DOUBLE-HARD DEBIASING: TAILORING WORD EMBEDDINGS FOR GENDER BIAS MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Gender bias in word embeddings has been widely investigated. However, recent work has shown that existing approaches, including the well-known Hard Debias algorithm (Bolukbasi et al., 2016) which projects word embeddings to a subspace orthogonal to an inferred gender direction, are insufficient to deliver gender-neutral word embeddings. In our work, we discover that semantic-agnostic corpus statistics such as word frequency are important factors that limit the debiasing performance. We propose a simple but effective processing technique, Double-Hard Debias, to attenuate the effect due to such noise. We experiment with Word2Vec and GloVe embeddings and demonstrate on several benchmarks that our approach preserves the distributional semantics while effectively reducing gender bias to a larger extent than previous debiasing techniques.

1 INTRODUCTION

Word embeddings are based on the distributional hypothesis that linguistic tokens, such as words, with similar distributions in the vector space have similar meanings. This assumption is the basis for *statistical semantics* that captures the frequency and order of recurrence of words in context along with their meaning. Widely used word embeddings Word2Vec (Mikolov et al., 2013b) and GloVe (Pennington et al., 2014) are known to exhibit gender bias (Bolukbasi et al., 2016) which gets further amplified when these biased embeddings are used in various downstream NLP tasks (Zhao et al., 2018a; Rudinger et al., 2018). To mitigate this effect of gender bias, most of the recent research has focused on identifying the gender dimensions in word embeddings and zeroing out the gender component in a post-hoc manner (Bolukbasi et al., 2016; Prost et al., 2019). However, the effectiveness of these efforts has been limited as argued in Gonen & Goldberg (2019).

We hypothesize that getting rid of the gender component from word embeddings is more complicated than it was expected. An important assumption towards the success of Hard Debias is we can effectively find the gender direction. However, this heavily relies on a set of gender word pairs which is manually defined. Although Bolukbasi et al. (2016) carefully derives 10 word pairs through crowdsourcing, it is hard to claim they can perfectly capture the gender direction in the embedding space. Even with proper pairs, as word embeddings are learned from text corpora through machine learning algorithms, statistical signals can be easily encoded into embeddings and further lead to an offset in gender direction. It has been shown (Gong et al., 2018; Mu & Viswanath, 2018) that *frequency* of words causes a twist of the geometry of word embeddings, which degrades the quality of them. We posit that frequency also contaminate the gender direction that we want to find. As a consequence, it constraints the debiasing ability of Hard Debias.

To this end, we propose a novel pre-processing algorithm called *Double-Hard Debias* that builds on the existing Hard Debias technique. The intuition behind our method is that word embeddings exhibit another more general type of bias in which words that have similar frequencies during training tend to be closer in the vector space even in cases when they do not have similar meanings – and this is a bias that closely intertwined with gender bias. Double-Hard Debias demonstrates that it is important to not just down weight the gender component as in previous works but also the frequency component in order to "purify" embeddings with respect to gender.

More concretely, inspired by (Mu & Viswanath, 2018), we conduct PCA(principal component analysis) on all word embeddings to get a set of principal components which we consider as candidate directions related with frequency. We then run kmeans clustering on male and female biased words

as a proxy experiment to pick a direction that is most helpful to alleviate the influence from frequency. We then finetune embeddings by projecting out the component along the selected direction. On those purified embeddings, Hard Debias manages to further reduce gender bias. This two-stage process, named as Double-Hard Debias , achieves better debiasing results and meanwhile, is capable of preserving distributional semantics.

2 BACKGROUND

Definitions. Let $\mathcal{W}$ be the vocabulary of a particular language. Following Bolukbasi et al. (2016), we assume there is a set of gender neutral words $\mathcal{N} \in \mathcal{W}$, such as "doctor" and "teacher", which by definition are not specific to any gender. We also assume a given set of female-male word pairs $\mathcal{P} \subset \mathcal{W} \times \mathcal{W}$, where the main difference between each pair of words is gender[1].

An embedding consists of a vector $\overrightarrow{w} \in \mathbb{R}^n$ for each word $w \in \mathcal{W}$. A subspace $\mathbf{B}$ is defined by k orthogonal unit vectors $\mathbf{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_k\} \in \mathbb{R}^d$. A subspace with $k = 1$ is also called a direction. We denote the projection of vector $\mathbf{v}$ on $\mathbf{B}$ by $\mathbf{v}_\mathbf{B} = \sum_{j=1}^{k} (\mathbf{v} \cdot \mathbf{b}_j)\mathbf{b}_j$ and the projection onto the orthogonal subspace $\mathbf{v}_{\perp\mathbf{B}} = \mathbf{v} - \mathbf{v}_\mathbf{B}$.

Hard Debias. Hard Debias is a debiasing algorithm defined in terms of word sets. The algorithm takes as input the set of words to neutralize N. and a family of equality sets $E = \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_m | \mathcal{E}_i \subseteq \mathcal{W}, i = 1, \ldots m\}$. First, it identifies a direction in the word embedding space that captures the bias. Then it transforms $\overrightarrow{w}$ such that every word $w \in N$ has zero projection in the bias direction and is equidistant to all words in each equality set.

Mathematically, the algorithm performs the following two steps.

1. Identify the bias subspace. Let $\mu_i := \sum_{w \in \mathcal{E}_i} \overrightarrow{w}/|\mathcal{E}_i|$. The bias subspace $\mathbf{B}$ is the first k (≥ 1) rows of SVD($\mathbf{C}$), where

$$\mathbf{C} := \sum_{i=1}^{m} \sum_{w \in \mathcal{E}_i} (\overrightarrow{w} - \mu_i)^T (\overrightarrow{w} - \mu_i)/|\mathcal{E}_i|. \tag{1}$$

2. Neutralize and equalize[2]. For each $\mathcal{E}_i \in E$, let $\nu_i := \mu_{i\perp\mathbf{B}}$. Then update $\overrightarrow{w}$'s by

$$\overrightarrow{w} := \begin{cases} \nu_i + \sqrt{(1 - \|\nu_i\|^2)} \overrightarrow{w}_{\perp\mathbf{B}} / \|\overrightarrow{w}_{\perp\mathbf{B}}\| & w \in \mathcal{E}_i, i = 1 \ldots m \\ \overrightarrow{w}_{\perp\mathbf{B}} / \|\overrightarrow{w}_{\perp\mathbf{B}}\| & w \in N \end{cases}. \tag{2}$$

For mitigating gender bias[3], we follows Bolukbasi et al. (2016) to set $k = 1$, $N = \mathcal{N}$ and $E = \mathcal{P}$, where each equality set is a female-male word pair.

Clustering Male and Female Biased Words. For each word $w \in \mathcal{W}$, we compute its *bias* score $\mathcal{B} = cos(\overrightarrow{w}, \overrightarrow{he}) - cos(\overrightarrow{w}, \overrightarrow{she})$. We then sort the vocabulary according to $\mathcal{B}$ and take top k words as male biased words set W_m and bottom k words as W_f. For $w_i \in W_m$, we create ground truth gender label $g_i = 0$. Similarly, for $w_i \in W_f$, annotate $g_i = 1$. We run Kmeans to cluster the embeddings of selected words and compute the accuracy score a based on the clustering output and the created ground truth gender labels:

$$a = \frac{1}{|W_m + W_f|} \sum\nolimits_{(w_i, y_i) \in W_m + W_f} \mathbb{1}[KMeans(\overrightarrow{w}_i) == g_i] \tag{3}$$

For comparison convenience, we set $a = \max(a, 1 - a)$. A higher a value shows stronger gender information in embedding which is captured by the clustering algorithm.

[1] $\mathcal{P}$ consists of "woman"&"man", "girl"&"boy", "she"&"he", "mother"&"father", "daughter"&"son", "gal"&"guy", "female"&"male", "her"&"his", "herself"&"himself", and "Mary"&"John" (Bolukbasi et al., 2016).

[2] Originally, Hard Debias normalizes embeddings. However, we found it is unnecessary in our experiments. This is also confirmed in Ethayarajh et al. (2019)

[3] Hard Debias is general and applicable to mitigating multi-class biases such as racial or religious bias.

3 DOUBLE-HARD DEBIASING

We recall that word embeddings are representations learned from massive text corpora to capture the relations between words. The attribute "gender" is induced from a set of gender specific words, such as "he" and "she", "man" and "woman". In the process of learning word embeddings, "gender" propagates to other words that interact with gender specific ones. Gender bias happens when a word is neutral by its definition but the embedding learned is strongly associated with gender. For example, "programmer" is closer to "male" than "female" in the embedding space because it appears more frequently with male words in the training corpus. To address this, Bolukbasi et al. (2016) proposes Hard Debias which aims to eliminate the effect from gender specific words. Specifically, Hard Debias projects word embeddings into a subspace orthogonal to a gender direction. To find the gender direction, it requires a collection of gendered word pairs which were obtained using crowdsourcing.

The method of Hard Debias should work well if it can effectively find the gender direction, however we realize that there may exist factors that can potentially lessen the effectiveness of Hard Debias. As observed in recent works (Mu & Viswanath, 2018; Gong et al., 2018), word embeddings also encode a seemingly more pervasive type of bias with respect to *word frequency*. Popular words and rare words cluster in different subregions of the embedding space. This further affects the semantic properties of word embeddings, as two words with similar semantic meanings may be far apart due to frequency. We posit that frequency also interferes with the gender direction. In fact, the selected ten word pairs typically used to compute the gender direction have significantly different frequencies, e.g. "guy" and "gal" are much less frequent than "he" and "she". Moreover, the context word of each words also occurs at various frequencies, ultimately affecting the embeddings for both words. On the other hand, as the geometry of embeddings can be affected by frequency, we speculate that the gender direction in some words is not well aligned with the direction generated by Hard Debias. This also constraints the Debias ability of Hard Debias.

Mu & Viswanath (2018) empirically found that the top D dominating directions of word embeddings encode word frequency to a significant degree. Hence, they propose a simple post-processing technique to alleviate the influence of different frequencies. In particular, they subtract the mean vector, then remove a few top dominating directions from the original word embeddings. The post-processed embeddings consistently perform better on multiple tasks, validating the effectiveness of this operation. They argue that the benefits emerge from obtaining new word embeddings that are more "isotropic", i.e. where the words are no longer represented mostly by a few dominant dimensions. As original embeddings are far from isotropic, by eliminating the frequency information, the proposed post-processing yields more isotropic embeddings. Inspired by this, we further look into the effect of this post-processing technique on debasing algorithms.

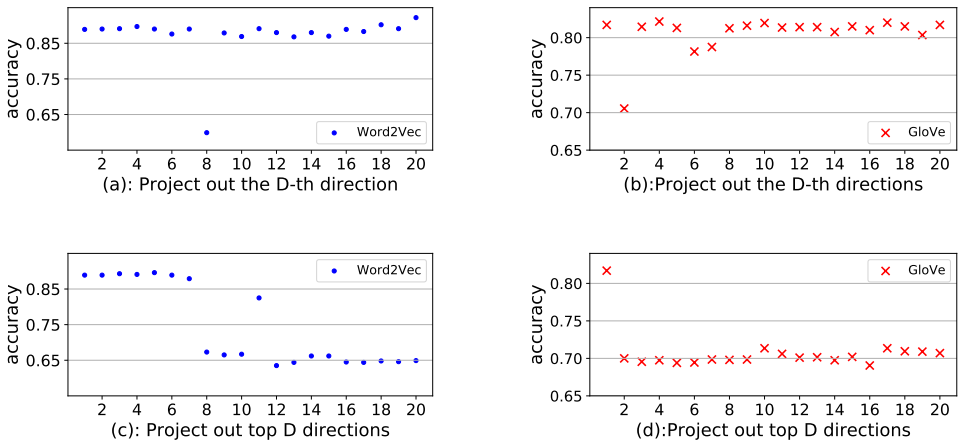


Figure 1: Clustering accuracy after projecting out D-th (a&b) or top D (c&d) dominating directions and Hard Debias. Lower accuracy indicates less bias.

We start with the aforementioned biased words clustering experiment for bias evaluation, where we cluster top biased words and measure the alignment accuracy with respect to gender. Following steps in Mu & Viswanath (2018), we first compute the PCA components of all word embeddings and then project embeddings away from the D_{th} dominating direction. Unsurprisingly, we observe little changes in terms of clustering accuracy. Why does it happen? First, it should be noticed that the post-processing method is supposed to only deal with frequency features implicitly; second, although word frequency becomes similar by the implicit way, it doesn't remove the gender specific words, which still exists in the context of the word. Therefore, the gender bias doesn't disappear and the clustering accuracy is maintained, but by removing the frequency factor on gender, it would only remain the gender direction to be eliminated.

To do so, we then apply Hard Debias on the post-processed embeddings. As we can see from Figure 1(a,b), projecting away the D_{th} (8 for Word2Vec and 2 for GloVe) components from word embeddings effectively helps Hard Debias on the gender evaluation task: Male and Female bias words clustering. This result is akin to experiments in Mu & Viswanath (2018). It clearly show that there exists certain direction that can largely improve the debias result of Hard Debias.[4] Furthermore, we also try to remove the top D components in Figure 1(c,d), clearly at $D = 8$ for Word2Vec it has a sudden drop, then it keeps smoothly no change or even worse which means only 8_{th} component works. This is different from Mu & Viswanath (2018), where they find that a good rule of thumb is to choose top $D = d/100$, where d is the dimension of a word embedding. We posit that while top components are useful for improving general word embeddings, for specific words that are affected by gender bias, certain directions are more helpful to eliminate the harmful effect due to frequency. More importantly, we propose to incorporate clustering as a proxy experiment to effectively decide the optimal direction. Through elaborate experiments in Section 4.2, we demonstrate that proper preprocessing on word embeddings improves debias results and at the same time, maintains the representational power of the original word embeddings. Combining this preprocessing with Hard Debias, we name our approach as Double-Hard Debias and our detailed algorithm is presented below:

Algorithm 1: Double-Hard Debias for word embeddings.

Input : Word embeddings $\{\overrightarrow{w} \in \mathbb{R}^d, w \in \mathcal{W}\}$, set of Male and Female bias words W_m and W_f

$S_{debias} = []$

Decentralize $\overrightarrow{w}$: $\mu \leftarrow \frac{1}{|\mathcal{V}|}\sum_{w\in\mathcal{V}} \overrightarrow{w}$, for each $\overrightarrow{w} \in \mathcal{W}$, $\tilde{w} \leftarrow \overrightarrow{w} - \mu$;

Compute the PCA components: $\{\mathbf{u}_1 \dots \mathbf{u}_d\} \leftarrow$ PCA($\{\tilde{w}, w \in \mathcal{W}\}$);

//discover the component on frequency

for $i = 1$ *to* d **do**

  $w'_m \leftarrow \tilde{w_m} - (\mathbf{u}_i^T w_m)\mathbf{u}_i$;

  $w'_f \leftarrow \tilde{w_f} - (\mathbf{u}_i^T w_f)\mathbf{u}_i$;

  $\hat{w}_m \leftarrow HardDebias(w'_m)$;

  $\hat{w}_f \leftarrow HardDebias(w'_f)$;

  $output = Kmeans([\hat{w}_m \hat{w}_f])$;

  score = eval(output, W_m, W_f);

  S_{debias}.append(score);

end

$k = \arg\min_i S_{debias}$;

//remove component on frequency

$w' \leftarrow \tilde{w} - (\mathbf{u}_k^T w)\mathbf{u}_k$;

//remove component on gender direction

$\hat{w} \leftarrow HardDebias(w')$;

Output: Debiased word embeddings $\{\hat{w}, \hat{w} \in \mathbb{R}^d\}$

[4] We show the top 20 components for readability.

4 EXPERIMENT

We use off-the-shelf Word2Vec embeddings trained on the Google News Dataset and GloVe embeddings trained on the wikidump dataset in our experiments.[5] We evaluate Double-Hard Debias across multiple bias measurements. Additionally, we benchmark it on word similarity, concept categorization, word analogy and coreference resolution. We show that Double-Hard Debias is able to reduce more gender bias and at the same time, maintain the quality of word embeddings.

4.1 DEBIAS RESULTS

Clustering Male and Female Biased Words. We first take the most biased words (500 male and 500 female) according to their cosine similarity with $\overrightarrow{he}$ and $\overrightarrow{she}$ in the original embedding space. We then run k-Means to cluster them into two clusters and compute their alignment accuracy with respect to gender, results are presented in Table 1. Using the original Word2Vec and GloVe embeddings, k-Means can accurately cluster selected words into a male group and a female group. Hard Debias is able to reduce bias in some way while GN-GloVe (Zhao et al., 2018b) appears to be less effective on this test. Double-Hard Debias , however, reaches an accuracy of 59.9 and 74.1 on Word2Vec and GloVe , indicating much less gender information after Debias. We also conduct tSNE (van der Maaten & Hinton, 2008) projection of the embeddings before/after Debias. As shown in Figure 2, from left to right, the male group and female group are closer to each other and confirming less gender cues after Double-Hard Debias .

Classifying Male and Female Biased Words. We follow the same setting in Gonen & Goldberg (2019), where the objective is to try to predict gender from the debiased word embeddings. We train an RBF-kernel SVM classifier with embeddings of 1000 biased words (500 male and 500 female) and test it on other 4000 embeddings (2000 male and 2000 female). The result is in line with the previous clustering experiment. As we can see in Table 1, Double-Hard Debias achieves the lowest accuracy on both Word2Vec and GloVe , suggesting a better Debias result. Note that this experiment explicitly finds common patterns between training data and test data, thus it is hard to reach an accuracy of as low as 50% and the absolute accuracy value may not be a good reflection of gender bias.

Embeddings	Clustering	Classification
Word2Vec	99.9	99.3
Hard Word2Vec	89.2	90.9
Double-Hard Word2Vec	**59.9**	**87.9**
GloVe	100.0	100.0
Hard GloVe	77.2	92.3
GN-GloVe	99.5	99.5
Double-Hard GloVe	**74.1**	**91.3**

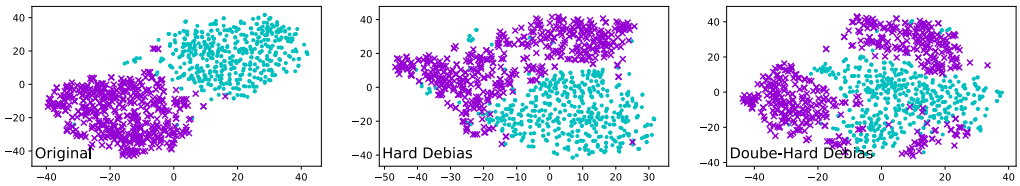
Table 1: Accuracy(x100) of clustering/classifying male and female words. Lower accuracy means less gender cues can be captured. Double-Hard Debias reaches the lowest accuracy on both tasks.

The Word Embeddings Association Test (WEAT). WEAT is a permutation test used to measure the bias in word embeddins. For more details, we refer the reader to Caliskan et al. (2017). We consider male names and females names as attribute sets and compute the differential association of the two sets of target words[6] and the gender attribute sets. We report effect sizes (d) and p-values (p) in Table2[7]. The effect size is a normalized measure of how separated the two distributions are. A higher value of effect size indicates larger bias between target words in regards to gender. A high p-value higher (more than 0.5) indicates the lack of bias. With different target words sets, Double-Hard Debias consistently outperforms other Debias methods.

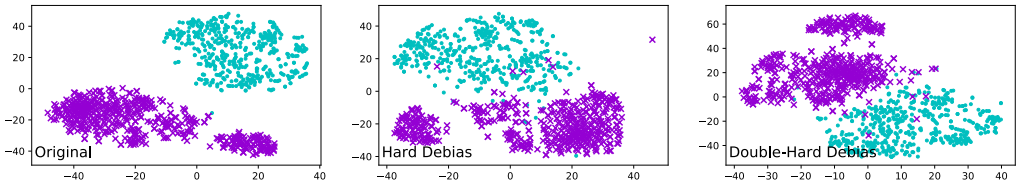
[5]The embeddings used are not centered and normalized to unit length as in Bolukbasi et al. (2016).

[6]All word lists are from Caliskan et al. (2017)

[7]We use lower cased names and replace "bill" with "tom" as we use uncased GloVe embeddings.



(a) Word2Vec



(b) GloVe

Figure 2: tSNE visualization of male and female words before/after Debias. Double-Hard Debiased embedding are more mixed up, showing less gender information encoded.

Embeddings	Career & Family		Math & Arts		Science & Arts	
	d	p	d	p	d	p
Word2Vec	1.89	0.0	1.82	0.0	1.57	$2e^{-4}$
Hard Word2Vec	1.81	0.0	1.62	0.0	0.91	0.03
Doubel-Hard Word2Vec	**1.73**	0.0	**1.51**	$5e^{-4}$	**0.68**	**0.09**
GloVe	1.80	0.0	0.66	0.07	0.94	0.03
Hard-GloVe	1.52	$8e^{-4}$	$5e^{-4}$	0.50	0.21	0.64
GN-GloVe	1.76	0.0	1.45	$2e^{-3}$	1.11	$6e^{-3}$
Double-Hard GloVe	**1.50**	$6e^{-4}$	$3e^{-3}$	0.50	0.14	0.60

Table 2: WEAT test of embeddings before/after Debias. The bias is significant when p-value, $p < 0.05$. Lower effective size (d) indicates less gender bias.

4.2 Retaining Desirable Properties

Word Similarity. The word similarity tasks evaluates if the similarities between word pairs in the embedding space are consistent with human judgements, in terms of Spearmans rank correlation. Original and debiased embeddings are evaluated on multiple datasets: WS353 (Finkelstein et al., 2001), RG65 (Rubenstein & Goodenough, 1965), the SimLex (Hill et al., 2015), MTurk (Radinsky et al., 2011), MEN (Bruni et al., 2014)), RW (Luong et al., 2013). The detailed performan is reported in Table3, where we see that embeddings after Double-Hard Debias preserve and at times increase the semantic information in the embeddings

Embeddings	WS353	RG65	SimLex	MTurk	RW	MEN
Word2Vec	**70.0**	76.1	44.2	68.4	53.4	77.1
Hard Word2Vec	69.9	76.5	44.3	68.4	53.4	76.9
Doubel-Hard Word2Vec	69.5	**77.3**	**44.5**	**68.6**	**54.1**	**78.1**

Table 3: Before/After Debias results (x100) on word similarity task on six datasets.

Concept Categorization. The goal of concept categorization is to cluster a set of words into different categorical subsets. For example, "sandwich" and "hotdog" are both food and "dog" and "cat" are both animals. The clustering performance is evaluated in terms of purity (Manning et al., 2008) - the fraction of the total number of the words that are correctly classified. Experiments are conducted on four datasets: the Almuhareb-Poesio (AP) dataset (Almuhareb, 2006); the ESSLLI 2008 (Baroni et al., 2008); the Battig 1969 set (Battig & Montague, 1969) and the BLESS dataset (Baroni

& Lenci, 2011). We run classical k-Means algorithm with fixed k. Again, across four datasets, the performance after Debias is on a par with original ones, full results are in Table4.

Word Analogy. Given three words A, B and C, the analogy tasks are to find word D such that "A is to B as C is to D". In our setting, D is the word that maximize the cosine similarity between D and $C - A + B$. We evaluate embeddings on MSR (Mikolov et al., 2013c) dataset which contains 8000 syntactic questions and Google (Mikolov et al., 2013a) which contains $19,544$ questions, including $8,869$ semantic and $10,675$ syntactic questions. The evaluation metric for the analogy tasks is the percentage of questions for which the correct answer is assigned the maximum score by the algorithm. Double-Hard Debias achieves comparable results for both syntactic and semantic parts as shown in Table4, indicating our trick is capable of preserving proximity among words.

Embeddings	**Analogy**			**Concept Categorization**			
	Google-Sem	Google-Syn	MSR	AP	ESSLI	Battig	BLESS
Word2Vec	**24.8**	**66.5**	73.6	**64.5**	75.0	46.3	**78.9**
Hard Word2Vec	23.8	66.3	73.5	62.7	75.0	**47.1**	77.4
Double-Hard Word2Vec	23.5	66.3	**74.0**	63.2	75.0	46.5	77.9
GloVe	80.5	**62.8**	**54.2**	55.6	**75.0**	49.0	81.0
Hard GloVe	80.3	62.5	54.0	57.1	70.5	48.9	78.0
GN-GloVe	77.7	61.6	51.9	57.6	72.7	**49.2**	**82.5**
Double-Hard GloVe	**80.9**	61.6	53.8	**59.6**	72.7	47.2	79.5

Table 4: Before/After Debias results (x100) on word analogy and concept categorization task.

Coreference Resolution. In addition, we examine Debias methods on a more complex downstream task. Coreference resolution aims at identifying noun phrases referring to the same entity. Zhao et al. (2018a) introduces a new benmark, WinoBias to certify gender bias in coreference resolution system. WinoBias provides pro-stereotype (PRO) and antistereotype (ANTI) subsets. In the PRO subset, gender pronouns refer to professions dominated by the same gender. For example, in the sentence "The physician hired the secretary because he was overwhelmed with clients.", the pronoun "he" refers to "physician", which is consistent with societal stereotype. On the other hand, the ANTI subset consists of the same set of sentences, but the opposite gender pronouns. As such, "he" is replaced by "she" in the aforementioned example. The hypothesis is that gender cues may distract the model. We consider a system to be gender biased if it performs more better in pro-stereotypical scenarios than in anti-stereotypical scenarios.

We train the end-to-end coreference resolution model (Lee et al., 2017) with different word embeddings on OntoNote training set and report their performance on the Ontonotes5.0 test set and the WinoBias dataset. We also include the average (Avg) and absolute difference (Diff) of F1 scores on PRO and ANTI for the WinoBias dataset. Note that a smaller Diff value indicates a less biased coreference system. Results in Table5 shows that thr proposed trick does not degrade the coreference performance but significantly reduce gender bias.

Embeddings	**OntoNotes-Test**	**PRO**	**ANTI**	**Avg**	**|Diff |**
Word2Vec	**67.2**	75.2	46.1	60.7	29.1
Hard Word2Vec	67.1	66.3	56.1	**61.2**	10.2
Double Hard Word2Vec	67.1	62.4	57.7	60.1	**4.7**
GloVe	**66.5**	76.2	46.0	61.1	30.2
Hard GloVe	66.2	70.6	54.9	**62.8**	15.7
GN-GloVe (w_a)	65.9	70.0	53.9	62.0	16.1
Double-Hard GloVe	66.2	63.6	59.0	61.3	**4.6**

Table 5: F1 score (%) of coreference systems trained on before/after debasing word embeddings. Smaller |Diff | value suggests the coreference system is less gender biased.

5 Related Work

Gender Bias in Word Embeddings. Word embeddings have been proven to carry social biases (e.g. gender and race). Bolukbasi et al. (2016) shows that Word2Vec (Mikolov et al., 2013b) embeddings trained on Google News dataset associate "programmer" with "man" closer than "woman". Such a bias will also propagate to downstream tasks, e.g. coreference systems (Zhao et al., 2018a) and machine translation (Stanovsky et al., 2019). More recently, researchers also observe significant gender bias in modern contextual word embeddings (Zhao et al., 2019; Kurita et al., 2019). To mitigate gender bais, Bolukbasi et al. (2016) introduces a post-processing method which zeros out the component along the gender direction of emebddings. Zhao et al. (2018b); Kaneko & Bollegala (2019) tackle this problem by proposing a new training procedure to explicitly restrict gender information in certain dimensions. While existing methods reduce gender bias in some degree, Gonen & Goldberg (2019) presents a series of experiments to show that they are far from delivering gender-netural embeddings. Our work builds on top of Bolukbasi et al. (2016). We propose a simple but effective preprocessing trick that helps to further reduce gender bias.

Word Embedding Learning. Our work also connects with recent research on word embedding learning. Gong et al. (2018) proposes to learn frequency-free embeddings through adversarial training. They found that words with various frequencies behave differently in the embedding space, which can potentially harm the semantic meaning. Similarly, Mu & Viswanath (2018) validates that dominating directions encode frequency and by eliminating them, performance gets improved on word embedding benchmarks. More generally, Wang et al. (2018) introduces a post-processing technique to normalize the variances of word embeddings instead of removing certain directions. We draw inspiration from these works and further transfer it to reducing gender bias in word embeddings. We provide extensive experiments to show that simple operation on word embeddings can advance existing debiasing algorithms.

6 Conclusion

Gender Bias has attracted lots of attentions and been widely studied in NLP. In this paper, we find out there is another important factor: words frequency, which is neglected in previous gender bias reduction works. Motivated by this observation, we proposed Double-Hard Debias method, which is composed of two stages: we first use PCA to obtain candidate components and adopt clustering on male and female biased words as a proxy method to choose the component on frequency; then we project out the frequency component and gender direction from the embeddings via Hard Debias. We experiment on several benchmarks and demonstrate that our Double-Hard Debias is more effective on gender bias reduction with semantics preserving. In future, we plan to study different way of finding the frequency component and also extend the proposed method for other types of bias reduction.

References

Abdulrahman Almuhareb. *Attributes in lexical acquisition*. PhD thesis, University of Essex, Colchester, UK, 2006. URL http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.428974.

Marco Baroni and Alessandro Lenci. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pp. 1–10, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-16-9. URL http://dl.acm.org/citation.cfm?id=2140490.2140491.

Marco Baroni, Stefan Evert, and Alessandro Lenci. Bridging the gap between semantic theory and computational simulations: Proceedings of the esslli workshop on distributional lexical semantics. 2008.

William F. Battig and William E. Montague. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, 80(3p2):1, 1969. doi: 10.1037/h0027577.

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

Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January 2014. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=2655713.2655714.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. ISSN 0036-8075. doi: 10.1126/science.aal4230.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1696–1705, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1166.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pp. 406–414, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: 10.1145/371920.372094. URL http://doi.acm.org/10.1145/371920.372094.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *NAACL-HLT*, 2019.

Chengyue Gong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Frage: Frequency-agnostic word representation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1334–1345. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7408-frage-frequency-agnostic-word-representation.pdf.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. doi: 10.1162/COLI_a_00237. URL https://www.aclweb.org/anthology/J15-4004.

Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. *CoRR*, abs/1906.00742, 2019. URL http://arxiv.org/abs/1906.00742.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *CoRR*, abs/1906.07337, 2019.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 188–197. Association for Computational Linguistics, 2017. ISBN 978-1-945626-83-8. URL https://www.aclweb.org/anthology/D17-1018/.

Thang Luong, Richard Socher, and Christopher Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3512.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008. ISBN 978-0-521-86571-5. URL http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 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*, pp. 746–751, Atlanta, Georgia, June 2013c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1090.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. Debiasing embeddings for fairer text classification. In *1st ACL Workshop on Gender Bias for Natural Language Processing*, 2019.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 337–346, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963455. URL http://doi.acm.org/10.1145/1963405.1963455.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October 1965. ISSN 0001-0782. doi: 10.1145/365628.365657. URL http://doi.acm.org/10.1145/365628.365657.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1164. URL https://www.aclweb.org/anthology/P19-1164.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Bin Wang, Fenxiao Chen, Angela Wang, and C.-C. Jay Kuo. Post-processing of word representations via variance normalization and dynamic embedding. *CoRR*, abs/1808.06305, 2018. URL http://arxiv.org/abs/1808.06305.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018a.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *EMNLP*, 2018b.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.