Bias Measurements in Word Embeddings with First-Order Co-occurrence

Anonymous Author(s)

Abstract

Word embedding models capture semantic relations, but also reflect cultural stereotypes and ethical biases. A precise method to measure such embedded biases is a crucial step towards addressing this issue. A common measure of the bias of a word's embedding towards a concept is its similarity to a vector from representative words for the concept, measuring whether the word and concept words share other words in their contexts. We argue that this second-order relationship introduces circularity and unrelated concepts into the measure, which causes an inaccurate measurement of the bias. We propose instead to measure bias using the direct co-occurrence between the word and the representative concept words, a first-order measure, by reconstructing the co-occurrence estimates inherent in the word embedding models into an explicitly interpretable vector. To evaluate our novel bias measurement method, we calculate the correlation of the estimated gender bias values to the actual gender bias statistics of the U.S. job market, provided by two recent collections. The results show a consistently higher correlation when using our method, with a variety of word embedding models, as well as a more severe degree of female bias in word embedding models. Finally, benefiting from the interpretability of the explicit vectors, we investigate the reasons behind the limitations of a recent debiasing algorithm.

1 Introduction

Word embedding models, widely used in language understanding tasks, provide low-dimensional semantic representations of words by exploiting their co-occurrence patterns. As shown in previous studies, such representations also capture cultural stereotypes, ethical biases, and historical prejudices towards certain social groups, from the provided text corpus (Molly and Lupyan, 2019; Garg et al., 2018; Caliskan et al., 2017; Bolukbasi et al., 2016). The existence of such biases indeed raise concerns about their effects on the decision making processes in down-stream tasks. On the other hand, capturing the patterns of language use in word embeddings provides an effective quantification tool for studying social dynamics. In both cases, accurately measuring the degrees of bias is crucial, which is the aim of this work.

The term bias in this work refers to demographic and/or corpus disparities that, based on ethical values, could be objectionable to use in decision making. To measure bias, we first need to measure the presence of a *concept* (e.g. female) in a word (e.g. 'nurse'), referred to as a *factor* (e.g. the female factor in 'nurse'). We consider a word to be biased towards a concept when a significant imbalance is observed between the factor of that concept and the factor of its counterpart concept (e.g. male).

To measure bias, previous studies calculate such factors using the similarity between the vector of a word and the representative vectors of the concepts (Gonen and Goldberg, 2019; Molly and Lupyan, 2019; Garg et al., 2018; Zhao et al., 2018b; Caliskan et al., 2017; Bolukbasi et al., 2016). For instance, the gender factors of the word 'nurse' is measured by using the cosine similarities of its embedding vector to the embedding vectors of 'she' and 'he', as the representative vectors of the concepts female and male. An embedding vector reflects the distribution of context words which its word co-occurs with, so this measure reflects context words which are shared between the target word and the concept word. For this reason, we refer to the approach using such factors as *second-order bias measurement*.

Figure 1 depicts the estimations the female factors of the words 'nurse' and 'physician' with a toy example. The example assumes that the dimensions of the vectors are fully interpretable, such that each represents a specific context word. We refer

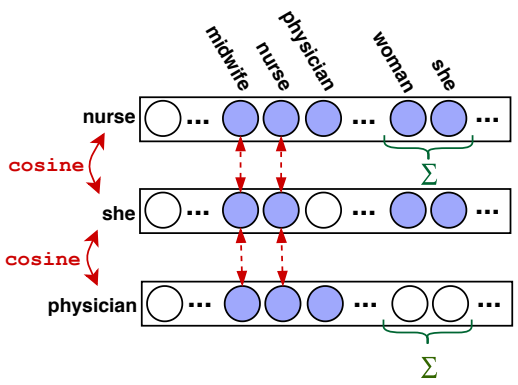


Figure 1: Word vectors with interpretable dimensions. Estimating bias with similarity between the vectors (second-order relations) also counts some unrelated concepts (red dashed lines). Instead, we propose to only use the co-occurrence/first-order relations (green summations).

to such high-dimensional word vectors as *explicit representations*, where each value of the vector represents the co-occurrence relation between the word and the corresponding context word, which we refer to as a *first-order relation*, discussed further below.

1.1 Problem Definition

The representative vectors in the second-order approach are intended to provide proxies for the concepts of interest. However, since the corresponding words of these vectors co-occur with various words, the representative vectors are composed of a mixture of different concepts, many unrelated to the expected ones. For instance, if in a corpus the word 'nurse' frequently co-occurs with female-related words such as 'she', its vector ($\boldsymbol{v}_{nurse}$) in an embedding model contains a high degree of the female concept. However, it also means that the concept 'nurse' is encoded in $\boldsymbol{v}_{she}$ — the representative vector of the female concept. We refer to this characteristic of word representation models as *circularity*. Considering this trait, given that $\boldsymbol{v}_{nurse}$ naturally contains the concept 'nurse', calculating the similarity between $\boldsymbol{v}_{she}$ and $\boldsymbol{v}_{nurse}$ (as the female factor of 'nurse') is mistakenly influenced by unrelated matched concepts, among which the 'nurse' concept. This issue is indicated in the toy example (in the explicit representation space) with the red dashed lines between the vectors of 'she' and 'nurse'.

Circularity is in fact a particular form of the *transitivity* characteristic in word representations. To explain it, we follow the toy example by assuming that in the corpus, 'physician' does not co-occur with female-related context words, while both 'physician' and 'she' co-occur with nurse-related context words. In this case, the female factor of 'physician', calculated with the second-order approach, is mistakenly affected by nurse-related context words, despite no direct co-occurrence relations between 'physician' and female-related context words. This issue in shown with the red dashed lines between the vectors of 'she' and 'physician'. Such matching of non-relevant concepts causes an imprecise measurement of bias.

1.2 Contributions

To address this issue, we propose measuring the factor of a concept in a word, by summing the normalized values of the first-order relations between the word and the concept's representative words. We refer to this method as *first-order bias measurement*, indicated with the green summation signs in Figure 1. This approach resolves the issue of the second-order bias measurement method caused by matching non-relevant concepts, by only taking into account the related concepts, resulting in a more accurate measurement of bias in language.

Calculating the first-order bias measurement requires obtaining the estimations of the co-occurrence relations from word representation models. In particular, we study the application of our approach on three families of representation models: the ones based on Pointwise Mutual Information (PMI), word2vec Skip-Gram (Mikolov et al., 2013), and GloVe (Pennington et al., 2014).

While the PMI-based representations explicitly define the co-occurrence relations based on the PMI measure, it is not immediately obvious how to extract the first-order relations from the word2vec Skip-Gram and GloVe model. In these two models, we reconstruct the co-occurrence relations using the word and context embedding vectors, resulting in high-dimensional explicit variations of the models, where the vectors are fully interpretable. Our approach is in the opposite direction of the methods such as Latent Semantic Indexing (Deerwester et al., 1990), or GloVe, where they start from a high-dimensional matrix and result in low-dimensional embeddings.

We use the bias measurement methods to study the gender bias of a set of occupations in the word representation models trained on a Wikipedia corpus. We evaluate the methods based on the cor-

relations of the gender bias of occupation words, measured using the discussed methods, to the actual statistics of gender bias in the U.S. job market, provided in two collections by Zhao et al. (2018a), and Garg et al. (2018). In all three word representation models and both collections, we observe consistently higher correlations using our proposed method in comparison to the second-order bias measurement. Benefiting from the interpretability of the explicit vectors, we diagnose the results of the second-order bias measurement method, showing several cases of the effects of circularity/transitivity in representation models. Overall, our results suggest the existence of a more severe degree of bias towards female in word embeddings, especially for some occupations, previously neglected by the second-order bias measurement.

Finally, in-line with Gonen and Goldberg (2019), we analyze the limitations of the debiasing method introduced by Bolukbasi et al. (2016), by diagnosing it using the explicit representations. Our observations show that despite a decrease in the effect of some gender-related context words, several other context words are still strong indicators of the underlying bias in the word embedding space.

The contribution of this work is three-fold:

- proposing a novel bias measurement method based on the first-order relations, including explicit variants of three word representation models
- extensive studies on the degree of gender bias in occupations
- investigating the reasons behind the limitations of a debiasing method, using the explicit vectors

In the remainder of the paper, Section 2 discusses related work, followed by the relevant previous methods in Section 3. Our bias measurement approach is introduced in Section 4. Section 5 describes the gender bias experiments, whose results are presented and discussed in Section 6. Finally, Section 7 presents our analysis on the debiasing method.

2 Related Work

Several pieces of work exploit word embeddings to study societal aspects. Garg et al. (2018) investigate the changes in gender- and race-related stereotypes over decades using historical text data. Caliskan et al. (2017) and more recently Molly and Lupyan (2019) study the patterns of language use, indicating accurate imprints of historical biases. Bolukbasi et al. (2016) show the reflection of gender stereotypes in word analogies derived from word embeddings. Our work directly contributes to these studies by proposing an more accurate approach for measuring bias.

A method for debiasing word embeddings is proposed by Bolukbasi et al. (2016), where a post-processing method subtracts an approximated gender direction vector from gender-neutral word vectors. The gender direction is created using the first principle component of several directional vectors, defined based on a set of gender definitional pairs. Zhao et al. (2018b) follow this direction by enforcing a debiasing criteria as regularization terms, added to the objective function of the GloVe model. Recently, Gonen and Goldberg (2019) point out the limitations of these debiasing methods. They show that learning a classifier or a clustering model on the debiased word vectors can easily retrieve the gender of the words assigned before debiasing. Our work further investigates this direction by analyzing the features of the classifier trained on explicit representations.

The existence of gender bias in statistical models is also studied in various downstream tasks, such as sentiment analysis (Kiritchenko and Mohammad, 2018), visual semantic role labeling (Zhao et al., 2017), and coreference resolution (Zhao et al., 2018a). In the context of debiasing, Elazar and Goldberg (2018) highlight the challenges in removing sensitive attributes from text-based classification tasks, where, in their approach, adversarial networks are used to enforce debiasing.

Regarding the interpretable word vectors, previous studies propose methods to increase sparsity (Faruqui et al., 2015; Sun et al., 2016), as in the sparse vectors, it becomes more clear which dimension might be referring to which concept. In contrast, each dimension of the explicit vectors represents a distinct context word. In this regard, the explicit representations become conceptually related to the Explicit Semantic Analysis method (Gabrilovich and Markovitch, 2009).

3 Background

3.1 Word Representation Models

word2vec Skip-Gram (SG): The model consists of two parameter matrices: word ($\boldsymbol{V}$) and context ($\boldsymbol{U}$) matrices, both of size $|\mathbb{W}| \times d$, where $\mathbb{W}$ is the set of words in the collection and d is

the embedding dimensionality. The matrices are joined with a linear hidden layer. Given the word c, appearing in a context of word w, the model calculates $p(y = 1|w, c)$, the probability that the co-occurrence of w and c come from a *genuine* distribution, defined as follows:

$$p(y = 1|w, c) = \sigma(\boldsymbol{v}_w \boldsymbol{u}_c^\top) \quad (1)$$

where $\boldsymbol{v}_w$ is the vector representation of w, $\boldsymbol{u}_c$ context vector of c, and σ denotes the sigmoid function. The SG model is optimized by maximizing the difference between $p(y = 1|w, c)$ and $p(y = 1|w, \check{c})$ for k negative samples $\check{c}$, randomly drawn from a *noisy* distribution $\mathcal{N}$.

GloVe: The model first defines an explicit matrix (size $|\mathbb{W}| \times |\mathbb{W}|$), where the corresponding co-occurrence value of each word and context word is set to $p(w|c) = p(w, c)/p(c)$. The probabilities are calculated based on the number of co-occurrences, such that $p(w|c) = \#\langle w, c\rangle / \#\langle \cdot, c\rangle$. We refer to these explicit representations as *initGloVe*.

The matrix of the initGloVe representations is then factorized to two matrices of size $|\mathbb{W}| \times d$. Using the same notation as SG, the factorization is done such that the dot products of the vectors of the matrices $\boldsymbol{V}$ and $\boldsymbol{U}$ estimate the log of the co-occurrence values, as defined in the following:

$$\boldsymbol{v}_w \boldsymbol{u}_c^\top \approx \log p(w|c) \quad (2)$$

The matrix factorization is done based on a weighted least squares regression model, where $\log p(w|c)$ is replaced with $\log \#\langle w, c\rangle$ plus two bias terms.

PMI-based Representations: The PMI representation is also defined in the explicit space using the count-based probabilities similar to those used in initGloVe. The co-occurrence relation between a word and a context word in the PMI representation is calculated by $\log\left(p(w, c)/p(w)p(c)\right)$. Positive PMI (PPMI) is a commonly-used variation, where negative values are replaced with zero.

Levy and Goldberg (2014) show an interesting relation between PMI and SG representations, i.e. when the dimension of the embedding vectors is very high (as in explicit representations), the optimal solution of the SG objective function is equal to PMI shifted by $\log k$. Based on this idea, they propose the Shifted Positive PMI (SPPMI) representation by subtracting $\log k$ from PMI vector representations and setting the negative values to zero.

3.2 Second-Order Bias Measurement

As mentioned in Section 1, measuring the bias of a word towards a concept requires an estimation of the factor of the concept in the vector representation of the word. To do this, first the concept z is defined with the set of *definitional words* $\mathbb{W}_z \in \mathbb{W}$. The representative vector of z, denoted as $\boldsymbol{v}_z$, is then defined as the average of the embeddings of the definitional words, shown as follows:

$$\boldsymbol{v}_z = \frac{\sum_{w \in \mathbb{W}_z} \boldsymbol{v}_w}{|\mathbb{W}_z|} \quad (3)$$

Given $\boldsymbol{v}_z$, the second-order bias measurement method defines the factor of the concept z in the word w, denoted with Λ_z^{2ND}, as follows:

$$\Lambda_z^{\text{2ND}}(w) = sim(\boldsymbol{v}_z, \boldsymbol{v}_w) \quad (4)$$

where sim refers to the similarity function, commonly measured by cosine. In addition, Garg et al. (2018) propose using *negative norm difference (nnd)*, defined as: $\text{nnd}(\boldsymbol{v}_z, \boldsymbol{v}_w) = -\left\|\boldsymbol{v}_z / \|\boldsymbol{v}_z\| - \boldsymbol{v}_w / \|\boldsymbol{v}_w\|\right\|$.

Using the measured factors, the bias is defined as $\Lambda_z^{\text{2ND}}(w) - \Lambda_{z'}^{\text{2ND}}(w)$, where z' is the counterpart concept of z.

4 Novel Bias Measurement

We first explain our approach to creating the explicit variations of the SG and GloVe vectors. We refer to these representations in the explicit space as *explicit Skip-Gram (eSG)* and *explicit GloVe (eGloVe)*. Using the explicit vectors, we then describe our first-order bias measurement method.

4.1 Explicit Representations

Explicit Skip-Gram (eSG) Revisiting the $p(y = 1|w, c)$ probability in the SG model, it measures the probability that the co-occurrence of two words w and c comes from the genuine co-occurrence distribution, derived from the training corpus. The model uses this probability to learn the embedding vectors, by separating these genuine co-occurrence relations from the sampled negative ones. We therefore use this estimation of the co-occurrence relations to define the vectors of the explicit SG representation, shown as follows:

$$e_{w:c} = p(y = 1|w, c) = \sigma(\boldsymbol{v}_w \boldsymbol{u}_c^\top),$$
$$\boldsymbol{e}_w = \sigma(\boldsymbol{v}_w \boldsymbol{U}^\top) \quad (5)$$

where $\boldsymbol{e}_w$ denotes the explicit vector representation of w with $|\mathbb{W}|$ dimensions, and $e_{w:c}$ is the value of the corresponding dimension to the context word c.

We should note that the eSG representation is considerably different from SPPMI. The SPPMI representation assumes very high embedding dimensions during the model training, while eSG draws the co-occurrence relations after the model is trained on low-dimensional embeddings, and is therefore smoother.

Explicit GloVe (eGloVe) Similar to eSG, the eGloVe representation estimates the co-occurrence relations using the word and context vectors, after training the GloVe model. Considering Eq. 2, we define the co-occurrence relations of the eGloVe representation as the dot product of the word and context vectors, shown as follows:

$$e_{w:c} = \boldsymbol{v}_w \boldsymbol{u}_c^\top, \quad \boldsymbol{e}_w = \boldsymbol{v}_w \boldsymbol{U}^\top \tag{6}$$

In fact, the eGloVe vector uses the word and context embedding vectors to estimate the log of the original explicit word vector, defined based on the count-based co-occurrence probabilities. In other words, the eGloVe vector can be seen as a smoothed variation of the original explicit word vector.

4.2 First-Order Bias Measurement

The difference between the first-order bias measurement method and the second-order approach is in the estimation of the factors of the concepts. The first-order bias measurement defines the factor related to the concept z as the sum of the co-occurrence values of the word w with the definitional words $\mathbb{W}_z$, normalized by the l_2 norm of the co-occurrence values of the word with all context words. Since the estimations of the co-occurrence relations are provided by the explicit vectors, this factor, denoted as $\Lambda^{1\text{ST}}$, is formulated as follows:

$$\Lambda_z^{1\text{ST}}(w) = \frac{\sum_{c \in \mathbb{W}_z} e_{w:c}}{\|\boldsymbol{e}_w\|} \tag{7}$$

As shown, $\Lambda_z^{1\text{ST}}$ only considers the context words related to the z concept. This potentially avoids the influence of other non-related concepts as in the second-order bias measurement.

Similar to the second-order bias measurement, the bias toward z in the first-order bias measurement method is calculated based on the differences between the factors: $\Lambda_z^{1\text{ST}}(w) - \Lambda_{z'}^{1\text{ST}}(w)$, where z' is the counterpart concept of z.

5 Gender Bias Experiment Design

We measure the gender bias in 504 occupations, among which 20 are female-specific (e.g. 'congresswoman'), 34 male-specific (e.g. 'congressman'), and the rest are gender neutral (e.g. 'nurse', 'dancer'). Regarding the definitional words of the genders' concepts, we create four lists with 2, 32, 64, and 156 words, referring to them as `Tiny`, `Small`, `Medium`, and `Large`, respectively. Each list contains an equal number of female- and male-definitional words, e.g. 'she', 'her', 'woman' for the female and 'he', 'his', 'man' for the male concept. The above-mentioned lists are compiled from the provided resources in previous studies (Bolukbasi et al., 2016; Garg et al., 2018).

We use two collections for evaluation of the bias measurement approaches, both containing the statistics of the gender bias of a set of occupations, obtained from the U.S. job market. The bias for each occupation in the collections is the percent of people in the occupation who are reported as female (e.g. 90% of nurses are women). We assume that these real gender distributions are highly correlated with the true degree of gender bias in the texts which refer to them, and therefore correlations between these distributions and gender bias measures applied to the words which refer to the occupations is a good indication of the accuracy of the gender bias measures. We compare the correlation between the $\Lambda^{1\text{ST}}$ and $\Lambda^{2\text{ND}}$ gender bias measures and the statistics of the gender bias given by the collections.

The first collection uses the data provided by Zhao et al. (2018a). The collection contains the statistics of 40 occupations, gathered from the U.S. Department of Labor. We refer to the collection as Labor Data. The second is provided by Garg et al. (2018) using the U.S. census data. From the provided data, we use the gender bias statistics of the year 2015 – the most recent year in the collection, resulting to a list of 96 occupations. We refer to this collection as Census Data. The correlation is calculated using the Spearman ρ and Pearson's r between the measured bias values and the statistics in the collections.

The word representation models are created on the English Wikipedia corpus of August 2017. We project all characters to lower case, and remove numbers and punctuation marks. For all models, we use the window size of 5, and filter the words with frequencies lower than 200, resulting in 197,549 unique words. The number of dimensions of the embeddings are set to 300. The rest of the parameters are set using the default parameter

Table 1: Correlation results of the gender bias values, calculated with word representations using the `Medium` set, to the statistics of the portion of women in occupations, provided by Zhao et al. (2018a) (Labor Data) and Garg et al. (2018) (Census Data). The best results in each section is shown in bold and the best overall results are indicated with underlines.

Representation	Method		Labor Data		Census Data	
			Spearman ρ	Pearson's r	Spearman ρ	Pearson's r
PMI-SVD	Λ^{2ND}	cosine	0.39	0.47	0.45	0.52
		nnd	0.40	0.48	0.46	0.52
PPMI-SVD	Λ^{2ND}	cosine	0.39	0.47	0.45	0.52
		nnd	0.40	0.48	0.46	0.52
SPPMI-SVD	Λ^{2ND}	cosine	0.29	0.34	0.40	0.43
		nnd	0.27	0.36	0.40	0.42
PMI	Λ^{1ST}	-	0.55	0.53	0.55	0.60
PPMI	Λ^{1ST}	-	**0.60**	**0.57**	**0.62**	**0.63**
SPPMI	Λ^{1ST}	-	0.45	0.47	0.48	0.46
GloVe	Λ^{2ND}	cosine	0.56	0.59	0.36	0.49
		nnd	0.54	0.58	0.35	0.47
initGloVe	Λ^{1ST}	-	0.43	0.44	0.49	0.55
eGloVe	Λ^{1ST}	-	**0.61**	**0.60**	**0.52**	**0.59**
SG	Λ^{2ND}	cosine	0.53	0.55	0.57	0.62
		nnd	0.54	0.54	0.57	0.61
eSG	Λ^{1ST}	-	**0.64**	**0.59**	**0.67**	**0.70**

setting of the word2vec Skip-Gram model in the Gensim library (Rehurek and Sojka, 2010), and the GloVe model in the provided tool by its authors. As Suggested by Levy and Goldberg (2014), we apply subsampling and context distribution smoothing (`cds`) on all PMI-based models with the same parameter values as the SG model.

We also create the low-dimensional representations of the PMI-based models using Singular Value Decomposition (SVD). We refer to these models as PMI-SVD, PPMI-SVD, and SPPMI-SVD.[1]

6 Measuring Gender Bias

The gender bias of the word w is defined as $\Lambda_f(w) - \Lambda_m(w)$, where Λ_f and Λ_m are the female and male factors, respectively, calculated using the Λ^{2ND} or Λ^{1ST} method. A positive bias value indicates the inclination towards female, and a negative value towards male.

In the following, we first present the evaluation results of the bias measurement methods, followed by a discussion on the issues of the second-order approach. Finally, we analyze the degree of gender bias of occupations.

6.1 Correlation to Gender Bias Statistics

We calculate the gender bias of the occupations, using Λ^{2ND} on the low-dimensional embeddings (PMI-based models with SVD, GloVe, and SG) and Λ^{1ST} on high-dimensional explicit representations (PMI-based models, initGloVe, eGloVe, and eSG). For the Λ^{2ND} method, we create the representative vectors of the female and male concepts, referred to as $\boldsymbol{v}_f$ and $\boldsymbol{v}_m$, and use both cosine and nnd similarity functions. We use the gender definitional words of the `Medium` list in both methods. We later in this section compare the effects of various gender definitional lists.

Table 1 shows the correlation results between the calculated gender bias, and the gender bias statistics, provided by the Labor Data and Census Data collections. Each section of the table is assigned to a family of the representation models, namely PMI, GloVe, and SG. The best results of each section are shown in bold, and the best overall results are

[1] We will publish our code, together with all resources.

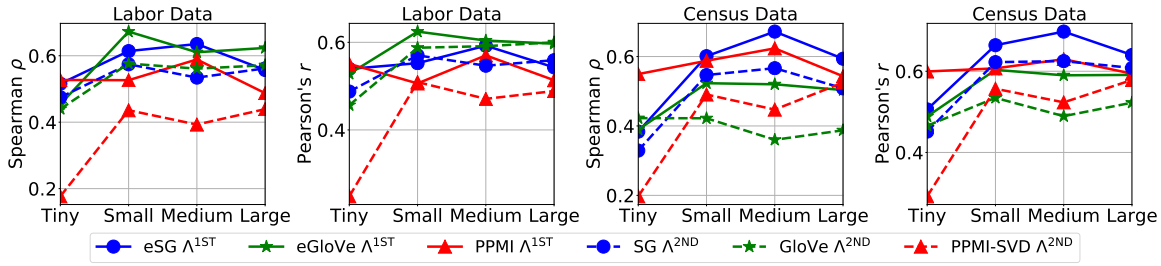


Figure 2: Evaluation results of the bias measurement methods to the gender bias statistics provided by the collections, using different gender definitional word lists

indicated with underlines.

In all three families of the models, the first-order bias measurement method (using Λ^{1ST}) shows consistently higher correlations than the results of the second-order measurement (using Λ^{2ND}), on both collections and correlation metrics. These results highlight the importance of using our proposed bias measurement method, since its results more accurately reflect the actual gender bias indicators.

Comparing the results, overall the eSG model shows the best performance across the representations. Between the PMI-based models, PPMI achieves the highest correlation results. Comparing eGloVe with initGloVe, we observe a large difference between the results, indicating the effect of drawing the explicit eGloVe vectors from the GloVe embeddings. Comparing the results across the low-dimensional embeddings, overall the SG model again shows higher correlations, specially on the Census Data collection. The cosine and nnd similarity functions perform similarly.

Figure 2 shows the sensitivity of the bias measurement methods to the choice of the gender definitional words. Apart from the `Tiny` set, the word sets perform similarly, while the `Medium` set has slightly higher performance than the others, especially with the best performing models. Overall, the SG and eSG representations show more stable and better performance across the collections.

In the following analysis, we therefore only focus on the eSG and SG models. For the Λ^{2ND} method with SG, we use the cosine function, due to its similar performance to nnd.

6.2 Diagnosis of Second-Order Measurement

To investigate the cause of the weaker performance of the second-order bias measurement, we recalculate the gender bias of the occupations with Λ^{2ND}, this time using the explicit vectors. The explicit representations enable the diagnosis of the results, particularly by looking at the context words with the highest contributions.

Table 2: Context words with the highest effects on bias towards female (F) and male (M) in the second-order measurement. Context words with unrelated concepts to gender (bold) mislead the bias measurement

Occupation: *nurse* F: *matron, midwife,* ***nurse****, Filipina, maternity* M: ***surgeon****,* ***enlisted****,* ***clerk****,* ***trained****,* ***sergeant***
Occupation: *physician* F: *midwife,* ***nurse****,* ***nursing****,* ***nurses****, maternity* M: *grandfather, grandson, nephew,* ***apprenticed****,* ***surgeon***
Occupation: *surgeon* F: ***nurse****, midwife, matron,* ***nursing****, maternity* M: ***apprenticed****, grandfather, grandson,* ***enlisted****, nephew*
Occupation: *housekeeper* F: *matron, maid, midwife, housewife, matriarch* M: *uncle, grandfather, nephew,* ***clerk****,* *gentleman*
Occupation: *CEO* F: *businesswoman, chairwoman,* ***chairperson****,* *businesswomen, michelle* M: *businessman,* ***billionaire****,* ***banker****,* *grandson,* ***entrepreneur***

We create the explicit variations of $\boldsymbol{v}_f$ and $\boldsymbol{v}_m$ of the SG model using Eq. 5, referred to as $\boldsymbol{e}_f$ and $\boldsymbol{e}_m$. Given the occupation o with the embedding vector $\boldsymbol{v}_o$, we also create its explicit variation, $\boldsymbol{e}_o$. The second-order bias measurements with cosine estimates gender bias with $\text{cosine}(\boldsymbol{v}_f, \boldsymbol{v}_o) - \text{cosine}(\boldsymbol{v}_m, \boldsymbol{v}_o)$. We calculate the gender bias with the explicit vectors, and provide its element-wise results by removing the summa-

(a) Second-order measurement (Λ^{2ND})

(b) First-order measurement (Λ^{1ST})

Figure 3: The female and male factors of the occupations, indicating their inclinations towards the genders. The ones in the gray area are considered as unbiased. The occupations, biased towards female and male are shown in purple and orange, respectively, where among them the gender-neutral ones have lighter colors.

tion of the cosine function, formulated as follows:

$$\boldsymbol{e}_{BIAS} = \frac{\boldsymbol{e}_f}{\|\boldsymbol{e}_f\|} \odot \frac{\boldsymbol{e}_o}{\|\boldsymbol{e}_o\|} - \frac{\boldsymbol{e}_m}{\|\boldsymbol{e}_m\|} \odot \frac{\boldsymbol{e}_o}{\|\boldsymbol{e}_o\|} \quad (8)$$

where $\odot$ denotes the element-wise product, and $\boldsymbol{e}_{BIAS}$ represents the gender bias results for the context word, corresponding to each dimension.

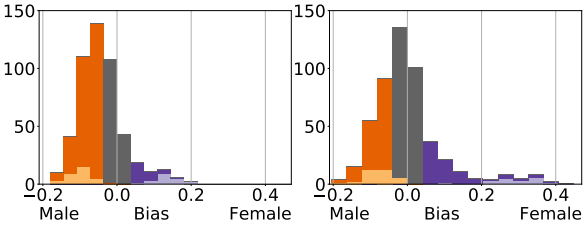
Table 2 shows the dimensions in $\boldsymbol{e}_{BIAS}$ with the largest influence on the results of the gender bias measurement for a representative sample of occupations, namely the top 5 context words with highest positive (bias towards female) and negative (bias towards male) values. The results show several cases of the effect of context words which are unrelated to gender (shown in bold). The second-order bias measurement perceives 'CEO' as male, since it highly co-occurs with 'billionaire', and assigns some degrees of the male factor to 'nurse' because it co-occurs with 'surgeon'! Such cases illustrate the circularity of second-order measurements, where the gender bias of a word is defined by its co-occurance with other gender-biased words. Even among gender-specific context words, some of them, like 'midwife' and 'businessman', are not purely representatives of gender but also contain other concepts such as an occupation. Such cases cause an inaccurate estimation of the gender factors and therefore gender bias, since the measurement of the factors are influenced by a mixture of related and unrelated concepts.

6.3 Visualization of Gender Bias Results

The female and male factors of the occupations, computed using the Λ^{2ND} method with SG and the Λ^{1ST} method with eSG are shown in Figures 3a and 3b, respectively. To make the results visually comparable, we apply Min-Max normalization to the factors of each approach, where the min/max values are calculated over the gender factors of all words.

In both bias measurement methods, we are interested in distinguishing between the words with significantly higher bias values and any random word with low bias values. To do this, we define a threshold for each plot, below which the words are considered as unbiased. To find such thresholds, since the number of biased words to a concept are limited, we assume that there is a high probability that any randomly sampled word is unbiased. We therefore define the threshold for each bias measurement method as the mean of the absolute bias values. These thresholds for Λ^{2ND} and Λ^{1ST} are 0.036 and 0.040, respectively. In each plot, the area where the distance from the line of zero bias (diagonal) is less than the corresponding threshold value, is referred to as the unbiased area, and shown in gray (■). The occupations located in the unbiased areas are considered as unbiased.

An occupation is considered to be biased to either female or male, when it is inclined towards the female/male factor, namely when it is located below/above the unbiased area, shown in dark



(a) Λ^{2ND} (b) Λ^{1ST}

Figure 4: Histograms of the gender bias of the occupations, measured using Λ^{2ND} and Λ^{1ST} methods. The positive values indicate the bias towards female, and negative values towards male

purple (■) and dark orange (■), repectively. The gender-specific occupations (e.g. 'actress', 'handyman') are shown in light colours, namely light purple (■) for female, and light orange (■) for male. As expected, all gender-specific occupation words are fall on the correct side of the diagonal for both measures. But for the gender-neutral occupation words, falling outside the unbiased area reflects the bias in language use.

Both figures show the existence of significant gender bias in several occupations. However, Figure 3a and 3b provide considerably different perspectives on the extents of the female and male factors in the occupations. In particular, the Λ^{1ST} method shows relatively larger degrees of bias towards female, specially for some gender-neutral occupations such as 'nurse', 'manicurist', and 'housekeeper'.

To have a better view on the distribution of the bias values, Figure 4 shows the histogram of the occupations over the range of the bias values, measured with the Λ^{2ND} and Λ^{1ST} methods. Similar to Figure 3, the gray color shows the number of unbiased occupations, and the purple and orange colors indicate the number of biased ones towards female and male in each bin, among which the the gender-neutral ones are shown with light colours.

As shown, in both measurement methods, a larger number of occupations are biased to male. However, the first-order bias measurement captures a larger degree of bias towards female, revealing a more severe degree of female bias, previously neglected by the second-order approach.

7 Analysis of Gender Debiasing

Gonen and Goldberg (2019) show that the debiasing algorithm, introduced by Bolukbasi et al. (2016)

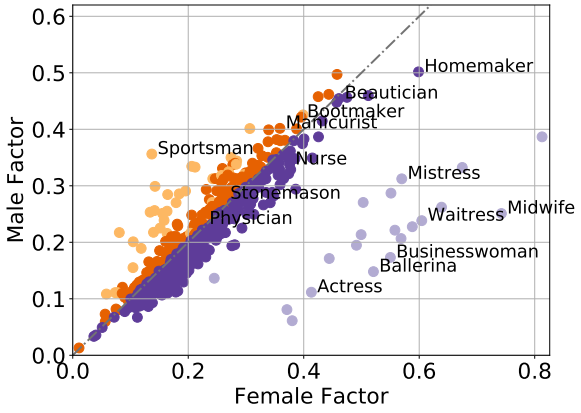


Figure 5: Female and male factors of the occupations, measured with Λ^{1ST}, after applying the debiasing algorithm proposed by Bolukbasi et al. (2016)

only partially reduces the existing gender bias of the word embeddings, as the perceived gender of the biased words can be retrieved after debiasing. In this section, we analyze in detail what the debiasing algorithm removes, and which remaining features of the vectors are still indicative of gender.

We apply the debiasing method on the SG model, and measure the female/male factors of the occupations in the resulting embeddings using the Λ^{1ST} method, shown in Figure 5.[2] The plot does not contain an unbiased area, since it is expected that the debiasing method removes the gender bias of the gender-neutral words. Comparing the results with Figure 3b, we observe a decrease in the gender bias of the gender-neutral occupations, such that the differences of the gender factors are reduced (points become closer to the identity line $y = x$). However, despite the decrease, the measured bias is not yet zero, namely the points are not located on the identity line. We suggest that it is because the proposed algorithm only uses one gender direction vector for all the words, despite the fact that different words contain different degrees of different gender concepts.

To identify the significant remaining gender-related concepts, we follow the experiments in Gonen and Goldberg (2019) using the explicit vectors. We train two classifiers, where the features are the eSG vectors of the gender-neutral occupations be-

[2]Since the debiasing algorithm should be applied on the normalized vectors but the Λ^{1ST} method is defined on the non-normalized ones, we first normalize the vectors and applied the debiasing method. We then multiply the calculated l_2-norms of the original vectors to the debiased ones, to return the vectors to their non-normalized forms.

Table 3: The context words with the highest contributions to the predictions of the perceived genders of the gender-neutral occupations. The classifiers are trained on the eSG vectors of the occupations before and after debiasing. The asterisks on words indicate named entities.

Female
Original: *matron, headmistress, feminist, midwife, housekeeper, suffragist*
Debiased: *matron, housekeeper, berkus*, vroman*, housekeeping, westrum**
Male
Original: *businessman, apprenticed, journeyman, engineer, entrepreneur, serjeant*
Debiased: *semon*, framer, businessperson, businessman, journeyman, fgs**

fore and after debiasing, and the labels are their perceived genders of the occupations, indicated by the Λ^{1ST} method using the vectors before debiasing. We use logistic regression classifiers, and measure the accuracy using 5-fold cross validation. The use of logistic regression is due to the interpretability of the models, but also they show slightly better performances than support vector classifiers with RBF kernels, used in Gonen and Goldberg (2019).

The results show similar observations to Gonen and Goldberg (2019), such that the classifiers achieve the accuracy values of 0.88, and 0.78 with the vectors before and after debiasing, given the baseline accuracy of 0.59 using the most-frequent label.[3] Exploiting the interpretability of the explicit vectors, we investigate which features of the classifiers contribute the most to the predictions of the genders. To do this, we retrain the classifiers on the complete data (all gender-neutral occupations). Since each feature corresponds to a context word, the absolute values of the positive/negative coefficients of the features indicate the degrees of the contributions of the corresponding context words to the prediction of female/male.

Table 3 reports the context words with the highest contributions, before and after debiasing. Comparing the results, we observe that the debiased model, in addition to having common features with the original model (e.g. 'matron', 'midwife', 'feminist'), also exploits other context words, such as named entities, to identify the perceived gender of the words. In fact, while the effect of some gender-related features are reduced through applying the debiasing method, other context words are still strong indicators of gender-related aspects, providing sufficient information to retrieve the perceived genders.

These observations highlight the challenges in gender debiasing of word embeddings. Our results show that the debiasing method does not completely remove the primary gender-related features. In addition to them, other features (such as the ones related to named-entities) can become effective in indicating the perceived social biases, which can be potentially more challenging for debiasing word embeddings.

8 Conclusion

Word representation models capture the social biases embedded in our language use. To accurately measure bias in word embedding, we propose a novel approach based on the first-order relation between words and their context words, reconstructed from the word embeddings as explicit vectors. Our approach corrects the essential issue of the commonly-used second-order bias measurement method, caused by the circularity and transitivity of bias in word representations. We study the application of our method on three families of representation models, namely word2vec Skip-Gram, GloVe, and the PMI-based ones. The measured gender bias values of a set of occupations with our proposed method shows significantly higher correlations to the gender bias statistics of the U.S. job market, provided by two recent collections, in all three representation families. These results highlight the benefits of using our proposed method for capturing bias from text, as it more accurately reflects the societal phenomena. In particular, our method reveals the existence of a more severe degree of bias towards female in text for some specific jobs. Finally, we analyze the causes of the limitations of the debiasing algorithm proposed by Bolukbasi et al. (2016) using the explicit representations. Our observations point out the challenges of debiasing word embeddings, in that despite a decrease in the effects of the primary gender-related context words through the debiasing algorithm, several other context words are still effective indicators of the socially-perceived gender bias of words.

[3] Conducting the experiment on the SG vectors results in the same accuracy values.

References

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*.

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

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, volume 1.

Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*.

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. *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Lewis Molly and Gary Lupyan. 2019. Language use shapes cultural stereotypes: Large scale evidence from gender. *PsyArXiv preprint*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC Workshop on New Challenges for NLP Frameworks*.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2016. Sparse word embeddings using l1 regularized online learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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 North American Chapter of the 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 Conference on Empirical Methods in Natural Language Processing*.