Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer

Jieyu Zhao[§] * **Subhabrata Mukherjee**[‡] **Saghar Hosseini**[‡]
Kai-Wei Chang[§] **Ahmed Hassan Awadallah**[‡]
[§]University of California, Los Angeles [‡]Microsoft Research AI
{jyzhao, kwchang}@cs.ucla.edu
{Subhabrata.Mukherjee, Saghar.Hosseini, hassanam}@microsoft.com

Abstract

Multilingual representations embed words from many languages into a single semantic space such that words with similar meanings are close to each other regardless of the language. These embeddings have been widely used in various settings, such as cross-lingual transfer, where a natural language processing (NLP) model trained on one language is deployed to another language. While the cross-lingual transfer techniques are powerful, they carry gender bias from the source to target languages. In this paper, we study gender bias in multilingual embeddings and how it affects transfer learning for NLP applications. We create a multilingual dataset for bias analysis and propose several ways for quantifying bias in multilingual representations from both the intrinsic and extrinsic perspectives. Experimental results show that the magnitude of bias in the multilingual representations changes differently when we align the embeddings to different target spaces and that the alignment direction can also have an influence on the bias in transfer learning. We further provide recommendations for using the multilingual word representations for downstream tasks.

1 Introduction

Natural Language Processing (NLP) plays a vital role in applications used in our daily lives. Despite the great performance inspired by the advanced machine learning techniques and large available datasets, there are potential societal biases embedded in these NLP tasks – where the systems learn inappropriate correlations between the final predictions and sensitive attributes such as gender and race. For example, Zhao et al. (2018a) and Rudinger et al. (2018) demonstrate that coreference resolution systems perform unequally on different gender groups. Other studies show that such bias is exhibited in various components of the NLP systems, such as the training dataset (Zhao et al., 2018a; Rudinger et al., 2018), the embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhou et al., 2019; Manzini et al., 2019) as well as the pre-trained models (Zhao et al., 2019; Kurita et al., 2019).

Recent advances in NLP require large amounts of training data. Such data may be available for resource-rich languages such as English, but they are typically absent for many other languages. Multilingual word embeddings align the embeddings from various languages to the same shared embedding space which enables transfer learning by training the model in one language and adopting it for another one (Ammar et al., 2016; Ahmad et al., 2019b; Meng et al., 2019; Chen et al., 2019). Previous work has proposed different methods to create multilingual word embeddings. One common way is to first train the monolingual word embeddings separately and then align them to the same space (Conneau et al., 2017; Joulin et al., 2018). While multiple efforts have focused on improving the models' performance on low-resource languages, less attention is given to understanding the bias in cross-lingual transfer learning settings.

In this work, we aim to understand the bias in multilingual word embeddings. In contrast to existing literature that mostly focuses on English, we conduct analyses in multilingual settings. We argue that the bias in multilingual word embeddings can be very different from that in English. One reason is that each language has its own properties. For example, in English, most nouns do not have grammatical gender, while in Spanish, all nouns do. Second, when we do the alignment to get the multilingual word embeddings, the choice of target space may cause bias. Third, when we do transfer learning based on multilingual word

*Most of the work was done while the first author was an intern at Microsoft Research.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2896–2907

embeddings, the alignment methods, as well as the transfer procedure can potentially influence the bias in downstream tasks. Our experiments confirm that bias exists in the multilingual embeddings and such bias also impacts the cross-lingual transfer learning tasks. We observe that the transfer model based on the multilingual word embeddings shows discrimination against genders. To discern such bias, we perform analysis from both the corpus and the embedding perspectives, showing that both contribute to the bias in transfer learning. Our contributions are summarized as follows:

- We build datasets for studying the gender bias in multilingual NLP systems.[1]
- We analyze gender bias in multilingual word embeddings from both intrinsic and extrinsic perspectives. Experimental results show that the pre-trained monolingual word embeddings, the alignment method as well as the transfer learning can have an impact on the gender bias.
- We show that simple mitigation methods can help to reduce the bias in multilingual word embeddings and discuss directions for future work to further study the problem. We provide several recommendations for bias mitigation in cross-lingual transfer learning.

2 Related Work

Gender Bias in Word Representations Word embeddings are widely used in different NLP applications. They represent words using low dimensional vectors. Bolukbasi et al. (2016) find that, in the embedding space, occupation words such as "professor" and "nurse" show discrepancy concerning the genders. Similarly, Caliskan et al. (2017) also reveal the gender stereotypes in the English word embeddings based on the Word Embedding Association Test (WEAT). However, both works only consider English and cannot be directly adapted to other languages such as Spanish. McCurdy and Serbetci (2017) reveal that bias exists in languages with grammatical gender while Zhou et al. (2019) and Lauscher and Glavaš (2019) show that there is bias in bilingual word embeddings. However, none of them consider the cross-lingual transfer learning which is an important application of the multilingual word embeddings. To mitigate the bias in word embeddings, various approaches have been proposed (Bolukbasi et al., 2016; Zhao et al., 2018b). In contrast to these methods in English embedding space, we propose to mitigate the bias from the multilingual perspectives. Comparing to Zhou et al. (2019), we show that a different choice of alignment target can help to reduce the bias in multilingual embeddings from both intrinsic and extrinsic perspectives.

Multilingual Word Embeddings and Cross-lingual Transfer Learning Multilingual word embeddings represent words from different languages using the same embedding space which enables cross-lingual transfer learning (Ruder et al., 2019). The model is trained on a labeled data rich language and adopted to another language where no or a small portion of labeled data is available (Duong et al., 2015; Guo et al., 2016). To get the multilingual word embeddings, Mikolov et al. (2013) learn a linear mapping between the source and target language. However, Xing et al. (2015) argue that there are some inconsistencies in directly learning the linear mapping. To solve those limitations, they constrain the embeddings to be normalized and enforce an orthogonal transformation. While those methods achieve reasonable results on benchmark datasets, they all suffer from the hubness problem which is solved by adding cross-domain similarity constraints (Conneau et al., 2017; Joulin et al., 2018). Our work is based on the multilingual word embeddings achieved by Joulin et al. (2018). Besides the commonly used multilingual word embeddings obtained by aligning all the embeddings to the English space, we also analyze the embeddings aligned to different target spaces.

Bias in Other Applications Besides the bias in word embeddings, such issues have also been demonstrated in other applications, including named entity recognition (Mehrabi et al., 2019), sentiment analysis (Kiritchenko and Mohammad, 2018), and natural language inferences (Rudinger et al., 2017). However, those analyses are limited to English corpus and lack the insight of multilingual situations.

3 Intrinsic Bias Quantification and Mitigation

In this section, we analyze the gender bias in multilingual word embeddings. Due to the limitations of the available resources in other languages, we analyze the bias in English, Spanish, German and

[1] Code and data will be available at https://aka.ms/MultilingualBias.

French. However, our systematic evaluation approach can be easily extended to other languages. We first define an evaluation metric for quantifying gender bias in multilingual word embeddings. Note that in this work, we focus on analyzing gender bias from the perspective of occupations. We then show that when we change the target alignment space, the bias in multilingual word embeddings also changes. Such observations provide us a way to mitigate the bias in multilingual word embeddings – by choosing an appropriate target alignment space.

3.1 Quantifying Bias in Multilingual Embeddings

We begin with describing inBias, our proposed evaluation metric for quantifying intrinsic bias in multilingual word embeddings from word-level perspective. We then introduce the dataset we collected for quantifying bias in different languages.

Bias Definition Given a set of masculine and feminine words, we define inBias as:

$$\text{inBias} = \frac{1}{N}\sum_{i=1}^{N} |dis(O_{M_i}, S_M) - dis(O_{F_i}, S_F)|, \tag{1}$$

where

$$dis(O_{G_i}, S) = \frac{1}{|S|}\sum_{s \in S}(1 - \cos(O_{G_i}, s)).$$

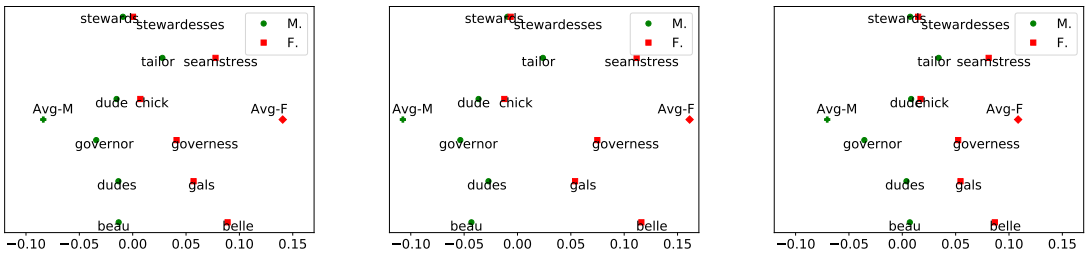
Here (O_{M_i}, O_{F_i}) stands for the masculine and feminine format of the i-th occupation word, such as ("doctor", "doctora"). S_M and S_F are a set of gender seed words that contain male and female gender information in the definitions such as "he" or "she".

Intuitively, given a pair of masculine and feminine words describing an occupation, such as the words "*doctor*" (Spanish, masculine doctor) and "*doctora*" (Spanish, feminine doctor), the only difference lies in the gender information. As a result, they should have similar correlations to the corresponding gender seed words such as "*él*" (Spanish, he) and "*ella*" (Spanish, she). If there is a gap between the distance of occupations and corresponding gender, (i.e., the distance between "doctor" and "él" against the distance between "doctora" and "ella"), it means such occupation shows discrimination against gender. Note that such metric can also be generalized to other languages without grammatical gender, such as English, by just using the same format of the occupation words. It is also worth noting that our metric is general and can be used to define other types of bias with slight modifications. For example, it can be used to detect age or race bias by providing corresponding seed words (e.g., "young" - "old" or names correlated with different races). In this paper we focus on gender bias as the focus of study. We provide detailed descriptions of those words in the dataset collection subsection.

Unlike previous work (Bolukbasi et al., 2016) which requires calculating a gender direction by doing dimensionality reduction, we do not require such a step and hence we can keep all the information in the embeddings. The goal of inBias is aligned to that of WEAT (Caliskan et al., 2017). It calculates the difference of targets (occupations in our case) corresponding to different attributes (gender). We use paired occupations in each language, reducing the influence of grammatical gender. Compared to Zhou et al. (2019), we do not need to separately generate the two gender directions, as in our definition, the difference of the distance already contains such information. In addition, we no longer need to collect the gender neutral word list. In multilingual settings, due to different gender assignments to each word (e.g., "spoon" is masculine is DE but feminine in ES), it is expensive to collect such resources which can be alleviated by the inBias metric.

Multilingual Intrinsic Bias Dataset To conduct the intrinsic bias analysis, we create the MIBs dataset by manually collecting pairs of occupation words and gender seed words in four languages: English (EN), Spanish (ES), German (DE) and French (FR). We choose these four languages as they come from different language families (EN and DE belong to the Germanic language family while ES and FR belong to the Italic language family) and exhibit different gender properties (e.g., in ES, FR and DE, there is grammatical gender).[2] We refer to languages with grammatical gender as GENDER-RICH languages; and otherwise, as GENDER-LESS languages. Among these three gender-rich languages, ES and FR only have feminine and masculine genders while in DE, there is also a neutral gender. We obtain the feminine and masculine words in EN from Zhao et al. (2018b) and extend them by manually adding other common occupations. The English gender seed words are from Bolukbasi et al.

[2] We also do analyses with Turkish where there is no grammatical gender and no gendered pronoun. Details are in Sec. 3.2.4.



(a) Original es embeddings. (b) In es-en embeddings. (c) In es-de embeddings.

Figure 1: Most biased occupations in ES projected to the gender subspace defined by the difference between two gendered seed words. Green dots are masculine (M.) occupations while the red squares are feminine (F.) ones. We also show the average projections of the gender seed words for male and female genders denoted by "Avg-M" and "Avg-F". Compared to EN, aligning to DE makes the distance between the occupation word and corresponding gender more symmetric.

(2016). For all the other languages, we get the corresponding masculine and feminine terms by using online translation systems, such as Google Translate. We refer to the words that have both masculine and feminine formats in EN (e.g., "waiter" and "waitress") as *strong gendered* words while others like "doctor" or "teacher" as *weak gendered* words. In total, there are 257 pairs of occupations and 10 pairs of gender seed words for each language. In the gender-rich languages, if the occupation only has one lexical format, (e.g., "prosecutor" in ES only has the format "fiscal"), we add it to both the feminine and the masculine lists.

3.2 Characterizing Bias in Multilingual Embeddings

As mentioned in Sec. 1, multilingual word embeddings can be generated by first training word embeddings for different languages individually and then aligning those embeddings to the same space. During the alignment, one language is chosen as target and the embeddings from other languages are projected onto this target space. We conduct comprehensive analyses on the MIBs dataset to understand: 1) how gender bias exhibits in embeddings of different languages; 2) how the alignment target affects the gender bias in the embedding space; and 3) how the quality of multilingual embeddings is affected by choice of the target language.

For the monolingual embeddings of individual languages and the multilingual embeddings that used English as the target language (*-en),[3] we use the publicly available fastText embeddings trained on 294 languages in Wikipedia (Bojanowski et al., 2017; Joulin et al., 2018). For all other embeddings aligned to a target space other than EN, we adopt the RCSLS alignment model (Joulin et al., 2018) based on the same hyperparameter setting (details are in Appendix).

[3]We refer to the aligned multilingual word embeddings using the format src-tgt. For example, "es-en" means we align the ES embeddings to the EN space. An embedding not following such format refers to a monolingual embedding.

Source	Target			
	EN	ES	DE	FR
EN	**0.0830**	0.0639*	0.0699*	0.0628*
ES	0.0889*	**0.0803**	0.0634*	0.0642*
DE	0.1124	0.0716*	**0.1079**	0.0805*
FR	0.1027	0.0768*	0.0782*	**0.0940**

Table 1: inBias score before and after alignment to different target spaces. Rows stands for the source languages while columns are the target languages. The diagonal values stand for the bias in the original monolingual word embeddings. Here * indicates the difference between the bias before and after alignment is statistically significant ($p < 0.05$).

3.2.1 Analyzing Bias before Alignment

We examine the bias using four languages mentioned previously based on all the word pairs in the MIBs. Table 1 reports the inBias score on this dataset. The diagonal values here stand for the bias in each language before alignment. Bias commonly exists across all the four languages. Such results are also supported by WEAT in Zhou et al. (2019), demonstrating the validity of our metric. What is more, comparing those four languages, we find DE and FR have stronger biases comparing to EN and ES.

Source	Target			
	EN	ES	DE	FR
EN	-	83.08	78.60	83.00
ES	86.40	-	72.40	87.27
DE	76.33	69.80	-	78.13
FR	84.27	84.80	75.53	-

Table 2: Performance (accuracy %) of the BLI task for the aligned embeddings. Row stands for the source language and column is the target language. The values in the first row are from Joulin et al. (2018).

3.2.2 How will the bias change when aligned to different languages?

Commonly used multilingual word embeddings align all languages to the English space. However, our analysis shows that the bias in the multilingual word embeddings can change if we choose a different target space. All the results are shown in Table 1. Specifically, when we align the embeddings to the gender-rich languages, the bias score will be lower compared to that in the original embedding space. In the other situation, when aligning the embeddings to the gender-less language space (i.e., EN in our case), the bias increases. For example, in original EN, the bias score is 0.0830 and when we align EN to ES, the bias decreases to 0.0639 with 23% reduction in the bias score. However, the bias in ES embeddings increases to 0.0889 when aligned to EN while only 0.0634 when aligned to DE.[4] In Fig. 1, we show the examples of word shifting along the gender direction when aligning ES to different languages. The gender direction is calculated by the difference of male gendered seeds and female gendered seeds. We observe the feminine occupations are further away from female seed words than masculine ones, causing the resultant bias. In comparison to using EN as target space, when aligning ES to DE, the distance between masculine and feminine occupations with corresponding gender seed words become more symmetric, therefore reducing the inBias score.

What words changed most after the alignment? We are interested in understanding how the gender bias of words changes after we do the alignment. To do this, we look at the top-15 most and least changed words. We find that in each language, the strongest bias comes from the strong gendered words; while the least bias happens among weak gendered words. When we align EN embeddings to gender-rich languages, bias in the strong gendered words will change most significantly; and the weak gendered words will change least significantly. When we align gender-rich languages to EN, we observe a similar trend. Among all the alignment cases, gender seed words used in Eq. (1) do not change significantly.

[4]We show the bias for all the 257 pairs of words in EN. In the appendix, we also show the bias for strong gendered words and weak gendered words separately.

3.2.3 Bilingual Lexicon Induction

To evaluate the quality of word embeddings after the alignment, we test them on the bilingual lexicon induction (BLI) task (Conneau et al., 2017) goal of which is to induce the translation of source words by looking at their nearest neighbors. We evaluate the embeddings on the MUSE dataset with the CSLS metric (Conneau et al., 2017).

We conduct experiments among all the pair-wise alignments of the four languages. The results are shown in Table 2. Each row depicts the source language, while the column depicts the target language. When aligning languages to different target spaces, we do not observe a significant performance difference in comparison to aligning to EN in most cases. This confirms the possibility to use such embeddings in downstream tasks. However, due to the limitations of available resources, we only show the result on the four languages and it may change when using different languages.

3.2.4 Languages of Study

In this paper, we mainly focus on four European languages from different language families, partly caused by the limitations of the currently available resources. We do a simplified analysis on Turkish (TR) which belongs to the Turkic language family. In TR, there is no grammatical gender for both nouns and pronouns, i.e., it uses the same pronoun "o" to refer to "he", "she" or "it". The original bias in TR is 0.0719 and when we align it to EN, the bias remains almost the same at 0.0712. When aligning EN to TR, we can reduce the intrinsic bias in EN from 0.0830 to 0.0592, with 28.7% reduction. However, the BLI task shows that the performance on such aligned embeddings drops significantly: only 53.07% when aligned to TR but around 80% when aligned to the other four languages. Moreover, as mentioned in Ahmad et al. (2019a), some other languages such as Chinese and Japanese cannot align well to English. Such situations require more investigations and forming a direction for future work.

Source	Target			
	ENDEB	ES	DE	FR
ENDEB	0.0501*	0.0458*	0.0524*	0.0441*
ES	0.0665*	0.0803	-	-
DE	0.0876*	-	0.1079	-
FR	0.0905	-	-	0.0940

Table 3: inBias score before and after alignment to ENDEB. * indicates statistically significant difference between the bias in original and aligned embeddings.

3.3 Bias after Mitigation

Researchers have proposed different approaches to mitigate the bias in EN word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018b). Although these approaches cannot entirely remove the bias (Gonen and Goldberg, 2019), they significantly reduce the bias in English embeddings. We refer to such embedding as *ENDEB*. We analyze how the bias changes after we align the embeddings to such ENDEB space. The ENDEB embeddings are obtained by adopting the method in Bolukbasi et al. (2016) on the original fastText monolingual word embeddings. Table 3 and 4 show the bias score and BLI performance when we do the alignment between ENDEB and other languages. Similar to Zhou et al. (2019), we find that when we align other embeddings to the ENDEB space, we can reduce the bias in those embeddings. What is more, we show that we can reduce the bias in ENDEB embeddings further when we align it to a gender-rich language such as ES while keeping the functionality of the embeddings, which is consistent with our previous observation in Table 1. Besides, comparing aligning to gender-rich languages and to ENDEB, the former one can reduce the bias more.

4 Extrinsic Bias Quantification and Mitigation

In addition to the intrinsic bias in multilingual word embeddings, we also analyze the downstream tasks, specifically in the cross-lingual transfer learning. One of the main challenges here is the absence of appropriate datasets. To motivate further research in this direction, we build a new dataset called MLBs. Experiments demonstrate that bias in multilingual word embeddings can also have an effect on models transferred to different languages. We further show how mitigation methods can help to reduce the bias in the transfer learning setting.

Source	Target			
	ENDEB	ES	DE	FR
ENDEB	-	84.07	79.13	83.27
Target	**Source**			
	ENDEB	ES	DE	FR
ENDEB	-	86.07	76.27	84.33

Table 4: Performance (accuracy %) on the BLI task using the aligned embeddings based on ENDEB embeddings. The top one is the result of aligning ENDEB to other languages while the bottom is to align other languages to ENDEB.

Language	EN	ES	DE	FR
#occupation	28	72	27	27
#instance	397,907	82,863	12,976	59,490

Table 5: Statistics of the MLBs for each language.

4.1 Quantifying Bias in Multilingual Models

In this section, we provide details of the dataset we collected for the extrinsic bias analysis as well as the metric we use for the bias evaluation.

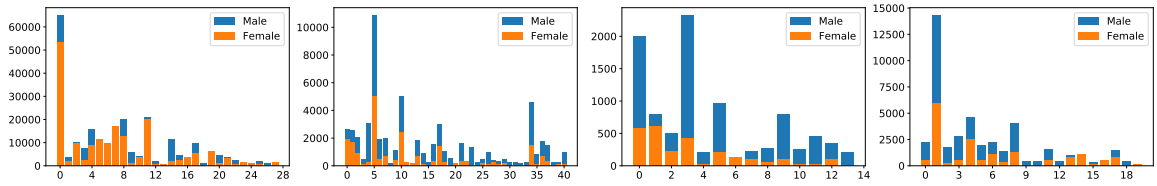
Multilingual BiosBias Datasets

De-Arteaga et al. (2019) built an English BiosBias dataset to evaluate the bias in predicting the occupations of people when provided with a short biography on the bio of the person written in third person. To evaluate the bias in cross-lingual transfer settings, we build the Multilingual BiosBias (MLBs) Dataset which contains bios in different languages.

Dataset Collection Procedure We collect a list of common occupations for each language and follow the data collection procedure used for the English dataset (De-Arteaga et al., 2019). To identify bio paragraphs, we use the pattern "NAME is an OCCUPATION-TITLE" where name is recognized in each language by using the corresponding Named Entity Recognition model from spaCy.[5] To control for the same time period for datasets across languages, we process the same set of Common Crawl dumps ranging from the year 2014 to 2018. For the occupations, we use both the feminine and masculine versions of the word in the gender-rich languages. For EN, we use the existing BiosBias dataset.

The number of occupations in each language is shown in Table 5. As the bios are written in third person, similar to De-Arteaga et al. (2019), we extract the binary genders based on the gendered pronouns in each language, such as "he" and "she".

[5]https://spacy.io/usage/models



(a) EN (b) ES (c) DE (d) FR

Figure 2: Gender statistics of MLBs dataset for different occupations where each occupation has at least 200 instances. X-axis here stands for the occupation index and y-axis is the number of instances for each occupation. Among all the languages, EN corpus is the most gender balanced one. All the corresponding occupations will be provided in the appendix.

Bias Evaluation

We follow the method in Zhao et al. (2018a) to measure the extrinsic bias: using the performance gap between different gender groups as a metric to evaluate the bias in the MLBs dataset. We split the dataset based on the gender attribute. A gender-agnostic model should have similar performance in each group. To be specific, we use the average performance gap across each occupation in the male and female groups aggregated across all occupations (|Diff| in Table 6) to measure the bias. However, as described in Swinger et al. (2019), people's names are potentially indicative of their genders. To eliminate the influence of names as well as the gender pronouns on the model predictions, we use a "scrubbed" version of the MLBs dataset by removing the names and some gender indicators (e.g., gendered pronouns and prefixes such as "Mr." or "Ms.").

To make predictions of the occupations, we adopt the model used in De-Arteaga et al. (2019) by taking the fastText embeddings as the input and encoding the bio text with bi-directional GRU units following by an attention mechanism. The predictions are generated by a softmax layer. We train such models using standard cross-entropy loss and keep the embeddings frozen during the training.

4.2 Characterizing Bias in Multilingual Models

In this section, we analyze the bias in the multilingual word embeddings from the extrinsic perspective. We show that bias exists in cross-lingual transfer learning and the bias in multilingual word embeddings contributes to such bias.

The gender distribution of the MLBs dataset is shown in Fig. 2. Among the three languages, EN corpus is most gender neutral one where the ratio between male and female instances is around 1.2 : 1. For all the other languages, male instances are far larger than female ones. In ES, the ratio between male and female is 2.7 : 1, in DE it is 3.53 : 1, and in FR, it is 2.5 : 1; all are biased towards the male gender.

MLBs	Emb.	Avg.	Female	Male	|Diff|
EN	en	82.82	84.69	80.70	**7.26**
	endeb	83.00	84.71	81.06	6.09 ↓
	en-es	83.43	85.14	81.51	6.72 ↓
	en-de	82.85	84.64	80.84	6.37 ↓
	en-fr	82.66	84.34	80.78	5.87 ↓
ES	es	63.83	64.47	63.56	6.56
	es-en	61.47	61.42	61.49	**7.13** ↑
	es-endeb	61.91	62.98	61.45	5.61 ↓
	es-de	61.61	62.82	61.11	5.51 ↓
	es-fr	62.91	63.31	62.73	4.32 ↓

Table 6: Results on scrubbed MLBs. "Emb." stands for the embeddings used in model training. "Avg.", "Female" and "Male" refer to the overall average accuracy (%), and average accuracy for different genders respectively. " |Diff|" stands for the average absolute accuracy gap between each occupation in the male and female groups aggregated across all the occupations. The results of FR and DE are in the appendix.

Bias in Monolingual BiosBias We first evaluate the bias in the MLBs monolingual dataset by predicting the occupations of the bios in each language.[6] From Table 6 we observe that: 1) Bias commonly exists across all languages ($|\text{Diff}| > 0$) when using different aligned embeddings, meaning that the model works differently for male and female groups. 2) When training the model using different aligned embeddings, it does not affect the overall average performance significantly ("Avg." column in the table). 3) The alignment direction influences the bias. On training the model based on the embeddings aligned to different target space, we find that aligning the embeddings to ENDEB

[6]The results of DE and FR are in the appendix.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→ES	en	es-en	41.68	42.29	41.42	2.83
	en-es	es	34.15	33.97	34.22	3.49
ES→EN	es	en-es	57.33	59.61	54.75	8.33
	es-en	en	57.05	59.32	54.47	10.13

Table 7: Results of transfer learning on the scrubbed MLBs. "Src." and "Tgt." stand for the embeddings in source model and fine tuning procedure respectively.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→ES	en	es-en	39.17	41.30	38.70	**7.97**
	en-es	es	35.66	36.11	35.47	4.53
	en-de	es-de	34.12	34.46	33.98	4.07
	en-fr	es-fr	37.63	38.75	37.16	4.87
ES→EN	es	en-es	58.41	61.78	54.60	9.03
	es-en	en	55.62	58.00	52.93	**9.52**
	es-de	en-de	57.98	60.47	55.17	9.13
	es-fr	en-fr	55.04	57.85	51.86	8.47

Table 8: Results of transfer learning on gender balanced scrubbed MLBs. The bias in the last column demonstrates that the bias in the multilingual word embeddings also influences bias in transfer learning.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→ES	endeb	es-endeb	37.44	39.90	36.40	5.93
ES→EN	es-endeb	endeb	52.51	54.45	50.03	9.06

Table 9: Bias mitigation results of transfer learning when we aligned the embeddings to the ENDEB space on gender balanced scrubbed MLBs.

or a gender-rich language reduces the bias in the downstream task. This is aligned with our previous observation in Section 3.

Bias in Transfer Learning Multilingual word embeddings are widely used in cross-lingual transfer learning (Ruder et al., 2019). In this section, we conduct experiments to understand how the bias in multilingual word embeddings impacts the bias in transfer learning. To do this, we train our model in one language (i.e., source language) and transfer it to another language based on the aligned embeddings obtained in Section 3.2. For the transfer learning, we train the model on the training corpus of the source language and randomly choose 20% of the dataset from the target language and use them to fine-tune the model.[7] Here, we do not aim at achieving state-of-the-art transfer learning performance but pay more attention to the bias analysis. Table 7 shows that the bias is present when we do the transfer learning regardless of the direction of transfer learning.

[7] As there are fewer examples in DE, we use the whole datasets for transfer learning.

MLBs	Avg.	Female	Male	|Diff|
EN	84.35	85.54	83.01	7.31
ES	67.93	65.79	68.82	4.16
DE	72.68	73.68	72.28	4.89
FR	79.18	78.80	79.35	8.75

Table 10: Bias in monolingual MLBs using M-BERT.

Trans.	Avg.	Female	Male	|Diff|
EN→ES	66.56	65.70	66.92	5.48
EN→DE	76.21	75.66	76.42	7.51
EN→FR	76.46	75.73	76.81	8.97

Table 11: Bias in MLBs using M-BERT when transferring from EN to other languages. Comparing to multilingual word embeddings, M-BERT achieves better transfer performance on the MLBs dataset across different languages. But the bias can be higher comparing to the multilingual word embeddings.

Bias from Multilingual Word Embeddings The transfer learning bias in Table 7 is a combined consequence of both corpus bias and the multilingual word embedding bias. To better understand the influence of the bias in multilingual word embeddings on the transfer learning, we make the training corpus gender balanced for each occupation by upsampling to approximately make the model free of the corpus bias. We then test the bias for different languages with differently aligned embeddings. The results are shown in Table 8. When we adopt the embeddings aligned to gender-rich languages, we could reduce the bias in the transfer learning, whereas adopting the embeddings aligned to EN results in an increased bias.

Bias after Mitigation Inspired by the method in Zhao et al. (2018a), we mitigate the bias in the downstream tasks by adopting the bias-mitigated word embeddings. To get the less biased multilingual word embeddings, we align other embeddings to the ENDEB space previously obtained in Section 3. Table 9 demonstrates that by adopting such less biased embeddings, we can reduce the bias in transfer learning. Comparing to Table 8, aligning the embeddings to a gender-rich language achieves better bias mitigation and, at the same time, remains the overall performance.

4.3 Bias Analysis Using Contextualized Embeddings

Contextualized embeddings such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2018) and XL-Net (Yang et al., 2019) have shown significant performance improvement in various NLP applica-

tions. Multilingual BERT (M-BERT) has shown its great ability for the transfer learning. As M-BERT provides one single language model trained on multiple languages, there is no longer a need for alignment procedure. In this section, we analyze the bias in monolingual MLBs dataset as well as in transfer learning by replacing the fastText embeddings with M-BERT embeddings. Similar to previous experiments, we train the model on the English dataset and transfer to other languages. Table 10 and 11 summarizes our results: comparing to results by fastText embeddings in Table 6, M-BERT improves the performance on monolingual MLBs dataset as well as the transfer learning tasks. When it comes to the bias, using M-BERT gets similar or lower bias in the monolingual datasets, but sometimes achieves higher bias than the multilingual word embeddings in transfer learning tasks such as the EN $\rightarrow$ ES (in Table 7).

5 Conclusion

Recently bias in embeddings has attracted much attention. However, most of the work only focuses on English corpora and little is known about the bias in multilingual embeddings. In this work, we build different metrics and datasets to analyze gender bias in the multilingual embeddings from both the intrinsic and extrinsic perspectives. We show that gender bias commonly exists across different languages and the alignment target for generating multilingual word embeddings also affects such bias. In practice, we can choose the embeddings aligned to a gender-rich language to reduce the bias.

However, due to the limitation of available resources, this study is limited to the European languages. We hope this study can work as a foundation to motivate future research about the analysis and mitigation of bias in multilingual embeddings. We encourage researchers to look at languages with different grammatical gender (such as Czech and Slovak) and propose new methods to reduce the bias in multilingual embeddings as well as in cross-lingual transfer learning.

Acknowledgments

This work was supported in part by NSF Grant IIS-1927554. We would like to thank Maria De-Arteaga and Andi Peng for the helpful discussion, and thank all the reviewers for their feedback.

References

Wasi Ahmad, Zhisong Zhang, Xuezhe Ma, Eduard Hovy, Kai-Wei Chang, and Nanyun Peng. 2019a. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2440–2452.

Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. 2019b. Cross-lingual dependency parsing with unlabeled auxiliary languages. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 372–382.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

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.

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

Xilun Chen, Ahmed Hassan, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112.

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

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 845–850.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *NAACL HLT 2018*, page 43.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 85–91.

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. pages 615–621.

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

Ninareh Mehrabi, Thamme Gowda, Fred Morstatter, Nanyun Peng, and Aram Galstyan. 2019. Man is to person as woman is to location: Measuring gender bias in named entity recognition. *arXiv preprint arXiv:1910.10872*.

Tao Meng, Nanyun Peng, and Kai-Wei Chang. 2019. Target language-aware constrained inference for cross-lingual dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1117–1128.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630.

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.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8–14.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311. ACM.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 629–634.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 15–20.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5279–5287.

A Appendices

A.1 Multilingual Word Embeddings Alignment

We use the default hyperparameters in the RCSLS alignment model (https://github.com/facebookresearch/fastText) but change batch size to 5000 and set "sgd" to true to make sure the batch size is used. The "maxsup" is set to the same as "maxneg" with 200000.

A.2 Intrinsic Bias Analysis

Category	Target			
	EN	ES	DE	FR
Strong-gendered	0.1138	0.0848	0.0935	0.0833
Weak-gendered	0.0477	0.0400	0.0430	0.0395

Table 12: Bias in EN before and after alignment to different languages for different word categories. For different situation, again we see the bias will reduce when we align the words to gender rich languages.

Category	Target			
	ENDEB	ES	DE	FR
Strong-gendered	0.0830	0.0683	0.0747	0.0685
Weak-gendered	0.0126	0.0201	0.0269	0.0162

Table 13: Bias in ENDEB before and after alignment to different languages for different word categories. When aligning to a gender rich language, the bias in those strong-gendered words reduces.

A.3 Transfer Learning Setting

For the transfer learning, we filter some occupations that commonly occur across all languages and manually make the distribution of each occupation similar in each language. For each corpus, we use 60% of the corpus for training, 20% for validation and 20% for testing.

MLBs	Emb.	Avg.	Female	Male	|Diff|
DE	de	55.4	59.87	53.63	10.42
	de-en	56.88	61.84	54.92	**15.41**
	de-endeb	54.09	55.26	53.63	6.54
	de-es	54.46	56.58	53.63	9.51
	de-fr	55.8	57.50	55.18	10.43
FR	fr	76.52	76.24	76.65	11.58
	fr-en	74.13	74.87	73.79	**12.96**
	fr-endeb	73.92	74.19	73.79	10.84
	fr-es	74.57	74.19	74.74	11.23
	fr-de	75.11	75.56	74.90	12.07

Table 14: Results on the scrubbed BiosBias dataset in DE and FR.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→DE	en	de-en	37.55	39.47	36.79	16.52
	en-de	es-de	34.57	32.89	35.23	13.58
DE→EN	de	en-de	42.47	45.76	38.77	6.46
	de-en	en	38.55	41.25	35.51	7.12

Table 15: Results of transfer learning between EN and DE on MLBs dataset.

A.4 Occupation Lists for MLBs Gender Statistics

We list all the occupations for each language in Fig. 2.

EN: professor, accountant, journalist, architect, photographer, psychologist, teacher, nurse, attorney, software_engineer, painter, physician, chiropractor, personal_trainer, surgeon, filmmaker, dietitian, dentist, dj, model, composer, poet, comedian, yoga_teacher, interior_designer, pastor, rapper, paralegal

ES: student, model, teacher, cook, musician, artist, painter, professor, administrator, scientist, writer, nurse, hotelier, lawyer, coach, computer_programmer, doctor, journalist, architect, soldier, pharmacist, poet, dancer, engineer, farmer, pianist, pilot, psychologist, surgeon, athlete, mechanic, driver, accountant, rapper, photographer, filmmaker, attorney, physician, dj, comedian, composer

DE: journalist, teacher, psychologist, attorney, dj, photographer, nurse, professor, pastor, architect, filmmaker, composer, painter, software_engineer

FR: filmmaker, teacher, composer, painter, journalist, physician, attorney, poet, photographer, pastor, rapper, architect, dj, comedian, psychologist, accountant, nurse, model, surgeon, dietitian

A.5 Extrinsic Bias Results in DE and FR

We show the bias in monolingual DE and FR datasets in Table 14 and in the transfer learning

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→FR	en	fr-en	41.43	41.03	41.62	5.96
	en-fr	en	43.12	44.96	42.26	8.33
FR→EN	fr	en-fr	57.81	62.02	51.94	9.79
	fr-en	fr	55.15	58.83	50.0	8.3

Table 16: Results of transfer learning between EN and FR on MLBs dataset.

between EN and them in Table 15 and 16 respectively.

Table 17 and 18 is the bias result of the transfer learning between EN and DE, FR when we manually make the gender ratio balanced for each occupation in the corpus. We also show the mitigation results when we align all the embeddings to the ENDEB space.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→DE	en	de-en	39.40	38.28	39.82	10.65
	endeb	de-endeb	33.51	31.37	34.42	8.9
	en-es	de-es	33.16	32.21	33.50	9.31
	en-de	de	33.96	31.02	35.03	9.13
	en-fr	de-fr	38.31	34.17	39.82	**11.04**
DE→EN	de	en-de	46.43	48.83	43.72	7.93
	de-en	en	50.48	53.91	46.58	**8.10**
	de-endeb	endeb	44.44	46.84	41.73	7.16
	de-es	en-es	44.04	47.54	40.09	7.29
	de-fr	en-fr	46.01	47.57	44.25	7.03

Table 17: Results of transfer learning between EN and DE on the scrubbed BiosBias dataset when we make the dataset gender balanced. The bias in the last column demonstrates that the bias in the multilingual word embeddings will also influence the bias in the transfer learning.

Trans.	Src.	Tgt.	Avg.	Female	Male	|Diff|
EN→FR	en	fr-en	36.66	36.24	36.85	**7.97**
	endeb	fr-endeb	34.86	32.82	35.82	5.44
	en-es	fr-es	34.82	34.19	35.11	6.77
	en-de	fr-de	33.51	33.85	33.36	5.78
	en-fr	fr	35.68	33.50	36.70	6.81
FR→EN	fr	en-fr	59.21	61.55	55.94	10.3
	fr-en	en	50.80	54.44	45.73	**11.42**
	fr-endeb	endeb	49.33	52.91	44.33	10.14
	fr-es	en-es	49.28	51.86	45.66	10.42
	fr-de	en-de	50.92	54.10	46.46	7.36

Table 18: Results of transfer learning between EN and FR on the scrubbed BiosBias dataset when we make the dataset gender balanced. The bias in the last column demonstrates that the bias in the multilingual word embeddings will also influence the bias in the transfer learning.