Examining Gender Bias in Languages with Grammatical Gender

Pei Zhou\textsuperscript{1,2}, Weijia Shi\textsuperscript{1}, Jieyu Zhao\textsuperscript{1}, Kuan-Hao Huang\textsuperscript{1}, Muhao Chen\textsuperscript{1,3}, Ryan Cotterell\textsuperscript{4}, Kai-Wei Chang\textsuperscript{4}

\textsuperscript{1}Department of Computer Science, University of California Los Angeles
\textsuperscript{2}Department of Computer Science, University of Southern California
\textsuperscript{3}Department of Computer and Information Science, University of Pennsylvania
\textsuperscript{4}Department of Computer Science, Johns Hopkins University

\{peiz@usc.edu, swj0419, jyzhao, khhuang, kwchang\}@cs.ucla.edu; muhao@seas.upenn.edu; ryan.cotterell@jhu.edu

Abstract

Recent studies have shown that word embeddings exhibit gender bias inherited from the training corpora. However, most studies to date have focused on quantifying and mitigating such bias only in English. These analyses cannot be directly extended to languages that exhibit morphological agreement on gender, such as Spanish and French. In this paper, we propose new metrics for evaluating gender bias in word embeddings of these languages and further demonstrate evidence of gender bias in bilingual embeddings which align these languages with English. Finally, we extend an existing approach to mitigate gender bias in word embedding under both monolingual and bilingual settings. Experiments on modified Word Embedding Association Test, word similarity, word translation, and word pair translation tasks show that the proposed approaches effectively reduce the gender bias while preserving the utility of the embeddings.

1 Introduction

Word embeddings are widely used in modern natural language processing tools. By virtue of their being trained on large, human-written corpora, recent studies have shown that word embeddings, in addition to capturing a word’s semantics, also encode gender bias in society (Bolukbasi et al., 2016; Caliskan et al., 2017). As a result of this bias, the embeddings may cause undesired consequences in the resulting models (Zhao et al., 2018a; Font and Costa-jussà, 2019). Therefore, extensive effort has been put toward analyzing and mitigating gender bias in word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018b; Dev and Phillips, 2019; Ethayarajh et al., 2019).

Existing studies on gender bias almost exclusively focus on English (EN) word embeddings. Unfortunately, the techniques used to mitigate bias in English word embeddings cannot be directly applied to languages with grammatical gender\textsuperscript{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: \textit{la buena enfermera} the good female nurse, \textit{el buen enfermero} the good male nurse) (Corbett, 1991, 2006). This is because 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). When the grammatical gender exists, such bias definition is problematic as masculine and feminine words naturally carry gender information. For example, “enfermero” (male nurse) and “enfermera” (female nurse) lean toward male and female, respectively. This is due to morphological agreement and should not be considered as a stereotype.

However, gender bias in the embeddings of languages with grammatical gender indeed exists. For example, when we align Spanish (ES) embeddings to English embeddings, the word “abogado” (male lawyer) is closer to “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 observation is also found in other languages, such as French (FR).

In this paper, we refer languages that have gender distinctions for all nouns (e.g., Spanish and German) as \textit{gendered languages} and others that do not mark grammatical gender as \textit{languages without grammatical gender} (e.g., English). We use Spanish and French as running examples and pro-

\textsuperscript{1}Grammatical gender is a complicated linguistic phenomenon; 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.
pose quantitative methods for evaluating bias in word embeddings of gendered languages and bilingual word embeddings that align a gendered language with English. We first define gender bias in the word embeddings of gendered languages by constructing two gender directions: the semantic gender direction and the grammatical gender direction. We then analyze gender bias in bilingual embeddings using a similar approach.

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 these two approaches effectively mitigate bias in Spanish and French word embeddings as well as ES-EN, FR-EN bilingual word embeddings. We also show through word similarity and word translation experiments that the utility of the original monolingual and bilingual word embeddings is preserved.

Our contributions are summarized in the following. (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 English also inherit the bias. (2) Based on the observation, we propose new definitions of gender bias by constructing two gender directions. (3) We propose methods to reduce gender bias for both monolingual and bilingual embeddings and show that they effectively mitigate bias while preserving the original utility of embeddings. Source code and data are available at https://github.com/shaoxia57/Bias_in_Gendered_Languages.

2 Gender Bias Analysis

This section discusses how to analyze bias in embeddings of gendered languages and bilingual word embeddings.

2.1 Bias in Gendered Languages

In gendered languages, all nouns are assigned a gender class. However, inanimate objects (e.g., water and spoon) do not carry the meaning of male or female. To address this issue, we define two gender directions: (1) semantic gender, which is defined by a set of gender definition words (e.g., man, male, waitress) and (2) grammatical gender, which is defined by a set of masculine and feminine nouns (e.g., water, table, woman). Most analysis of gender bias assumes that the language (e.g., English) only has the former direction. However, when analyzing languages like Spanish and French, considering the second is necessary. We discuss these two directions in detail as follows.

Semantic Gender Following Bolukbasi et al. (2016), we collect a set of gender-definition pairs (e.g., “mujer” (woman) and “hombre” (man) in Spanish). Then, the gender direction is derived by conducting principal component analysis (PCA) (Jolliffe, 2011) over the differences between male- and female-definition word vectors. Similar to the analysis in English, we observe that there is one major principal component carrying the meaning of gender in French and Spanish and we define it as the semantic gender direction $\vec{d}_{PCA}$.

Grammatical Gender The number of grammatical gender classes ranges from two to several tens (Corbett, 1991). To simplify the discussion, we focus on noun class systems where two major gender classes are feminine and masculine. However, the proposed approach can be generalized to languages with multiple gender classes (e.g., German). To identify grammatical gender direction, we collect around 3,000 common nouns that are grammatically masculine and 3,000 nouns that are feminine. Since most nouns (e.g., water) do not have a paired word in the other gender class, we do not have pairs of words to represent different grammatical genders. Therefore, instead of applying PCA, we learn the grammatical gender direction $\vec{d}_g$ by Linear Discriminant Analysis (LDA) (Fisher, 1936), which is a standard approach for supervised dimension reduction. 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.

Comparing $\vec{d}_{PCA}$ with $\vec{d}_g$, the cosine similarity between them is 0.389, indicating these two directions are overlapped to some extent but not identical. This is reasonable because words such as “mujer (woman),” “doctor (female doctor)” are both semantically and grammatically marked as feminine. To better distinguish between these two directions, we project out the grammatical gender component in the computed gender direction to make the semantic gender direction $\vec{d}_s$ orthogonal to the grammatical gender direction:

$$\vec{d}_s = \vec{d}_{PCA} - \frac{\langle \vec{d}_{PCA}, \vec{d}_g \rangle}{\vec{d}_g} \vec{d}_g.$$
Grammatical Gender Direction

where \( \langle \hat{x}, \hat{y} \rangle \) represents the inner product of two vectors.

**Visualizing and Analyzing Bias in Spanish** We take Spanish as an example and analyze gender bias in Spanish fastText word embeddings (Bojanowski et al., 2017) pre-trained on Spanish Wikipedia. For simplicity, we assume that the embeddings contain all gender forms of the words. To show bias, we randomly select several pairs of gender-definition and occupation words as well as inanimate nouns in Spanish and visualize them on the two gender directions defined above.

Figure 1 shows that the inanimate nouns lie near the origin point on the semantic gender direction, while the masculine and feminine forms of the occupation words are on the opposite sides for both directions. However, while projections of occupation words on the grammatical gender direction are symmetric with respect to the origin point, they are asymmetric when projected on the semantic gender direction. Along the semantic gender direction, female occupation words incline more to the feminine side than the male occupation words to the masculine side. This discrepancy shows that the embeddings carry unequal information for two genders.

**Quantification of Gender Bias** To quantify gender bias in English embeddings, Caliskan et al. (2017) propose Word Embedding Association Test (WEAT), which measures the association between two sets of target concepts and two sets of attributes. Let \( X \) and \( Y \) be equal-sized sets of target concept embeddings (e.g., words related to mathematics and art) and let \( A \) and \( B \) be sets of attribute embeddings (e.g., male and female definition words such as “man”, “girl”). Let \( \cos(\vec{a}, \vec{b}) \) denote the cosine similarity between vectors \( \vec{a} \) and \( \vec{b} \). The test statistic is a difference between sums over the respective target concepts,

\[
\mathcal{s}(X, Y, A, B) = \sum_{x \in X} s(\vec{x}, A, B) - \sum_{y \in Y} s(\vec{y}, A, B),
\]

where \( s(\vec{w}, A, B) \) is the difference between average cosine similarities of the respective attributes,

\[
s(\vec{w}, A, B) = \frac{1}{|A|} \sum_{\vec{a} \in A} \cos(\vec{w}, \vec{a}) - \frac{1}{|B|} \sum_{\vec{b} \in B} \cos(\vec{w}, \vec{b}).
\]

In the following, we extend the definition to quantify individual words in gendered languages.

We consider two types of words. One is *inanimate nouns*, like “agua” (water, feminine). They are assigned as either a masculine or feminine noun. The other is *animate nouns* that usually have two gender forms, like “doctor” (male doctor) and “doctora” (female doctor).

For inanimate nouns that should not be semantically leaning towards one gender, we can quantify the gender bias for word \( w \) simply using WEAT,

\[
b_w = |s(\vec{w}, A, B)|,
\]

which measures the association strength of the word \( w \) with the gender concepts. The larger \( b_w \) is, the stronger its association with the gender concepts.

For nouns with two gender forms, we test if their masculine (\( \vec{w}_m \)) and feminine (\( \vec{w}_f \)) forms are symmetrical with respect to gender definition terms.

\[
b_w = ||s(\vec{w}_m, A, B)| - |s(\vec{w}_f, A, B)||. \tag{1}
\]

The larger the value is, the larger gender bias presents. For example, if “doctor” (male doctor) leans more toward to masculine (i.e., close to masculine definition words like “el” (he) and far away

\[\footnote{Note that some animate nouns in certain languages, like “m´edecin” (male doctor) in French, only have one gender form. We omit such cases in this paper for the sake of simplicity.} \]

5277
Figure 2: Projections of occupation words on the semantic gender direction in two gender forms (blue for male and red for female) in Spanish, the original English word (grey), and the bias-mitigated version of the English word (black with “*”). We find that feminine occupations are farther away from the English counterparts as well as the neutral position (indicated by the vertical line) compared to masculine words.

from feminine definition words like “ella” (she)) than “doctora” leans toward feminine, then embeddings of “doctor” and “doctora” are biased. We will discuss detailed quantitative analysis using this metric in Section 4.

2.2 Bias in Bilingual Word Embeddings

Bilingual word embeddings are used widely in tasks such as multilingual transfer. It is essential to quantify and mitigate gender bias in these embeddings to avoid them from affecting downstream applications. We focus on bilingual word embeddings that align a grammatical gender language like Spanish with a language without grammatical gender like English. The definitions of the two gender directions are similar: we construct the grammatical gender direction using the same sets of words in the gendered language since the other language does not mark grammatical gender. We combine the gender-definition words for both languages and remove the grammatical gender component to get the semantic gender direction similarly.

We use bilingual word embeddings from MUSE (Conneau et al., 2018) that align English and Spanish fastText embeddings together in a single vector space. Figure 2 shows the projections of 5 occupation words on the semantic gender direction. First, we notice that the embeddings of some English occupations are much closer to one gender form than the other in Spanish. For example, “nurse” in English is closer to “enfermera” (female nurse) than “enfermero” (male nurse) in Spanish. We further separate out the bias in English embedding in the analysis by mitigating the English embeddings using “hard-debiasing” approach proposed in Bolukbasi et al. (2016) and aligning Spanish embedding with the bias-reduced English embedding (words annotated with “*”). Results in Figure 2 demonstrate another type of gender unequal in bilingual embedding: masculine occupation words are closer to the corresponding English occupation words than the feminine counterparts. This result aligns with the observation in Figure 1.

More quantitative analyses about gender bias in bilingual embedding will be discussed in Section 4.3.

3 Mitigation Methods

In the following, we first describe the optimization objectives for mitigating gender bias. Then, we present mitigation methods under two settings: (1) monolingual setting, where we only have access to word embeddings of the gendered language, and (2) bilingual setting, where the embeddings of a language without grammatical gender (e.g. English) aligned in the same space are also provided. Specifically, for the bilingual setting, we can use English word vectors to facilitate the bias mitigation.

3.1 Mitigation Objectives

For inanimate nouns, we project out their semantic gender component as they should not carry semantic gender. Equivalently, we minimize the inner product

\[ \langle \vec{w}, \vec{d}_s \rangle \] (2)

between the inanimate word vector \( \vec{w} \) and the semantic gender direction \( \vec{d}_s \).

For nouns with both gender forms, we aim at removing the gender inequality on the semantic gender direction as shown in Figure 1 and Figure 2. Specifically, we minimize the difference between the distance of feminine form of the word \( \vec{w}_f \) and the masculine form of the word \( \vec{w}_m \) to an anchor point (e.g., the origin point) on the semantic gender direction \( \vec{d}_s \):

\[ \| \langle \vec{w}_m, \vec{d}_s \rangle - \langle \vec{w}_f, \vec{d}_s \rangle \| - \| \langle \vec{w}_f, \vec{d}_s \rangle - \langle \vec{w}_a, \vec{d}_s \rangle \| \] (3)

Note that for most cases, \( \vec{w}_m \) and \( \vec{w}_f \) lie on the opposite sides of the anchor point on the semantic
gender axis. Therefore, we can simplify Eq. (3) as
\[
\left| \langle \vec{w}_m, \vec{d}_s \rangle + \langle \vec{w}_f, \vec{d}_s \rangle - 2 \langle \vec{w}_a, \vec{d}_s \rangle \right|. \tag{4}
\]
Intuitively, this measures how much we need to move the word pair on the semantic gender direction so that they are symmetric with respect to the anchor point. If the anchor point is the origin point (i.e., \( \langle \vec{w}_a, \vec{d}_s \rangle = 0 \)), this objective is the absolute value of the sum of two projections.

3.2 Monolingual Setting

**Shifting Along Semantic Gender Direction (Shift)** We mitigate bias by minimizing Eq. (4) and Eq. (2) in a post-processing manner. Effectively, for inanimate nouns, we remove the semantic gender component from the embeddings. For nouns with two gender forms (e.g., occupations), we shift the two forms along the semantic gender direction so that they have the same distance to the anchor point. For monolingual setting, we use the origin point on the gender direction as the anchor point and call the method Shift\_Ori.

Note that Gonen and Goldberg (2019) show that mitigating gender bias in word embeddings by moving on the gender direction is not sufficient and words with gender bias still tend to group together. Similarly, our approach might not entirely remove the gender bias. However, the approach in Gonen and Goldberg (2019) to test the remainder bias cannot be applied directly for gendered languages, because grouping of the embeddings of masculine and feminine words does not always indicate bias due to grammatical gender.

3.3 Bilingual Setting

**Mitigating Before Alignment (De-Align)** As gender bias can be observed when aligning gendered languages with language without grammatical gender, we consider using English to facilitate mitigating bias. Specifically, we first apply the “hard-debasing” approach Bolukbasi et al. (2016) to English embedding and then align gendered language with the bias-reduced English. Because most words with both gender forms align to the same word in English (e.g., “enfermero” and “enfermera” in Spanish both align to “nurse” in English), the alignment places the two gender forms of the words in a more symmetric positions in the vector space after the alignment.

**Shifting Along Semantic Gender Direction (Shift)** We consider using English words as anchor positions when mitigating bias using Eq (4). We call this approach Shift\_EN. For example, when applying Eq. (4) to re-position “enfermero” and “enfermera” in Spanish we take English word “nurse” aligned in the same space as the anchor position.

**Hybrid Method (Hybrid)** We also consider a hybrid method that integrates the aforementioned two approaches. In particular, we first mitigate English embeddings, align English embeddings with the embeddings of gendered languages, and then shift words in languages with grammatical gender along the semantic gender direction. We consider two variants based on how the anchor positions are chosen: (1) Hybrid\_Ori uses the origin point as the anchor position; and (2) Hybrid\_EN uses the corresponding bias-reduced English word as an anchor position when shifting a pair of words with different gender forms.

4 Experiments

We evaluate gender bias mitigation methods in both monolingual and bilingual settings and test the utility of the bias-reduced embeddings in word-level translation tasks Conneau et al. (2018). Following other analyses on gender bias (Bolukbasi et al., 2016; Caliskan et al., 2017), we quantify the gender bias based on a set of animate nouns including occupations and gender-definition words. These words are representative to the major challenge when quantifying and mitigating gender bias. Different from previous analyses in English, most occupation words in gendered languages have more than one forms and their embeddings are inherently different from each other due to morphological agreement.

4.1 Data and Configuration

We apply the proposed mitigation methods on Spanish-English and French-English bilingual embeddings from MUSE (Conneau et al., 2018) that align the fastText monolingual embeddings pretrained on Wikipedia corpora together in a single vector space. We collect 58 occupation words in Spanish from the occupation list provided by Font and Costa-jussà (2019) and 23 in French (some occupations in French do not distinguish between two forms) for the bias mitigation experiments. We use WEAT word lists from Caliskan et al. (2017) for
Table 1: Analyses on Spanish and French monolingual embeddings before and after bias mitigation. Results show that the original Spanish and French embeddings exhibit strong bias and Hybrid_Ori significantly reduces the bias in the embedding to an insignificant level (p-value > 0.05).
### Table 2: Results on word translation and word pair translation based on Spanish-English and French-English bilingual embeddings. We find that the original bilingual embeddings have a large discrepancy between the two genders, indicated by the average cosine similarity difference (ASD) and mean reciprocal ranks (MRR) difference. After applying the mitigation methods, both ASD and MRR difference drop.

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>Original</th>
<th>Shift_Ori</th>
<th>Shift_EN</th>
<th>De-Align</th>
<th>Hybrid_Ori</th>
<th>Hybrid_EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN→ES</td>
<td>P@1/P@5</td>
<td>79.2/89.0</td>
<td>80.7/90.3</td>
<td>80.7/90.3</td>
<td>76.5/88.9</td>
<td>80.7/90.3</td>
<td>80.7/90.3</td>
</tr>
<tr>
<td>ES→EN</td>
<td>P@1/P@5</td>
<td>79.2/89.0</td>
<td>79.2/89.0</td>
<td>79.2/89.0</td>
<td>80.1/90.7</td>
<td>79.2/89.0</td>
<td>79.2/89.0</td>
</tr>
<tr>
<td>EN→FR</td>
<td>P@1/P@5</td>
<td>78.2/89.4</td>
<td>79.9/91.1</td>
<td>79.9/91.1</td>
<td>74.3/87.8</td>
<td>79.9/91.1</td>
<td>79.9/91.1</td>
</tr>
<tr>
<td>FR→EN</td>
<td>P@1/P@5</td>
<td>76.1/88.1</td>
<td>76.1/88.1</td>
<td>76.1/88.1</td>
<td>74.4/87.2</td>
<td>76.1/88.1</td>
<td>76.1/88.1</td>
</tr>
</tbody>
</table>

#### Word Translation

<table>
<thead>
<tr>
<th>Task</th>
<th>ASD</th>
<th>F_MRR</th>
<th>M_MRR</th>
<th>MRR Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN→ES</td>
<td>0.1082</td>
<td>0.0961</td>
<td>0.0961</td>
<td>0.0827</td>
</tr>
<tr>
<td></td>
<td>0.0755</td>
<td>0.0772</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2073</td>
<td>0.2507</td>
<td>0.2507</td>
<td>0.2919</td>
</tr>
<tr>
<td></td>
<td>0.3450</td>
<td>0.3150</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6940</td>
<td>0.6766</td>
<td>0.6766</td>
<td>0.6775</td>
</tr>
<tr>
<td></td>
<td>0.6398</td>
<td>0.6696</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4867</td>
<td>0.4259</td>
<td>0.4259</td>
<td>0.3856</td>
</tr>
<tr>
<td></td>
<td>0.2949</td>
<td>0.3546</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN→FR</td>
<td>0.1208</td>
<td>0.1048</td>
<td>0.1082</td>
<td>0.0892</td>
</tr>
<tr>
<td></td>
<td>0.0735</td>
<td>0.0805</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1663</td>
<td>0.2101</td>
<td>0.1943</td>
<td>0.2679</td>
</tr>
<tr>
<td></td>
<td>0.3128</td>
<td>0.2975</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.6549</td>
<td>0.6313</td>
<td>0.6419</td>
<td>0.6610</td>
</tr>
<tr>
<td></td>
<td>0.6393</td>
<td>0.6467</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4886</td>
<td>0.4212</td>
<td>0.4476</td>
<td>0.3931</td>
</tr>
<tr>
<td></td>
<td>0.3265</td>
<td>0.3492</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Word Pair Translation

Results Table 2 shows the results. Results on word translation shows that the mitigation approaches do not harm the utility of bilingual word embedding and words in different languages still align well. From the results of word pair translation, the original bilingual embedding exhibits strong bias and the gap between two genders are around 0.49 in MRR in both ES-EN and FR-EN bilingual embeddings and the average cosine similarity difference is larger than 0.1. Hybrid_Ori results in smallest difference for cosine similarity and MRR gap between two gender forms. However, the discrepancy between two forms is not completely removed.

To visualize the effect of bias mitigation, we
project the bias-reduced embeddings of occupation words on the semantic gender direction as shown in Figure 2. Figure 3 shows that after debiasing using Hybrid_Ori, the two gender forms of occupation words are more symmetric with respect to the corresponding English embedding.

4.4 Bias Correlation between Languages

Finally, we compute the WEAT/MWEAT scores\(^3\) for each occupation and show the commonality and difference of gender bias between English and Spanish embedding. Results in Figure 4 shows that occupations (“nurse” and “mathematician”) lean towards the same gender in both languages. However, some occupations (e.g., “chef”) lean to different genders in different languages. In general, there is a strong correlation between gender bias in English and Spanish embedding (Spearman correlation=0.45 with p-value=0.0004).

5 Related Work

Previous work has proposed to quantify bias in English Embedding definitions for gender bias in English word embeddings. Besides the aforementioned studies (Bolukbasi et al., 2016; Caliskan et al., 2017) in quantifying bias in English embedding, McCurdy and Serbeti (2017) examine grammatical gender bias in word embeddings of gendered languages by computing the WEAT association score (Caliskan et al., 2017) between gendered object nouns (e.g., moon-sun) and gender-definition words. They propose to mitigate bias by applying lemmatization to remove gender information from the training corpus. However, their focus is different from us as our approaches aim at keeping the grammatical gender information and only removing the bias in semantic genders. A few recent studies focus on measuring and reducing gender bias in contextualized word embeddings (Zhao et al., 2019; May et al., 2019; Basta et al., 2019). However, they only focus on English embeddings in which the gender is mostly only expressed by pronouns (Stahlberg et al., 2007). An interesting future direction is to extend their analyses to language with grammatical gender.

Regarding bias mitigation approaches, Zhao et al. (2018b) mitigate bias by saving one dimension of the word vector for gender. Bordia and Bowman (2019) propose 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 only English 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.

Bias in word embedding may affect the downstream applications (Zhao et al., 2018a; Rudinger et al., 2018; 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).

6 Conclusion and Future Work

We analyze gender bias in the embeddings of languages with grammatical gender and bilingual word embeddings that align such a language with a language without grammatical gender. We propose new methods to evaluate and mitigate gender bias for languages with grammatical gender and bilingual word embeddings and results show that our methods can mitigate gender bias while preserving the quality of original embeddings.

Directions for future work include testing monolingual and bilingual word embeddings on downstream tasks like machine translation to measure bias and also test the performance for mitigation methods. Moreover, the number of noun classes for other languages with grammatical gender ranges from two to several tens (Corbett, 1991) and future work can extend methods proposed in this paper to address grammatical gender in languages with more gender forms.

7 Acknowledgements

This work was supported in part by DARPA (HR0011-18-9-0019). Ryan Cotterell was supported by a Facebook Fellowship. We acknowledge fruitful discussions with Fred Morstatter, Nanyun Peng, Ninareh Mehrab, Shunjie Wang, Yizhou Yao, and Aram Galstyan. We also thank anonymous reviewers for their comments.

---

\(^3\)We use WEAT score for English and MWEAT for Spanish. To show the direction of bias, we take out the outer absolute function in Eq. (1).
References


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


