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

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Kai-Wei Chang

University of California, Los Angeles

UCLA ENGINEERING Computer Science

Overview

We extend analysis of gender bias in English (EN) word embeddings to **languages with grammatical genders** like Spanish (ES) and French (FR). We also analyze bias **in bilingual word embeddings** that align a genderless language like English (EN) with a gendered language.

Gender bias indeed exists in gendered languages and in bilingual word embeddings. We propose new definitions and quantification methods of gender bias by constructing two gender directions.

We propose new approaches to evaluate and mitigate gender bias in gendered languages. Results show that our methods **effectively mitigate bias while preserving the utility of embeddings**.

Problem Statement

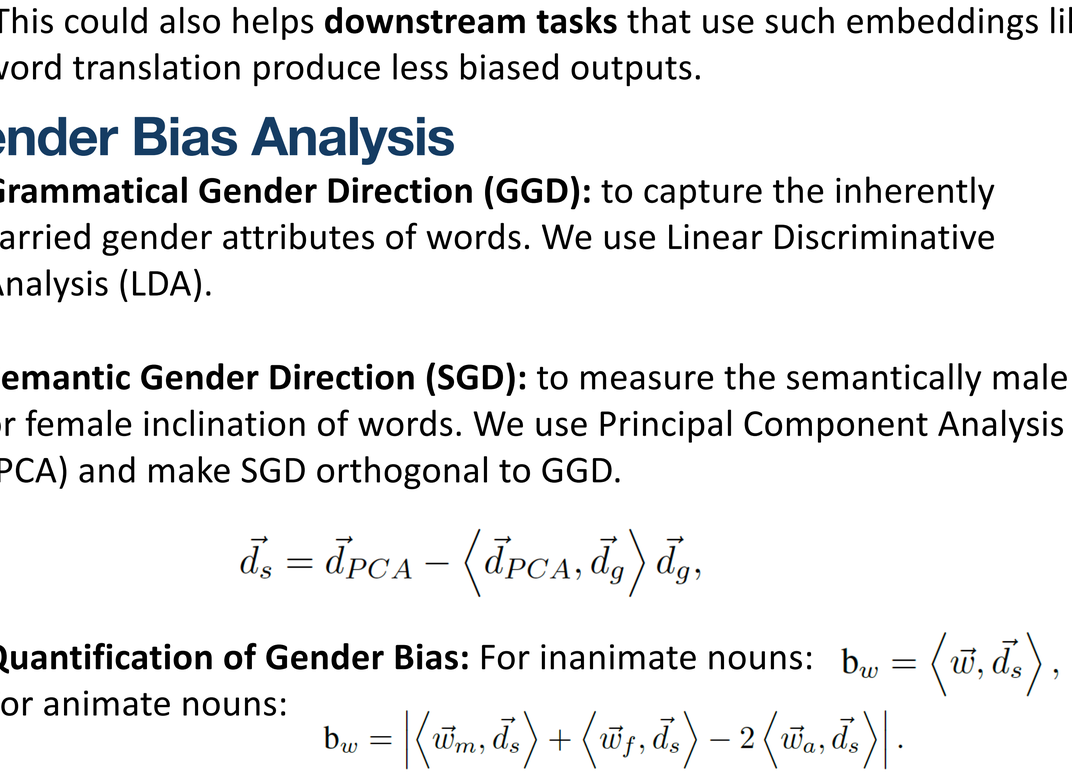
- In gendered languages, all nouns are assigned a gender class and the dependent words have to follow morphological agreement.
- For example, in Spanish, **la buena enfermera:** the good female nurse, **el buen enfermero:** the good male nurse.
- Previous definition of bias based on embedding projection in gender direction becomes problematic, since the inclination could be caused by grammatical genders.
- We also consider **bilingual embeddings** and use words from the mitigated genderless language to better "debias" gendered languages.
- This could also helps **downstream tasks** that use such embeddings like word translation produce less biased outputs.

Gender Bias Analysis

- **Grammatical Gender Direction (GGD):** to capture the inherently carried gender attributes of words. We use Linear Discriminative Analysis (LDA).
- **Semantic Gender Direction (SGD):** to measure the semantically male or female inclination of words. We use Principal Component Analysis (PCA) and make SGD orthogonal to GGD.

$$\vec{d}_s = \vec{d}_{PCA} - \left\langle \vec{d}_{PCA}, \vec{d}_g \right\rangle \vec{d}_g,$$

- **Quantification of Gender Bias:** For inanimate nouns: $\mathrm{b}_w = \left\langle \vec{w}, \vec{d}_s \right\rangle$, For animate nouns: $\mathrm{b}_w = \left| \left\langle \vec{w}_m, \vec{d}_s \right\rangle + \left\langle \vec{w}_f, \vec{d}_s \right\rangle - 2 \left\langle \vec{w}_a, \vec{d}_s \right\rangle \right|$.



Mitigation Methods

- **Monolingual Setting:** Shifting along the semantic gender direction with respect to an anchor point (Shift).
- **Bilingual Setting:** Mitigating English before alignment (De-Align), and hybrid of De-Align and Shift (Hybrid).
- We consider two types of anchor points: origin point of SGD (_Ori) and the position of mitigated English words (_EN)

Experiment Results

- **Monolingual Experiments:**
  - **Modified Word Embedding Association Tests (MWEAT)**

$$\left| \left| \sum_{x \in X} s(x, A, B) \right| - \left| \sum_{y \in Y} s(y, A, B) \right| \right| \qquad s(w, A, B) = \mathrm{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \mathrm{mean}_{b \in B} \cos(\vec{w}, \vec{b}).$$

  - **Word Similarity**

Monolingual	Original	Shift_Ori	Shift_EN	De-Align	Hybrid_Ori	Hybrid_EN
ES–MWEAT–Diff	3.6918	0.3090	0.3324	3.5748	**0.3090**	2.2494
ES–MWEAT–p-value	0.0000	0.1130	0.0010	0.0010	**0.7330**	0.0020
FR–MWEAT–Diff	2.3437	0.2446	0.3882	2.3436	**0.2446**	1.1758
FR–MWEAT–p-value	0.0000	0.1470	0.0010	0.0020	**0.5290**	0.0910
ES–Word Similarity	**0.7392**	0.7363	0.7359	**0.7392**	0.7358	0.7356
FR–Word Similarity	**0.7294**	0.7218	0.7218	0.7156	0.7218	0.7218

- **Cross-lingual Experiments:**
  - **Cross-lingual Analogy Task:**

hospital - doctor = hospital - ___

```
0.7499 - doctor
0.5555 - doctora
0.5337 - médico
0.5263 - doctors
0.5063 - dr
0.5047 - cirujano
0.5013 - hospital
0.5010 - enfermera
0.4930 - psiquiatra
0.4847 - doctores
```

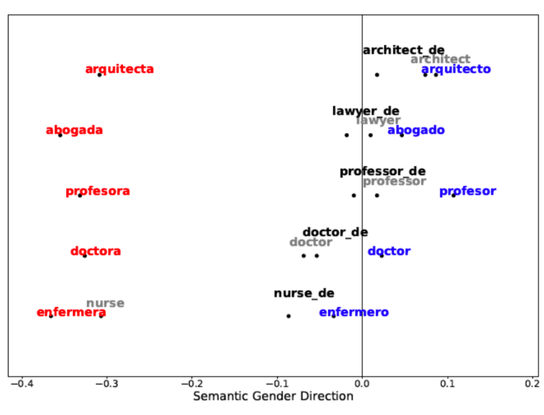
university - professor = universidad - ___

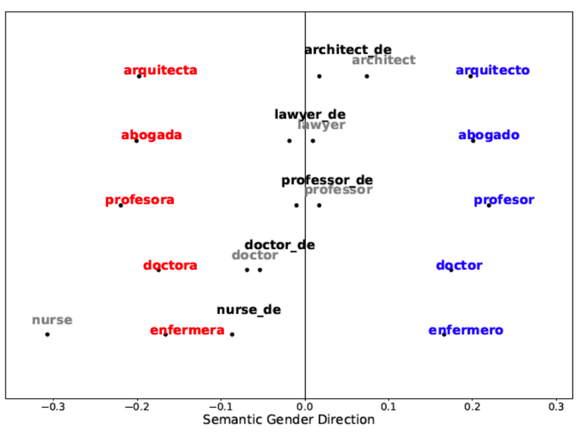
```
0.7976 - profesor
0.7261 - catedrático
0.6797 - catedrática
0.6768 - universidad
0.6557 - profesora
0.6308 - doctorado
0.6194 - investigador
0.6119 - doctoró
0.6113 - facultad
0.6046 - emérito
```

  - **Word Translation**

Bilingual	Original	Shift_Ori	Shift_EN	De-Align	Hyrid_Ori	Hyrid_EN
ES-EN–CLAT–ASD	0.1082	0.0961	0.0961	0.0827	**0.0755**	0.0772
ES-EN–CLAT–F_MRR	0.2073	0.2507	0.2507	0.2919	0.3450	0.3150
ES-EN–CLAT–M_MRR	0.6940	0.6766	0.6766	0.6775	0.6398	0.6696
ES-EN–CLAT–MRR Diff	0.4867	0.4259	0.4259	0.3856	**0.2949**	0.3546
FR-EN-CLAT-ASD	0.1208	0.1048	0.1082	0.0892	**0.0735**	0.0805
FR-EN–CLAT–F_MRR	0.1663	0.2101	0.1943	0.2679	0.3128	0.2975
FR-EN–CLAT–M_MRR	0.6549	0.6313	0.6419	0.6610	0.6393	0.6467
FR-EN–CLAT–MRR Diff	0.4886	0.4212	0.4476	0.3931	**0.3265**	0.3492
EN-ES–WT–P@1/5	79.2/89.0	**80.7/90.3**	**80.7/90.3**	76.5/88.9	**80.7/90.3**	**80.7/90.3**
ES-EN–WT–P@1/5	79.2/89.0	79.2/89.0	79.2/89.0	**80.1/90.7**	79.2/89.0	79.2/89.0
EN-FR–WT–P@1/5	78.2/89.4	**79.9/91.1**	**79.9/91.1**	74.3/87.8	**79.9/91.1**	**79.9/91.1**
FR-EN–WT–P@1/5	**76.1/88.1**	**76.1/88.1**	**76.1/88.1**	74.4/87.2	**76.1/88.1**	**76.1/88.1**

Hybrid with origin as anchor points can mitigate most effectively. We present others as ablation studies. Word similarity and Word translation shows that our methods can even increase the performances of the original mono-/bilingual embeddings.





Future Works

- Apply on downstream tasks like machine translation.
- This work focuses on gendered languages with feminine and masculine forms, but languages with more noun classes exist.