Learning Bias-reduced Word Embeddings Using Dictionary Definitions

Haozhe An[*]
University of Maryland, College Park
haozhe@umd.edu

Xiaojiang Liu and **Jian Zhang**
Apple
{xiaojiang_liu, donald_zhang}
@apple.com

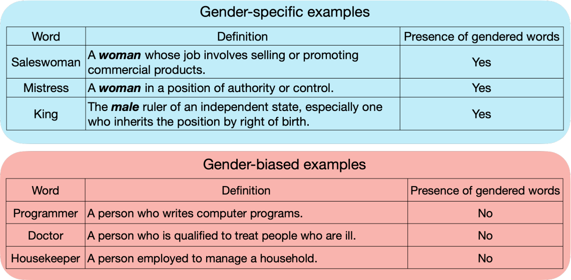
Abstract

Pre-trained word embeddings, such as GloVe, have shown undesirable gender, racial, and religious biases. To address this problem, we propose DD-GloVe, a train-time debiasing algorithm to learn word embeddings by leveraging dictionary definitions. We introduce dictionary-guided loss functions that encourage word embeddings to be similar to their relatively neutral dictionary definition representations. Existing debiasing algorithms typically need a pre-compiled list of seed words to represent the bias direction, along which biased information gets removed. Producing this list involves subjective decisions and it might be difficult to obtain for some types of biases. We automate the process of finding seed words: our algorithm starts from a single pair of initial seed words and automatically finds more words whose definitions display similar attributes traits. We demonstrate the effectiveness of our approach with benchmark evaluations and empirical analyses. Our code is available at https://github.com/haozhe-an/DD-GloVe.

1 Introduction

Word embeddings can meaningfully capture semantic and syntactic similarities between words. Popular embeddings are Word2Vec (Mikolov et al., 2013b), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017). Although contextual word embeddings, like BERT embeddings (Devlin et al., 2019) and ELMo (Peters et al., 2018), gain increasing popularity, some recent research keeps using static word embeddings as input to their state-of-the-art algorithms in downstream natural language processing and computer vision applications (Guan et al., 2021; Gao et al., 2021).

Despite the effectiveness of word embeddings, biases in them reflect undesirable association between some concepts. Bolukbasi et al. (2016) first identify that the distance between $\overrightarrow{\text{man}}$ and $\overrightarrow{\text{woman}}$ is close to that between $\overrightarrow{\text{programmer}}$ and $\overrightarrow{\text{homemaker}}$. Similar phenomena in word embeddings lead to biased interpretations in the word analogy task, associating certain words with gender, racial, and religious stereotypes (Manzini et al., 2019). Deploying such biased word embeddings in downstream tasks would cause allocational and representational harms (Blodgett et al., 2020). It is important to learn bias-reduced word embeddings.



Gender-specific examples

Word	Definition	Presence of gendered words
Saleswoman	A ***woman*** whose job involves selling or promoting commercial products.	Yes
Mistress	A ***woman*** in a position of authority or control.	Yes
King	The ***male*** ruler of an independent state, especially one who inherits the position by right of birth.	Yes

Gender-biased examples

Word	Definition	Presence of gendered words
Programmer	A person who writes computer programs.	No
Doctor	A person who is qualified to treat people who are ill.	No
Housekeeper	A person employed to manage a household.	No

Figure 1: Definitions of example gender-specific and gender-biased words. Gender-specific words typically contain gendered words in their definitions, whereas gender-biased words tend to have neutral definitions.

Dictionary definitions, however, are a neutral source for mitigating biases in word embeddings. The objective, impartial, and concise definitions of words in a dictionary could be unbiased reference points. We propose to encourage word embeddings to be similar to their relatively neutral representations in a dictionary for bias reduction. We simultaneously train and debias the word embeddings from a new initialization point, so as to learn distributional representations and mitigate biases using dictionary definitions concurrently. In addition, several gender-debiasing algorithms rely on a list of pre-compiled seed words to approximate the gender direction, along which the vector component is removed for bias mitigation. We find that, given one pair of the initial seed words, dictionary definitions can help automatically search relevant seed words. Thus, the compilation of seed

[*] Work done during an internship at Apple.

Findings of the Association for Computational Linguistics: ACL 2022, pages 1139 - 1152
May 22-27, 2022 ©2022 Association for Computational Linguistics

words becomes automated. We also find that the automatically generated seed words better capture the notion of gender in the word embedding space.

Our contributions Leveraging the advantages of dictionary definitions, we propose DD-GloVe, a train-time debiasing algorithm to learn bias-reduced GloVe word embeddings. In summary, we make the following contributions:

1. We propose four dictionary-guided loss functions that encourage word embeddings to contain less biased information and richer semantic knowledge by referencing to their relatively neutral dictionary definition representations. (Sec. 3.1)
2. DD-GloVe automatically approximates the bias direction given only one pair of initial seed words. This method finds the most attribute-specific definitions by computing the definition embeddings' projection onto the difference of the initial seed words' definition embeddings. We average the embeddings of the most attribute-specific words to approximate the bias direction. (Sec. 3.2)
3. We empirically demonstrate that DD-GloVe effectively learns bias-reduced word embeddings as we achieve state-of-the-art results in WEAT. Also, our experiments show that debiasing is achieved without sacrificing semantic meanings. (Sec. 4)

2 Motivations

We analyze the limitations in current debiasing algorithms for word embeddings and present our corresponding solutions.

Debiasing algorithms Existing mainstream gender-debiasing algorithms are projection-based post-processing (Bolukbasi et al., 2016; Wang et al., 2020). They need a list of manually selected words (e.g. "she" and "he", "girl" and "boy", "woman" and "man") to compute a gender direction in the word embedding space. They then project the pre-trained word embeddings onto the gender direction and remove the vector component living in this direction. The resultant word vectors preserve useful semantic meanings but contain less gender information. However, these algorithms do not consider the possible usage of additional knowledge like dictionary definitions. Furthermore, there is a limitation in this projective post-processing approach. The manually compiled list to approximate the bias direction might be difficult to obtain for other types of biases. It would be helpful to find an alternative that involves less human labor.

Our approach: using dictionary definitions Using dictionary definitions to train bias-reduced word embeddings could address the above limitation and gives us additional advantages.

(1) Dictionary definitions provide a source of unbiased word representations for debiasing. We define *gender-specific words* as words that are supposedly associated with a particular gender by their definitions. Some examples of gender-specific words are "countryman", "countrywoman", "fraternal", and "sororal." We define *gender-biased words* as words that could refer to a person of any gender but tend to be stereotypically recognized as one gender due to human biases. For example, "nurse", "cashier", and "driver" are gender-biased words. *Gendered words* are a list of 1,441 words compiled by Wang et al. (2020) that explicitly define or describe a gender. Examples of gendered words are like "man", "woman", "he", and "she." In a dictionary, gender-specific words typically contain gendered words in their definitions, whereas gender-biased words tend to have neutral definitions. Example words and their definitions from Oxford online dictionary[1] are shown in Fig. 1. We further obtain 379 gender-specific words, compiled by Wang et al. (2020), and 40 words of gender-biased occupations, compiled by Zhao et al. (2018a), to verify if this trend is general. For each definition of the words, we check whether any gendered words are present. We find that gendered words are absent from 39 out of 40 gender-biased occupations. This result shows dictionary definitions are almost bias-free. In contrast, gendered words are present in 327 out of 379 gender-specific words' definitions. This shows that if a definition contains a gendered word, it is highly likely that the word defined is gender-specific. Dictionary definitions can thus act as a reliable guidance for bias mitigation.

(2) Dictionary definitions could automate the process of finding seed words that approximate the bias direction. We compare definition similarities to find words that commonly associate with some attribute. It is relatively easy to obtain one pair of seed words that describe two opposite concepts associated with a protected attribute (e.g. "she" and "he" for gender). We then look into the definitions of these initial seed words, and find other words whose definitions are similar to theirs. As a measure of similarity, we compute the projection onto

[1] https://www.lexico.com/

the difference between the definition embedding of one initial seed word and the definition embedding of the other. Detailed algorithm is described in Sec. 3.2. This method avoids using manually compiled words to approximate the bias direction.

(3) Dictionary definitions offer additional semantic knowledge. Researchers improve word embeddings using dictionary definitions (Faruqui et al., 2015; Tissier et al., 2017). These works primarily enhance semantic meanings of word embeddings rather than reduce biases in them. Nevertheless, their successes indicate the possibility to preserve, or even enhance, the semantic meaning representations of word embeddings as we use dictionary definitions to debias them.

Existing dictionary debiasing algorithm A recent work makes the first attempt to debias word embeddings using dictionary definitions via post-processing (Kaneko and Bollegala, 2021). They compute a weighted average of pre-trained word vectors as the definition embeddings. They assume these definition embeddings are the "neutral" reference points for word embeddings. However, this is a major flawed assumption in post-processing debiasing. Due to the biases in pre-trained word vectors, the definition embeddings also contain biases. Partially owing to this flawed assumption, their resultant embeddings show limited effectiveness in several benchmark evaluations like the Word Embedding Association Test (Caliskan et al., 2017).

Our approach: training from scratch Training from scratch addresses the problem of biased definition embeddings computed from pre-trained, biased word vectors. As word embeddings are initialized randomly, they contain virtually no biases. Correspondingly, the definition embeddings obtained at this point will contain minimal biases. As training proceeds, the debiasing algorithm can continuously apply corrections, so as to learn distributional semantics and reduce biased information simultaneously. In Sec. 5.1, we empirically demonstrate that training from scratch could produce substantially more neutral definition embeddings that lead to improved debiasing.

3 DD-GloVe

We propose four dictionary-guided loss functions, namely (1) **orthogonal loss**, which mitigates general biases by diminishing the redundant component in word vectors that disagree with their definition embeddings, (2) **projection loss**, which directly reduces a specific type of bias by minimizing the difference between word vectors' projection and definitions' projection onto the bias direction, (3) **definition loss**, which injects semantic meanings from definitions into word embeddings, and (4) **bias-aware GloVe loss**, which dynamically adjusts weights of co-occurrences for bias reduction.

In addition, we introduce a novel algorithm that automatically searches seed words for bias direction approximation with only one pair of initial seed words as the input.

Notations We use $w \in \mathbb{R}^d$ to denote word vectors with dimension d. We overload the symbol w to represent a word in some contexts. $s(w)$ denotes the definition embedding of word w. A word can have multiple definitions in a dictionary. Since GloVe does not distinguish word meanings, we choose to use all available definitions for w when computing $s(w)$. Previous works compute definition embeddings by smoothed inverse frequency (Arora et al., 2017; Kaneko and Bollegala, 2021). We propose a simpler but empirically effective method that averages the definitional words. Therefore, our definition embedding is

$$s(w) = \frac{1}{K}\sum_{i=1}^{K} h(w)_i \tag{1}$$

where h is the function that returns all definitional words (excluding stop words) of w, and $K = |h(w)|$ is the number of definitional words.

3.1 Dictionary-guided Loss Functions

Orthogonal loss for general debiasing The definition embedding $s(w)$ reflects the redundant encoding in w, which is defined as

$$\phi\left(w, s(w)\right) = w - \frac{w \cdot s(w)}{s(w) \cdot s(w)} s(w) \tag{2}$$

where $(\cdot)$ is the dot product of vectors. $\phi\left(w, s(w)\right)$ represents the unnecessary, and likely biased, meaning encoded in the word vector w, because $\phi\left(w, s(w)\right)$ is the component in w that lives in the subspace orthogonal to $s(w)$.

We minimize the squared dot product between $\phi\left(w, s(w)\right)$ and w by

$$J_{ortho}(w) = \left(\phi\left(w, s(w)\right) \cdot w\right)^2 . \tag{3}$$

This loss term is ignored if a word does not have definitions in the dictionary. The orthogonal loss mitigates almost all general types of biases because it signals word embeddings to drop any information that is absent from their definition embeddings.

Projection loss for specific debiasing We design a projection-based loss to further enhance the debiasing effectiveness for a specific type of bias. The type of bias depends on use cases. With the definition embedding $s(w)$ as an unbiased reference for w, we want the projection of w onto the bias direction g (g is explained in Sec. 3.2) to be similar to that of $s(w)$. Thus,

$$J_{proj}(w) = \left\| \frac{w \cdot g}{g \cdot g} g - \frac{s(w) \cdot g}{g \cdot g} g \right\|_1 . \quad (4)$$

If the dictionary does not define w, we assume w should be a neutral word and $s(w) \cdot g = 0$. Dictionary definitions would indicate if a word vector should express the meaning associated with a protected attribute. This loss function thus avoids human intervention or using an additional classifier to decide what word to debias.

Definition loss for semantic meaning This loss function aims to inject the semantic meaning represented in dictionary definitions into word embeddings. The definition loss encourages a word vector to be similar to its definition embedding. As a result, it signals word embeddings about what to keep and what is lacking in their semantic meaning representations. We propose to minimize the $l1$-norm difference between w and its definition embedding $s(w)$ via definition loss

$$J_{def}(w) = \|w - s(w)\|_1 . \quad (5)$$

If a word is not defined in the dictionary, we skip its gradient update for this loss term.

Bias-aware GloVe loss The original GloVe loss is a log-bilinear regression of word co-occurrences. Each co-occurrence composes a word and its context word $(w, \tilde{w})$. It is evident that if the training corpus has more balanced word co-occurrences over the protected attributes, the trained word embeddings show a smaller extent of bias (Hall Maudslay et al., 2019; Lu et al., 2020). For example, if "nurse" occurs equally likely with gendered words like "she" and "he", the embedding of "nurse" would be more neutral with respect to genders. To equivalently create more balanced word co-occurrences, we introduce the bias-aware Glove loss. Different from static co-occurrence weights in the original Glove, bias-aware Glove loss adjusts co-occurrence weights according to the bias of a word and its context word.

What co-occurrences should be assigned new weights? If either w or $\tilde{w}$ is biased, we modify its weight, so that the number of co-occurrences containing biased words are equivalently modified. To decide if w (similarly for $\tilde{w}$) is biased in training, we quantify its genderedness by

$$u(w) = \frac{w \cdot v_1}{\|w\| \|v_1\|} - \frac{w \cdot v_2}{\|w\| \|v_2\|} \quad (6)$$

where v_1, v_2 are initial seed words like "she" and "he" (explained in Sec. 3.2). We then compare $u(w)$ with its neutral reference point $s(w)$. Hence, the bias of a word is

$$d(w) = |u(w) - u(s(w))| . \quad (7)$$

Increase or decrease the weights? If a biased w and $\tilde{w}$ are associated with opposite genders (i.e. $u(w)$ and $u(\tilde{w})$ have opposite signs), we assign a higher weight, equivalently increasing such co-occurrences; if a biased w and $\tilde{w}$ are associated with the same gender (i.e. $u(w)$ and $u(\tilde{w})$ have the same sign), we assign a lower weight, equivalently decreasing such co-occurrences.

By how much? The magnitude of the weight change is proportional to the maximum extent of bias in a given co-occurrence pair, which is computed by $\max(d(w), d(\tilde{w}))$.

The proposed weight for a co-occurrence pair is

$$f'(w, \tilde{w}) = \\ 1 - \alpha \cdot \mathrm{sgn}(u(w)) \cdot \mathrm{sgn}(u(\tilde{w})) \cdot \max(d(w), d(\tilde{w})) \quad (8)$$

where we multiply a constant α to keep $f'(w, \tilde{w})$ within a reasonable range, about $[0.9, 1.1]$, for stable performance. The modified GloVe loss is

$$J_{G-bias} = \sum_{i,j=1}^{|V|} f'(w_i, \tilde{w_j}) f(X_{ij}) (w_i^T \tilde{w}_j \\ + b_i + \tilde{b}_j - \log X_{ij})^2 \quad (9)$$

where V is the set of vocabulary, and $b, \tilde{b}$ are scalar bias terms. f is a function that assigns weights to co-occurrence pairs based on their frequency (introduced in GloVe). If a co-occurrence pair contains at least one word that is not defined, we set $f' = 1$.

DD-GloVe loss function Putting all the proposed loss functions together, we have the loss function

$$J = J_{G-bias} + \beta J_{ortho} + \gamma J_{proj} + \lambda J_{def} \quad (10)$$

where β, γ, λ are hyperparameters.

3.2 Approximating the Bias Direction g

Algorithm 1 approximates the bias direction g with a single pair of initial seed words. Let a pair of attribute-specific words be (v_1, v_2) such that word vector difference $v_1 - v_2$ is similar to the true bias direction associated with the protected attributes $\mathcal{A}_1$ and $\mathcal{A}_2$. For example, (v_1, v_2) could be "she" and "he" for gender debiasing, and the corresponding $\mathcal{A}_1$ and $\mathcal{A}_2$ are female and male respectively. We find two sets of most attribute-specific definitions $Q_{\mathcal{A}_1}$ and $Q_{\mathcal{A}_2}$ along $s(v_1) - s(v_2)$ by looking at definition embeddings' projection onto this direction. The sizes of $Q_{\mathcal{A}_1}$ and $Q_{\mathcal{A}_2}$ are determined empirically based on the availability of words associated with a certain concept. For instance, in our experiment that focuses on gender-debiasing, we set $N = 30$. One can run Algorithm 1 once at the beginning of training to obtain a set of seed words that will be used throughout the training, or run Algorithm 1 multiple times to update seed words periodically. We find that the former works better with attributes that have a large number of words associated with them, such as gender. The latter tends to fit attributes that have a smaller number of associated words, such as races.

4 Experiments

We present two settings for DD-GloVe. (1) In DD-GloVe$_{gender}$, we mainly mitigate gender bias, thus using "she" and "he" as the initial seed words. (2) DD-GloVe$_{race}$, we focus on reducing racial bias. The initial seed words are "black" and "white".

For each word in the vocabulary of Glove, we try to find its definitions from the Oxford online dictionary. If the word has multiple definitions, we simply concatenate them into one definition. Stopwords are removed for pre-processing. We average the definitional words to obtain $s(w)$ by following Eqn. 1. Words that are not present in the Oxford dictionary are skipped. In total, we have 92,140 words with definitions.

We run GloVe (Pennington et al., 2014), Double Hard Debias (DHD) (Wang et al., 2020), dictionary-based debiasing (Dict Debias) (Kaneko and Bollegala, 2021), and GN-GloVe (Zhao et al., 2018b) as baselines for comparison. The detailed experimental setup is described in the appendix (A.1).

Algorithm 1 Find seed words automatically and approximate the bias direction

```
Input: Initial seed words $(v_1, v_2)$, desired total number of seed words $N$ for each attribute
Output: Two sets of seed words $Q_{\mathcal{A}_1}, Q_{\mathcal{A}_2}$, the approximated bias direction $g$
  $Q_{\mathcal{A}_1} \leftarrow \{v_1\}, Q_{\mathcal{A}_2} \leftarrow \{v_2\}, R \leftarrow \emptyset$
  ▷ Get each word's definition projection onto the difference between the definition embeddings of $v_1, v_2$ i.e. projection along $s(v_1) - s(v_2)$.
  for all $w \in V$ do
      $r(w) \leftarrow \frac{s(w) \cdot s(v_1)}{\|s(w)\| \|s(v_1)\|} - \frac{s(w) \cdot s(v_2)}{\|s(w)\| \|s(v_2)\|}$
      $R \leftarrow R \cup \{(w, r(w))\}$
  end for
  ▷ Find top $N$ most attribute-specific words and approximate the bias direction.
  $R_{sorted} \leftarrow$ Sort $R$ by $r(w)$ in descending order
  for $n \in \{1, 2, \ldots, N\}$ do
      $w_1, r(w_1) \leftarrow R_{sorted}[n]$
      $w_2, r(w_2) \leftarrow R_{sorted}[|R_{sorted}| - n]$
      $Q_{\mathcal{A}_1} \leftarrow Q_{\mathcal{A}_1} \cup \{w_1\}, Q_{\mathcal{A}_2} \leftarrow Q_{\mathcal{A}_2} \cup \{w_2\}$
  end for
  $g \leftarrow \frac{1}{|Q_{\mathcal{A}_1}|} \sum_{w \in Q_{\mathcal{A}_1}} w - \frac{1}{|Q_{\mathcal{A}_2}|} \sum_{w \in Q_{\mathcal{A}_2}} w$
```

4.1 WEAT

To evaluate bias in word embeddings, researchers commonly use Word Embedding Association Test (WEAT) (Caliskan et al., 2017). This test quantifies the strength of association between a set of target words (such as science and arts) and a set of attribute words (such as male and female names). The test result produces effect size d and p-value. If there exist strong associations between target and attribute words, d would be large and p-value would be small. Bias-reduced word embeddings should ideally have low d and high p-values.

We report WEAT results in Table 1. We observe that DD-GloVe$_{gender}$ outperforms all the baselines in gender-related tests. DD-GloVe$_{race}$ performs as effectively as the state-of-the-art dictionary-based debiasing algorithm in racial association test. DD-GloVe$_{race}$ also shows some effects of gender debiasing in Gender-2 test and produces the best result in the nature test. It is evident that DD-GloVe can reduce multiple types of biases simultaneously with an emphasis on the bias we want to mitigate to the greatest extent. This phenomenon benefits

Embeddings	Gender-1		Gender-2		Race		Age		Nature	
	$d\downarrow$	$p\uparrow$	$d\downarrow$	$p\uparrow$	$d\downarrow$	$p\uparrow$	$d\downarrow$	$p\uparrow$	$d\downarrow$	$p\uparrow$
GloVe	1.74	0.00	1.07	0.013	1.18	0.0029	1.03	0.0090	1.15	0.0029
DHD	1.38	0.0014	0.45	0.19	1.06	0.0076	0.88	0.023	1.22	0.0017
Dict Debias	1.68	0.00	1.15	0.0081	0.82	0.033	**0.62**	**0.086**	1.27	0.0012
GN-GloVe	1.80	0.00	1.18	0.0063	1.01	0.010	0.96	0.014	1.21	0.0018
DD-GloVe$_{gender}$	**1.25**	**0.0029**	**0.083**	**0.44**	1.01	0.011	0.94	0.017	1.01	0.0088
DD-GloVe$_{race}$	1.75	7.8e-5	0.77	0.063	**0.80**	**0.037**	0.64	0.078	**0.99**	**0.0099**

Table 1: WEAT results for various word embeddings. The gender attribute set contains male and female names. Gender-1 tests gender v.s. career & family. Gender-2 tests gender v.s. math & arts. The race set consist of European American names and African American names. The age set contains stereotypically young and old names (Nosek et al., 2002). The nature set composes flower and insects vocabulary (Greenwald et al., 1998). Attributes sets of race, age, and nature are tested against pleasant and unpleasant words (Caliskan et al., 2017). For GN-GloVe, we exclude the gender dimension in word embeddings for these tests.

Embeddings	Pro	Anti	Avg	Diff
GloVe	67.03	55.96	61.50	11.07
DHD	60.56	57.99	59.28	2.57
Dict Debias	66.30	57.22	61.76	9.08
GN-GloVe	64.67	60.78	62.73	3.89
DD-GloVe	65.53	57.59	61.56	7.94

Table 2: Coreference resolution F1-score (%) using models trained with different embeddings. We also report the average F1-score (Avg) and the difference (Diff) between pro-stereotype and anti-stereotype subsets in WinoBias. We use all dimensions in GN-GloVe embeddings in this experiment.

from our design of loss functions: orthogonal loss reduces general types of biases while projection loss mitigates the chosen type of bias along g.

4.2 Coreference Resolution

We verify the effects of bias-reduced word embeddings on a downstream task – coreference resolution. WinoBias (Zhao et al., 2018a) is a dataset tailored to measure a model's gender bias when clustering the denotative noun phrases referring to the same entity. It consists of pro-stereotype and anti-stereotype sentences. Every sentence in pro-stereotype subset has a counterpart in the anti-stereotype subset with the gendered pronoun replaced with the opposite one. Models should ideally have similar performance in these two subsets. We train the end-to-end coreference resolution model proposed by Lee et al. (2017) with OntoNotes 5.0 (Weischedel et al., 2012) using various word embeddings. The coreference resolution model is implemented using AllenNLP (Gardner et al., 2017). We evaluate each model using WinoBias Type 1 set.

Model F1-scores are shown in Table 2 and training F1-scores are reported in the appendix. Compared to post-processing dictionary-based debiasing, DD-GloVe produces a lower F1-score difference, indicating less biased information is used to make coreference resolution predictions. DHD outperforms DD-GloVe in terms of F1-score difference, but DD-GloVe enjoys overall higher average. GN-GloVe performs the best in this task, likely because the occupations in WinoBias are found in their manually compiled male and female words. Their model could easily force these words to be completely neutral, whereas DD-GloVe would depend on dictionary definitions to decide the genderedness of words. The occasional noise in definitions may cause DD-GloVe to not outperform.

4.3 Semantic Meaning Preservation

We conduct experiments in word analogy and concept categorization to ensure semantic meaning of word embeddings are well preserved after bias mitigation. The word analogy task tests "A is to B as C is to what?" We find a word vector w that is nearest to $w_A - w_B + w_C$ as the solution. We use Google word analogy (Mikolov et al., 2013a) and MSR (Mikolov et al., 2013c) for evaluation. Concept categorization aims to group words into various categories based on their semantic meanings. The metric for this task is purity (Schütze et al., 2008). We evaluate various embeddings with Almuhareb-Poesio (AP) (Almuhareb, 2006), ESSLLI (Baroni et al., 2008), Battig (Battig and Montague, 1969), and BLESS (Baroni and Lenci, 2011).

Embeddings	Word analogy (%)				Concept categorization (%)			
	G-Sem	G-Syn	G-Total	MSR	AP	ESSLI	Battig	BLESS
GloVe	79.26	63.19	70.48	**54.10**	57.71	66.91	49.42	**83.50**
DHD	79.77	61.65	69.87	53.25	59.20	67.00	46.57	79.50
Dict Debias	79.46	**63.22**	70.59	53.89	**60.95**	66.91	**53.31**	83.00
GN-GloVe	77.11	61.88	68.79	50.55	57.96	60.47	46.68	81.00
DD-GloVe	**80.27**	62.67	**70.66**	53.69	58.71	**67.78**	48.06	76.00

Table 3: Experiments to verify semantic meaning preservation of debiased word embeddings. G-Sem, G-Syn, and G-Total refer to Google-Semantic subset accuracy, Google-Syntactic subset accuracy, and Google word analogy total accuracy respectively.

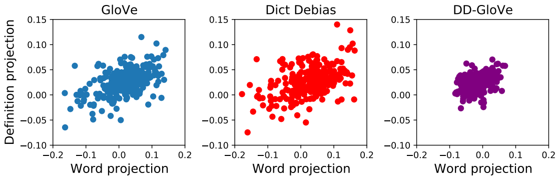


Figure 2: Scatter plots of definition embedding projections against word embedding projections for gender-neutral profession vocabularies. Both the definition embeddings and word embeddings in DD-GloVe consistently have closer-to-zero projection values.

KMeans clustering is run for categorization.

We obtain the top-1 accuracy for word analogy task and purity for concept categorization shown in Table 3. We see that there is minimal degradation in performance in most datasets we have tested. Sometimes, DD-GloVe achieves marginally higher top-1 accuracy or purity than the baseline GloVe. Two reasons lead to the improvement: it is partially due to the trend that using additional knowledge to train word vectors enhances their semantic meaning representations; also, reducing biased information enables fairer predictions in these tasks.

In addition to these experiments, we conduct more extrinsic evaluations for semantic meaning preservation in the appendix (A.2). We find that DD-GloVe preserves useful semantic meanings that help models to perform well in a variety of downstream tasks such as coreference resolution, sentiment analysis, and document classification.

5 Discussion

5.1 Benefit of Training from Scratch

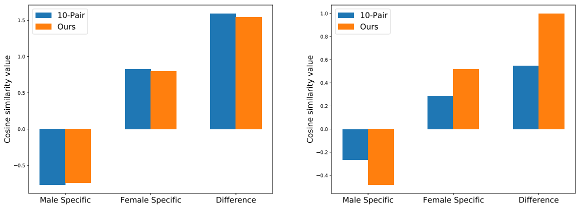
Training from scratch plays a key role in DD-GloVe because it significantly reduces the biases in definition embeddings, which are used as reference points for word embedding debiasing. We use the gender-neutral profession words provided by Bolukbasi et al. (2016). We project their definition embedding and word embedding onto the direction $\overrightarrow{\text{he}} - \overrightarrow{\text{she}}$. We present the scatter plots for three embeddings in Fig. 2. We fix the scale for both axes for easy comparison. In GloVe, a more biased occupation word tends to have a more biased definition embedding. This trend is visible from the strong linear correlation between definition embedding projections and word embedding projections ($p = 1.16 \times 10^{-18}$). Due to the biases in definition embeddings, using the GloVe definition embeddings as the optimization objective in post-processing would not effectively mitigate word embedding biases. Consequently, Dict Debias exhibits a similar trend in its definition embeddings and word embeddings. However, training from scratch allows word vectors to learn semantic meanings from a new random initialization, at which word vectors do not contain meaningful biased information. The definition embeddings will thus contain negligible biases. During training, these more neutral definition embeddings can consistently function as relatively neutral reference points for word embeddings to drop redundant information and keep useful semantic meanings. Shown in Fig. 2, DD-GloVe generates more neutral word and definition embeddings.

5.2 Bias Direction Approximation

We present part of the word list produced by Algorithm 1 in Table 4. Most choices are interpretable by human as they specifically refer to or describe a particular gender. We also quantitatively evaluate the quality of gender direction approximation. Similar to Antoniak and Mimno (2021)'s argument, a good gender direction should have large magnitude in cosine similarity with gender specific words while the signs are opposite for the two gen-

Female	ex-wife, girl, jane, woman, wife, witch, women, she, pilipinas, heroine, maids, hens, dona, wives
Male	he, son, brother, brothers, boys, sons, boy, businessman, yang, gentleman, wizard, headmaster, statesman

Table 4: Sample words chosen by our dictionary-guided algorithm (Algorithm 1) to approximate the gender direction. The full list can be found in the appendix (A.3)



(a) GloVe embeddings

(b) DD-GloVe embeddings

Figure 3: Average cosine similarities between gender specific words and gender directions. “10-Pair” refers to the gender direction computed using the 10 pairs of seed words provided by Bolukbasi et al. (2016). We normalize the cosine values so that their mean is 0 and standard deviation is 1.

ders. This phenomenon would imply that the male-specific words and female-specific words are far apart from the other set when they are projected onto the gender direction.

We borrow 190 male-specific words and 177 female specific words used by Wang et al. (2020) and compute their average cosine similarities with different gender directions. Fig. 3a shows that gender-specific words have similar cosine similarities with both the gender direction used by Bolukbasi et al. (2016) and the gender direction found by our Algorithm 1. This indicates that, in the GloVe embedding space, our gender direction is as effective as the baseline to capture the notion of gender. In DD-GloVe embeddings, our gender direction has greater magnitude of average cosine similarities for both genders. Consequently, the difference between male and female cosine similarity is larger, indicating a clearer manifestation of gender.

5.3 Choice of Initial Seed Words

Initial seed	G-Sem (%)	$d\downarrow$	$p\uparrow$
she-he	80.47	1.25	0.0029
herself-himself	79.63	1.30	0.0012
her-his	80.25	1.50	7.8e-5
girl-boy	81.18	1.38	0.0011
mother-father	80.81	1.71	7.8e-5
woman-man	80.20	1.69	7.8e-5

Table 5: Performance of DD-GloVe on Google-Sem (%) and WEAT gender tests with different initial seed words. We finetune the hyper-parameter for each setting.

We conduct experiments to understand if different initial seed words affect the performance of DD-GloVe. We report our results in Table. 5. While all settings show similarly good semantic meaning preservation, we see that the choice of initial seed words gives rise to varying debiasing results. This is mainly due to the fact that some words have more diverse definitions than others. For example, definition of “he” contains mainly gendered words like *“man”, “boy”, and “male”*, whereas the definition of “man” can be far more general, where it has definitions like *“a human being of either sex; a person.”* As a result, the gender direction approximated by Algorithm. 1 may suffer from the noisy definitional words, leading to less effective debiasing results.

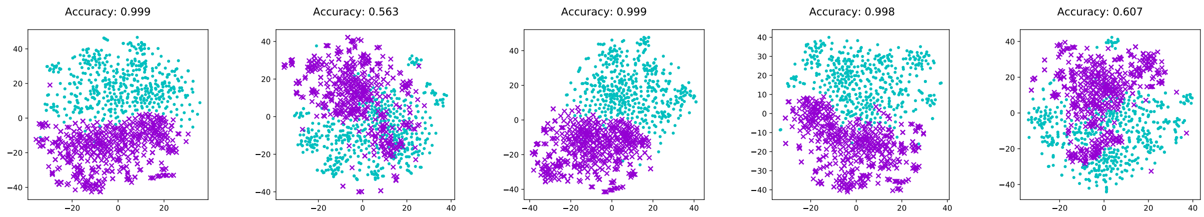
5.4 Does DD-GloVe Simply Hide Biases?

We use the neighborhood metric (Gonen and Goldberg, 2019) to evaluate if the debiased word embeddings actually reduce biases. We cluster these most biased words using the classical KMeans algorithm for different embeddings. We expect effective bias-mitigated word embeddings to achieve a classification accuracy close to 0.5, which indicates word embeddings do not encode any useful information regarding the protected attributes in these words and the clustering algorithm can only make random guesses. Fig. 4 illustrates tSNE projections of the word embeddings of top 500 most gender-biased words in GloVe. The visualization shows that DD-GloVe$_{gender}$ mixes up the embeddings in a similar fashion as Double Hard Debias. In contrast, using dictionary definitions for post-processing debiasing and GN-GloVe tend to hide biases since the two clusters remain easily separable.

5.5 Ablation Study

We carry out an ablation study to better understand the role of each loss in DD-GloVe. Detailed discussions are in the appendix (A.4). We summarize our findings from the ablation study here.

J_{ortho} contributes to both semantic meaning enhancement and general bias reduction in word embeddings when its weight is small. Nonetheless,



(a) GloVe (b) DHD (c) Dict Debias (d) GN-GloVe (e) DD-GloVe$_{gender}$

Figure 4: tSNE projections of word vectors for neighborhood metric evaluation. The most biased words in GloVe are found by projecting word vectors onto the difference between $\overrightarrow{\text{boy}}$ and $\overrightarrow{\text{girl}}$.

this loss term reduces biases at the expense of semantic meaning preservation as its weight gets higher. Hence, the weight for J_{ortho} should be kept relatively low. We also find that J_{ortho} is not the most effective component for bias mitigation but it is still a crucial part for reducing general biases. J_{proj} is essential for effective bias reduction. We find the projection-based loss function largely contributes to debiasing. J_{def} enhances semantic meaning representation but does not help much in bias mitigation. J_{G-bias} further mitigates bias, suggesting that adjusting word co-occurrence weights could help learn bias-reduced word embeddings.

6 Related Work

6.1 Biases in Word Embeddings

Biases in embeddings can cause harms in downstream tasks. Gender bias is found in coreference resolution (Rudinger et al., 2018; Zhao et al., 2018a), dialogue systems (Henderson et al., 2018) and machine translation models (Escudé Font and Costa-jussà, 2019). Researchers also find pretrained word embeddings exhibit racial and religious biases (Manzini et al., 2019).

6.2 Debiasing Word Embeddings

Algorithms to debias word embeddings can be classified into projection-based post-processing, dictionary-based post-processing, and train-time algorithms. **Projection-based post-processing** subtracts a word vector's projection onto the bias direction. Bolukbasi et al. (2016), Wang et al. (2020), Ravfogel et al. (2020), Kumar et al. (2020), Kaneko and Bollegala (2019), Dev and Phillips (2019), and Karve et al. (2019)'s works fall into this category. **Dictionary definitions** have been largely overlooked by debiasing algorithms. Kaneko and Bollegala (2021) uses dictionary definitions via post-processing, but its effectiveness is limited due to using biased definition embeddings as reference points. **Train-time algorithms** either introduce bias-decreasing objectives (Zhao et al., 2018b) or counter-factually augment training data (Lu et al., 2020; Hall Maudslay et al., 2019).

6.3 Using Additional Knowledge

Researchers have attempted to learn word embeddings with resources outside the training corpora. Faruqui et al. (2015); Mrkšić et al. (2017); Tissier et al. (2017); Bosc and Vincent (2018); Zhang et al. (2020) are successful in enhancing semantic meaning representations with the aid of semantic relationships in word graphs or dictionaries. However, these works do not mitigate biases. In DD-GloVe, we specifically design loss functions that utilize dictionary definitions for bias alleviation.

7 Conclusion

In this paper, we propose DD-GloVe, a train-time debiasing algorithm to learn word embeddings leveraging dictionary definitions. We achieve effective debiasing results while preserving semantic meanings. The bias direction in DD-GloVe is automatically approximated using our dictionary-guided algorithm given a single pair of initial seed words. Our current implementation is based on GloVe, but the idea of using dictionary definitions to mitigate biases can be generalized to other word embeddings since our dictionary-guided losses are orthogonal to word embedding objectives. It is also likely that incorporating dictionary definitions can alleviate biases in contextualized word embeddings. This is out of the scope of this paper and remains an open research problem.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. We also thank Professor Rachel Rudinger for her helpful discussion when we prepared the manuscript.

References

Abdulrahman Almuhareb. 2006. *Attributes in lexical acquisition*. Ph.D. thesis, University of Essex.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online. Association for Computational Linguistics.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium. 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(6334):183–186.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. 2021. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3064–3073.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Chaoyu Guan, Xin Wang, and Wenwu Zhu. 2021. Autoattend: Automated attention representation search. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3864–3874. PMLR.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Dictionary-based debiasing of pre-trained word embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 212–223, Online. Association for Computational Linguistics.

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

Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. 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.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Julien Tissier, Christophe Gravier, and Amaury Habrard. 2017. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Copenhagen, Denmark. Association for Computational Linguistics.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Online. Association for Computational Linguistics.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D Hwang, Claire Bonial, et al. 2012. Ontonotes release 5.0.

Yichi Zhang, Yinpei Dai, Zhijian Ou, Huixin Wang, and Junlan Feng. 2020. Improved learning of word embeddings with word definitions and semantic injection. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4253–4257. ISCA.

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, New Orleans, Louisiana. Association for Computational Linguistics.

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

A Appendix

A.1 Experimental Set-up

We give a more detailed description of our experimental set-up in this section.

We use Wikipedia dump available on Hugging Face[2] as our training corpora. We follow the same pre-processing procedure in the original GloVe implementation. We build a vocabulary of 400,000 most frequently occurring words. We set the dimension of word vector to be 300. Although the baseline GloVe is trained with 100 iterations, we find that training about 40 iterations yields excellent debiasing result while keeping the quality of word embeddings in other semantic tasks. We clip the the values in word vectors to be within $[-1, 1]$ to avoid numerical difficulties.

In the setting of DD-GloVe$_{gender}$, we place major emphasis on minimizing gender bias while mitigating other types of biases. We use one pair of initial seed words, “she” and “he”. We run Algorithm 1 once at the beginning with $N = 30$. We then use the same set of seed words throughout. Gender direction is approximated once in each iteration. We choose the hyperparameter values in Eqn. 10 to be $\beta = 1\times10^{-4}, \gamma = 0.2, \lambda = 1\times10^{-4}$. Note that the difference in the magnitude is caused by the trend that definition loss and orthogonal loss have considerably larger values because the losses are not normalized by the vector dimension. We set α in Eqn. 8 to be 0.4.

We also conduct experiments that targets to mitigate racial bias In this experiment DD-GloVe$_{race}$, we find seed words using Algorithm 1 in the first 5 iterations and update them every 10 iterations. The initial seed words are “black“ and “white.” We choose the hyperparameter values $\beta = 1 \times 10^{-4}, \gamma = 0.05, \lambda = 1 \times 10^{-4}$. α in Eqn. 8 remains 0.4.

We run GloVe (Pennington et al., 2014), Double Hard Debias (DHD) (Wang et al., 2020), dictionary-based debiasing (Dict Debias) (Kaneko and Bollegala, 2021), and GN-GloVe (Zhao et al., 2018b) as baselines for comparison. When reproducing the baselines, we follow the default hyperparameter settings in their released code. Each baseline algorithm represents a major debiasing

[2] https://huggingface.co/datasets/wikipedia

Embeddings	OntoNotes 5.0
GloVe	60.50
DHD	59.61
Dict Debias	60.66
GN-GloVe	60.78
DD-GloVe	60.44

Table 6: Coreference resolution F1-score (%) using models trained with different embeddings. These results show that Dd-GloVe keeps useful semantic meanings in embeddings since the F1-score on OntoNotes 5.0 is similar to the baseline and its counterparts.

Word Embeddings	Sentiment Analysis	Document Classification
GloVe	87.94	74.16
DD-GloVe	88.34	74.45

Table 7: F-1 score (%) of models in two downstream tasks. These results show that DD-GloVe well preserve semantic meaning of word vectors after debiasing.

technique: DHD uses projective correction via post-processing; Dict Debias uses dictionary definitions in post-processing. GN-GloVe trains GloVe from scratch with new objectives for debiasing.

A.2 Additional Experimental Results

We report coreference resolution models' F1-score on the training set OntoNotes 5.0 in Table 6. These results indicate that DD-GloVe is able to preserve useful semantic meanings that help train coreference resolution models.

We conduct additional experiments to evaluate model F-1 scores in downstream tasks. We train an LSTM model with pre-trained word embeddings for sentiment analysis on an IMDB dataset[3]. We also train a CNN model with pre-trained word embeddings for document classification using the 20 Newsgroups data set[4]. We report F-1 scores of models in both tasks' test set in Table. 7. We see that DD-GloVe performs marginally better than the baseline GloVe in these two tasks. These results demonstrate that DD-GloVe preserves semantic meanings in the debiased word embeddings.

Female	ex-wife, girl, jane, woman, wife, witch, women, she, pilipinas, heroine, maids, hens, dona, wives, fiancee, goddess, bint, sheila, hostess, hen, nun, sisters, girls, waitress, doe, sister, actress, businesswoman, chairwoman, goddesses
Male	he, son, brother, brothers, boys, sons, boy, businessman, yang, gentleman, wizard, headmaster, statesman, nobleman, policeman, salesman, bahadur, stallion, fiance, manny, englishman, beau, widower, chicano, workmen, councilman, stallions, schoolmaster, scotsman, horseman

Table 8: Full lists of words chosen by our dictionary-guided algorithm (Algorithm 1) to approximate the gender direction.

A.3 Full List of Seed Words

We report the full list of chosen seed words by running Algorithm 1 for approximating gender direction in Table. 8.

A.4 Ablation Study

To understand the role of each dictionary-guided loss in DD-GloVe, we conduct an ablation study that only uses one of the proposed losses, and an experiments that avoid using one of the losses but optimizes the other two in Table. 10. We have made the following observations.

J_{ortho} contributes to both semantic meaning preservation and general bias reduction Both word analogy accuracy and WEAT results improve as the weight of J_{ortho} increases from $1e-5$ to 0.01, as shown in Table. 10. However, if the weight of J_{ortho} gets large, it debiases word embeddings at the expense of semantic meaning representations. We should keep its weight low for both semantic meaning preservation and bias mitigation. We see that J_{ortho} is not the most effective component for bias mitigation because the debiasing effect does not suffer a significant drop when J_{ortho} is not used to train DD-GloVe, shown in Table. 10. However, J_{ortho} remains an important component in the loss function because of its ability to reduce general types of biases. In Table. 9, we report the WEAT results of DD-GloVe without using J_{ortho} and com-

[3] https://www.kaggle.com/lakshmi25npathi/sentiment-analysis-of-imdb-movie-reviews/data

[4] http://qwone.com/∼jason/20Newsgroups/

Setting	Gender-1		Gender-2		Race		Age		Nature	
	$d \downarrow$	$p \uparrow$	$d \downarrow$	$p \uparrow$	$d \downarrow$	$p \uparrow$	$d \downarrow$	$p \uparrow$	$d \downarrow$	$p \uparrow$
GloVe	1.74	0.00	1.07	0.013	1.18	0.0029	1.03	0.0090	1.15	0.0029
All losses	1.25	0.0029	0.083	0.44	1.01	0.011	0.94	0.017	1.01	0.0088
w/o J_{ortho}	1.22	0.0037	0.025	0.48	1.17	0.0035	1.09	0.0061	1.06	0.0064

Table 9: WEAT results when orthogonal loss is not used, compared with GloVe and DD-GloVe trained with all proposed loss terms. Without orthogonal loss, DD-GloVe can still mitigate gender bias but non-gender WEAT tests show similar results as the original GloVe. These results indicate that J_{ortho} can reduce general types of biases.

Setting	Weight	G-Sem (%)	$d \downarrow$	$p \uparrow$
References				
GloVe		79.26	1.74	0.00
DHD		79.77	1.38	0.0014
DD-GloVe$_{gender}$		80.27	1.25	0.0029
Only using one of the losses				
J_{ortho} only	0.001	80.56	1.75	0.0
	0.005	80.93	1.73	0.0
	0.01	81.50	1.73	7.8e-5
	0.1	76.89	1.71	0.0
	0.2	71.61	1.68	7.8e-5
J_{proj} only	0.2	79.96	1.40	8.6e-4
	0.25	79.69	1.26	0.0023
	0.3	79.10	1.03	0.017
	0.35	78.93	1.13	0.010
	0.4	79.39	0.99	0.021
J_{def} only	1e-5	80.09	1.77	7.8e-5
	1e-4	80.22	1.76	0.0
	0.001	80.54	1.74	0.0
	0.005	81.29	1.78	0.0
Without using one of the losses				
w/o J_{ortho}		79.60	1.22	0.0037
w/o J_{proj}		80.29	1.76	0.0
w/o J_{def}		79.78	1.23	0.0044
w/o J_{G-bias}		80.35	1.39	7.8e-4

Table 10: Ablation study to understand the effects of each loss in DD-GloVe. The table shows the performance of DD-GloVe in Google-sem word analogy (G-Sem) and WEAT Gender-1 test (effect size d and p-value). In the experiment without J_{G-bias}, we replace J_{G-bias} with the original GloVe loss function.

pare them with the baseline GloVe and DD-GloVe with all losses used. It is evident that the absence of J_{ortho} causes race, age, and nature WEAT test to have worse results.

J_{proj} is essential for effective bias reduction Table. 10 shows that WEAT results improve significantly as we increase the weight of J_{proj}. When the projection loss is not used, there is a significant degradation in debiasing performance in Table. 10.

J_{def} enhances semantic meaning representation In Table. 10, we see that the word analogy task enjoys higher accuracy when the weight of J_{def} increases. This benefits from the additional semantic meaning injected from dictionary definitions. In terms of debiasing, J_{def} does not help much as illustrated in Table. 10. This finding explains why simply doing retrofitting with dictionary definitions does not mitigate biases.

J_{G-bias} further mitigates bias We find that when J_{G-bias} is replaced with the original GloVe loss function, there remains evidence of debiasing but it is less effective, as shown in Table. 10. This suggests that adjusting co-occurrence weights according to the word bias and context word bias can learn more neutral word embeddings.