Fairness in Natural Language Processing (NLP) Systems ---Examining Gender Bias in Embeddings of Languages with Grammatical Gender

IRI Graduation Presentation

Pei Zhou
May. 30th

Topic/Research Area

- Numerous NLP applications use word embeddings/vectors.
- Reports have shown **gender biases in word embeddings in English** and they have caused deteriorated effects on **downstream tasks** like sentiment analysis.
- This project aims to extend bias analysis and mitigation methods to **other languages as well as bilingual Word Embeddings.**

Ideas

- Analysis of bias in English word embeddings cannot be applied to languages with **grammatical gender** such as Spanish.
- Propose new definitions and mitigation approaches of gender bias in languages with grammatical gender.
- Also naturally extend to bilingual case.
- Help NLP applications produce **less biased results** by making the embeddings fairer.

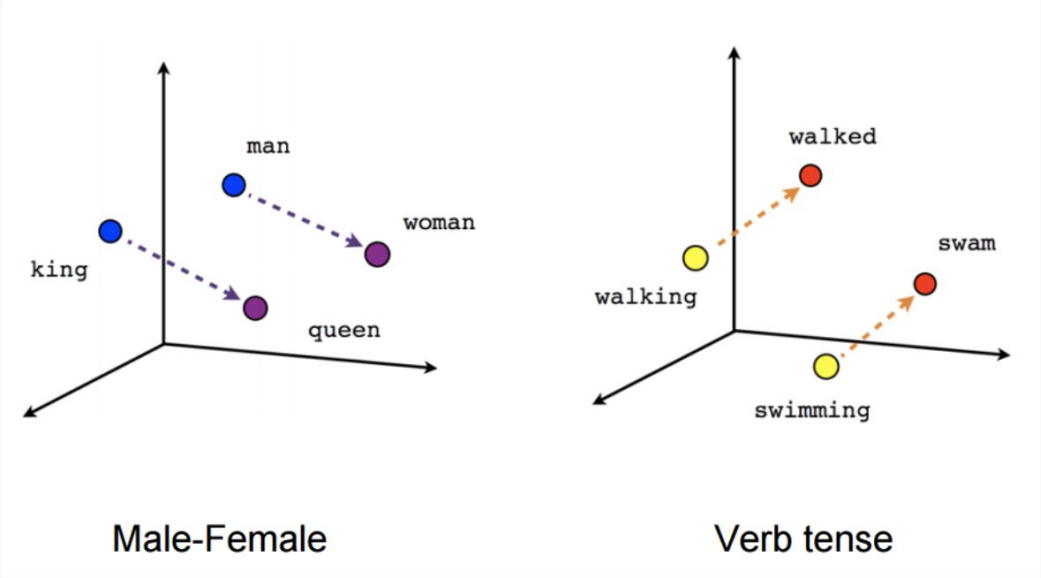
Research Method

- Define two directions: **semantic gender** and **grammatical gender** to quantify bias using Linear Discriminative Analysis (LDA).
- Propose two methods to mitigate gender bias in languages with grammatical gender and bilingual word embeddings..
- Evaluate on two metrics to quantify the gender bias and other metrics to test the utility of the embeddings

Findings/Future Work

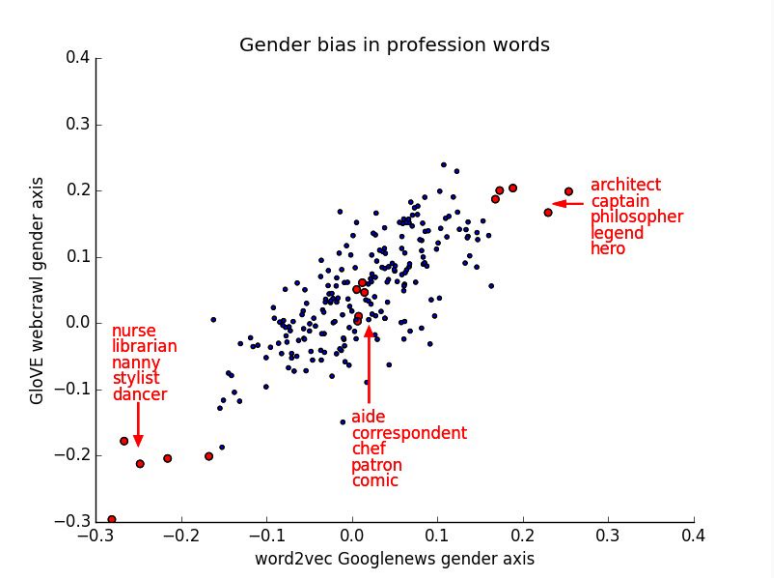
- Found that gender bias indeed exists in other languages and bilingual embeddings.
- The defined gender directions are able to capture the two types of gender information.
- Evaluation results show that proposed methods can **effectively mitigate gender bias** while **preserving the utility** of the embeddings.
- Future work can focus on bias analysis for gender-agnostic languages like Turkish.

Background: Bias in Word Embeddings

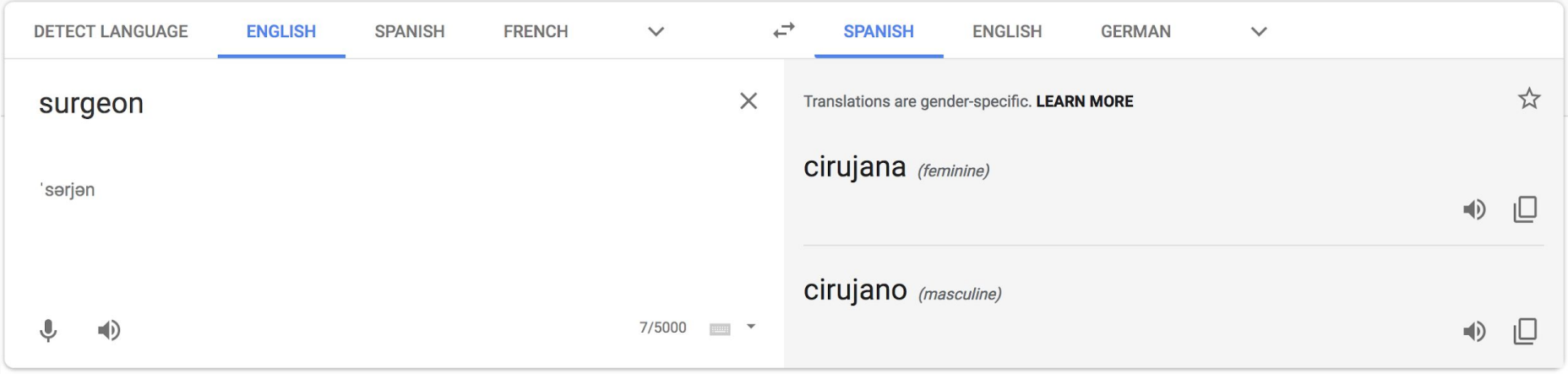


Male-Female

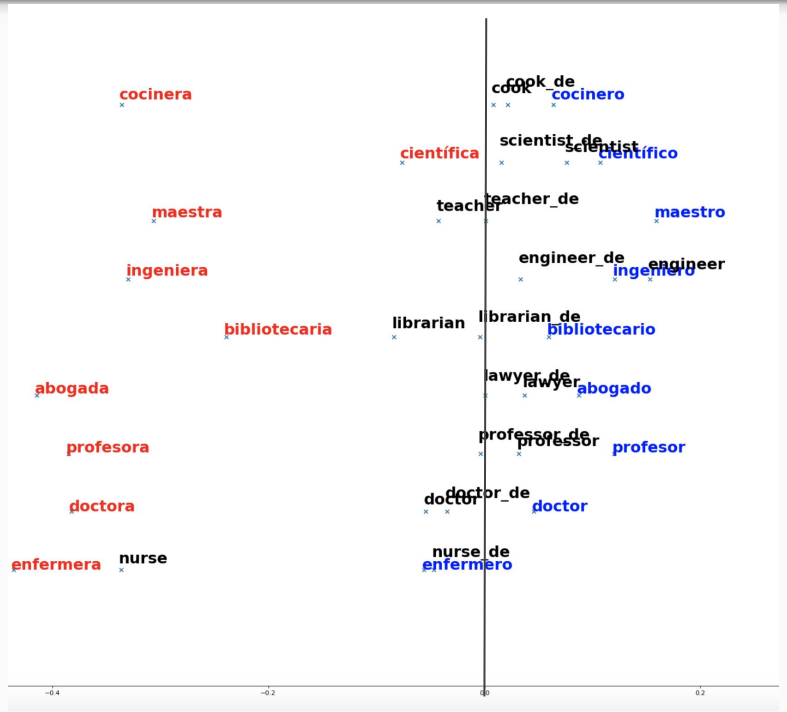
Verb tense



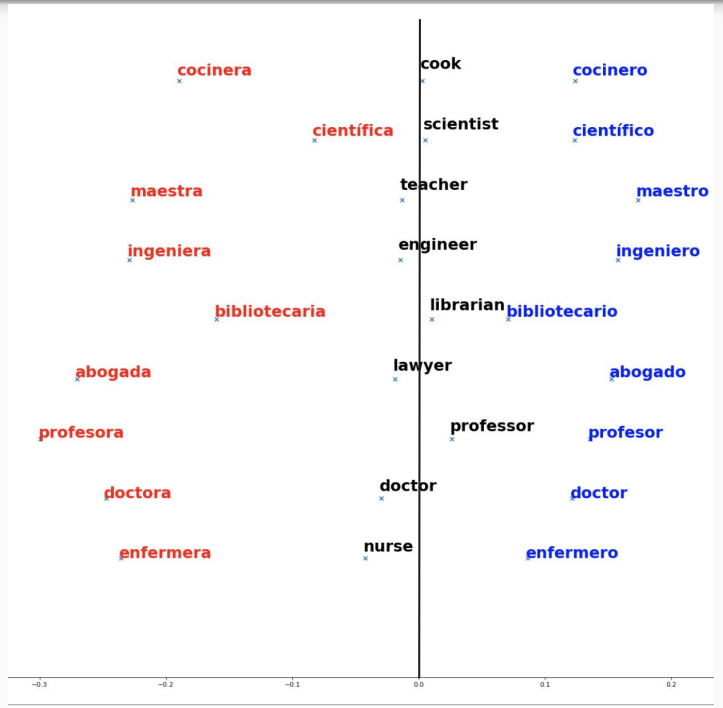
Novelty: Languages with Grammatical Gender



Projections of Spanish and English Occupation Words on Semantic Gender Direction



Debias



Cross-Lingual Analogy Tests

hospital - doctor = hospital - ____

```
[('doctor', 0.62945422730374645),
 ('dr', 0.48846381285697832),
 ('psiquiatra', 0.44350407901682343),
 ('doctores', 0.4413784086390321),
 ('médico', 0.43021179463164205),
 ('cirujano', 0.41413434873805194),
 ('doctora', 0.41265042568597021),
 ('medico', 0.39072843127340701),
 ('enfermero', 0.38469410637211388),
 ('neurólogo', 0.38102243679938841),
```

(debiased)

```
[('doctor', 0.71337390151945113),
 ('doctora', 0.58878033522567863),
 ('dr', 0.52914145082276209),
 ('doctores', 0.5202312767730457),
 ('dra', 0.48144637990355704),
 ('psiquiatra', 0.44874802604434361),
 ('doctors', 0.43686882756405893),
 ('honoris', 0.43046121187846353),
 ('enfermera', 0.41241463285558572),
 ('tardis', 0.40956702226946973),
```

reliable - engineer = fiable - ____

```
[('ingeniero', 0.67290869902080463),
 ('ingenieros', 0.55207964801265874),
 ('ingeniera', 0.54440290722120077),
 ('ingeniería', 0.53668305009389816),
 ('constructor', 0.4764584117892337),
 ('mecánico', 0.47195333758564678),
 ('diseñador', 0.47095740909166506),
 ('mezclador', 0.470548986511937),
 ('contratista', 0.46140857570078719),
 ('engineering', 0.45197427415757452),
```

(debiased)

```
[('ingeniero', 0.7624931127803396),
 ('ingeniera', 0.71463607090874359),
 ('ingeniería', 0.62249029965855818),
 ('ingenieros', 0.58735455765790445),
 ('agrónomo', 0.51416807851659785),
 ('ingenierías', 0.49474552702043062),
 ('mecánico', 0.48187978627474926),
 ('mezclador', 0.47108353589505847),
 ('electricista', 0.4650644801549213),
 ('diseñador', 0.45496389974761386),
```

Quantitative Evaluation

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**

Final Deliverables

1. Research paper already accepted at workshops at top conferences: ICML AI for Social Good and ACL Student Research Workshop and I will be presenting our paper in June and July.
2. Submitted an extended version to EMNLP 2019 (top tier NLP conference).
3. Data and code will be released on Github.

Thanks for your time