$, $\boldsymbol{\gamma}_1 \in \mathcal{R}^{s_1\times p}$, $\boldsymbol{\gamma}_2 \in \mathcal{R}^{m\times p}$, $\boldsymbol{\theta}_1 \in \mathcal{R}^{m\times d}$, and $\boldsymbol{\theta}_2 \in \mathcal{R}^{p\times d}$ are parameters. We can proceed as before by intervening on Z, that is, by setting $Z = z'$. In this case, we want to cancel the remaining information from D to Y by intervening on Z: Figure 2 illustrates this procedure. The non-gender-specific word matrix $\widehat{\mathbf{Y}}$ with unresolved bias removed is

$$\widehat{\mathbf{Y}} = \mathbf{Z}\widehat{\boldsymbol{\theta}}_1. \tag{4}$$

This debiasing procedure does not retain any information of $\mathbf{Y}_{\Delta^\perp}$. Therefore we restore the information from $\mathbf{Y}_{\Delta^\perp}$ by taking a similar way to the previous procedure.

Algorithm 2: (U-DeSIP) Removing unresolved bias.

Input: $\mathbf{D}$, $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{Y}$.

1: Solve $\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta}_1 + \mathbf{X}\boldsymbol{\theta}_2$ by PLS to get $(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$

2: Compute $\widehat{\mathbf{Y}} = \mathbf{Z}\widehat{\boldsymbol{\theta}}_1$

3: Compute $\widehat{\mathbf{Y}}_{\Delta^\perp} = \mathbf{Y} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{Y}$

4: Compute $\widehat{\mathbf{Y}}_{\text{U-DeSIP}} = \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}}_{\Delta^\perp}$

Output: $\widehat{\mathbf{Y}}_{\text{U-DeSIP}}$ as debiased word matrix.

Experiments

In this section, we compare the proposed methods against other debiasing algorithms in a set of comprehensive experiments. Our results show that the proposed methods not only reduce bias in various evaluation tasks, but also enhance the performance of word embeddings in semantic evaluation tasks. Our debiasing methods outperform in downstream part-of-speech (POS) tagging, POS chunking, and named-entity recognition tasks.

We apply the proposed debiasing methods to 300-dimensional GloVe embeddings pre-trained on English Wikipedia data with 322,636 unique words (Pennington, Socher, and Manning 2014). As baselines, we also compare our results against previous state-of-the-art debiasing methods, including the hard-debiasing method (Hard) (Bolukbasi et al. 2016), the gender-preserving debiasing method (GP) (Kaneko and Bollegala 2019), word vector learning method (GN) (Zhao et al. 2018), and the half-sibling regression debiasing method (HSR) (Yang and Feng 2020). For a fair comparison, we utilize the other authors' implementations. [1]

To separate the words in the following experiments, we manually pick 11 pairs of pure gender words such as (*he*,

[1] See the accompanying appendix for links to these implementations.

she) and (*him*, *her*)[2]. We form $\mathbf{D}$ using the differences between the vector embeddings corresponding to these word pairs. We similarly compute $\mathbf{P}$ using the gendered word pairs listed in Zhao et al. (2018). The words represented in $\mathbf{P}$ contain significant non-gender-related information and gender-related information, e.g., *bride* and *bridegroom*. We choose the 50,000 most frequent words in GloVe to form $\mathbf{Y}$, which contains the words to be debiased, following the evaluation procedure in Gonen and Goldberg (2019); $\mathbf{X}$ is formed using the remaining words. In all of the below experiments, we use a fixed screening parameter $\gamma_n = 0.92$ in P-DeSIP and $\gamma_n = 0.80$ in U-DeSIP.

Quantitative Evaluation for Bias Tasks

Throughout this section, the top N gender-biased words are chosen by evaluating dot products with the gender direction $\overrightarrow{he} - \overrightarrow{she}$ in the original word embedding (i.e. GloVe) and choosing the most positive and negative values as the most male- and female-biased words, respectively.

Bias-by-projection Task. Bias-by-projection uses the dot product between the gender direction $\overrightarrow{he} - \overrightarrow{she}$ and the word to be tested. We compute and average the absolute projection bias of the top 50,000 most frequent words.

The first column of Table 2 shows that our methods achieve very good results. Its performance is just below that of Hard-GloVe, which can be explained by the fact that Hard-Glove is trained by removing projections along the gender direction.

Sembias Analogy Task. The SemBias test was first introduced in Zhao et al. (2018) as a set of word analogy tests. The task is to find the word pair in best analogy to the pair (*he*, *she*) among four options: a gender-specific word pair, e.g., (*waiter*, *waitress*); a gender-stereotype word pair, e.g., (*doctor*, *nurse*); and two highly-similar, bias-free word pairs, e.g. (*dog*, *cat*). The dataset contains 440 instances, of which 40 instances, denoted by SemBias(subset), are not used during training. We report accuracy in identifying gender-specific word pairs.

The second and third columns of Table 2 quantify accuracy in identifying gender-specific word pairs. Our P-DeSIP methods achieve very good performance in both tasks. Specifically, in the subset test, P-DeSIP outperforms GloVe by almost 40%.

Clustering Male- and Female-biased Words. As noted in Gonen and Goldberg (2019), biased words tend to cluster together. Even some debiased embeddings were unable to escape from this phenomenon. Here we take the top 500 male-biased words and the top 500 female-biased words and partition them via K-means clustering (K=2) (Hartigan and Wong 1979). Accuracy in splitting the 1,000 words into male and female clusters is presented in Table 3. Our methods achieve the best performance among all other methods.

Correlation between Bias-by-projection and Bias-by Neighbors. Taking again the top 50,000 most frequent words as targets, we compute the Pearson correlation coefficient between the bias-by-projection and bias-by-neighbor results. The latter is computed using the neighborhood metric, which counts the percentage of male- and female-biased words within the K-nearest neighbors of each target word (Gonen and Goldberg 2019; Wang et al. 2020). Here, we take $K = 100$. Referring to the second column of Table 3, our methods generally achieve the best performance.

[2]See the accompanying appendix for details of word list

	Bias-by-projection	SemBias	SemBias (subset)
GloVe	0.0375	0.8023	0.5750
Hard	**0.0007**	0.8250	0.3250
GP	0.0366	0.8432	<u>0.6500</u>
GN	0.0555	**0.9773**	0.7500
HSR	0.0218	0.8591	0.1000
P-DeSIP	<u>0.0038</u>	<u>0.9523</u>	**0.9750**
U-DeSIP	<u>0.0038</u>	0.9090	0.5000

Table 2: Gender-direction-related task performance. In each column, the best and second-best results are boldfaced and underlined, respectively.

Bias-by-neighbors for Profession Words. In this task, we assess the effect of debiasing by calculating the correlation between bias-by-neighbor measures before and after debiasing. We use the neighborhood metric, as in the previous task, but we restrict our targets to the list of professional words in Bolukbasi et al. (2016) and Zhao et al. (2018). Results, in the third column of Table 3, show that our methods outperform GloVe and are comparable to HSR-GloVe.

Classifying Previously Female- and Male-biased Words. After selecting the top 2,500 biased words for each gender, for each baseline model we train a support vector machine (SVM) model using 1,000 randomly sampled words. This classifier is then applied to the remaining 4,000 words to predict gender bias direction. Prediction accuracy is shown in the last column of Table 3: a lower accuracy indicates the trained model is unable to capture gender-related information from the original embedding and thus, that the debiasing method is superior. Again, both of our methods outperform the other methods.

	Clustering	Correlation	Profession	Classification
GloVe	1.0000	0.7727	0.8200	0.9980
Hard	0.8050	0.6884	0.7161	0.9068
GP	1.0000	0.7700	0.8102	0.9978
GN	0.8560	0.7336	0.7925	0.9815
HSR	0.9410	<u>0.6422</u>	**0.6804**	0.9055
P-DeSIP	**0.7910**	0.6431	0.7096	**0.8547**
U-DeSIP	<u>0.7920</u>	**0.6421**	<u>0.7060</u>	<u>0.8550</u>

Table 3: Gender bias word relation task performance. In each column, the best and second-best results are boldfaced and underlined, respectively.

Word Embedding Association Test (WEAT) The WEAT test (Caliskan, Bryson, and Narayanan 2017) is a

	Task1		Task2		Task3	
	p	d	p	d	p	d
GloVe	0.090*	0.704	0.000	1.905	0.026	0.987
Hard	0.363*	**0.187**	0.000	1.688	0.583*	-0.104
GP	0.055*	0.832	0.000	1.909	0.025	0.997
GN	0.157*	0.541	0.074*	**0.753**	0.653*	-0.222
HSR	0.265*	0.340	0.000	1.555	0.410*	0.122
P-DeSIP	0.755*	-0.373	0.001	<u>1.459</u>	0.486*	<u>0.019</u>
U-DeSIP	0.732*	<u>-0.335</u>	0.001	1.462	0.491*	**0.012**

Table 4: WEAT test result. In each column of p-value, * indicates statistically **non**-significant compare with $\alpha = 0.05$; In each column of d, the best and second-best results are boldfaced and underlined, respectively.

permutation-based test that measures bias in word embeddings. We report effect sizes (d) and p-values (p) in our results. The effect size is a normalized measure of how separated two distributions are. Here, a higher value indicates a larger bias between target words with respect to attribute words. The p-values denote whether the bias is significant or not.

We conduct three tests using the Pleasant & Unpleasant (Task 1), Career & Family (Task 2), and Science & Art (Task 3) word sets. We consider male and female names as attribute sets.[3]. As shown in Table 4, we achieve results comparable to those for other methods. In two out of three tasks, the p-value is not significant. We also achieve a reasonably small effect size in all three tasks.

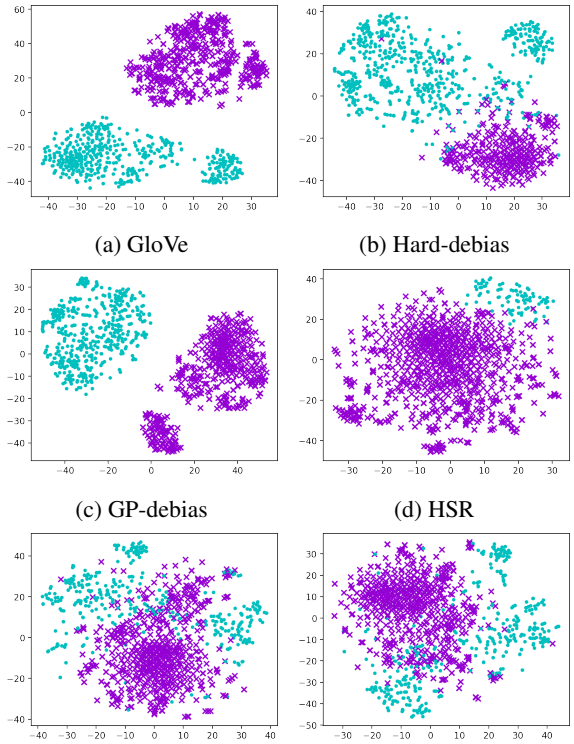
Visualization

In order to visually illustrate that our proposed methods effectively reduce gender bias, we took the top 500 male- and female-biased embeddings and generated a t-SNE projection (Hinton and Roweis 2002) for all of the baseline embeddings. In Figure 3, the two colors in the graphs indicate male- and female-biased embeddings. We can see our two methods more effectively mix up the male- and female-biased embeddings.

Word Similarity Tasks

Another important aspect of word embedding is its ability to encode words' semantic information. While bias removal is our main goal, it is unacceptable to disregard how semantic information is influenced by the debiasing process. We next implement several word similarity tests to evaluate our algorithms against existing baseline methods. We consider the following tasks: RG65 (Rubenstein and Goodenough 1965), WordSim-353 (Finkelstein et al. 2001), Rarewords (Luong, Socher, and Manning 2013), MEN (Bruni, Tran, and Baroni 2014), MTurk-287 (Radinsky et al. 2011), and MTurk-771 (Halawi et al. 2012). *SimLex-999* (Hill, Reichart, and Korhonen 2015), and *SimVerb-3500* (Gerz et al. 2016). These

[3]All word lists are from Caliskan, Bryson, and Narayanan (2017). Because GloVe embeddings are uncased, we use lower case words.



(a) GloVe (b) Hard-debias (c) GP-debias (d) HSR (e) P-DeSIP (f) U-DeSIP

Figure 3: t-SNE visualization.

datasets associated with each task contain word pairs and a corresponding human-annotated similarity score.

As an evaluation measure, we compute Spearman's rank correlation coefficient between these two ranks. Results are shown in Table 6 and 7. We see that our methods have the leading performance for most of the tasks.

Downstream Task Utility Evaluation

In order to demonstrate that our de-biased word embeddings still retain good downstream utility and performance, we follow the CoNLL2003 shared task (Sang and De Meulder 2003) and use POS tagging, POS chunking, and named-entity recognition as the evaluation tasks. Following Manzini et al. (2019) we evaluate each task in two ways: embedding matrix replacement and model retraining.

In embedding matrix replacement, we first train the task model using the original biased GloVe vectors and then calculate test data performance differences when using the original biased GloVe embeddings versus other debiased embeddings. Table 5 suggests constant performance degradation for all debiasing methods relative to the original embedding. Despite this, our methods outperform all the other tasks (in the sense of minimizing degradation) by a large margin across all the tasks and evaluation metrics (i.e., F1 score, precision, and recall). Furthermore, we even achieve a small improvement in precision on the named-entity recognition task.

In model retraining, we first train two task models, one

Embedding Matrix Replacement									
	POS Tagging			POS Chunking			Named Entity Recognition		
	Δ F1	Δ Precision	Δ Recall	Δ F1	Δ Precision	Δ Recall	Δ F1	Δ Precision	Δ Recall
Hard	-0.0776	-0.0736	-0.2079	-0.0653	-0.1500	-0.1009	-0.0118	-0.0187	-0.0238
GP	-0.1021	-0.1910	-0.2068	-0.0702	-0.1385	-0.1301	-0.0353	-0.0366	-0.0871
GN	-0.0987	-0.1001	-0.2554	-0.0702	-0.1269	-0.1401	-0.0294	-0.0610	-0.0472
HSR	-0.0666	-0.0589	-0.1820	-0.0377	-0.0753	-0.0689	-0.0055	-0.0068	-0.0128
P-DeSIP	-0.0133	-0.0006	-0.0471	**-0.0108**	**-0.0036**	-0.0346	-0.0014	**0.0002**	-0.0052
U-DeSIP	**-0.0107**	**0.0033**	**-0.0405**	-0.0110	-0.0073	**-0.0324**	**-0.0007**	0.0013	**-0.0035**
Model Retraining									
	POS Tagging			POS Chunking			Named Entity Recognition		
	Δ F1	Δ Precision	Δ Recall	Δ F1	Δ Precision	Δ Recall	Δ F1	Δ Precision	Δ Recall
Hard	-0.0194	0.0078	-0.0741	-0.0106	**0.0075**	-0.0438	-0.0050	**0.0013**	-0.0179
GP	-0.0071	0.0011	-0.0264	-0.0069	0.0043	-0.0278	-0.0013	-0.0014	-0.0030
GN	-0.0027	**0.0089**	-0.0174	0.0000	-0.0074	0.0067	-0.0011	-0.0254	**0.0189**
HSR	-0.0055	-0.0009	-0.0192	0.0002	-0.0089	0.0084	-0.0017	-0.0011	-0.0050
P-DeSIP	-0.0018	0.0002	-0.0068	-0.0005	-0.0041	0.0016	0.0002	-0.0007	0.0011
U-DeSIP	**-0.0010**	0.0000	**-0.0036**	**0.0032**	-0.0009	**0.0125**	**0.0005**	0.0008	0.0013

Table 5: Result of downstream tasks, positive value means the task has better performance than using Original GloVe. In each column, the best and second-best results are boldfaced and underlined, respectively.

	RG65	WS	RW	MEN
GloVe	0.7540	0.6199	0.3722	0.7216
Hard	0.7648	0.6207	0.3720	0.7212
GP	0.7546	0.6003	0.3450	0.6974
GN	0.7457	0.6286	**0.3989**	0.7446
HSR	0.7764	0.6554	0.3868	0.7353
P-DeSIP	**0.7794**	**0.6856**	0.3970	**0.7484**
U-DeSIP	0.7734	0.6828	0.3956	0.7478

Table 6: Word similarity task performance 1. In each column, the best and second-best results are boldfaced and underlined, respectively.

	MT-287	MT-771	SimLex	SimVerb
GloVe	0.6480	0.6486	0.3474	0.2038
Hard	0.6468	0.6504	0.3501	0.2034
GP	0.6418	0.6391	0.3389	0.1877
GN	**0.6617**	0.6619	0.3700	0.2219
HSR	0.6335	0.6652	**0.3971**	**0.2635**
P-DeSIP	0.6452	**0.6741**	0.3765	0.2286
U-DeSIP	0.6455	0.6731	0.3756	0.2273

Table 7: Word similarity task performance 2. In each column, the best and second-best results are boldfaced and underlined, respectively.

using the original biased GloVe embeddings and the other using debiased embeddings. We then calculate differences in test performance. Table 5 again suggests that our methods have the closest performance to the model trained and tested using the original GloVe embeddings. Our method also displays the most consistent and comparable performance across the three tasks.

Conclusion

In this paper, we develop two causal inference methods for removing biases in word embeddings. We show that using the differences between vectors corresponding to paired gender-specific words can better represent and eliminate gender bias. We find an intuitive and effective way to better represent gender information that needs to be removed and use this approach to achieve oracle-like retention of semantic and lexical information. We also show that our methods outperform other debiasing methods in downstream NLP tasks. Furthermore, our methods easy accommodate situations where other kinds of bias exist, such as social, racial, or class biases.

There are several important directions for future work. First, we only consider the linear relationship among the proposed casual inference frameworks. Further investigation is warranted to extend these frameworks to incorporate the non-linear causal relationship (Hoyer et al. 2008). Second, when P are not attainable, we select the resolving variables Z to contain the adjectives and nouns correlated to gender bias variables D. This selection method is rather heuristic. If prior knowledge about resolving variable was introduced, it would surely improve the performance of the unresolved bias removal. Third, we introduce a residual block to restore the information not retained from the debiasing procedure. The construction of it is rather intuitive and requires more rigorous justification. Finally, although our methods facilitate easy accommodations for situations where other kinds of bias exist, how the proxy and resolving variables as well as the bias variables are properly pre-specified may require non trivial efforts.

References

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29: 4349–4357.

Bruni, E.; Tran, N.-K.; and Baroni, M. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49: 1–47.

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

Dev, S.; and Phillips, J. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 879–887. PMLR.

Douglas, L. 2017. AI is not just learning our biases; it is amplifying them. *Medium, December*, 5.

Fan, J.; and Lv, J. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, 406–414.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Gerz, D.; Vulić, I.; Hill, F.; Reichart, R.; and Korhonen, A. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *NAACL-HLT*.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Halawi, G.; Dror, G.; Gabrilovich, E.; and Koren, Y. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1406–1414.

Hartigan, J. A.; and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1): 100–108.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4): 665–695.

Hinton, G.; and Roweis, S. T. 2002. Stochastic neighbor embedding. In *NIPS*, volume 15, 833–840. Citeseer.

Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; Schölkopf, B.; et al. 2008. Nonlinear causal discovery with additive noise models. In *NIPS*, volume 21, 689–696. Citeseer.

Kaneko, M.; and Bollegala, D. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742*.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*.

Luong, M.-T.; Socher, R.; and Manning, C. D. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the seventeenth conference on computational natural language learning*, 104–113.

Manzini, T.; Lim, Y. C.; Tsvetkov, Y.; and Black, A. W. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *NAACL*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Radinsky, K.; Agichtein, E.; Gabrilovich, E.; and Markovitch, S. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, 337–346.

Rubenstein, H.; and Goodenough, J. B. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10): 627–633.

Sang, E. F.; and De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Shin, S.; Song, K.; Jang, J.; Kim, H.; Joo, W.; and Moon, I.-C. 2020. Neutralizing gender bias in word embedding with latent disentanglement and counterfactual generation. *arXiv preprint arXiv:2004.03133*.

Vinzi, V. E.; Chin, W. W.; Henseler, J.; and Wang, H. 2010. Handbook of Partial Least Squares: Concepts, Methods and Applications. *Springer*.

Wang, T.; Lin, X. V.; Rajani, N. F.; McCann, B.; Ordonez, V.; and Xiong, C. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. *arXiv preprint arXiv:2005.00965*.

Yang, Z.; and Feng, J. 2020. A causal inference method for reducing gender bias in word embedding relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 9434–9441.

Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.