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

Pei Zhou [1] **Weijia Shi** [1] **Jieyu Zhao** [1] **Kuan-Hao Huang** [1] **Muhao Chen** [1] **Kai-Wei Chang** [1]

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) 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 & 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 , 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, 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 embeddings.

To address this unresolved yet critical issue, we propose analysis methods for bias in Spanish word embeddings and English-Spanish bilingual word embeddings. We first quantify gender bias in Spanish by constructing two gender directions:[1] the semantic gender direction and the grammatical gender direction. After projecting occupation words with two gender forms on the directions, we find that masculine and feminine forms are mostly in a symmetric position on the grammatical gender direction but are asymmetric on the semantic gender direction.

Then, we propose two types of methods to mitigate gender bias in Spanish embeddings and EN-ES bilingual word embeddings: (1) mitigating English first and then align the embedding space and (2) shifting along the semantic gender direction of Spanish word embeddings directly. Results show that the combination of the two approaches is able to effectively mitigate bias in Spanish word embeddings as well as EN-ES bilingual word embeddings.

2. Related Work

Previous work has proposed several different approaches to define and mitigate gender bias in English word embeddings. Bolukbasi et al. (2016) define bias in English embeddings

[1]Department of Computer Science, University of California, Los Angeles. Correspondence to: Kai-Wei Chang <kwchang@cs.ucla.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

[1]In this paper, we follow the literature and address only binary gender.

being that one word that is not gender-specific shows different inclinations of genders. They define a gender direction using the difference between male- and female-definition word embeddings and show that occupational words have different distance to "male" or "female" in this direction. This is appropriate for English as it does not distinguish between the masculine and feminine forms for most nouns. However, in Spanish, all nouns have grammatical gender including those inanimate objects such as "table" and "apple". The gender information in such words does not necessarily indicate the words to be biased towards male or female.

As for mitigation methods for gender bias in English, Zhao et al. (2018b) mitigate bias by saving one dimension of the word vector for gender. Bordia & Bowman (2019) proposes a regularization loss term for word-level language models. Zhang et al. (2018) uses an adversarial network to mitigate bias in word embeddings. All these approaches adopt definition for bias from Bolukbasi et al. (2016), so they still cannot be applied to Spanish and bilingual word embeddings easily. Moreover, Gonen & Goldberg (2019) show that mitigation methods based on gender directions are not sufficient, since socially-biased words still cluster together in high dimensional space.

McCurdy & Serbeti (2017) examine grammatical gender in word embeddings by computing the WEAT association score (Caliskan et al., 2017) between gendered object nouns and the corresponding gender attribute words and find that the association is larger than topical gender bias shown in Caliskan et al. (2017). They also mitigate bias by lemmatizing to remove gender information in corpora. However, we argue that the large association between gendered objects nouns with gender attributes should not be considered as gender bias since the association could be caused by the grammatical form instead of stereotypes. Mitigation by removing gender information is also implausible as too much information will be lost.

3. Gender Bias Analysis and Mitigation

3.1. Bias in Spanish Embeddings

Gender Directions in Spanish In Spanish word embeddings, we propose a new way to define gender directions to evaluate the bias. Specifically, we define two gender directions, one is for *grammatical gender*, which is used to capture the inherently carried gender value of the word and the other for *semantic gender*, which is used to measure the semantically male or female inclination of this word. We claim that the semantic gender direction is enough for English since it does not have grammatical gender. But for Spanish, grammatical gender leads to another type of gender information besides semantic gender so we need two directions. We also constrain that the two directions

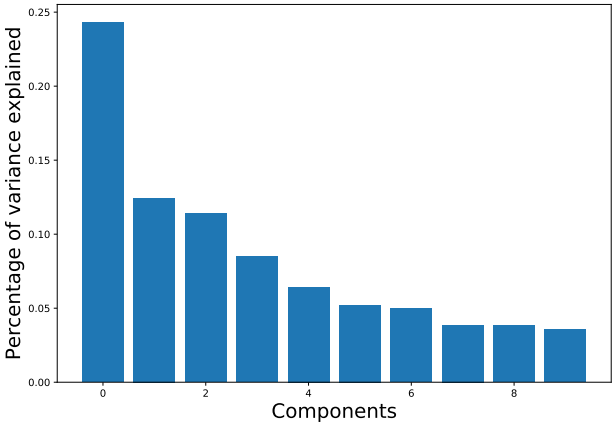

Figure 1. Percentage of variance explained in PCA of vector differences for gender-definition pairs when constructing *semantic gender* direction.

are orthogonal to each other to better distinguish the two types of gender information. For all nouns in Spanish, we do not take the different inclinations along the grammatical gender direction as bias and only focus on the bias shown in the semantic gender. We define that there is gender bias in Spanish if two forms of the same occupation word are far from symmetric on the gender direction with respect to an anchor point. The anchor point represents the gender-neural position along the gender directions.

Grammatical Gender Since all nouns in Spanish are either grammatically masculine or feminine, we cannot follow the previous approach to collect pairs of words and directly capture the grammatical gender using principal component analysis (PCA) (Jolliffe, 2011). We instead collect two sets of words in Spanish that only have one common gender form for masculine and feminine nouns. We get the centroid of the male and female clusters and use the difference between these two as the *grammatical gender* direction.

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 10 pairs of gender-definition nouns in Spanish. Figure 1 shows that there is one main principal component of gender direction in Spanish. To better distinguish between *grammatical gender* and *semantic gender*, we remove the *grammatical gender* component in the computed gender direction to make the semantic gender direction orthogonal to the grammatical gender direction. Along this direction, masculine and feminine forms of one occupation word should have similar distances to male concepts and female concepts respectively, i.e., symmetric with respect to the gender-neutral position, otherwise they shows bias.

Visualizing and Analyzing Bias in Spanish We use Spanish fastText (Bojanowski et al., 2017) embeddings pre-

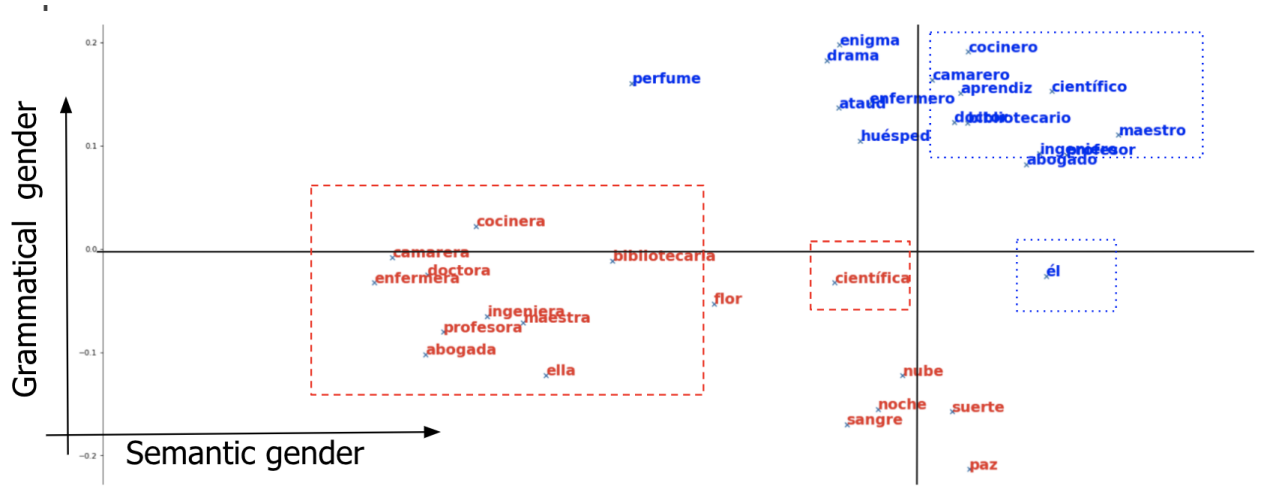

Figure 2. Projections of selected occupation words (enclosed in dotted lines) and common nouns in Spanish word embeddings on grammatical and semantic directions with masculine nouns in blue and feminine nouns in red.

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. Figure 2 shows that the Spanish word embeddings are biased. 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, and the rest of the words are common nouns that are not used to describe people. 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, masculine) and "flor" (flower, 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.

For occupation words, the masculine and feminine forms are on the opposite sides for both directions. However, while their projections on 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 compared to masculine forms on 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.

3.2. Mitigation Methods

Mitigating English Before Alignment Although mitigation method for Spanish word embeddings is 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 and we also get a less biased bilingual word embeddings.

Shifting Along Semantic Gender Direction The second method mitigates bias as post-processing and extends the "hard-debiasing" approach from 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 direction and the projection of the mitigated English word using "hard-debiasing" approach in the bilingual word

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

Table 1. Results for different debiasing methods on two types of evaluation metrics. "CLAT" stands for cross-lingual analogy task, "WEAT–Male" is the association of male occupation words with male-definition terms subtracting that with female-definition, similarily for "WEAT–Female", "Shift (Ori)" is the debiasing method of shifting along the semantic gender direction with the zero point as anchors, similarily 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.

embeddings. Although Gonen & 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 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 bilingual 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) and the corresponding Spanish word, the task is to predict the missing Spanish occupation word. Based on this, we can evaluate how differently the masculine and feminine occupation words perform in this task. We will calculate the ranking difference between two versions 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 measuring 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, where each addend is the difference between mean cosine similarities of the respective attributes, $s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})$. (May et al., 2019)

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 associate 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 gender direction (Shift) is better at reducing the discrepancy in the WEAT association for two gender forms. Overall the results suggest that a combination of mitigating English gender bias before alignment and post-processing (De-Shift (Ori)) can effectively mitigate the gender bias in ES or bilingual word embeddings according to the two tasks we consider.

5. Conclusion

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.

References

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

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

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

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

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

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

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

Font, J. E. and Costa-jussà, M. R. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116*, 2019.

Gonen, H. and Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

Jolliffe, I. *Principal component analysis*. Springer, 2011.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*, 2019.

McCurdy, K. and Serbeti, O. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *WiNLP*, 2017.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340. ACM, 2018.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018a.

Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*, 2018b.