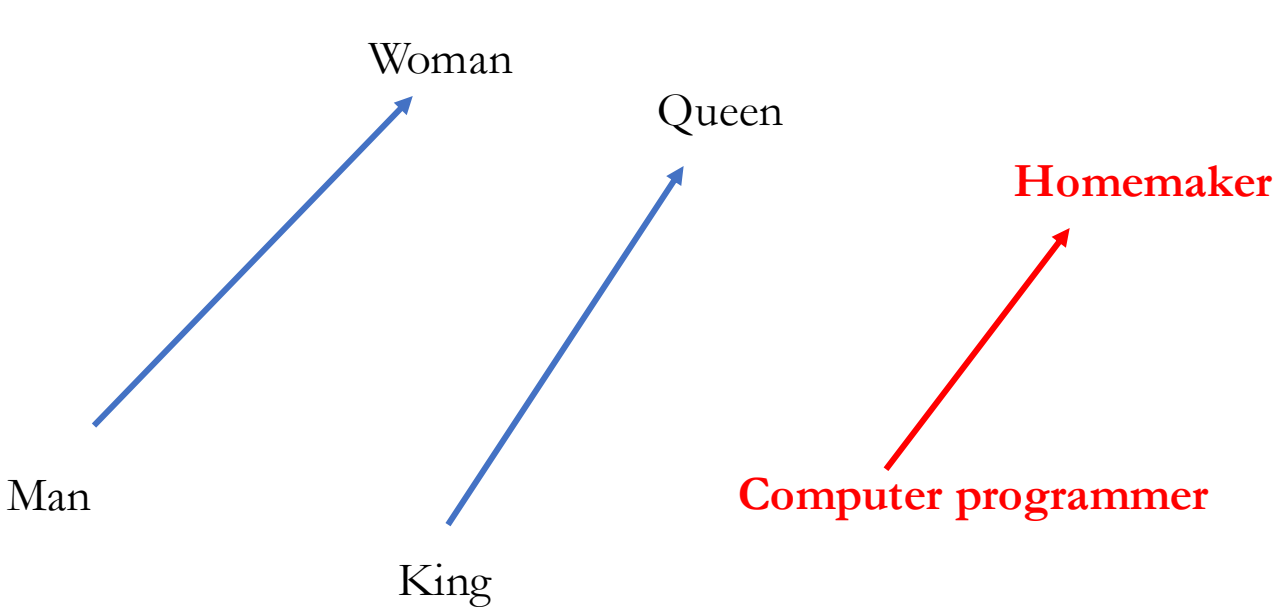
Bias in NLP: Word Embeddings

Word embeddings reflect gender, age, sexual orientation and other biases



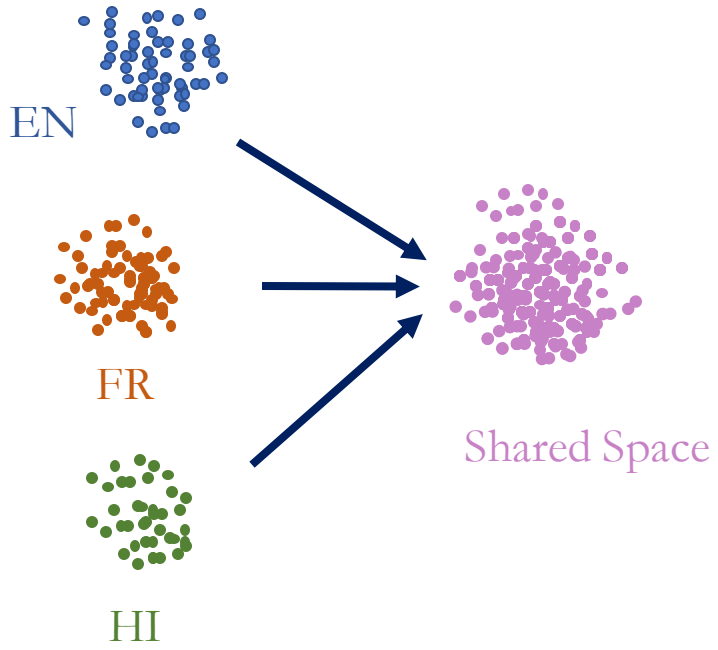
Bolukbasi et. al., Man is to computer programmer as woman is to homemaker? Debiasing Word Embeddings. Neurips, 2016

Microsoft

UCLA NLP

Multilingual Word Embeddings

- Align embeddings → Shared space
- Widely used in cross-lingual transfer
  - E.g., transfer a down-stream model from source to target languages



Microsoft

UCLA NLP

Analyzing Bias in Multilingual Embeddings

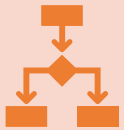
- Does the language property affect bias exhibition?
  - e.g. different grammatical gender systems in English (EN) & Spanish (ES)
- Does the target alignment space for generating multilingual embeddings affect the bias?
- Does cross-lingual transfer learning based on multilingual embedding inherit the bias?

Microsoft

UCLA NLP

Outline

Intrinsic Bias Analysis



Extrinsic Bias Analysis

Bias Mitigation

Intrinsic Bias in Multilingual Embeddings

- Quantify bias in multilingual word vectors
  - We study English (EN), Spanish (ES), German (DE), French (FR)
  - Embeddings obtained from fastText

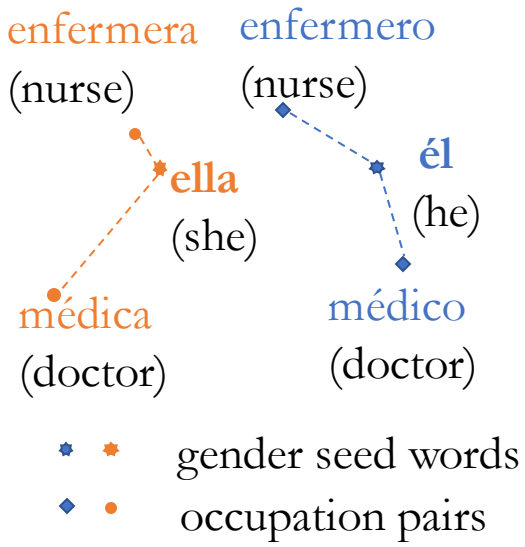
- inBias: averaged distance(targets, attributes) gap between different groups

occupations

gender seeds

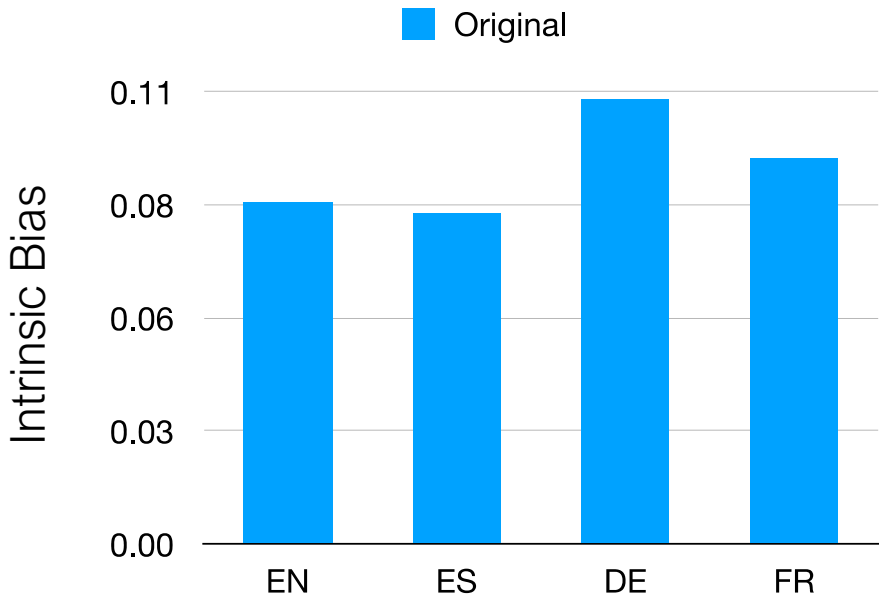
$$\text{inBias} = \frac{1}{N}\sum_{i=1}^{N} |dis(O_{M_i}, S_M) - dis(O_{F_i}, S_F)|$$

- Over 200 pairs of occupations and 10 pairs of gender seeds for each language



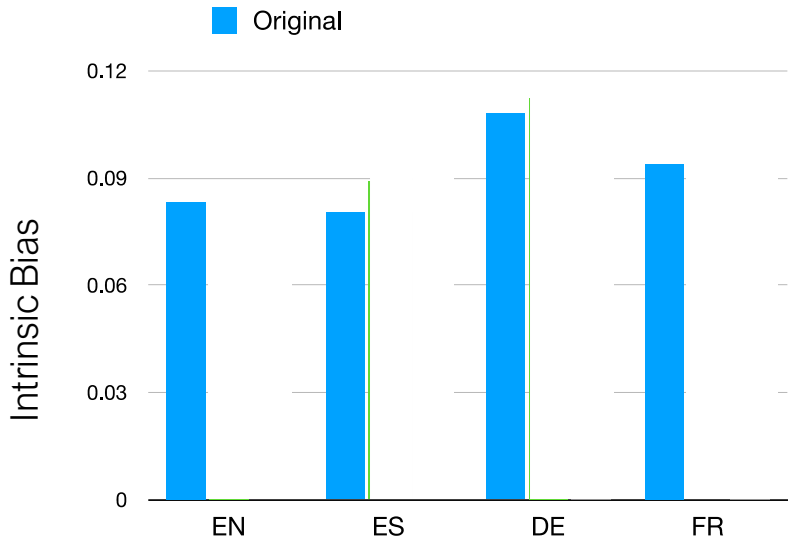
How gender bias exhibits in each language?

- Bias commonly exists across all languages (Original)
- On average bias in DE, FR is stronger than EN, ES



How the alignment target affects gender bias?

- Bias changes when choosing different target space
  - Aligning to English, bias increases (-EN)
  - Aligning to Spanish , bias decreases (-ES)



Extrinsic Bias in Multilingual Embeddings

- Analyze how bias in multilingual embedding affects cross-lingual transfer
- We create BiosBias, a multilingual biography dataset for the study
- Follow [1] to identify the bio pattern

"NAME is an OCCUPATION-TITLE. Other descriptions… "

NER model

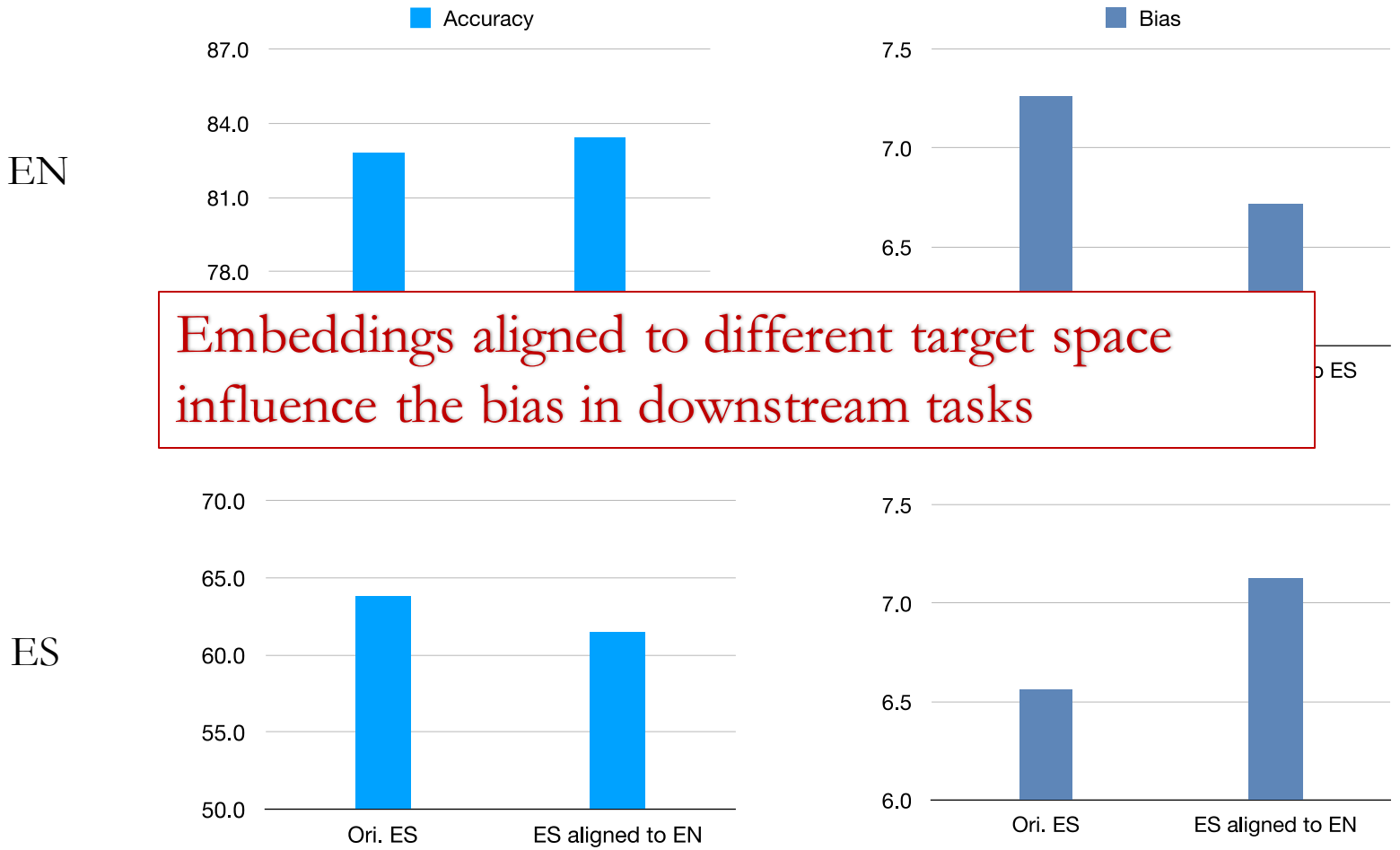
Feminine and masculine versions

- Extract binary genders based on gendered pronouns in each language
- Predict occupations based on the descriptions
- Bias evaluation: averaged |performance gap| between different gender groups for all occupations

[1] De-Arteaga et al. Bias in bios: A case study of semantic representation bias in a high-stakes setting. FAT 2019

Bias in Occupation Prediction Task

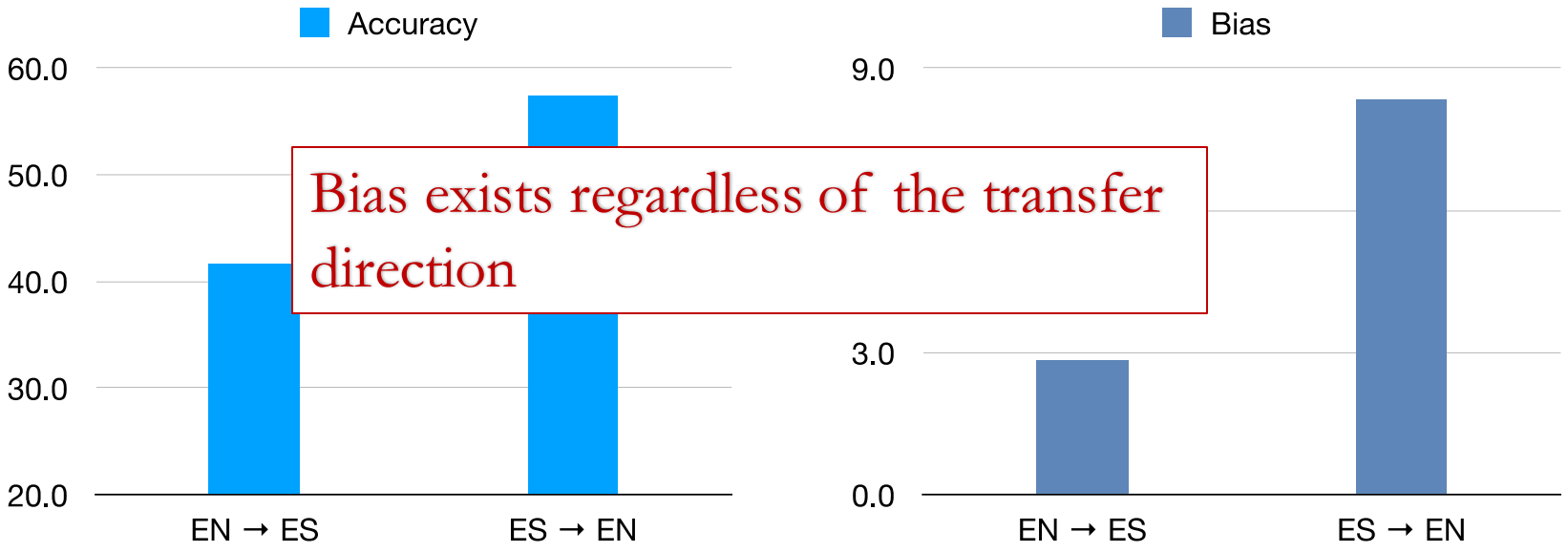
- Train and evaluate in EN and ES portion of BiosBias respectively



Embeddings aligned to different target space influence the bias in downstream tasks

Extrinsic Bias in Cross-lingual Transfer Learning

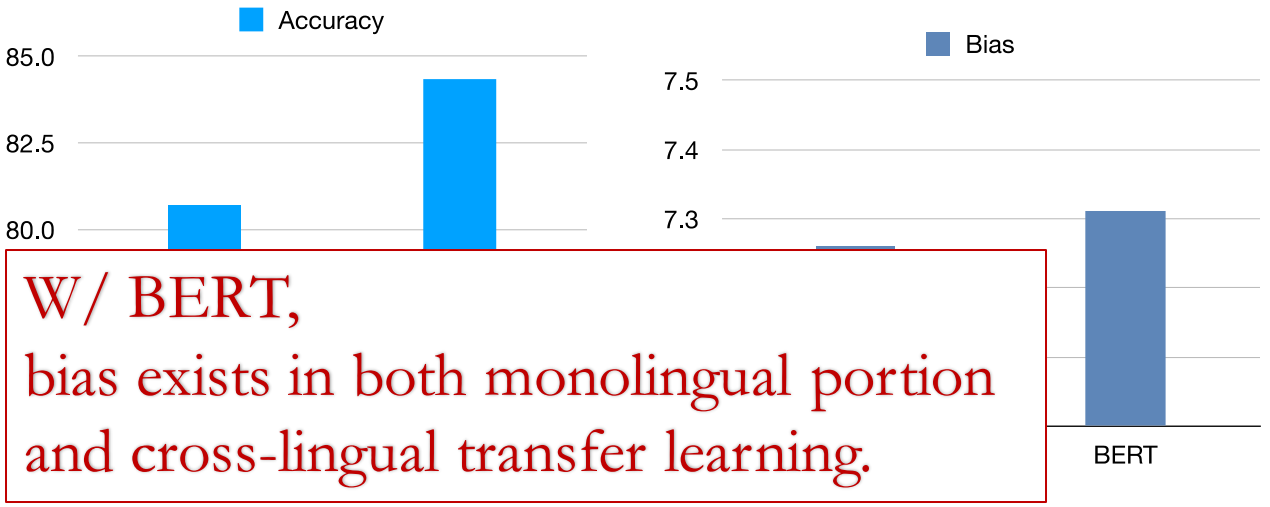
- Transfer a model trained in English (EN) to Spanish (ES) and vice versa by multi-lingual embedding.



Please refer to the paper for more analysis.

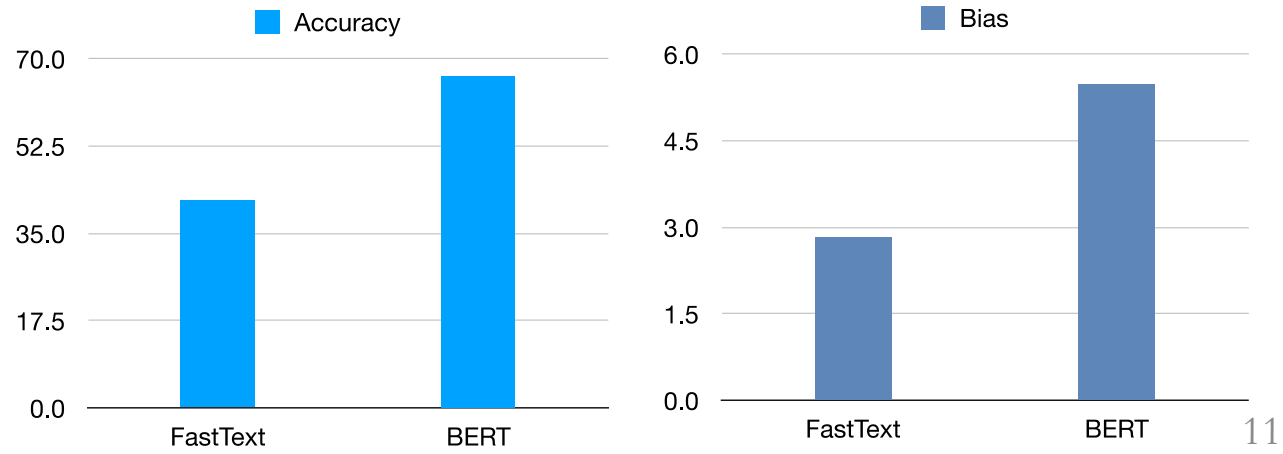
Bias Using Contextualized Embeddings

- In EN BiosBias



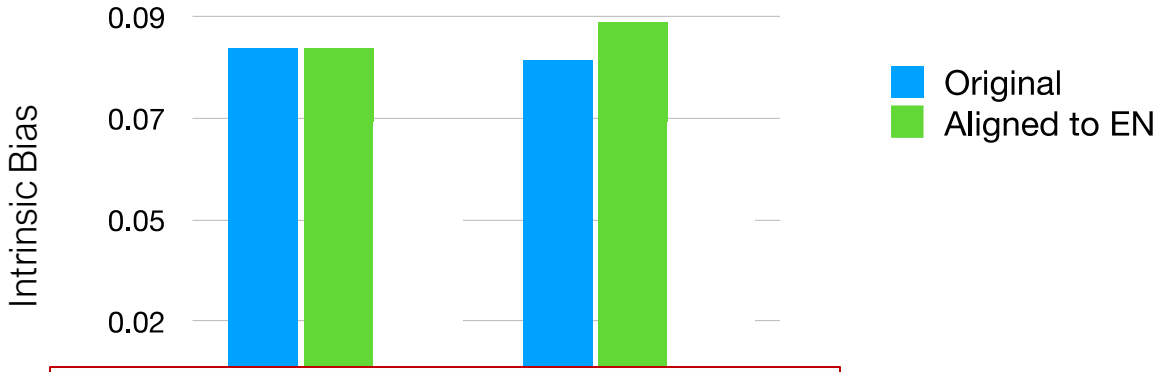
W/ BERT,
bias exists in both monolingual portion and cross-lingual transfer learning.

- Cross-lingual Transfer Learning
  EN → ES



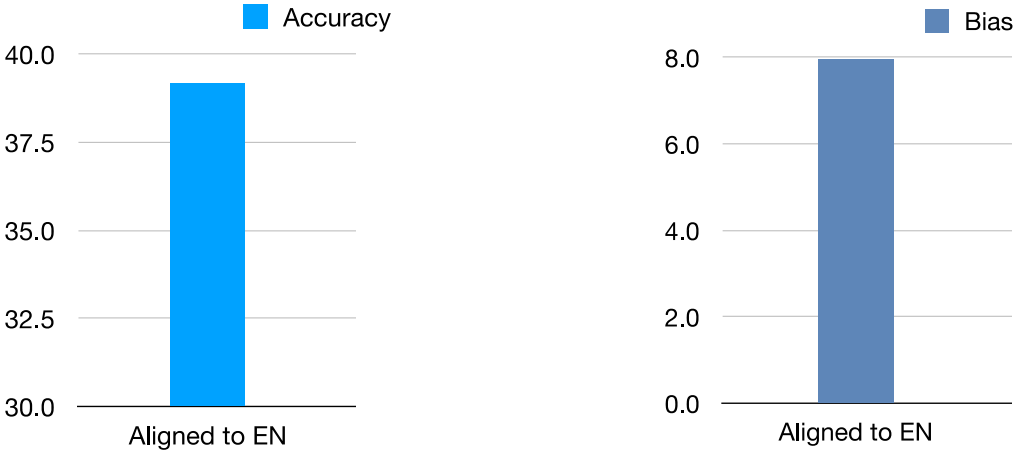
Bias Mitigation - Aligning to a Debiased Space

- Intrinsic Bias



Reduces the bias but cannot completely remove that.

- Extrinsic Bias (EN→ ES)



Microsoft

UCLA NLP

Conclusion

- Goal: To understand bias in multilingual word embeddings
- Resources: New datasets for intrinsic and extrinsic bias analysis
- Key Takeaways:
  - Bias commonly exists in different languages
  - Different alignment target spaces affect the bias in both intrinsic and extrinsic perspectives
  - Using BERT, improve performance but intrinsic and extrinsic biases still exist
- Bias Mitigation
  - Choosing a specific alignment target space (e.g. ES, ENDEB) helps
  - But new methods are wanted

- https://github.com/MSR-LIT/MultilingualBias