A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces

Anne Lauscher,[1] **Goran Glavaš,**[1] **Simone Paolo Ponzetto,**[1] **Ivan Vulić**[2]
[1]Data and Web Science Research Group
University of Mannheim
{anne, goran, simone}@informatik.uni-mannheim.de
[2]Language Technology Lab
University of Cambridge
iv250@cam.ac.uk

Abstract

Distributional word vectors have recently been shown to encode many of the human biases, most notably gender and racial biases, and models for attenuating such biases have consequently been proposed. However, existing models and studies (1) operate on under-specified and mutually differing bias definitions, (2) are tailored for a particular bias (e.g., gender bias) and (3) have been evaluated inconsistently and non-rigorously. In this work, we introduce a general framework for debiasing word embeddings. We operationalize the definition of a bias by discerning two types of bias specification: explicit and implicit. We then propose three debiasing models that operate on explicit or implicit bias specifications and that can be composed towards more robust debiasing. Finally, we devise a full-fledged evaluation framework in which we couple existing bias metrics with newly proposed ones. Experimental findings across three embedding methods suggest that the proposed debiasing models are robust and widely applicable: they often completely remove the bias both implicitly and explicitly without degradation of semantic information encoded in any of the input distributional spaces. Moreover, we successfully transfer debiasing models, by means of cross-lingual embedding spaces, and remove or attenuate biases in distributional word vector spaces of languages that lack readily available bias specifications.

Introduction

Distributional word vectors have been recently shown to encode prominent human biases related to, e.g., gender or race (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Manzini et al. 2019). Such biases are observed across languages and embedding methods (Lauscher and Glavaš 2019), both in static and contextualized word embeddings (Zhao et al. 2019). While this issue requires remedy, the finding itself is hardly surprising: we project our biases, in terms of biased word co-occurrences, into the texts we produce. Consequently, this is propagated to embedding models, both static (Mikolov et al. 2013; Pennington, Socher, and Manning 2014; Bojanowski et al. 2017) and contextualized (Peters et al. 2018) alike, by virtue of the distributional hypothesis (Harris 1954).[1] While biases may be useful for diachronic or sociological analyses (Garg et al. 2018), they (1) raise ethical issues, since biases are amplified by machine learning models using embeddings as input (Zhao et al. 2017), and (2) impede tasks like coreference resolution (Zhao et al. 2018a; Rudinger et al. 2018) and abusive language detection (Park, Shin, and Fung 2018).

A number of methods for attenuating and eliminating human-like biases in word vectors have been proposed recently (Bolukbasi et al. 2016; Zhao et al. 2018a; 2018b; Dev and Phillips 2019). While they address the same types of bias – primarily the gender bias – they start from different bias "specifications" and either lack proper empirical evaluation (Bolukbasi et al. 2016) or employ different evaluation procedures, both hindering a direct comparison of the methods' "debiasing abilities" (Zhao et al. 2019; Dev and Phillips 2019; Manzini et al. 2019). What is more, the most prominent debiasing models (Bolukbasi et al. 2016; Zhao et al. 2018b) have been criticized for merely masking the bias instead of removing it (Gonen and Goldberg 2019). To resolve inconsistencies in the current debiasing research and evaluation, in this work we propose a general debiasing framework DEBIE (**DEB**iasing embeddings **I**mplicitly and **E**xplicitly), which operationalizes bias specifications, groups models according to the bias specification type they operate on, and evaluates models' abilities to remove biases both explicitly and implicitly (Gonen and Goldberg 2019).

We first define two types of bias specifications – *implicit* and *explicit* – and propose a method of augmenting bias specifications with the help of embeddings specialized for semantic similarity (Mrkšić et al. 2017; Ponti et al. 2018). We then introduce the main contributions of this work as follows. First, we present three novel debiasing models. (1) We adjust the linear projection method of Dev and Phillips (2019), an extension of the debiasing model of Bolukbasi et al. (2016), to operate on the augmented implicit bias specifications. (2) We then propose an alternative model that projects the embedding space to itself using the term sets from implicit bias specification as the projection signal. (3) Finally, we propose a simple

[1]Borrowing the famous example (Bolukbasi et al. 2016), *man* will be found more often in the same context with *programmer*, and *woman* with *homemaker* in any sufficiently large corpus.

and effective neural debiasing model, which is, to the best of our knowledge, the first debiasing model that operates on an explicit bias specification. All three models perform *post-hoc* debiasing: they can be applied to any pretrained distributional word vector space.[2] As another contribution, we combine existing bias metrics with newly proposed ones and assemble an evaluation suite that tests word vectors for explicit biases, implicit biases, and (preservation of) semantic quality. Finally, by coupling the proposed debiasing models with the cross-lingual embedding spaces (Ruder, Vulić, and Søgaard 2019; Glavaš et al. 2019), we facilitate cross-lingual debiasing transfer: we successfully debias embedding spaces in target languages without bias specifications in those languages. We hope that our work will lead to standardization of preprocessing and evaluation procedures in debiasing research and to increased comparability of debiasing models.[3]

General Debiasing Framework

In what follows, we first formalize two bias specifications – implicit and explicit. We then introduce new debiasing models: two operate on the implicit bias specification and the third on the explicit bias specification. Finally, we show how to debias word embeddings in a variety of target languages via cross-lingual embeddings.

Bias Specifications

An *implicit bias specification* $B_I = (T_1, T_2)$ consists of two sets of *target* terms with respect to which a bias is expected to exist in the embedding space. For example, two sets of science and art terms, $T_1 = \{$*physics*, *chemistry*, *experiment*$\}$ and $T_2 = \{$*poetry*, *dance*, *drama*$\}$ constitute an implicit specification of the gender bias. Strictly speaking, B_I does not specify a bias directly – it merely specifies two categories of concepts for which we *implicitly* assume that there exists some set of reference terms A (e.g., male terms *man*, *father* and/or female terms like *woman*, *girl*) with respect to which T_1 and T_2 exhibit differences. Most existing debiasing models (Bolukbasi et al. 2016; Zhao et al. 2018b; Dev and Phillips 2019; Manzini et al. 2019) operate on $B_I = (T_1, T_2)$, i.e., not requiring reference terms A.

An *explicit bias specification* B_E defines, in addition to sets T_1 and T_2, one or more reference *attribute* sets. We consider an explicit bias specification with a single attribute set, $B_E = (T_1, T_2, A)$ (as employed by our DEBIASNET model),[4] and also with two (opposing) attribute sets, $B_E = (T_1, T_2, A_1, A_2)$, as used in WEAT tests (Caliskan, Bryson, and Narayanan 2017).

Augmentation of Bias Specifications. The initial bias specification (B_I or B_E) commonly contains only a handful of words in each target and attribute set. These are commonly the most representative words of a category (e.g., *man*, *boy*, *father* to represent the category *male*). However, in order to provide a finer-grained bias specification, we propose to augment each term set with synonyms and semantically similar words of the initial terms. We therefore extract nearest neighbours of initial terms from an embedding space specialized to accentuate true semantic similarity and attenuate other types of semantic association (Faruqui et al. 2015; Vulić et al. 2018; Glavaš and Vulić 2018, *inter alia*). For the augmentation process we rely on the recent state-of-the-art similarity specialization method of Ponti et al. (2018): for more details see the original work.

Given B_I or B_E and a similarity-specialized word vector space $\mathbf{X_{sim}}$, we augment each of the term sets in the specification by retrieving the top k most (cosine-)similar terms from $\mathbf{X_{sim}}$ for each of the initial terms.[5] Extending bias specification sets using a similarity-specialized word vector space – as opposed to a regular distributional space – reduces the noisy augmentation stemming from semantic relatedness instead of true semantic similarity.[6] Table 1 illustrates the initial bias specification and the corresponding augmentation (showing $k = 2$ nearest neighbors, without the initial terms) for one explicitly defined gender bias.

Debiasing Models

We present three debiasing models, two of which operate on $B_I = (T_1, T_2)$ and one on the explicit bias specification $B_E = (T_1, T_2, A)$.

Generalized Bias-Direction Debiasing (GBDD) focuses on B_I as a generalization of the linear projection model proposed by Dev and Phillips (2019), itself, in turn, an extension of the hard-debiasing model of Bolukbasi et al. (2016).

The model of Dev and Phillips (2019) requires a stricter bias specification than our B_I: it requires T_1 and T_2 to be ordered lists of equal length L, so that the so-called equivalence pairs $\{(t_1^l, t_2^l)\}_{l=1}^{L}$ can be created. For instance, $T_1 = \{$*man*, *father*, *boy*$\}$ and $T_2 = \{$*woman*, *mother*, *girl*$\}$ give rise to the following equivalence pairs: (*man*, *woman*), (*father*, *mother*), and (*boy*, *girl*). For each equivalence pair (t_1^l, t_2^l) they compute the *bias direction vector* $\mathbf{b_l}$ by subtracting the vector of term t_2^l from the vector of term t_1^l. We find this bias specification overly restrictive: it requires an additional effort to create true equivalence pairs from T_1 and T_2 *and* it produces only L partial bias direction vectors. In contrast, we propose to create one bias direction vector $\mathbf{b_{ij}}$ for each pair (t_1^i, t_2^j), $t_1^i \in T_1$, $t_2^j \in T_2$. If T_1 and T_2 truly specify categories that are opposite in some regard (e.g., gender-wise), then any pair (t_1^i, t_2^j) should induce a meaningful partial bias direction vector. This way we also obtain a much larger number of

[2] In contrast, debiasing models like GN-GloVe (Zhao et al. 2018b) integrate debiasing constraints into objectives of embedding models like GloVe (Pennington, Socher, and Manning 2014), and thus cannot be directly ported to other embedding models.

[3] The code is available at https://github.com/umanlp/DEBIE.

[4] The attribute set A can be any set of attributes towards which the bias is to be removed. In our experiments, we joined the WEAT test specification attribute sets A_1 and A_2.

[5] We discard nearest neighbors initially present in other sets of the same bias specification: e.g., if we retrieve an augmentation candidate *woman* for an initial T_1 term *man*, *woman* will not be added to T_1 if it exists in T_2 (or in A-s).

[6] We also considered using clean lexical knowledge from WordNet (Miller 1995) directly, but this resulted in much lower recall as well as less accurate augmentation candidates.

Initial T_1	*science*, *technology*, *physics*, *chemistry*, *Einstein*, *NASA*, *experiment*, *astronomy*
Initial T_2	*poetry*, *art*, *Shakespeare*, *dance*, *literature*, *novel*, *symphony*, *drama*
Initial A_1	*brother*, *father*, *uncle*, *grandfather*, *son*, *he*, *his*, *him*
Initial A_2	*sister*, *mother*, *aunt*, *grandmother*, *daughter*, *she*, *hers*, *her*
Augmentation T_1	*automation*, *radiochemistry*, *test*, *biophysics*, *learning*, *electrodynamics*, *biochemistry*, *astrophysics*, *astrometry*
Augmentation T_1	*orchestra*, *artistry*, *dramaturgy*, *poesy*, *philharmonic*, *craft*, *untried*, *hop*, *poem*, *dancing*, *dissertation*, *treatise*
Augmentation A_1	*beget*, *buddy*, *forefather*, *man*, *nephew*, *own*, *himself*, *theirs*, *boy*, *crony*, *cousin*, *grandpa*, *granddad*
Augmentation A_2	*niece*, *girl*, *parent*, *grandma*, *granny*, *woman*, *theirs*, *sire*, *auntie*, *sibling*, *herself*, *jealously*, *stepmother*, *wife*

Table 1: Initial and augmented gender bias specifications. Test T8 from WEAT (Caliskan, Bryson, and Narayanan 2017).

partial bias direction vectors (e.g., L^2 if T_1 and T_2 are of the same length L): this should result in a more reliable *general bias direction vector*, computed as follows. We stack all of the obtained bias direction vectors $\mathbf{b_{ij}}$ corresponding to pairs (t_1^i, t_2^j), $t_1^i \in T_1$, $t_2^j \in T_2$ to form a bias direction matrix $\boldsymbol{B}$. We then obtain the *global* bias direction vector $\mathbf{b}$ as the top singular vector of $\boldsymbol{B}$, i.e., as the first row of matrix $\boldsymbol{V}$, where $\boldsymbol{U\Sigma V}^\top$ is the singular value decomposition of $\boldsymbol{B}$. Let $\mathbf{x}$ be the ℓ_2-normalized d-dimensional vector from a biased input vector space. Its debiased version is then computed as:

$$\text{GBDD}(\mathbf{x}) = \mathbf{x} - \langle \mathbf{x}, \mathbf{b} \rangle \mathbf{b} \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes a dot product. In other words, the closer the vector $\mathbf{x}$ is to the global bias direction $\mathbf{b}$, the more it is bias-corrected (i.e., the larger portion of $\mathbf{b}$ is subtracted from $\mathbf{x}$). Vectors orthogonal to the bias direction $\mathbf{b}$ remain unchanged (zero dot-product with the bias vector $\mathbf{b}$).

Bias-Alignment Model (BAM). An alternative to computing a bias direction vector $\mathbf{b}$ is to use target-term pairs (t_1^i, t_2^j), $t_1^i \in T_1$, $t_2^j \in T_2$ to learn a projection of the biased embedding space $\boldsymbol{X} \in \mathbb{R}^d$ to itself that (approximately) aligns T_1 and T_2. The idea behind this model stems from the research on projection-based cross-lingual word embeddings (CLWEs), where an orthogonal mapping between monolingual embedding spaces is learned from a set of word translations (Smith et al. 2017; Glavaš et al. 2019).[7]

Here, we use pairs (t_1^i, t_2^j) to learn the debiasing projection of $\boldsymbol{X}$ with respect to itself. Let $\boldsymbol{X}_{T_1}$ and $\boldsymbol{X}_{T_2}$ be the matrices obtained by stacking (biased) vectors of left and right words of pairs (t_1^i, t_2^j), respectively. We then learn the orthogonal map $\boldsymbol{W_X} = \boldsymbol{UV}^\top$, where $\boldsymbol{U\Sigma V}^\top$ is the singular value decomposition of $\boldsymbol{X}_{T_2}\boldsymbol{X}_{T_1}^\top$. Since $\boldsymbol{W_X}$ is orthogonal, the projection $\boldsymbol{X'} = \boldsymbol{XW_X}$ is isomorphic to the original space $\boldsymbol{X}$, and thus equally biased. However, the transformation (specified by $\boldsymbol{W_X}$) defines the angle and direction of debiasing. We obtain the debiased space by averaging the original space $\boldsymbol{X}$ and the projected space $\boldsymbol{X'}$:

$$\text{BAM}(\mathbf{X}) = \frac{1}{2}\left(\mathbf{X} + \mathbf{X}\boldsymbol{W}_X\right). \tag{2}$$

[7]Note that a self-consistent linear mapping W is the one offering consistent mapping from one space to the other and back, $x = \boldsymbol{W}^\top \boldsymbol{W} x$, i.e., $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}$, thus W is orthogonal; an orthogonal projection $W(\boldsymbol{X'} = \boldsymbol{WX})$ preserves all distances in $\boldsymbol{X}$, making $\boldsymbol{X'}$ isomorphic to $\boldsymbol{X}$.

Explicit Neural Debiasing (DEBIASNET). The final model, dubbed DEBIASNET, is a neural model that operates on the explicit bias specification B_E. It is inspired by the work on semantic specialization of word embeddings (Vulić et al. 2018; Glavaš and Vulić 2018): but instead of using linguistic constraints (e.g., synonyms), we "specialize" the vector space by leveraging debiasing constraints.

Given a biased input space $\boldsymbol{X}$ and the specification $B_E = (T_1, T_2, A)$, we learn a debiasing function $\text{DBN}(\boldsymbol{X}; \theta)$ that transforms $\boldsymbol{X}$ to a debiased space $\boldsymbol{X'}$. We aim for the terms from both sets T_1 and T_2 to be similarly close to the terms from A in $\boldsymbol{X'}$. For simplicity, we execute $\text{DBN}(\boldsymbol{X}; \theta)$ as a feed-forward neural network with non-linear activations. The training set for learning the parameters θ consists of triples $(t_1 \in T_1, t_2 \in T_2, a \in A)$. It is obtained as a full Cartesian product $T_1 \times T_2 \times A$. Let $\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{a}$ be the respective vectors of t_1, t_2, and a from the input biased space $\boldsymbol{X}$, and let $\mathbf{t}_1'$, $\mathbf{t}_2'$ and $\mathbf{a}'$ be their "debiased" transformations: $\mathbf{t}_1' = \text{DBN}(\mathbf{t}_1; \theta)$, $\mathbf{t}_2' = \text{DBN}(\mathbf{t}_2; \theta)$, and $\mathbf{a}' = \text{DBN}(\mathbf{a}; \theta)$. For a training instance (t_1, t_2, a), we then minimize the following loss function L_D:

$$L_D = \left(\cos_d\left(\mathbf{t}_1', \mathbf{a}'\right) - \cos_d\left(\mathbf{t}_2', \mathbf{a}'\right)\right)^2. \tag{3}$$

$\cos_d(\cdot, \cdot)$ refers to the cosine distance. The objective pushes the terms from the two target sets T_1 and T_2 to be equidistant to the terms from the attribute set A. That is, it is designed to specifically remove the explicit bias. By minimizing L_D as the only objective, the model would remove the bias, but it would also destroy the useful semantic information in the input space. We thus couple the objective L_D with the regularization L_R that prevents the debiased vectors to deviate too much from their original estimates:

$$L_R = \cos_d(\mathbf{t}_1, \mathbf{t}_1') + \cos_d(\mathbf{t}_2, \mathbf{t}_2') + \cos_d(\mathbf{a}, \mathbf{a}') \tag{4}$$

The final loss is then $J = L_D + \lambda L_R$, with λ as the regularization weight. The learned function is then applied to the full input space: $\boldsymbol{X'} = \text{DBN}(\boldsymbol{X}; \theta)$.

Composing Debiasing Models. The presented models can be seamlessly composed with one another. For example, given an explicit specification B_E, we can first explicitly debias a distributional space $\mathbf{X}$ using DEBIASNET. We can then apply either GBDD or BAM on the resulting vector space by deriving B_I from B_E (i.e., by considering only T_1 and T_2): e.g., $\boldsymbol{X'} = \text{GBDD}(\text{DBN}(\boldsymbol{X}))$.

Cross-Lingual Transfer of Debiasing

Cross-lingual word embeddings have been shown to be a viable solution for zero-shot language transfer of NLP models

(Ruder, Vulić, and Søgaard 2019; Glavaš et al. 2019). Conceptually, given a source language L_1 with its monolingual distributional space $\boldsymbol{X}_{L1}$ and a target language L_2 with the space $\boldsymbol{X}_{L2}$, we can apply any $L1$ model trained on $\boldsymbol{X}_{L1}$ on the instances from $L2$, given a matrix $\boldsymbol{W}_{CL}$ that projects $\boldsymbol{X}_{L2}$ to $\boldsymbol{X}_{L1}$. From the plethora of cross-lingual word embedding models (Smith et al. 2017; Conneau et al. 2018; Artetxe, Labaka, and Agirre 2018, *inter alia*), we opt for a supervised projection-based model (Smith et al. 2017) that obtains $\boldsymbol{W}_{CL}$ by solving the Procrustes problem (Schönemann 1966) on the set of word translation pairs.[8] We select this approach due to its simplicity and competitive zero-shot language transfer performance on other NLP tasks (Glavaš et al. 2019). With the cross-lingual projection matrix $\boldsymbol{W}_{CL}$ in place, the debiasing of the space $\boldsymbol{X}_{L2}$ amounts to composing the projection with the debiasing model in $L1$: e.g., for GBDD, $\boldsymbol{X}'_{L2} = \text{GBDD}_{L1}(\boldsymbol{X}_{L2}\boldsymbol{W}_{CL})$.

Evaluation and Experimental Setup

We now introduce the metrics for testing different aspects of debiased embedding spaces, and then outline two datasets used in our experiments.

Evaluation Aspects

We use three diverse tests to measure the presence of explicit bias, and two tests that focus on the presence of implicit bias. Finally, we test the debiased spaces for their ability to preserve the initial semantic information.

Word Embedding Association Test (WEAT). Introduced by Caliskan, Bryson, and Narayanan (2017), WEAT tests the embedding space for the presence of an explicit bias defined as B_E=(T_1, T_2, A_1, A_2). It computes the differential association between T_1 and T_2 based on their mean similarity with terms from the attribute sets A_1 and A_2:

$$s(B_E) = \sum_{t_1 \in T_1} s(t_1, A_1, A_2) - \sum_{t_2 \in T_2} s(t_2, A1, A2) \quad (5)$$

The association s of term $t \in T_i$ is computed as:

$$s(t,A_1,A_2) = \frac{1}{|A_1|}\sum_{a_1 \in A_1} \cos(\mathbf{t}, \mathbf{a_1}) - \frac{1}{|A_2|}\sum_{a_2 \in A_2} \cos(\mathbf{t}, \mathbf{a_2}) \quad (6)$$

The significance of the statistic is computed by comparing $s(B_E)$ with the scores $s(B^*_E)$ obtained with all permutations $B^*_E = (T^*_1, T^*_2, A_1, A_2)$, where T^*_1 and T^*_2 are equally sized partitions of $T_1 \cup T_2$. The p-value of the test is the probability of $s(B^*_E) > s(B_E)$. The "amount" of bias, the so-called *effect size*, is then a normalized measure of separation between association distributions:

$$\frac{\mu(\{s(t_1, A_1, A_2)\}_{t_1 \in T_1}) - \mu(\{s(t_2, A_1, A_2)\}_{t_2 \in T_2})}{\sigma\left(\{s(t, A_1, A_2)\}_{t \in T_1 \cup T_2}\right)} \quad (7)$$

where μ is the mean and σ is the standard deviation.

[8]Note that we obtain the cross-lingual projection $\boldsymbol{W}_{CL}$ in the similar way as debiasing projection $\boldsymbol{W}_X$ in BAM; but now the aligned matrices contain vectors (each from respective language) corresponding to word translation pairs (not pairs created from bias target sets as in BAM).

Embedding Coherence Test (ECT). It quantifies the amount of explicit bias B_E=$\{T_1, T_2, A\}$ (Dev and Phillips 2019). Unlike WEAT, it compares vectors of target sets T_1 and T_2 (averaged over the constituent terms) with vectors from a single attribute set A. ECT first computes the mean vectors for the target sets T_1 and T_2: $\mu_1 = \frac{1}{|T_1|}\sum_{t_1 \in T_1} \mathbf{t}_1$ and $\mu_2 = \frac{1}{|T_2|}\sum_{t_2 \in T_2} \mathbf{t}_2$. Next, for both μ_1 and μ_1 it computes the (cosine) similarities with vectors of all $\mathbf{a} \in A$. The two resultant vectors of similarity scores, $\mathbf{s}_1$ (for T_1) and $\mathbf{s}_2$ (for T_2) are used to obtain the final ECT score. It is the Spearman's rank correlation between the rank orders of $\mathbf{s}_1$ and $\mathbf{s}_2$ – the higher the correlation, the lower the bias.

Bias Analogy Test (BAT). Based on the observation of (Bolukbasi et al. 2016) that in a biased vector space $programmer - homemaker \approx man - woman$, Dev and Phillips (2019) proposed an analogy-based bias test: Embedding Quality Test (EQT). However, EQT depends on WordNet to extend the bias definition with synonyms and plurals of bias specification terms. In contrast, we propose an alternative Bias Analogy Test (BAT) that relies only on the specification $B_E = (T_1, T_2, A_1, A_2)$.

We first create all possible biased analogies $\mathbf{t}_1 - \mathbf{t}_2 \approx \mathbf{a}_1 - \mathbf{a}_2$ for $(t_1, t_2, a_1, a_2) \in T_1 \times T_2 \times A_1 \times A_2$. We then create two query vectors from each analogy: $\mathbf{q}_1 = \mathbf{t}_1 - \mathbf{t}_2 + \mathbf{a}_2$ and $\mathbf{q}_2 = \mathbf{a}_1 - \mathbf{t}_1 + \mathbf{t}_2$ for each 4-tuple (t_1, t_2, a_1, a_2). We then rank the vectors in the vector space $\boldsymbol{X}$ according to the Euclidean distance with each of the query vectors. In a biased space, we expect the vector $\mathbf{a}_1$ to be ranked higher for the query $\mathbf{q}_1$ than the vectors of terms from the opposing attribute set A_2 (e.g., for a gender-biased space we expect *woman* to be ranked higher than *father* or *boy* for the query *man - programmer + homemaker*). Also, $\mathbf{a}_2$ is expected to be more similar to $\mathbf{q}_2$ than vectors of A_1 terms . The BAT score is the percentage of cases where: (1) a_1 is ranked higher than a term $a'_2 \in A_2 \setminus \{a_2\}$ for $\mathbf{q}_1$ and (2) a_2 is ranked higher than a term $a'_1 \in A_1 \setminus \{a_1\}$ for $\mathbf{q}_2$.

Implicit Bias Tests. Gonen and Goldberg (2019) recently suggested that the two sets of target terms can still be clearly distinguished (with KMeans clustering, or in a supervised manner with an SVM classifier) from one another after applying debiasing procedures of (Bolukbasi et al. 2016) and (Zhao et al. 2018b). We adopt their approach and test the debiased spaces for the presence of implicit bias by clustering terms from T_1 and T_2 with KMeans++, and by classifying them using an SVM with the RBF kernel: it is trained on the vectors of terms from the augmentations of target sets. For each debiasing model, we average the clustering and classification scores over 20 independent runs.

Semantic Quality. Debiasing procedures change the topology of the input vector space; we thus have to verify that debiasing does not occur at the expense of the encoded semantic information. We test the debiased embedding spaces on two standard word similarity/relatedness benchmarks: SimLex-999 (Hill, Reichart, and Korhonen 2015) and WordSim-353

(Finkelstein et al. 2002).

Evaluation Datasets

Our proposed framework is versatile as it enables debiasing models to operate on any bias specified in the B_I or B_E format. To demonstrate this, we evaluate the debiasing models from the previous section on two different bias specifications: tests T1 and T8 from the WEAT dataset (Caliskan, Bryson, and Narayanan 2017). WEAT tests are given as explicit bias specifications B_E= (T_1, T_2, A_1, A_2).

WEAT T8: Gender Bias Test. WEAT T8, shown in Table 1, encodes a type of a gender bias in relation to affinities towards science and art. T_1 contains terms from the areas of science and technology, whereas T_2 contains art terms. Attribute sets contain male (A_1) and female (A_2) terms. In a gender-biased vector space the scientific targets are expected to be more strongly associated with male attributes, and artistic targets with female terms.

WEAT 1: Flowers vs. Insects. WEAT T1 specifies another bias type: the difference in *sentiment* humans attach to *insects* as opposed to *flowers*. Target sets contain different flowers (T_1) and insect species (T_2), and attribute sets contain universally positive (A_1) and negative (A_2) terms. The full bias specification of WEAT T1 is available in the supplementary.

XWEAT. For evaluating the language transfer setup, we use bias specifications in target languages as our test data. We use tests T1 and T8 from XWEAT, created by Lauscher and Glavaš (2019) by translating the English (EN) WEAT tests to six languages: German (DE), Spanish (ES), Italian (IT), Russian (RU), Croatian (HR), and Turkish (TR).

Preprocessing and Training Setup

Augmented Bias Specifications. We first augment the bias specifications using a similarity-specialized embedding space produced by Ponti et al. (2018)[9] based on the EN fastText embeddings (Bojanowski et al. 2017). For WEAT T8, we augment the target and attribute lists with $k = 4$ nearest neighbours of each term. As the initial lists of WEAT T1 are longer than those of T8, we use $k = 2$ with T1. We train all debiasing models using bias specifications containing *only* the augmentation terms (i.e., without the initial bias specification terms); we use the initial terms for testing.

Input Word Embeddings. We test the robustness of debiasing models on three different word embedding models trained on Wikipedia: CBOW (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), and fastText (FT) (Bojanowski et al. 2017). For cross-lingual transfer, we induce a multilingual space spanning seven languages (EN + 6 targets) by projecting FT vectors of each target to the EN space. Following an established procedure (Glavaš et al. 2019), we learn projections $\boldsymbol{W}_{CL}$ using automatically compiled translations of the 5K most frequent EN words.

[9]Available at: https://tinyurl.com/y273cuvk.

Training Setup. For GBDD and BAM there is a deterministic closed-form solution for any given bias specification. On the other hand, the hyper-parameters of DEBIASNET are optimized via grid search and cross-validation on the training set. The final DEBIASNET model uses 5 hidden layers with 300 units each and the weight λ is fixed to 0.2.

Results and Analysis

We first report debiasing results on three EN distributional spaces, for the individual models as well as for three composite models: GBDD ∘ BAM = GBDD(BAM($\boldsymbol{X}$)), BAM ∘ GBDD, and GBDD ∘ DEBIASNET.[10] We then show the results for the cross-lingual debiasing transfer. Finally, we analyze the topology of debiased spaces.

Main Evaluation

Biases of Distributional Spaces. The main results are summarized in Table 2. All three input distributional spaces generally exhibit explicit and implicit biases, with CBOW spaces displaying the lowest biases, both according to the WEAT tests (e.g., the effect size is even insignificant with $p < 0.05$ for the gender bias test T8) and the implicit bias tests of Gonen and Goldberg (2019). Interestingly – according to our BAT test, and despite the original claims and examples from Bolukbasi et al. (2016) – the encoded biases do not reflect strongly in the analogy tests. Nonetheless, our debiasing methods in most test settings manage to affect the input vector spaces by further reducing BAT scores.

Comparison of Debiasing Models. While the results vary across the two WEAT tests and evaluation metrics, GBDD emerges as the most robust model on average. It attenuates the explicit bias while being the most successful in removing the bias implicitly: the spaces debiased with GBDD completely confuse the KM clustering and SVM classifier. It also fully retains the useful semantic information: we do not observe drops on SL and WS compared to the input distributional spaces. While GBDD outperforms BAM and DEBIASNET (DBN) on average according to ECT and BAT measures, it is not able to fully remove the explicit gender bias (T8) according to the WEAT test.

Despite operating on an implicit specification B_I, BAM removes the explicit biases much better than the implicit ones. DBN seems even better than BAM in removing the explicit biases. This is not a surprise, since DBN is trained on an explicit bias specification. However both DBN and BAM are unsuccessful in removing the implicit biases. Moreover, DBN distorts the input space more than BAM, yielding substantial drops on SL and WS.

The complementarity of debiasing effects between GBDD, and BAM/DBN are confirmed by the performance of their compositions. All composition models robustly remove both explicit and implicit biases, also showing that there is no "one

[10]BAM and DEBIASNET display similar results and so does their composition. For brevity, we thus omit the scores of BAM ∘ DEBIASNET. We also do not report the scores with DEBIASNET ∘ GDBB as its scores were similar to its inverse composition GDBB ∘ DEBIASNET in our preliminary tests.

		WEAT T8 (gender bias, science vs. art)							WEAT T1 (sentiment, flowers vs. insects)						
		Explicit			Implicit		SemQ		Explicit			Implicit		SemQ	
	Model	WEAT	ECT	BAT	KM	SVM	SL	WS	WEAT	ECT	BAT	KM	SVM	SL	WS
FT	Distributional	1.30	73.5	41.0	100	100	38.2	73.8	1.67	46.2	56.1	95.7	100	38.2	73.0
	GBDD	0.96	84.7	33.9	62.9	**50.0**	**38.4**	**73.8**	0.08*	96.2	41.7	56.0	53.1	38.1	72.9
	BAM	0.10*	71.8	38.4	99.8	100	37.7	70.4	1.57	50.3	56.0	95.7	100	37.4	71.5
	DBN	**0.05***	79.1	**33.6**	99.8	100	34.1	65.1	0.18*	79.8	45	95.7	100	35.09	68.6
	GBDD ∘ BAM	0.18*	**94.4**	38.7	65.1	65.3	37.7	70.2	0.42*	89.3	48.1	75.0	91.4	37.3	71.3
	BAM ∘ GBDD	0.57*	90.3	34.6	**60.1**	**50.0**	36.4	72.6	**0.07***	94.4	42.4	56.9	**51.3**	37.9	68.4
	GBDD ∘ DBN	0.11*	81.5	37.4	65.8	50.3	33.9	64.6	-0.08*	95.9	**41.9**	**54.6**	**52.0**	34.9	68.4
CBOW	Distributional	0.81*	-24.0	45.6	90.6	93.4	34.7	59.4	1.13	78.1	50.2	62.6	93.9	**34.7**	**59.4**
	GBDD	0.38*	50.9	43.4	59.5	**50.0**	**34.8**	**59.8**	-0.07*	90.7	**41.1**	55.7	51.9	**34.7**	**59.4**
	BAM	0.14*	36.8	51.1	95.1	89.4	33.4	59.2	0.44*	82.4	50.7	60.9	94.4	34.4	59.3
	DBN	0.45*	4.7	57.5	97.4	98.4	33.9	52.2	0.60	82.5	46	85.7	90.8	33.4	53.4
	GBDD ∘ BAM	**0.00***	**69.4**	50.3	**52.7**	68.8	33.4	59.3	**-0.04***	**91.3**	48.7	60.7	68.1	34.5	59.2
	BAM ∘ GBDD	0.09*	65.6	42.7	62.6	**50.0**	33.2	56.9	-0.17*	89.2	45.3	55.6	**51.1**	33.2	57
	GBDD ∘ DBN	0.38*	-3.5	57.6	61.9	50.3	34.0	52.1	-0.15*	90.5	41.3	**55.4**	52.6	33.4	53.3
GloVe	Distributional	1.28	84.1	36.3	100	100	**36.9**	**60.5**	1.38	76.2	40.5	94.1	100	**36.9**	60.5
	GBDD	0.95	89.7	29.1	57.4	50.6	**36.9**	59.6	0.44*	**92.4**	32.7	55.6	54.5	36.8	**60.7**
	BAM	1.08	89.7	27.8	96	100	36.2	59.5	0.96	82.1	39.2	90.7	100	34.4	56.4
	DBN	0.83*	81.5	30.8	100	100	35.9	58.6	0.55	77.6	34.8	95.3	100	36.7	59.1
	GBDD ∘ BAM	0.98	94.7	**25.8**	63.6	79.1	36.6	59.3	0.40*	90.7	36.5	57.7	76.5	34.2	56.4
	BAM ∘ GBDD	0.78*	97.1	36.9	**53.9**	**50.0**	36.3	59.2	0.65	87.3	44.1	**55.5**	**51.2**	35.5	58.6
	GBDD ∘ DBN	**0.51***	**97.4**	28.2	59.5	**50.0**	35.8	58.4	**-0.03***	89.7	**30.3**	57.4	52.1	36.5	59.1

Table 2: Main results on two bias test sets, WEAT T8 and T1 for three EN distributional spaces debiased with three models – GBDD, BAM, and DebiasNet (DBN) – and their compositions. We quantify the explicit bias (Explicit): WEAT, ECT, and BAT evaluation measures; implicit bias (Implicit): clustering with KMeans++ (KM) and classification with SVM; and the preservation of semantic quality (SemQ): word similarity scores on SimLex (SL) and WordSim-353 (WS). Asterisks (*) indicate insignificant ($\alpha = 0.05$) bias effects for the WEAT evaluation measure.

		DE			ES			IT			RU			HR			TR		
	Model	W	KM	SL	W	KM	SL	W	KM	SL	W	KM	SL	W	KM	SL	W	KM	SL
FT	Distributional	**0.05***	98.3	40.7	1.16	99.8	–	0.10*	99.8	**29.8**	0.37*	62	25.6	0.13*	98.6	**32.7**	1.72	99.3	–
	GBDD	0.15*	55.4	40.7	0.41*	60	–	-0.28*	**56.1**	**29.8**	0.73*	62.4	**25.8**	0.54*	**59.9**	32.5	1.41	64.3	–
	BAM	-0.97	97.4	40.7	0.11*	99.0	–	-0.70*	99.6	29	-0.41*	74.4	25.1	**-0.01***	93.5	32	1.49	98.8	–
	DBN	-0.15*	97.4	36.2	0.76*	100	–	-1.05	100	25.4	**0.31***	77.9	20.7	0.25*	99.9	25.3	1.54	100	–
	GBDD ∘ BAM	0.35*	57.6	35.9	0.78*	**52.4**	–	-0.64*	60.1	25.0	0.77*	61.9	20.7	0.67*	67.5	25.1	1.29	62.5	–
	BAM ∘ GBDD	-0.12*	56.3	**40.8**	**0.05***	58	–	-0.62*	57.9	29	0.34*	**56.8**	24.8	0.52*	60.8	31.7	**0.99**	**56.9**	–
	GBDD ∘ DBN	-0.09*	**54.4**	37.3	0.11*	56.6	–	**-0.05***	58.9	27.1	0.59*	61.6	25.4	0.68*	75.4	29.4	1.27	62.4	–

Table 3: Results for cross-lingual debiasing transfer on XWEAT T8 for six languages: DE, ES, IT, RU, HR, and TR. Input word embeddings are fastText (FT) for all target languages.W=WEAT; KM=KMeans++; SL=SimLex.

model rules them all" solution to various debiasing aspects. GBDD ∘ DBN most effectively removes the biases, but it inherits the undesirable semantic distortions of DBN. On the other hand, BAM ∘ GBDD offers solid bias removal while for the most part retaining the semantic quality of the space.

Differences between Evaluation Measures. The three aspects of evaluation complement each other: they all inform the selection of the most appropriate debiasing model w.r.t. the desired application-specific criteria.[11] However, results of WEAT, ECT, and BAT are not always aligned. For example, the CBOW space is unbiased according to the WEAT test, but extremely biased (negative correlation!) according to ECT. In contrast, GloVe vectors are biased according to WEAT but not according to ECT (correlation of 0.84). These findings point to different bias aspects, accentuating the need for multiple, mutually complementary, bias measures.

[11] E.g., Note that for some bias specifications, one might not want to reduce/remove the implicit bias. WEAT T1 can be seen as an example of such bias: while we may want to make *insects* similarly *good/bad* as *flowers*, we do not want to make them indistinguishable from *flowers* in the vector space.

Cross-Lingual Transfer

The results in the cross-lingual debiasing transfer are shown in Table 3. For brevity, we show only the results on XWEAT T8 (gender bias wrt. science vs. art) and for a subset of evaluation measures (one for each evaluation aspect): WEAT

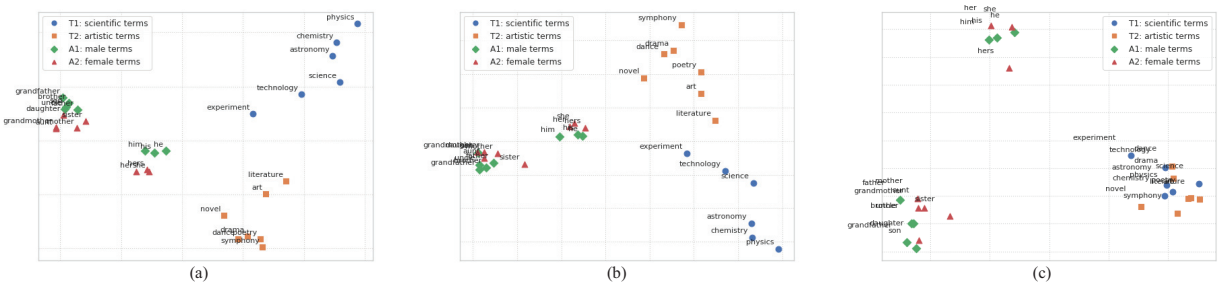

Figure 1: The topology of a vector space before and after debiasing. Terms from WEAT T8 test: T_1 – *science* terms (blue), T_2 – *art* terms (orange), A_1 – *male* terms (green), and A_2 – *female* terms (red). 2D projection with PCA. (a) Distributional EN FT vectors; (b) Debiased using BAM; (c) Debiased using GBDD.

(W), KMeans++ (KM), and SimLex (SL).[12,13]

We first confirm the results from Lauscher and Glavaš (2019): DE, IT, RU, and HR fastText vectors do not exhibit significant explicit gender bias (wrt. science vs. art), according to the WEAT test. The explicit bias is, however, significant in ES and TR distributional vectors. Implicit bias is clearly present in all distributional spaces except RU. Debiasing models display similar properties as before: DBN reduces the explicit bias more effectively than BAM and GBDD, but it semantically distorts the vectors; and only GBDD successfully removes the implicit bias. None of the models fully removes the explicit bias for TR (the lowest bias effect of 0.99 for BAM ∘ GBDD is still significant). We suspect that this is a result of the lower-quality cross-lingual TR→EN projection, which is in line with the bilingual lexicon induction results from Glavaš et al. (2019).

For DE and IT, BAM and DBN *invert* the direction of the bias: negative WEAT scores mean that *sciences* are more correlated with *female* attributes and *arts* with *male* attributes. We believe that this is the result of applying a (strong) bias correction learned on a biased EN space on the (explicitly) unbiased DE and IT spaces. The BAM ∘ GBDD composition seems most robust in the cross-lingual transfer setting – it successfully removes both the explicit (if they exist) and implicit biases, while preserving useful semantic information (SL). These results indicate that we can attenuate or remove biases in distributional vectors of languages for which (1) we do not require the initial bias specification and (2) we do not even need similarity-specialized word embeddings used to augment the bias specifications for the target language.

Topology of Debiased Spaces. Finally, we qualitatively analyze the debiasing effects suggested by evaluation measures. We project the input and the debiased embeddings into 2D with PCA, and show the constellation of words from the initial bias specification of WEAT T8 (Table 1) in Figure 1.[14]

In the distributional space, the two target sets (*science* vs *art*) are clearly distinguishable from one another (implicit bias), and so are the *male* and *female* attributes. The *science* terms are notably closer to the *male* terms and *art* terms to the *female* terms (explicit bias). The space produced by BAM intertwines the *male* and *female* terms and makes the *science* and *art* terms roughly equidistant to the gender terms (explicit bias removed), but the *science* terms are still clearly distinguishable from *art* terms (implicit bias still present). In the space produced by GBDD, both biases are removed: *science* and *art* terms cannot be clearly separated and are roughly equidistant to gender terms.

Conclusion

We have introduced a general framework for debiasing distributional word vector spaces by 1) formalizing the differences between implicit and explicit biases, 2) proposing new debiasing methods that deal with the two different bias specifications, and 3) designing a comprehensive evaluation framework for testing the (often complementary) effects of debiasing. While the proposed framework offers a systematized view on human biases encoded in word embeddings, the main results indicate that our debiasing methods can effectively attenuate biases in arbitrary input distributional spaces and can also be transferred to a variety of target languages.

References

Artetxe, M.; Labaka, G.; and Agirre, E. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, 789–798.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the ACL* 5:135–146.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of NIPS*, 4356–4364.

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

[12]We provide the full results, with all evaluation measures, and also on the XWEAT T1 test in the supplemental material.

[13]We evaluate word similarities for DE, IT, RU, and HR on their respective SimLex datasets (Leviant and Reichart 2015; Mrkšić et al. 2017); there is no ES and TR SimLex.

[14]We show only the input space and the spaces debiased with GBDD and BAM. We provide similar illustrations for other debiasing models in the supplementary material.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2018. Word translation without parallel data. In *Proceedings of ICLR*.

Dev, S., and Phillips, J. 2019. Attenuating bias in word vectors. In *Proceedings of AISTATS*.

Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E.; and Smith, N. A. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, 1606–1615.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems* 20(1):116–131.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS* 115(16):3635–3644.

Glavaš, G., and Vulić, I. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of ACL*, 34–45.

Glavaš, G.; Litschko, R.; Ruder, S.; and Vulić, I. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, 710–721.

Gonen, H., and Goldberg, Y. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, 609–614.

Harris, Z. S. 1954. Distributional structure. *Word* 10(23):146–162.

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

Lauscher, A., and Glavaš, G. 2019. Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. In *Proceedings of *SEM*, 85–91.

Leviant, I., and Reichart, R. 2015. Separated by an uncommon language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Manzini, T.; Chong, L. Y.; Black, A. W.; and Tsvetkov, Y. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of NAACL-HLT*, 615–621.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*, 3111–3119.

Miller, G. A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11):39–41.

Mrkšić, N.; Vulić, I.; Ó Séaghdha, D.; Leviant, I.; Reichart, R.; Gašić, M.; Korhonen, A.; and Young, S. 2017. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the ACL* 5:309–324.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing gender bias in abusive language detection. In *Proceedings of EMNLP*, 2799–2804.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, 1532–1543.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, 2227–2237.

Ponti, E. M.; Vulić, I.; Glavaš, G.; Mrkšić, N.; and Korhonen, A. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP*, 282–293.

Ruder, S.; Vulić, I.; and Søgaard, A. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research* 65:569–631.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender bias in coreference resolution. In *Proceedings of NAACL-HLT*, 8–14.

Schönemann, P. H. 1966. A generalized solution of the orthogonal Procrustes problem. *Psychometrika* 31(1):1–10.

Smith, S. L.; Turban, D. H.; Hamblin, S.; and Hammerla, N. Y. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Vulić, I.; Glavaš, G.; Mrkšić, N.; and Korhonen, A. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of the NAACL-HLT*, 516–527. New Orleans, Louisiana: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of EMNLP*, 2979–2989.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*, 15–20.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018b. Learning gender-neutral word embeddings. In *Proceedings of EMNLP*, 4847–4853.

Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender bias in contextualized word embeddings. In *Proceedings of NAACL-HLT*, 629–634.