Analyzing and Mitigating Gender Bias in Languages with Grammatical Gender and Bilingual Word Embeddings

Pei Zhou **Weijia Shi** **Jieyu Zhao**
{zpcs03, swj0419}@g.ucla.edu, jyzhao@cs.ucla.edu
Kuan-Hao Huang **Muhao Chen** **Kai-Wei Chang**
{khhuang, muhaochen, kwchang}@cs.ucla.edu
University of California, Los Angeles

Abstract

Word embeddings have been shown to contain gender bias that is inherited from their training corpora. However, existing work focuses on quantifying and mitigating such bias in English, and the analysis cannot be directly applied in language with grammatical gender, such as Spanish. In this paper, we propose new definitions of gender bias for languages with grammatical gender and apply bilingual word embeddings to analyze and mitigate the bias. Experimental results on cross-lingual analogy test and Word Embedding Association Test show that the proposed methods can effectively mitigate the multifaceted gender bias.

1 Introduction

Although word embeddings (Mikolov et al., 2013; Pennington et al., 2014) are widely used in many NLP tasks, recent work has shown that such embeddings derived from text corpora reflect gender biases in society (Bolukbasi et al., 2016; Caliskan et al., 2017) and cause deteriorated effects in downstream tasks (Zhao et al., 2018a; Font and Costa-jussà, 2019). Hence, extensive efforts have been put to mitigate the bias in word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018b).

Previous work focuses on gender bias in English (EN) word embeddings. However, these methods for measuring and mitigating bias in English are not able to address gender bias in languages that contain grammatical gender[1], where all nouns are assigned a gender class and the corresponding dependent articles, adjectives, and verbs must agree in gender with the noun (e.g. in Spanish: *la buena enfermera*–the good female nurse, *el buen enfermero*–the good male nurse) (Corbett, 1991, 2006). Most existing approaches define bias in word embeddings based on the projection of a word on a gender direction (e.g. "nurse" in English is biased because its projection on the gender direction inclines towards female but there is no gender information in its definition). When grammatical gender exists, such bias definition is problematic as masculine and feminine words naturally contain gender information from morphological agreement, e.g. the definitions of "enfermero" (male nurse) and "enfermera" (female nurse) are gendered, but this should not be considered as a stereotype.

However, gender bias in the embeddings of languages with grammatical gender indeed exists. When we align bias-mitigated English embeddings with Spanish (ES) embeddings, the word "lawyer" is closer to "abogado" (male lawyer) than "abogada" (female lawyer). This observation implies a discrepancy in semantics between the masculine and feminine forms of the same occupation in Spanish word embeddings. A similar discrepancy is also found in French (FR) word embeddings.

In this paper, we refer to languages with grammatical gender as *gendered languages* and languages that does not mark grammatical gender as *genderless languages*. We use Spanish as a running example and propose new methods for quantifying bias in word embeddings of gendered languages and bilingual word embeddings that align a gendered language with a genderless language like English[2]. We first define gender bias in the word embedding of gendered languages by constructing two gender directions: the semantic gender direc-

[1] Grammatical gender is a complicated linguistic reality. Many languages contain more than two gender classes. In this paper, we focus on languages with masculine and feminine classes. For gender in semantics, we follow the literature and address only binary gender.

[2] Although English has distinct male and female pronouns, it has no distinction of grammatical gender for most nouns. https://en.wikipedia.org/wiki/Genderless_language

tion and the grammatical gender direction. We then analyze gender bias in bilingual embeddings using similar approaches.

To mitigate gender bias in these embeddings, we propose two approaches. One is shifting words along the semantic gender direction with respect to an anchor point and the other is mitigating English first and then aligning the embedding spaces. Results show that a hybrid of the two approaches is able to effectively mitigate bias in Spanish word embeddings as well as EN-ES bilingual word embeddings.

We summarize our contributions as below. (1) We show that word embeddings of gendered languages such as Spanish and French contain gender bias and bilingual word embeddings aligning these languages to a genderless language like English also inherit the bias. (2) Based on our observation, we propose new definitions of gender bias by constructing two gender directions as those for English word embeddings fail to adopt to gendered languages directly. (3) We propose new metrics to evaluate gender bias and new methods to mitigate it for both monolingual and bilingual embeddings and show that our methods effectively mitigate bias.

2 Related Work

Previous work has proposed definitions for gender bias in English word embeddings. Bolukbasi et al. (2016) define bias in English embeddings being that one word without gender information in its definition shows an inclination towards one gender. They define a gender direction using the difference between male- and female-definition word embeddings and show that occupational words have different distances to “male” or “female” on this direction. This is appropriate for English as it does not distinguish between the masculine and feminine forms for most nouns but not applicable for gendered languages as mentioned earlier.

McCurdy and Serbeti (2017) examine grammatical gender bias in word embeddings by computing the WEAT association score (Caliskan et al., 2017) between gendered object nouns (e.g. moon-sun) and gender-definition words. They also mitigate bias by lemmatization to remove gender information in corpora. However, we argue that the association between gendered object nouns with gender attributes should not be considered as stereotypical bias since the association could be caused by the morphological agreement instead of stereotypes. Mitigation by completely removing gender information is also implausible as too much information will be lost by lemmatization.

Others have also worked on measuring and reducing gender bias in contextualized word embeddings (Zhao et al., 2019; May et al., 2019; Basta et al., 2019), but they also focus on the English monolingual embeddings in which the gender is only expressed by the pronouns (Stahlberg et al., 2007) while in morphologically rich languages, nouns are assigned with a gender form such as feminine and masculine (Corbett, 1991, 2006).

To mitigate gender bias in English, Zhao et al. (2018b) mitigate bias by saving one dimension of the word vector for gender. Bordia and Bowman (2019) proposes a regularization loss term for word-level language models. Zhang et al. (2018) use an adversarial network to mitigate bias in word embeddings. All these approaches consider the definition of gender bias from Bolukbasi et al. (2016) and they still focus on English word embeddings. Moreover, Gonen and Goldberg (2019) show that mitigation methods based on gender directions are not sufficient, since the embeddings of socially-biased words still cluster together.

When adopting the word embeddings in downstream taks, the output can also be biased (Zhao et al., 2018a; Font and Costa-jussà, 2019). Besides the gender bias in word embeddings, implicit stereotypes have been shown in other real world applications, such as online reviews (Wallace and Paul, 2016), advertisement (Sweeney, 2013) and web search (Kay et al., 2015). Dataset bias (Zhao et al., 2017; Rudinger et al., 2017) and word embedding bias (Bolukbasi et al., 2016; Caliskan et al., 2017) both contribute to this problem while this work focuses on the later part, specifically for gendered languages.

3 Gender Bias Analysis and Mitigation

As mentioned earlier, gender information in Spanish is far different from that in English, so we cannot directly adopt the gender direction from English embeddings based on previous work. In this section, we will first describe our definition about the gender directions in Spanish and by projecting the masculine and feminine forms of same occupation words to these directions, we show there is gender bias in Spanish embeddings. In the end, we

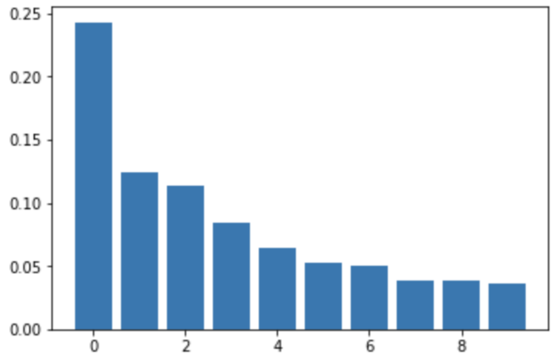

Figure 1: Percentage of variance explained in PCA of vector differences for gender-definition pairs when constructing *semantic gender* direction.

propose methods to mitigate the bias.

3.1 Bias in Spanish Embeddings

Gender Directions in Spanish We define two directions for gendered languages. One is for *semantic gender*, which is used to measure the semantically male or female inclination of the word. The other is for *grammatical gender*, which is used to capture the inherently carried gender attribute of the word. We claim that the semantic gender direction is enough for measuring bias in English embeddings since English does not have grammatical gender. But for languages like Spanish and French, the second type of gender direction is necessary. We also constrain that the two directions are orthogonal to each other to better distinguish between the two types of gender information. For all nouns in gendered languages, we do not take the inclinations along the grammatical gender direction as bias and only focus on the bias shown in the semantic gender.

Grammatical Gender Most or all nouns in gendered languages are assigned with one gender class. The number of grammatical gender classes ranges from two to several tens (Corbett, 1991). We focus on noun class systems where feminine and masculine grammatical gender exist but claim that our method can be generalized to languages where multiple gender classes exist like German. Since most nouns are assigned only one gender class, we cannot follow the previous approach to collect pairs of gendered words (e.g., "she" and "he") and capture the grammatical gender direction using principal component analysis (PCA) (Jolliffe, 2011). We instead collect around 3000 common object nouns in the gendered language that are grammatically masculine and 3000 that are feminine (data included in the supplementary materials). We use Linear Discriminant Analysis (LDA) (Fisher, 1936), a standard approach for supervised dimension reduction, to learn the grammatical gender direction $\vec{d_g}$ from the collected words. The model achieves an average accuracy of 0.92 for predicting the grammatical gender in Spanish and 0.83 in French with 5-fold cross-validation. We also verify the computed direction by using another classifier: linear SVM, and find that the direction from SVM has a cosine similarity of 0.99 with the direction from LDA.

Semantic Gender Similar to Bolukbasi et al. (2016), we first define a gender direction by the difference between male- and female-definition word embeddings. We conduct PCA using gender-definition pairs in gendered languages, e.g. "mujer" (woman) and "hombre" (man) for Spanish. Figure 1 shows the percentage of variance explained by each component for Spanish fastText (Bojanowski et al., 2017) word embeddings pre-trained on Wikipedia. The cosine similarity between these two gender directions ($\vec{d}_{PCA}$, $\vec{d_g}$) is 0.389, indicating these two directions are overlapped to some extent. This result is reasonable because some pairs of gender-definition words are also marked with grammatical gender, e.g. "mujer" is grammatically feminine and "hombre" is grammatically masculine. Therefore, some grammatical gender information is also included in the semantic gender direction. To better distinguish between these two directions, we remove the grammatical gender component in the computed gender direction to make the semantic gender direction $\vec{d_s}$ orthogonal to the grammatical gender direction:

$$\vec{d_s} = \vec{d}_{PCA} - \left\langle \vec{d}_{PCA}, \vec{d_g} \right\rangle \vec{d_g},$$

where $\langle \vec{x}, \vec{y} \rangle$ represents the inner product of two vectors.

Quantification of Gender Bias With the grammatical gender and the semantic gender directions defined, we can formally quantify gender bias in gendered languages. We consider two types of words in gendered languages. One is *inanimate nouns* that have only one assigned grammatical gender, like "agua" (water, feminine). The other is *animate nouns* that have two grammatical gender

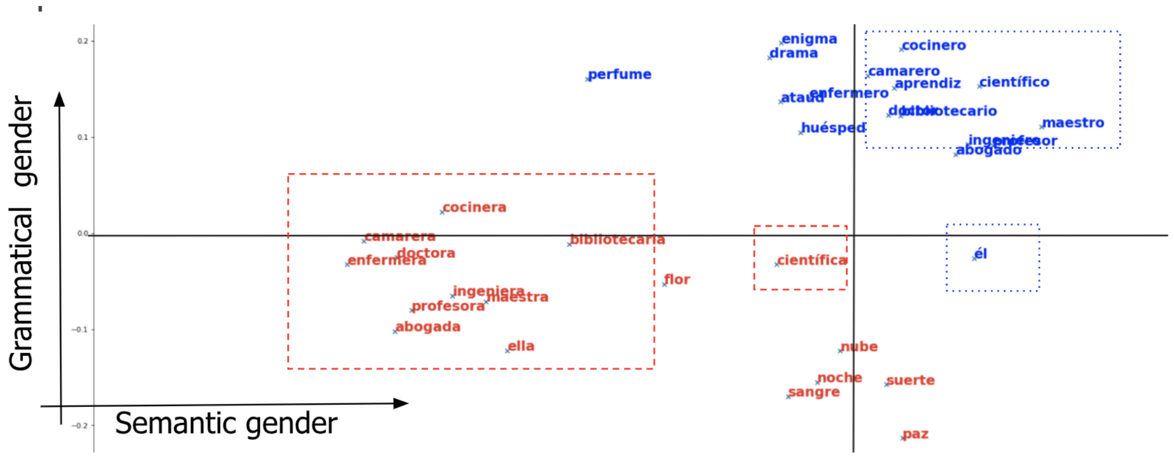

Figure 2: Projections of selected occupation words (enclosed in dotted lines) and common nouns in Spanish word embeddings on grammatical and semantic directions with masculine nouns in blue and feminine nouns in red.

forms, like "doctor" (male doctor) and "doctora" (female doctor). For inanimate nouns, we follow the previous work's approach to define the gender bias as the component of the word on the calculated semantic gender direction. One important difference from the previous approach is that we use the semantic gender direction which is orthogonal to the grammatical gender to remove the effect of grammatical gender when quantifying bias. We define the bias for inanimate nouns as:

$$\mathrm{b}_w = \left\langle \vec{w}, \vec{d_s} \right\rangle,$$

where $\vec{w}$ is embedding for the target word. For animate nouns, most of which are used to describe people, we propose a new metric to quantify gender bias. We define that there is gender bias in the embeddings if two forms of the same word are far from symmetric on the semantic gender direction with respect to an anchor point. The anchor point represents the gender-neutral position on the gender directions and a simple choice could be the origin point on the axis defined by the semantic gender direction. Let $\vec{w_f}$ be the word in feminine form, $\vec{w_m}$ be the word in masculine form, and $\vec{w_a}$ be the anchor point. We define bias as:

$$\mathrm{b}_w = \left| \left\langle \vec{w_m}, \vec{d_s} \right\rangle + \left\langle \vec{w_f}, \vec{d_s} \right\rangle - 2 \left\langle \vec{w_a}, \vec{d_s} \right\rangle \right|.$$

Note that for most pairs, $\left\langle \vec{w_m}, \vec{d_s} \right\rangle$ and $\left\langle \vec{w_f}, \vec{d_s} \right\rangle$ have opposite signs, meaning that the two forms lie on the opposite sides of the semantic gender axis.

Intuitively, this measures how much we need to move the pair of the words on the semantic gender direction so that they are symmetric with respect to the anchor point (gender-neutral position). If we use the origin as anchor point, i.e. $\left\langle \vec{w_a}, \vec{d_s} \right\rangle = 0$, the bias is just the absolute value of the sum of two projections.

Visualizing and Analyzing Bias in Spanish We use Spanish fastText (Bojanowski et al., 2017) embeddings pre-trained on Spanish Wikipedia and bilingual word embeddings from MUSE (Conneau et al., 2017) that aligns English and Spanish fastText embeddings together in a single vector space. To show bias in Spanish, we take the masculine and feminine pairs of several occupational words and project them on the gender directions we defined above. We also project some other common nouns with one gender form on the directions. We enclose masculine and feminine forms of the occupation words as well as the Spanish word for "he" ("él") and "she" ("ella") by dotted blue and red lines, respectively. Figure 2 shows that the Spanish word embeddings are biased from the following analysis.

The masculine and feminine forms of the occupation words are on the opposite sides for both directions. However, while their projections on the grammatical gender direction are symmetric with respect to the x-axis that indicates the neutral grammatical gender position, but are largely asymmetric with respect to the y-axis, i.e. the neutral *semantic gender* position. Along the *semantic gender* direction, occupation words in feminine forms incline to the feminine more than the inclination of masculine forms to the opposite side.

This discrepancy shows the difference in the gender information carried by the two forms of the same words and conforms our definition for gender bias in ES. Besides, we also find some interesting cases like “él” (he) and “científica” (feminine scientist) that are different from rest of the words in their group. We speculate that their extreme frequencies (too high or too low) lead to this phenomena.

As for common nouns, we find that most common nouns lie in the middle on the semantic gender direction, but words with different grammatical gender are on different sides when projected on the grammatical gender direction. Two exceptions are “perfume” (perfume, grammatically masculine) and “flor” (flower, grammatically feminine), which are leaning towards the feminine semantic gender. This shows that the two directions are able to distinguish between *grammatical gender* and *semantic gender* in Spanish and provide a way to measure two types of gender information.

3.2 Mitigation Methods

Mitigating English Before Alignment Although mitigation methods for Spanish word embeddings are underexplored, many approaches have been proposed for English and they could be helpful for mitigating Spanish word embeddings. The alignment for constructing bilingual word embeddings is based on EN-ES seed-lexicon (Conneau et al., 2017). The intuition of mitigating gender bias in English before alignment is that it could potentially align the Spanish words with the less biased English embeddings and thus fix the two gender forms of the Spanish terms in more symmetric positions in the vector space. After alignment, we can treat Spanish words in bilingual word embeddings as our mitigated Spanish word embeddings.

Shifting Along the Semantic Gender Direction The second method mitigates bias as a post-processing step and extends the “hard-debiasing” approach proposed by Bolukbasi et al. (2016). For words that have two gender forms like occupation terms, instead of zeroing the projection of gender-neutral words on the gender direction, we want them to be symmetric along the semantic gender direction on opposite sides. We find an *anchor* point that represents the gender-neutral position and shift the two forms along the semantic gender direction so that they have the same distance to the anchor position. We consider two types of anchor position: the zero point of the gender axis and the projection of the mitigated English word using “hard-debiasing” approach in the bilingual word embeddings. Although Gonen and Goldberg (2019) show that mitigating by moving on the gender direction is not sufficient because words with gender bias still tend to group together, we argue that for languages with grammatical gender, grouping of masculine and feminine words does not necessarily indicate bias and shifting words on the semantic gender direction is able to reduce gender bias.

4 Experiments

4.1 Evaluation Methods

Cross-lingual Analogy Task (CLAT) To better evaluate the bias in Spanish, we propose a cross-lingual word analogy task. The task follows the format “a:b = c:?”. Specifically, given a pair of English words (one can be either noun, adjective or verb, and the other is an occupation word that has been “debiased” in English word embeddings) and the corresponding Spanish word, the task is to predict the missing Spanish occupation word. This task compares the masculine and feminine forms of the occupation word in Spanish with the word in English to see whether the embedding of one form is closer to that in the mitigated English embeddings. We test around 50 analogy pairs and collect the ranking difference between two forms as well as the similarity scores. A larger gap between the two versions shows stronger bias in this occupation.

Word Embedding Association Test (WEAT) WEAT is developed by Caliskan et al. (2017) to measure the association between two sets of target concepts and two sets of attributes. Let X and Y be equal-size sets of target concept embeddings and let A and B be sets of attribute embeddings. Let $\cos(\vec{a}, \vec{b})$ denote the cosine similarity for vectors $\vec{a}$ and $\vec{b}$. The test statistic is a difference between sums over the respective target concepts,

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where each addend is the difference between mean cosine similarities of the respective at-

	Original	Shift (Ori)	Shift (EN)	De-Align	De-Shift (Ori)	De-Shift (EN)
CLAT–ASD	0.1244	0.1024	0.0978	0.0735	0.0642	**0.0586**
CLAT–ARD	17.8413	15.3968	14.8413	1.6984	**1.6191**	**1.6191**
WEAT–Male	0.4633	0.9245	**0.9010**	0.4633	0.9254	0.5699
WEAT–Female	1.3339	0.87272	**0.8962**	1.3339	0.8718	1.2273

Table 1: Results for different bias mitigation methods on two types of evaluation metrics. "CLAT" stands for cross-lingual analogy task, "ASD" is the average similarity difference for masculine and feminine words, "ARD" is the average ranking difference, "WEAT–Male" is the association of male occupation words with male-definition terms subtracting that with female-definition, similarly for "WEAT–Female", "Shift (Ori)" is the debiasing method of shifting along the semantic gender direction with the zero point as anchors, similarly for "Shift (EN)", which treats the debiased EN counterparts as anchors, "De-Align" is first debias EN and then align, and "De-Shift (Ori)" is the method combining first debiasing then align and shifting along semantic direction with the origin as anchor as post-processing.

tributes (May et al., 2019),

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}).$$

Since ES contains grammatical gender, masculine words should be associated with male-definition terms more than female-definition terms and vice versa. Thus, we modify WEAT and compare the association scores for masculine and feminine occupation words with male and female attribute words. We treat $\sum_{x \in X} s(x, A, B)$ as the association of target concept X with the attribute and compare the absolute values for masculine and feminine occupation words. If the difference is large, then one set of words in one gender form associates with that gender more than the other, indicating the gap in gender information carried by two forms.

4.2 Results

This section analyzes our experimental results on the two evaluation methods before and after using our mitigation approaches. We test "Mitigating-First" and "Shifting" approaches introduced before. We also test the combination of the above two approaches, i.e., we first mitigate English, align English and Spanish, and shift words along *semantic gender* direction as a post-processing step. We consider both zero and mitigated English words the neutral *anchor* position.

From Table 1, we can see that mitigating before alignment (De-Align) can significantly shorten the gap between two gender forms for the cross-lingual analogy task, while shifting along the semantic gender direction (Shift) is better at reducing the discrepancy in the WEAT association for two gender forms. This is probably because that aligning Spanish words with the mitigated English words will make the two gender forms have similar distances to the corresponding English word. While shifting after alignment forces the embeddings of the two gender forms associate with each gender concepts more equally. Overall the results suggest that a combination of the two approaches can benefit from both and (De-Shift (Ori)) can effectively mitigate the gender bias in ES or bilingual word embeddings according to the two tasks we consider.

5 Conclusion and Future Work

We conduct analysis and mitigation of gender bias in Spanish and English-Spanish bilingual word embeddings. We introduce new definitions to measure and quantify bias in Spanish, analyze phenomena for both grammatical and semantic gender, and design methods to mitigate bias. We show that the proposed method of combining mitigating before alignment and post-processing by shifting along the semantic gender direction efficiently closes the gap between the two gender forms in Spanish as well as English-Spanish bilingual word embeddings.

Several directions for future work include testing Spanish and English-Spanish bilingual word embeddings on downstream tasks to measure bias and also test the performance for mitigation methods. Moreover, one can extend our approach to other languages with grammatical gender like French or German.

References

Christine Basta, Marta R Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783*.

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

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035*.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press.

Greville G Corbett. 2006. *Agreement*, volume 109. Cambridge University Press.

Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Joel Escudé Font and Marta R Costa-jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*.

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. *arXiv preprint arXiv:1903.03862*.

Ian Jolliffe. 2011. *Principal component analysis*. Springer.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Human Factors in Computing Systems*, pages 3819–3828.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Katherine McCurdy and Oguz Serbeti. 2017. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *WiNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social communication*, pages 163–187.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue*, 11(3):10.

Byron C Wallace and Michael J Paul. 2016. jerk or judgemental? patient perceptions of male versus female physicians in online reviews.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *NAACL (short)*.

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. *arXiv preprint arXiv:1707.09457*.

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

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.