Gender-preserving Debiasing for Pre-trained Word Embeddings

Masahiro Kaneko
Tokyo Metropolitan University, Japan
kaneko-masahiro@ed.tmu.ac.jp

Danushka Bollegala
University of Liverpool, UK
danushka@liverpool.ac.uk

Abstract

Word embeddings learnt from massive text collections have demonstrated significant levels of discriminative biases such as gender, racial or ethnic biases, which in turn bias the down-stream NLP applications that use those word embeddings. Taking gender-bias as a working example, we propose a debiasing method that preserves non-discriminative gender-related information, while removing stereotypical discriminative gender biases from pre-trained word embeddings. Specifically, we consider four types of information: *feminine*, *masculine*, *gender-neutral* and *stereotypical*, which represent the relationship between gender vs. bias, and propose a debiasing method that (a) preserves the gender-related information in feminine and masculine words, (b) preserves the neutrality in gender-neutral words, and (c) removes the biases from stereotypical words. Experimental results on several previously proposed benchmark datasets show that our proposed method can debias pre-trained word embeddings better than existing SoTA methods proposed for debiasing word embeddings while preserving gender-related but non-discriminative information.

1 Introduction

Despite the impressive success stories behind word representation learning (Devlin et al., 2018; Peters et al., 2018; Pennington et al., 2014; Mikolov et al., 2013c,a), further investigations into the learnt representations have revealed several worrying issues. The semantic representations learnt, in particular from social media, have shown to encode significant levels of racist, offensive and discriminative language usage (Bolukbasi et al., 2016; Zhao et al., 2018b; Elazar and Goldberg, 2018; Rudinger et al., 2018; Zhao et al., 2018a). For example, Bolukbasi et al. (2016) showed that word representations learnt from a large (300GB) news corpus to amplify unfair gender biases. Microsoft's AI chat bot *Tay* learnt abusive language from Twitter within the first 24 hours of its release, which forced Microsoft to shutdown the bot (The Telegraph, 2016). Caliskan et al. (2017) conducted an implicit association test (IAT) (Greenwald et al., 1998) using the cosine similarity measured from word representations, and showed that word representations computed from a large Web crawl contain human-like biases with respect to gender, profession and ethnicity.

Given the broad applications of pre-trained word embeddings in various down-stream NLP tasks such as machine translation (Zou et al., 2013), sentiment analysis (Shi et al., 2018), dialogue generation (Zhang et al., 2018) etc., it is important to debias word embeddings *before* they are applied in NLP systems that interact with and/or make decisions that affect humans. We believe that no human should be discriminated on the basis of demographic attributes by an NLP system, and there exist clear legal (European Union, 1997), business and ethical obligations to make NLP systems unbiased (Holstein et al., 2018).

Despite the growing need for unbiased word embeddings, debiasing pre-trained word embeddings is a challenging task that requires a fine balance between removing information related to discriminative biases, while retaining information that is necessary for the target NLP task. For example, profession-related nouns such as *professor*, *doctor*, *programmer* have shown to be stereotypically male-biased, whereas *nurse*, *homemaker* to be stereotypically female-biased, and a debiasing method must remove such biases. On the other hand, one would expect[1], *beard* to be associated with male nouns and *bikini* to be associ-

[1]This indeed is the case for pre-trained GloVe embeddings

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1641–1650
Florence, Italy, July 28 - August 2, 2019.

ated with female nouns, and preserving such gender biases would be useful, for example, for a recommendation system (Garimella et al., 2017). As detailed later in section 2, existing debiasing methods can be seen as embedding word embeddings into a subspace that is approximately orthogonal to a gender subspace spanned by gender-specific word embeddings. Although unsupervised, weakly-supervised and adversarially trained models have been used for learning such embeddings, they primarily focus on the male-female gender direction and ignore the effect of words that have a gender orientation but not necessarily unfairly biased.

To perform an extensive treatment of the gender debiasing problem, we split a given vocabulary $\mathcal{V}$ into four mutually exclusive sets of word categories: (a) words $w_f \in \mathcal{V}_f$ that are female-biased but non-discriminative, (b) words $w_m \in \mathcal{V}_m$ that are male-biased but non-discriminative, (c) words $w_n \in \mathcal{V}_n$ that are gender-neutral, and (d) words $w_s \in \mathcal{V}_s$ that are stereotypical (i.e., unfairly[2] gender-biased). Given a large set of pre-trained word embeddings and small seed example sets for each of those four categories, we learn an embedding that (i) preserves the feminine information for the words in $\mathcal{V}_f$, (ii) preserves the masculine information for the words in $\mathcal{V}_m$, (iii) protects the neutrality of the gender-neutral words in $\mathcal{V}_n$, while (iv) removing the gender-related biases from stereotypical words in $\mathcal{V}_s$. The embedding is learnt using an encoder in a denoising autoencoder, while the decoder is trained to reconstruct the original word embeddings from the debiased embeddings that do not contain unfair gender biases. The overall model is trained end-to-end to dynamically balance the competing criteria (i)-(iv).

We evaluate the bias and accuracy of the word embeddings debiased by the proposed method on multiple benchmark datasets. On the SemBias (Zhao et al., 2018b) gender relational analogy dataset, our proposed method outperforms previously proposed *hard-debiasing* (Bolukbasi et al., 2016) and *gender-neural Global Vectors* (GN-GloVe) (Zhao et al., 2018b) by correctly debiasing stereotypical analogies. Following prior work, we evaluate the loss of information due to debiasing on benchmark datasets for semantic similarity and word analogy. Experimental results show that the proposed method can preserve the semantics of the original word embeddings, while removing gender biases. This shows that the debiased word embeddings can be used as drop-in replacements for word embeddings used in NLP applications. Moreover, experimental results show that our proposed method can also debias word embeddings that are already debiased using previously proposed debiasing methods such as GN-GloVe to filter out any remaining gender biases, while preserving semantic information useful for downstream NLP applications. This enables us to use the proposed method in conjunction with existing debiasing methods.

[2]We use the term *unfair* as used in *fairness-aware machine learning.*

2 Related Work

To reduce the gender stereotypes embedded inside pre-trained word representations, Bolukbasi et al. (2016) proposed a post-processing approach that projects gender-neutral words to a subspace, which is orthogonal to the gender dimension defined by a list of gender-definitional words. They refer to words associated with gender (e.g., *she*, *actor*) as gender-definitional words, and the remainder gender-neutral. They proposed a *hard-debiasing* method where the gender direction is computed as the vector difference between the embeddings of the corresponding gender-definitional words, and a *soft-debiasing* method, which balances the objective of preserving the inner-products between the original word embeddings, while projecting the word embeddings into a subspace orthogonal to the gender definitional words. They use a seed set of gender-definitional words to train a support vector machine classifier, and use it to expand the initial set of gender-definitional words. Both hard and soft debiasing methods ignore gender-definitional words during the subsequent debiasing process, and focus only on words that are *not* predicted as gender-definitional by the classifier. Therefore, if the classifier erroneously predicts a stereotypical word as a gender-definitional word, it would not get debiased.

Zhao et al. (2018b) proposed Gender-Neutral Global Vectors (GN-GloVe) by adding a constraint to the Global Vectors (GloVe) (Pennington et al., 2014) objective such that the gender-related information is confined to a sub-vector. During optimisation, the squared ℓ_2 distance between gender-

related sub-vectors are maximised, while simultaneously minimising the GloVe objective. GN-GloVe learns gender-debiased word embeddings from scratch from a given corpus, and cannot be used to debias pre-trained word embeddings. Moreover, similar to hard and soft debiasing methods described above, GN-GloVe uses pre-defined lists of feminine, masculine and gender-neutral words and *does not* debias words in these lists.

Debiasing can be seen as a problem of *hiding* information related to a *protected* attribute such as gender, for which adversarial learning methods (Xie et al., 2017; Elazar and Goldberg, 2018; Li et al., 2018) have been proposed in the fairness-aware machine learning community (Kamiran and Calders, 2009). In these approaches, inputs are first encoded, and then two classifiers are trained – a *target task predictor* that uses the encoded input to predict the target NLP task, and a *protected-attribute predictor* that uses the encoded input to predict the protected attribute. The two classifiers and the encoder is learnt jointly such that the accuracy of the target task predictor is maximised, while minimising the accuracy of the protected-attribute predictor. However, Elazar and Goldberg (2018) showed that although it is possible to obtain chance-level development-set accuracy for the protected attribute during training, a post-hoc classifier, trained on the encoded inputs can still manage to reach substantially high accuracies for the protected attributes. They conclude that adversarial learning alone does not guarantee invariant representations for the protected attributes.

Gender biases have been identified in several tasks in NLP such as coreference (Rudinger et al., 2018; Zhao et al., 2018a) resolution and machine translation (Prates et al., 2018). For example, rule-based, feature-based as well as neural coreference resolution methods trained on biased resources have shown to reflect those biases in their predictions (Rudinger et al., 2018). Google Machine Translation, for example, provides male and female versions of the translations[3], when the gender in the source language is ambiguous.

3 Gender-Preserving Debiasing

3.1 Formulation

Given a pre-trained set of d-dimensional word embeddings $\{\boldsymbol{w}_i\}_{i=1}^{|\mathcal{V}|}$, over a vocabulary $\mathcal{V}$, we consider the problem of learning a map $E : \mathbb{R}^d \rightarrow \mathbb{R}^l$ that projects the original pre-trained word embeddings to a debiased l-dimensional space. We do not assume any knowledge about the word embedding learning algorithm that was used to produce the pre-trained word embeddings given to us. Moreover, we do not assume the availability or access to the language resources such as corpora or lexicons that might have been used by the word embedding learning algorithm. Decoupling the debiasing method from the word embedding learning algorithm and resources increases the applicability of the proposed method, enabling us to debias pre-trained word embeddings produced using different word embedding learning algorithms and using different types of resources.

We propose a debiasing method that models the interaction between the values of the protected attribute (in the case of *gender* we consider *male*, *female* and *neutral* as possible attribute values), and whether there is a stereotypical bias or not. Given four sets of words: *masculine* ($\mathcal{V}_m$), *feminine* ($\mathcal{V}_f$), *neutral* ($\mathcal{V}_n$) and *stereotypical* ($\mathcal{V}_s$), our proposed method learns a projection that satisfies the following four criteria:

(i) for $w_f \in \mathcal{V}_f$, we protect its feminine properties,

(ii) for $w_m \in \mathcal{V}_m$, we protect its masculine properties,

(iii) for $w_n \in \mathcal{V}_n$, we protect its gender neutrality, and

(iv) for $w_s \in \mathcal{V}_s$, we remove its gender biases.

By definition the four word categories are mutually exclusive and the total vocabulary is expressed by their disjunction $\mathcal{V} = \mathcal{V}_m \cup \mathcal{V}_f \cup \mathcal{V}_n \cup \mathcal{V}_s$. A key feature of the proposed method that distinguishes it from prior work on debiasing word embeddings is its ability to differentiate between undesirable (stereotypical) biases from the desirable (expected) gender information in words. The procedure we follow to compile the four word-sets is described later in subsection 4.1, and the words that belong to each of the four categories are shown in the supplementary material.

To explain the proposed gender debiasing method, let us first consider a *feminine* regressor $C_f : \mathbb{R}^l \rightarrow [0, 1]$, parameterised by $\boldsymbol{\theta_f}$, that predicts the degree of femininess of the word w. Here, highly feminine words are assigned values

[3] https://bit.ly/2B0nVHZ

close to 1. Likewise, let us consider a *masculine* regressor $C_m : \mathbb{R}^l \rightarrow [0, 1]$, parametrised by $\boldsymbol{\theta_m}$, that predicts the degree of masculinity of w. We then learn the debiasing function as the encoder $E : \mathbb{R}^d \rightarrow \mathbb{R}^l$ of an autoencoder (parametrised by $\boldsymbol{\theta}_e$), where the corresponding decoder (parametrised by $\boldsymbol{\theta}_d$) is given by $D : \mathbb{R}^l \rightarrow \mathbb{R}^d$.

For feminine and masculine words, we require the encoded space to retain the gender-related information. The squared losses, L_f and L_m, given respectively by (1) and (2), express the extent to which this constraint is satisfied.

$$L_f = \sum_{w \in \mathcal{V}_f} ||C_f(E(\boldsymbol{w})) - 1||_2^2 + \sum_{w \in \mathcal{V} \setminus \mathcal{V}_f} ||C_f(E(\boldsymbol{w}))||_2^2 \tag{1}$$

$$L_m = \sum_{w \in \mathcal{V}_m} ||C_m(E(\boldsymbol{w})) - 1||_2^2 + \sum_{w \in \mathcal{V} \setminus \mathcal{V}_m} ||C_f(E(\boldsymbol{w}))||_2^2 \tag{2}$$

Here, for notational simplicity, we drop the dependence on parameters.

For the stereotypical and gender-neutral words, we require that they are embedded into a subspace that is orthogonal to a gender directional vector, $\boldsymbol{v}_g$, computed using a set, Ω, of feminine and masculine word-pairs $(w_f, w_m)(\in \Omega)$ as given by (3).

$$\boldsymbol{v}_g = \frac{1}{|\Omega|} \sum_{(w_f, w_m) \in \Omega} (E(\boldsymbol{w}_m) - E(\boldsymbol{w}_f)) \tag{3}$$

Prior work on gender debiasing (Bolukbasi et al., 2016; Zhao et al., 2018b) showed that the vector difference between the embeddings for male-female word-pairs such as *he* and *she* accurately represents the gender direction. When training, we keep $\boldsymbol{v}_g$ fixed during an epoch, and re-estimate $\boldsymbol{v}_{\mathrm{g}}$ between every epoch. We consider the squared inner-product between $\boldsymbol{v}_g$ and the debiased stereotypical or gender-neutral words as the loss, L_g, as given by (4).

$$L_g = \sum_{w \in \mathcal{V}_n \cup \mathcal{V}_s} (\boldsymbol{v}_g{}^\top \boldsymbol{w})^2 \tag{4}$$

It is important that we preserve the semantic information encoded in the word embeddings as much as possible when we perform debiasing. If too much information is removed from the word embeddings, not limited to gender-biases, then the debiased word embeddings might not be sufficiently accurate to be used in downstream NLP applications. For this purpose, we minimise the reconstruction loss, L_r, for the autoencoder given by (5).

$$L_r = \sum_{w \in \mathcal{V}} ||D(E(\boldsymbol{w})) - \boldsymbol{w}||_2^2 \tag{5}$$

Finally, we define the total objective as the linearly-weighted sum of the above-defined losses as given by (6).

$$L = \lambda_f L_f + \lambda_m L_m + \lambda_g L_g + \lambda_r L_r \tag{6}$$

Here, the coefficients $\lambda_f, \lambda_m, \lambda_g, \lambda_r$ are nonnegative hyper-parameters that add to 1. They determine the relative importance of the different constraints we consider and can be learnt using training data or determined via cross-validation over a dedicated validation dataset. In our experiments, we use the latter approach.

3.2 Implementation and Training

C_f and C_m are both implemented as feed forward neural networks with one hidden layer and the sigmoid function is used as the nonlinear activation. Increasing the number of hidden layers beyond one for C_f and C_m did not result in a significant increase in accuracy. Both the encoder E and the decoder D of the autoencoder are implemented as feed forward neural networks with two hidden layers. Hyperbolic tangent is used as the activation function throughout the autoencoder.

The objective (6) is minimised w.r.t. the parameters $\boldsymbol{\theta}_f$, $\boldsymbol{\theta}_m$, $\boldsymbol{\theta}_e$ and $\boldsymbol{\theta}_d$ for a given pre-trained set of word embeddings. During optimisation, we used dropout with probability 0.01 and use stochastic gradient descent with initial learning rate set to 0.1. The hyper-parameters $\lambda_f, \lambda_m, \lambda_g, \lambda_r$ are estimated using a separate validation dataset as described later in subsection 4.1.

Note that it is possible to pre-train C_f and C_m separately using $\mathcal{V}_f$ and $\mathcal{V}_m$ prior to training the full objective (6). In our preliminary experiments, we found that initialising $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_m$ to the pre-trained versions of C_f and C_m to be helpful for the optimisation process, resulting in early convergence to better solutions compared to starting from random initialisations for $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_m$. For pre-training C_f and C_m we used Adam optimiser (Kingma and Ba, 2015) with initial learning rate set to 0.0002 and a mini-batch size of 512. Autoencoder is also pre-trained using a randomly

selected 5000 word embeddings and dropout regularisation is applied with probability 0.05.

We note that $\mathcal{V}_f$ and $\mathcal{V}_m$ are separate word sets, not necessarily having corresponding feminine-masculine pairs as in Ω used in (4). It is of course possible to re-use the words in Ω in $\mathcal{V}_f$ and $\mathcal{V}_m$, and we follow this approach in our experiments, which helps to decrease the number of seed words required to train the proposed method. Moreover, the number of training examples across the four categories $\mathcal{V}_f, \mathcal{V}_m, \mathcal{V}_n, \mathcal{V}_s$ were significantly different, which resulted in an imbalanced learning setting. We conduct one-sided undersampling (Kubat and Matwin, 1997) to successfully overcome this data imbalance issue. The code and the debiased embeddings are publicly available[4].

4 Experiments

4.1 Training and Development Data

We use the feminine and masculine word lists (223 words each) created by Zhao et al. (2018b) as $\mathcal{V}_f$ and $\mathcal{V}_m$, respectively. To create a gender-neutral word list, $\mathcal{V}_n$, we select gender-neutral words from a list of 3000 most frequent words in English[5]. Two annotators independently selected words and subsequently verified for gender neutrality. The final set of $\mathcal{V}$ contains 1031 gender-neutral words. We use the stereotypical word list compiled by Bolukbasi et al. (2016) as $\mathcal{V}_s$, which contains 166 professions that are stereotypically associated with one type of a gender. The four sets of words used in the experiments are shown in the supplementary material.

We train GloVe (Pennington et al., 2014) on 2017 January dump of English Wikipedia to obtain pre-trained 300-dimensional word embeddings for 322636 unique words. In our experiments, we set both d and l to 300 to create 300-dimensional de-biased word embeddings. We randomly selected 20 words from each of the 4 sets $\mathcal{V}_f$, $\mathcal{V}_m$, $\mathcal{V}_n$ and $\mathcal{V}_s$, and used them as a development set for pre-training C_f and C_m and to estimate the hyperparameters in (6). The optimal hyperparameter values estimated on this development dataset are: $\lambda_f = \lambda_m = \lambda_g = 0.0001$, and $\lambda_r = 1.0$. In our preliminary experiments we observed that increasing λ_f, λ_m and λ_g relative to λ_r results in higher reconstruction losses in the autoencoder. This shows that the ability to accurately reconstruct the original word embeddings is an important requirement during debiasing.

4.2 Baselines and Comparisons

We compare our proposed method against several baselines.

GloVe: is the pre-trained GloVe embeddings described in subsection 4.1. This baseline denotes a non-debiased version of the word embeddings.

Hard-GloVe: We use the implementation[6] of hard-debiasing (Bolukbasi et al., 2016) method by the original authors and produce a debiased version of the pre-trained GloVe embeddings.[7]

GN-GloVe : We use debiased GN-GloVe embeddings released by the original authors[8], without retraining ourselves as a baseline.

AE (GloVe): We train an autoencoder by minimising the reconstruction loss defined in (5) and encode the pre-trained GloVe embeddings to a vector space with the same dimensionality. This baseline can be seen as surrogated version of the proposed method with $\lambda_f = \lambda_m = \lambda_g = 0$. **AE (GloVe)** does *not* perform debiasing and shows the amount of semantic information that can be preserved by autoencoding the input embeddings.

AE (GN-GloVe): Similar to **AE (GloVe)**, this method autoencodes the debiased word embeddings produced by **GN-GloVe**.

GP (GloVe): We apply the proposed *gender-preserving* (**GP**) debiasing method on pre-trained GloVe embeddings to debias it.

GP (GN-GloVe): To test whether we can use the proposed method to further debias word embeddings that are already debiased using other methods, we apply it on GN-GloVe.

4.3 Evaluating Debiasing Performance

We use the SemBias dataset created by Zhao et al. (2018b) to evaluate the level of gender bias in word embeddings. Each instance in SemBias consists of four word pairs: a gender-definition word pair (**Definition**; e.g. "waiter - waitress"),

[4] https://github.com/kanekomasahiro/gp_debias

[5] https://bit.ly/2SvBINY

[6] https://github.com/tolga-b/debiaswe

[7] Bolukbasi et al. (2016) released debiased embeddings for word2vec only and for comparison purposes with GN-GloVe, we use GloVe as the pre-trained word embedding and apply hard-debiasing on GloVe

[8] https://github.com/uclanlp/gn_glove

Embeddings	SemBias			SemBias-subset		
	Definition ↑	Stereotype ↓	None ↓	Definition ↑	Stereotype ↓	None ↓
GloVe	80.2	10.9	8.9	57.5	20	22.5
Hard-Glove	84.1	9.5	6.4	25	47.5	27.5
GN-GloVe	97.7	1.4	0.9	75	15	10
AE (GloVe)	82.7	8.2	9.1	$62.5^{\dagger}$	$17.5^{\dagger}$	20
AE (GN-GloVe)	$98.0^{\dagger *}$	$1.6^{\dagger *}$	$\mathbf{0.5}^{\dagger *}$	77.5	$17.5^{\dagger}$	$\mathbf{5}^{\dagger *}$
GP (GloVe)	84.3^{*}	8.0	7.7^{*}	$65^{\dagger}$	$15^{\dagger}$	20
GP (GN-GloVe)	$\mathbf{98.4}^{\dagger *}$	$\mathbf{1.1}^{\dagger *}$	$\mathbf{0.5}^{\dagger *}$	$\mathbf{82.5}^{\dagger *}$	$\mathbf{12.5}^{\dagger *}$	$\mathbf{5}^{\dagger *}$

Table 1: Prediction accuracies for gender relational analogies. $*$ and † indicate statistically significant differences against respectively **GloVe** and **Hard-GloVe**.

a gender-stereotype word pair (**Stereotype**; e.g., "doctor - nurse") and two other word-pairs that have similar meanings but not a gender relation (**None**; e.g., "dog - cat", "cup - lid"). SemBias contains 20 gender-stereotype word pairs and 22 gender-definitional word pairs and use their Cartesian product to generate 440 instances. Among the 22 gender-definitional word pairs, 2 word-pairs are not used as the seeds for training. Following, Zhao et al. (2018b), to test the generalisability of a debiasing method, we use the subset (SemBias-subset) of 40 instances associated with these 2 pairs. We measure relational similarity between (he, she) word-pair and a word-pair (a, b) in SemBias using the cosine similarity between the $\overrightarrow{he} - \overrightarrow{she}$ gender directional vector and $\boldsymbol{a} - \boldsymbol{b}$ using the word embeddings under evaluation. For the four word-pairs in each instance in SemBias, we select the word-pair with the highest cosine similarity with $\overrightarrow{he} - \overrightarrow{she}$ as the predicted answer. In Table 1, we show the percentages where a word-pair is correctly classified as **Definition**, **Stereotype**, or **None**. If the word embeddings are correctly debiased, we would expect a high accuracy for **Definitions** and low accuracies for **Stereotypes** and **Nones**.

From Table 1, we see that the best performances (highest accuracy on **Definition** and lowest accuracy on **Stereotype**) are reported by **GP (GN-GloVe)**, which is the application of the proposed method to debias word embeddings learnt by **GN-GloVe**. In particular, in both SemBias and SemBias-subset, **GP (GN-GloVe)** statistically significantly outperforms **GloVe** and **Hard-Glove** according to Clopper-Pearson confidence intervals (Clopper and Pearson, 1934). Although **GN-GloVe** obtains high performance on SemBias, it does not generalise well to SemBias-subset. However, by applying the proposed method, we can further remove any residual gender biases from **GN-GloVe**, which shows that the proposed method can be applied in conjunction with **GN-GloVe**. We see that **GloVe** contains a high percentage of stereotypical gender biases, which justifies the need for debiasing methods. By applying the proposed method on **GloVe** (corresponds to **GP (GloVe)**) we can decrease the gender biases in **GloVe**, while preserving useful gender-related information for detecting definitional word-pairs. Comparing corresponding **AE** and **GP** versions of **GloVe** and **GN-GloVe**, we see that autoencoding alone is insufficient to consistently preserve gender-related information.

4.4 Preservation of Word Semantics

It is important that the debiasing process removes only gender biases and preserve other information unrelated to gender biases in the original word embeddings. If too much information is removed from word embeddings during the debiasing process, then the debiased embeddings might not carry adequate information for downstream NLP tasks that use those debiased word embeddings.

To evaluate the semantic accuracy of the debiased word embeddings, following prior work on debiasing (Bolukbasi et al., 2016; Zhao et al., 2018a), we use them in two popular tasks: semantic similarity measurement and analogy detection. We recall that we do *not* propose novel word embedding learning methods in this paper, and what is important here is whether the debiasing process preserves as much information as possible in the

Embeddings	sem	syn	total	MSR	SE
GloVe	80.1	62.1	70.3	53.8	38.8
Hard-GloVe	80.3	**62.7**	**70.7**	**54.4**	39.1
GN-GloVe	77.8	60.9	68.6	51.5	39.1
AE (GloVe)	**81.0**	61.9	70.5	52.6	38.9
AE (GN-GloVe)	78.6	61.3	69.2	51.2	39.1
GP (GloVe)	80.5	61.0	69.9	51.3	38.5
GP (GN-GloVe)	78.3	61.3	69.0	51.0	**39.6**

Table 2: Accuracy for solving word analogies.

Datasets	#Orig	#Bal
WS	353	366
RG	65	77
MTurk	771	784
RW	2,034	2,042
MEN	3,000	3,122
SimLex	999	1,043

Table 3: Number of word-pairs in the original (**Orig**) and balanced (**Bal**) similarity benchmarks.

original word embeddings.

4.4.1 Analogy Detection

Given three words a, b, c in analogy detection, we must predict a word d that completes the analogy "a is b as c is to d". We use the CosAdd (Levy and Goldberg, 2014) that finds $\boldsymbol{d}$ that has the maximum cosine similarity with $(\boldsymbol{b}-\boldsymbol{a}+\boldsymbol{c})$. We use the semantic (**sem**) and syntactic (**syn**) analogies in the Google analogy dataset (Mikolov et al., 2013b) (in **total** contains 19,556 questions), **MSR** dataset (7,999 syntactic questions) (Mikolov et al., 2013d) and SemEval dataset (**SE**, 79 paradigms) (Jurgens et al., 2012) as benchmark datasets. The percentage of correctly solved analogy questions is reported in Table 2. We see that there is no significant degradation of performance due to debiasing using the proposed method.

4.4.2 Semantic Similarity Measurement

The correlation between the human ratings and similarity scores computed using word embeddings for pairs of words has been used as a measure of the quality of the word embeddings (Mikolov et al., 2013d). We compute cosine similarity between word embeddings and measure Spearman correlation against human ratings for the word-pairs in the following benchmark datasets: Word Similarity 353 dataset (**WS**) (Finkelstein et al., 2001), Rubenstein-Goodenough dataset (**RG**) (Rubenstein and Goodenough, 1965), **MTurk** (Halawi et al., 2012), rare words dataset (**RW**) (Luong et al., 2013), **MEN** dataset (Bruni et al., 2012) and **SimLex** dataset (Hill et al., 2015).

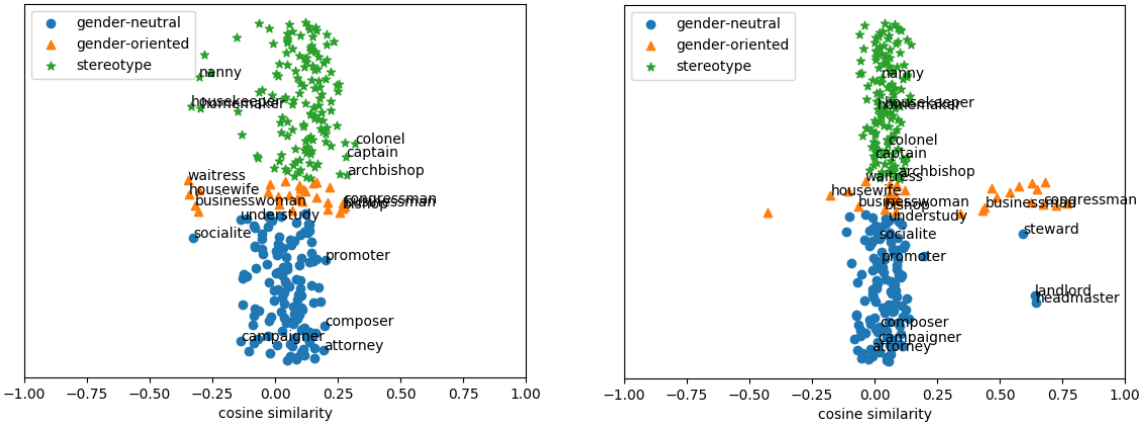
Unfortunately, existing benchmark datasets for semantic similarity were not created considering gender-biases and contain many stereotypical examples. For example, in **MEN**, the word *sexy* has high human similarity ratings with *lady* and *girl* compared to *man* and *guy*. Furthermore, masculine words and *soldier* are included in multiple datasets with high human similarity ratings, whereas it is not compared with feminine words in any of the datasets. Although prior work studying gender bias have used these datasets for evaluation purposes (Bolukbasi et al., 2016; Zhao et al., 2018a), we note that high correlation with human ratings can be achieved with biased word embeddings.

To address this issue, we *balance* the original datasets with respect to gender by including extra word pairs generated from the opposite sex with the same human ratings. For instance, if the word-pair (*baby*, *mother*) exists in the dataset, we add a new pair (*baby*, *father*) to the dataset. Ideally, we should re-annotate this balanced version of the dataset to obtain human similarity ratings. However, such a re-annotation exercise would be costly and inconsistent with the original ratings. Therefore, we resort to a proxy where we reassign the human rating for the original word-pair to its derived opposite gender version. Table 3 shows the number of word-pairs in the original (**Orig**) and balanced (**Bal**) similarity benchmarks.

As shown in Table 4, **GP (GloVe)** and **GP (GN-**

Embeddings	WS		RG		MTurk		RW		MEN		SimLex	
	Orig	Bal	Orig	Bal	Orig	Bal	Orig	Bal	Orig	Bal	Orig	Bal
GloVe	61.6	62.9	75.3	75.5	64.9	63.9	37.3	37.5	73.0	72.6	34.7	35.9
Hard-GloVe	61.7	63.1	76.4	76.7	65.1	64.1	37.4	37.4	72.8	72.5	35.0	36.1
GN-GloVe	62.5	63.7	74.1	73.7	66.2	65.5	40.0	40.1	74.9	74.5	37.0	38.1
AE (GloVe)	61.3	62.6	**77.1**	**76.8**	64.9	64.1	35.7	35.8	71.9	71.5	34.7	35.9
AE (GN-GloVe)	61.3	62.6	73.0	74.0	66.3	65.5	38.7	38.9	73.8	73.4	36.7	37.7
GP (GloVe)	59.7	61.0	75.4	75.5	63.9	63.1	34.7	34.8	70.8	70.4	33.9	35.0
GP (GN-GloVe)	**63.2**	**64.3**	72.2	72.2	**67.9**	**67.4**	**43.2**	**43.3**	**75.9**	**75.5**	**38.4**	**39.5**

Table 4: Spearman correlation between human ratings and cosine similarity scores computed using word embeddings for the word-pairs in the original and balanced versions of the benchmark datasets.



(a) GloVe

(b) GN (GloVe)

(c) Hard-Glove

(d) GP (GloVe)

Figure 1: Cosine similarity between gender, gender-neutral, stereotypical words and the gender direction.

GloVe) obtain the best performance on the balanced versions of all benchmark datasets. Moreover, the performance of **GP (GloVe)** on both original and balanced datasets is comparable to that of **GloVe**, which indicates that the information encoded in GloVe embeddings are preserved in the debiased embeddings, while removing stereotypical gender biases. The autoencoded versions report similar performance to the original input embeddings.

Overall, the results on the analogy detection and semantic similarity measurement tasks show that our proposed method removes only gender-biases and preserve other useful gender-related information.

4.5 Visualising the Effect of Debiasing

To visualise the effect of debiasing on different word categories, we compute the cosine similarity between the gender directional vector $\overrightarrow{he} - \overrightarrow{she}$,

and selected *gender-oriented* (female or male), *gender-neutral* and stereotypical words. In Figure 1, horizontal axises show the cosine similarity with the gender directional vector (positive scores for masculine words) and the words are alphabetically sorted within each category.

From Figure 1, we see that the original **GloVe** embeddings show a similar spread of cosine similarity scores for gender-oriented as well as stereotypical words. When debiased by hard-debias (**Hard-GloVe**) and **GN-GloVe**, we see that stereotypical and gender-neutral words get their gender similarity scores equally reduced. Interestingly, **Hard-GloVe** shifts even gender-oriented words towards the masculine direction. On the other hand, **GP (GloVe)** decreases gender bias in the stereotypical words, while almost preserving gender-neutral and gender-oriented words as in **GloVe**.

Considering that a significant number of words in English are gender-neutral, it is essential that debiasing methods do not adversely change their orientation. In particular, the proposed method's ability to debias stereotypical words that carry unfair gender-biases, while preserving the gender-orientation in feminine, masculine and neutral words is important when applying the debiased word embeddings in NLP applications that depend on word embeddings for representing the input texts

5 Conclusion

We proposed a method to remove gender-specific biases from pre-trained word embeddings. Experimental results on multiple benchmark datasets demonstrate that the proposed method can accurately debias pre-trained word embeddings, outperforming previously proposed debiasing methods, while preserving useful semantic information. In future, we plan to extend the proposed method to debias other types of demographic biases such as ethnic, age or religious biases.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145. Association for Computational Linguistics.

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

C. J. Clopper and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404 – 413.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proc. of EMNLP*.

European Union. 1997. Treaty of amsterdam (article 13).

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 406–414, New York, NY, USA. ACM.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2275–2285, Copenhagen, Denmark. Association for Computational Linguistics.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwatz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1406–1414, New York, NY, USA. ACM.

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

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need?

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: 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 (SemEval 2012), pages 356–364. Association for Computational Linguistics.

Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Proc. of International Conference on Computer, Control and Communication (IC4)*, pages 1–6.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Miroslav Kubat and Stan Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML 1997*, pages 179 – 186.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *CoNLL*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30. Association for Computational Linguistics.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, and Jeffrey Dean. 2013a. Efficient estimation of word representation in vector space. In *ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continous space word representations. In *NAACL-HLT*, pages 746 – 751.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013d. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Jeffery Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: global vectors for word representation. In *EMNLP*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL-HLT*.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2018. Assessing Gender Bias in Machine Translation – A Case Study with Google Translate.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14. Association for Computational Linguistics.

Bei Shi, Zihao Fu, Lidong Bing, and Wai Lam. 2018. Learning domain-sensitive and sentiment-aware word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2494–2504. Association for Computational Linguistics.

The Telegraph. 2016. Microsoft deletes 'teen girl' ai after it became a hitlter-loving sex robot within 24 hours. `https://goo.gl/mE8p3J`.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Proc. of NIPS*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In *Proc. of EMNLP*, pages 4847–4853.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proc. of EMNLP'13*, pages 1393 – 1398.