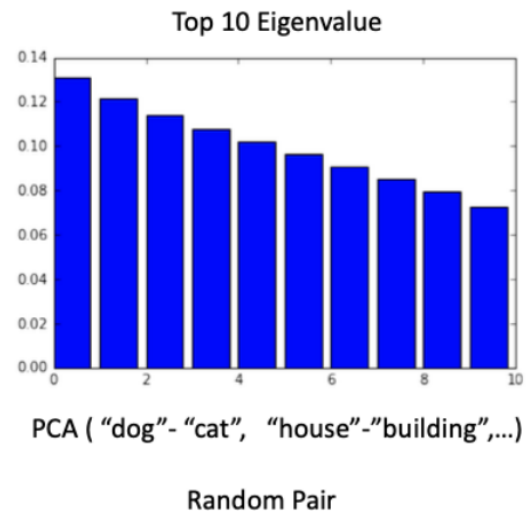
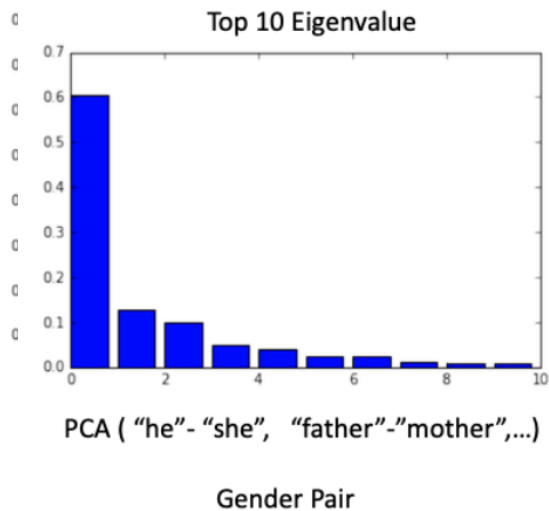
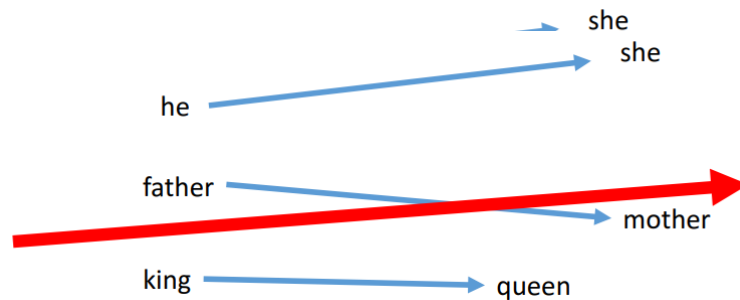


Fairness and Bias in NLP- Part 1

Dr. Debashish Das

What We Will Cover?

- A Cartoon of ML (NLP) Pipeline
- Motivate Example: Conference Resolution
- Wino-Bias Data
- Gender bias in Coref System
- Misrepresentation and Bias Stereotypes
- Bias in Wikipedia
- Bias in Language Generation
- Representational Harm in NLP
- Implicit association test (IAT)
- Word Embedding Association Test (WEAT)
- Beyond Gender & Race/Ethnicity Bias
- Linear Discriminative Analysis (LDA)
- Unequal Treatment of Gender
- Biases in NLP Classifiers/Taggers
- Control Biases: Debiasing, Data Augmentation



[illegible]

Can we Extend the Analysis beyond Binary Gender?

Beyond Gender & Race/Ethnicity Bias

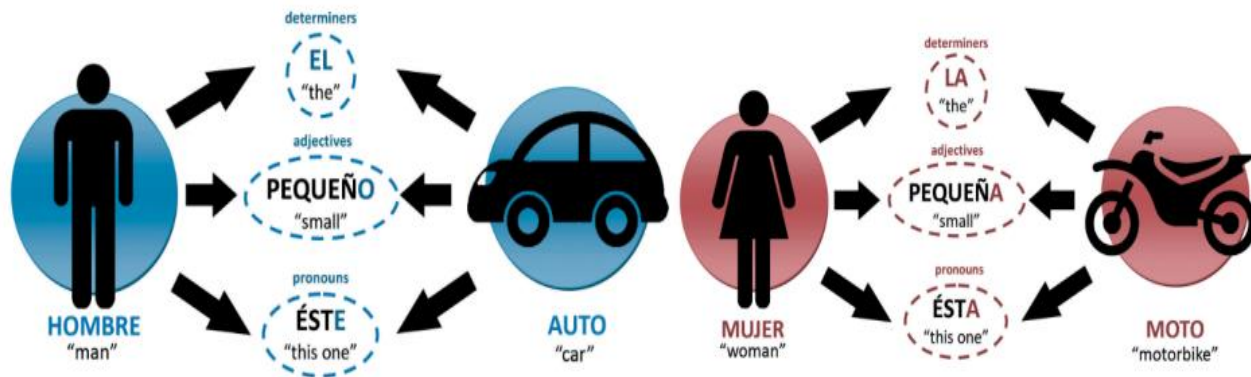
| Racial Analogies | |
|----------------------------|----------------------------|
| black → homeless | caucasian → servicemen |
| caucasian → hillbilly | asian → suburban |
| asian → laborer | black → landowner |
| Religious Analogies | |
| jew → greedy | muslim → powerless |
| christian → familial | muslim → warzone |
| muslim → uneducated | christian → intellectually |

Biases in word embeddings trained on the Reddit data from US users.

How about other Embedding?

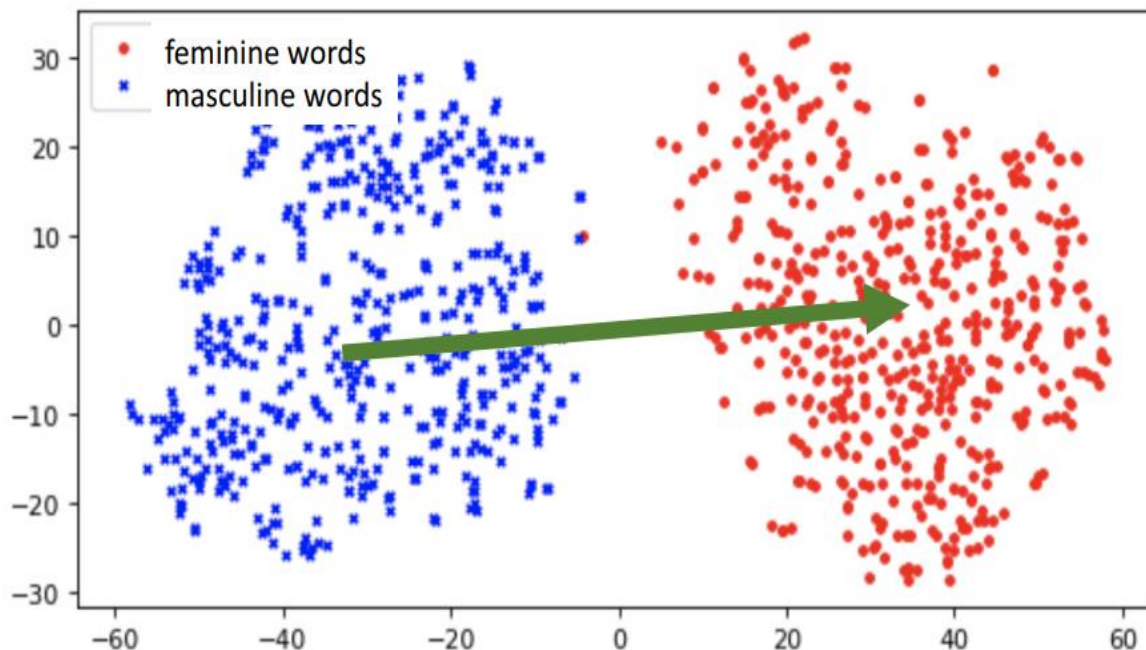
How about other Embedding?

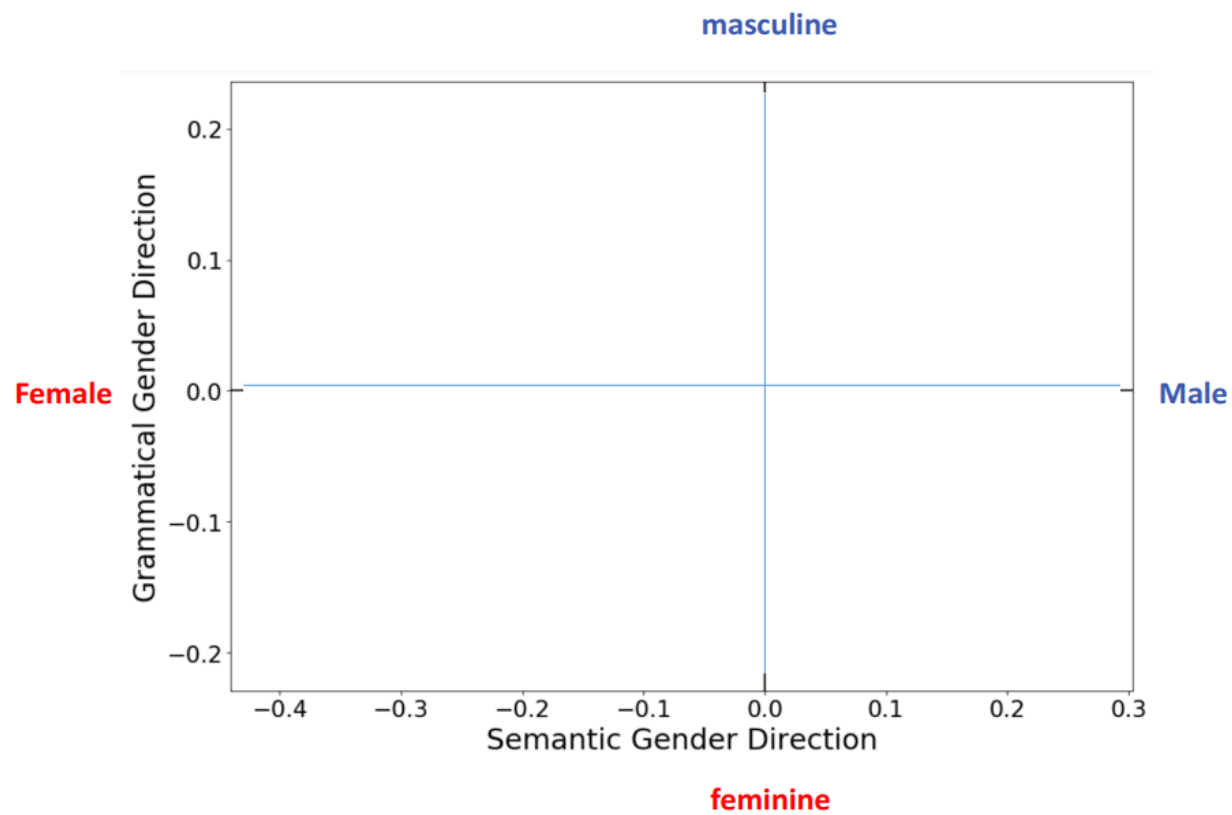
- Language with grammatical gender
- Morphological agreement

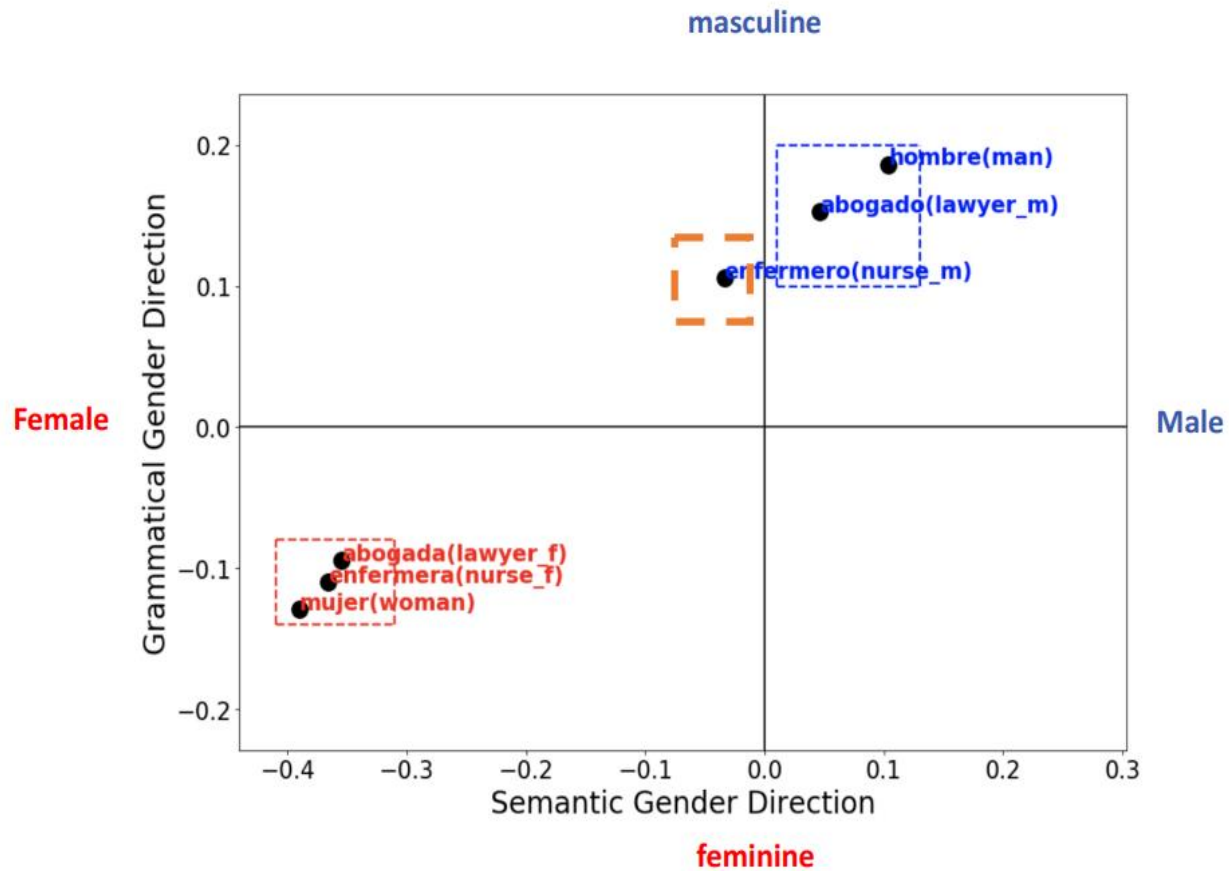


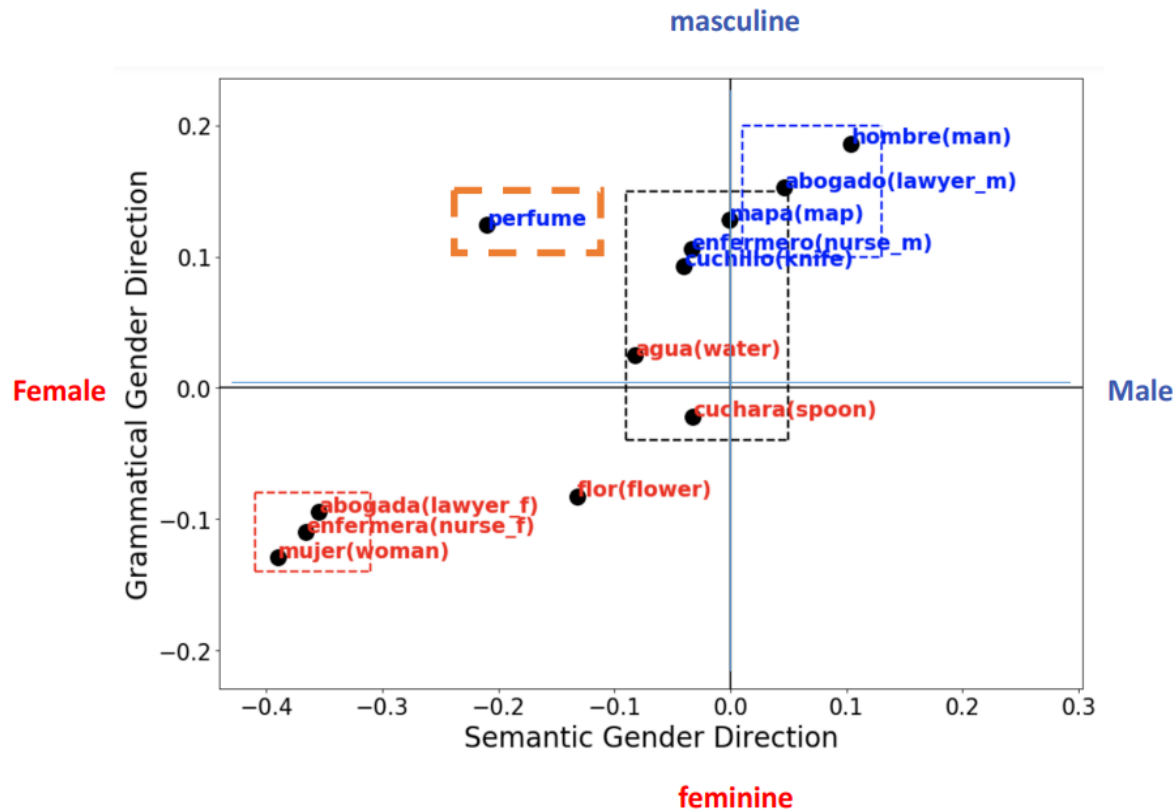
Linear Discriminative Analysis (LDA)

- Identify grammatical gender direction

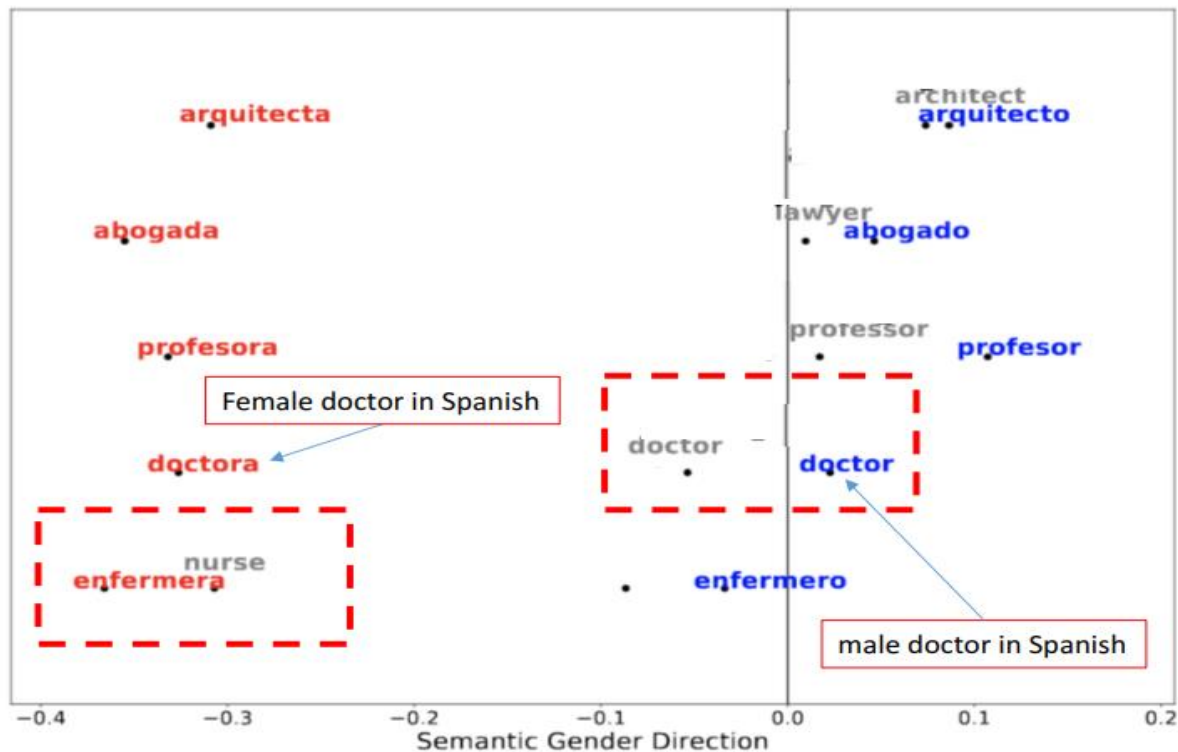








How about bilingual embedding?



(Zhou et al, EMNLP 2019)

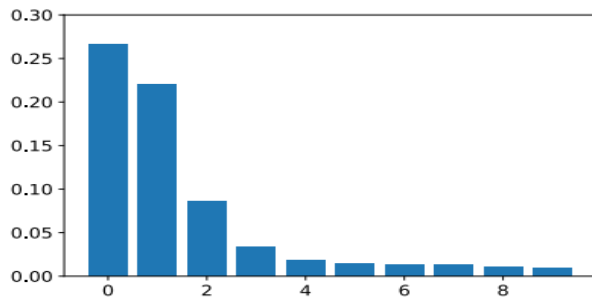
How about Contextualized Representation?

- Gender Bias in Contextualized Word Embeddings

(Feminine) The driver stopped the car at the hospital because she was paid to do so

(Masculine) The driver stopped the car at the hospital because he was paid to do so


gender direction: $\text{ELMo}(\text{driver}) - \text{ELMo}(\text{driver})$

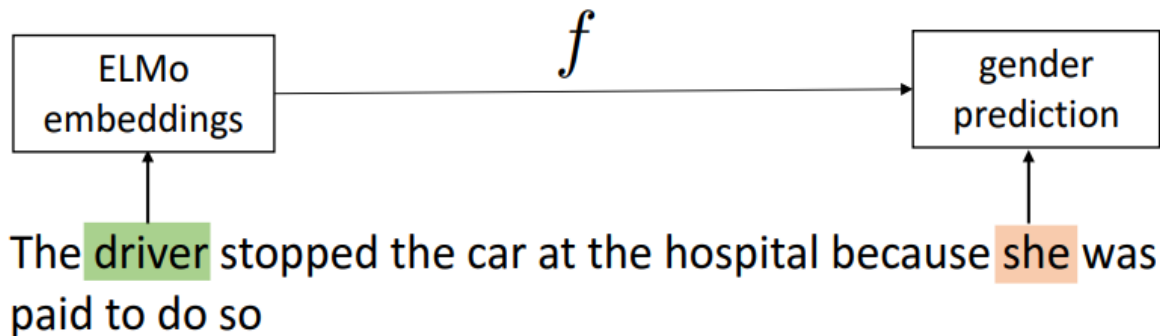


(Zhou et al, EMNLP 2019)

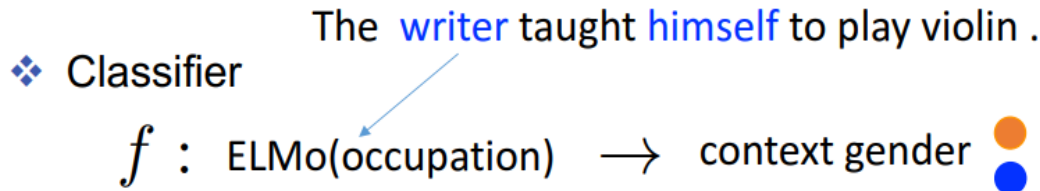
Unequal Treatment of Gender

- Classifier

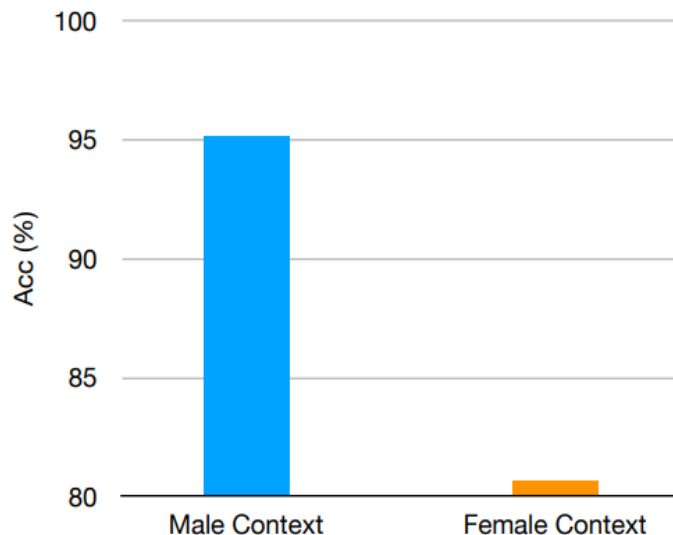
$$f : \text{ELMo}(\text{occupation}) \rightarrow \text{context gender}$$




Unequal Treatment of Gender (Cont.)

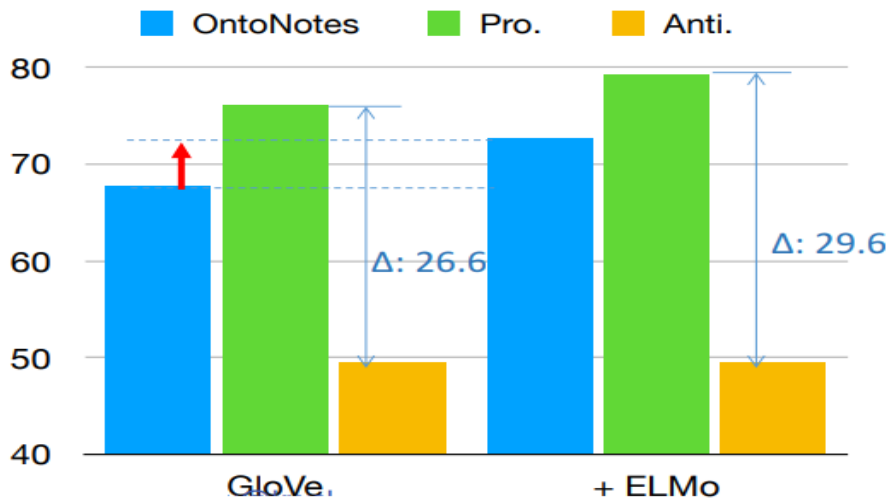


- ELMo propagates gender information to other words
- Male information is 14% more accurately propagated than female



Coreference with contextualized embedding

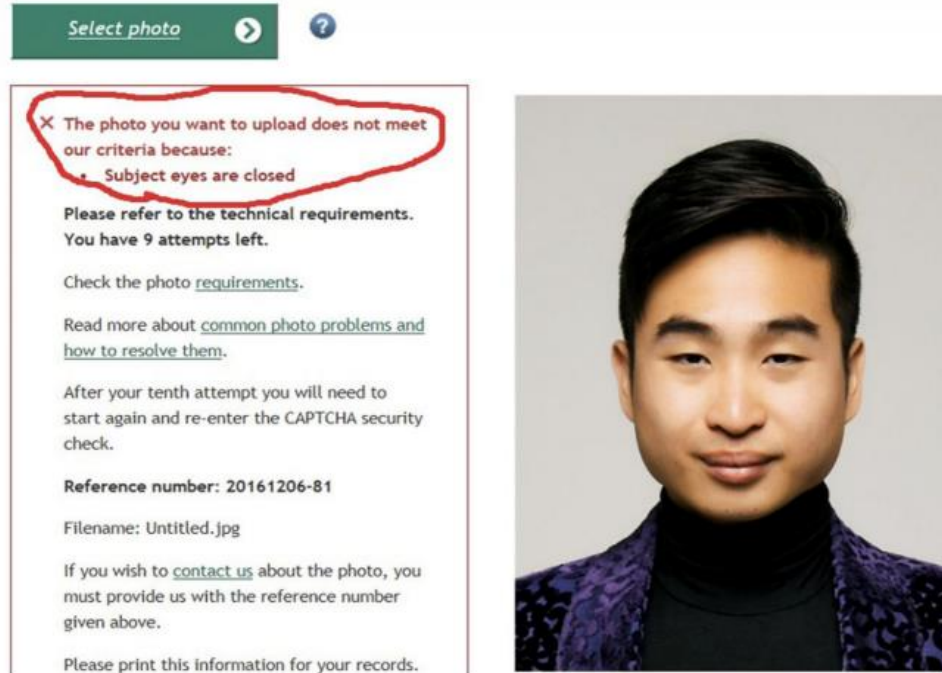
- ELMo boosts the performance
- However, enlarge the bias (Δ)



Does such Bias do “Harm” Certain People?

Biases in NLP Classifiers/Taggers

- ❖ Gender Bias in Coreference resolution
 - ❖ Zhao, Jieyu, et al. **Gender bias in coreference resolution: Evaluation and debiasing methods**. *NAACL* (2018)
 - ❖ Webster, Kellie, et al. **Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns**. *TACL* (2018)
- ❖ Gender, Race, and Age Bias in Sentiment Analysis
 - ❖ Svetlana and Mohammad. **Examining gender and race bias in two hundred sentiment analysis systems**. arXiv (2018)
 - ❖ Díaz, et al. **Addressing age-related bias in sentiment analysis**. CHI Conference on Human Factors in Comp. Systems. (2018)
- ❖ LGBTQ identity terms bias in Toxicity classification
 - ❖ Dixon, et al. **Measuring and mitigating unintended bias in text classification**. AIES. (2018)
- ❖ Gender Bias in Occupation Classification
 - ❖ De-Arteaga et al. **Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting**. FAT* (2019)
- ❖ Gender bias in Machine Translation
 - ❖ Prates, et al. **Assessing gender bias in machine translation: a case study with Google Translate**. Neural Computing and Applications (2018)

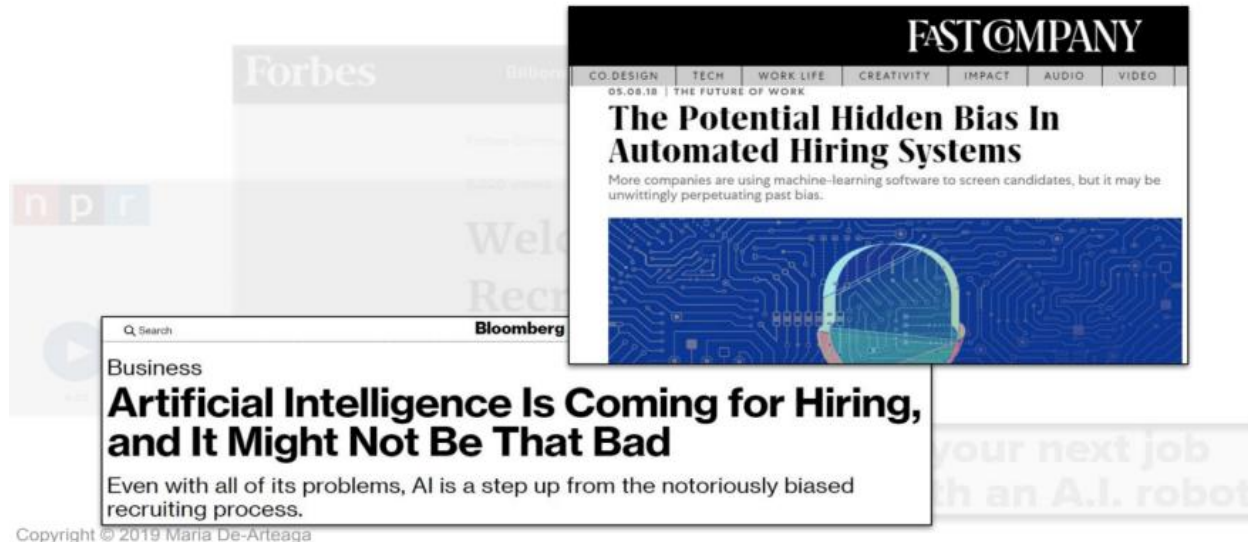


A screenshot of New Zealand man Richard Lee's passport photo rejection notice, supplied to Reuters December 7, 2016. Richard Lee/Handout via REUTERS

Towards Inclusive AI

Examples of Harm from NLP Bias

An artificially intelligent headhunter?



Copyright © 2019 Maria De-Arteaga

Prevent Allocative Harm in Sensitive Applications

Can we ~~Remove~~/Control these biases?

Towards Debiasing

1. Identify gender subspace: B
2. Identify gender-definitional (S) and gender-neutral words (N)
3. Apply transform matrix (T) to the embedding matrix (W)
 - a. Project away the gender subspace B from the gender- neutral words N
 - b. But, ensure the transformation doesn't change the embeddings too much

$$\min_T \underbrace{\|(TW)^T(TW) - W^TW\|_F^2}_{\text{Don't modify embeddings too much}} + \lambda \underbrace{\|(TN)^T(TB)\|_F^2}_{\text{Minimize gender component}}$$

T - the desired debiasing transformation
W - embedding matrix
matrix of gender neutral words

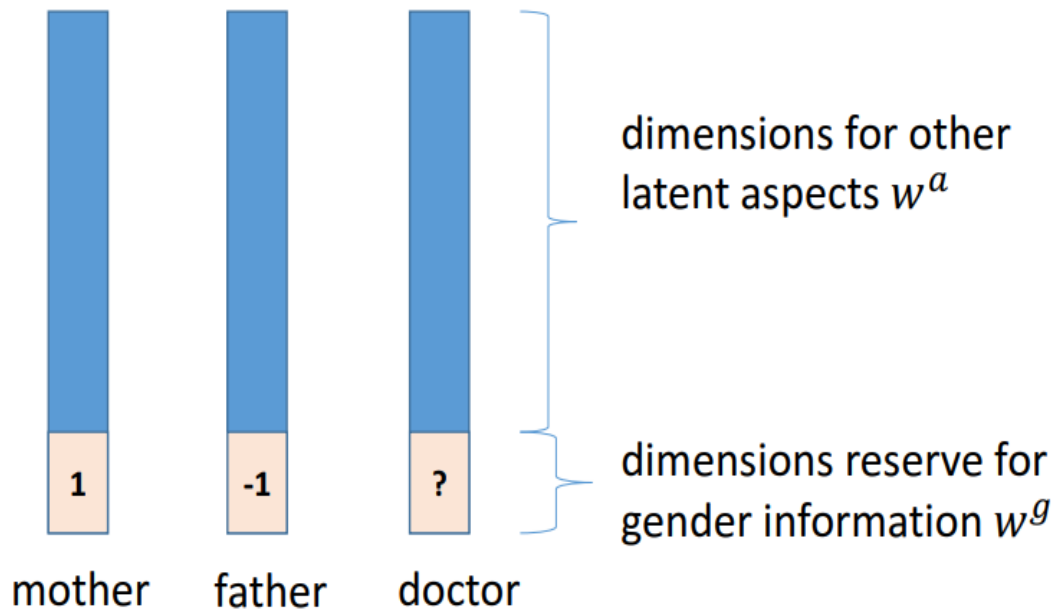
B - biased space

N - embedding

Bolukbasi et al. (2016)

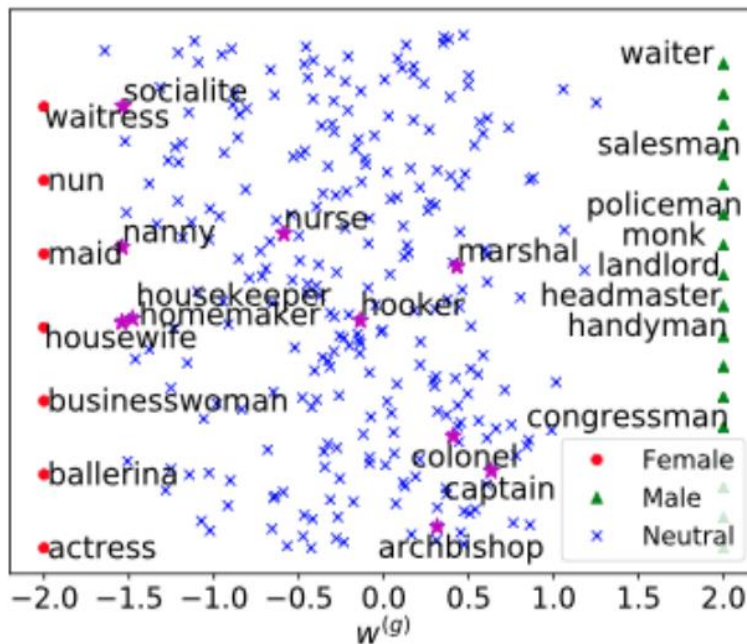
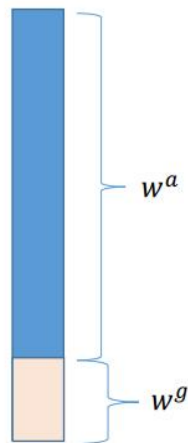
Make Gender Information Transparent in Word Embedding

- Learning Gender-Neutral Word Embeddings [Zhao et al; EMNLP18]



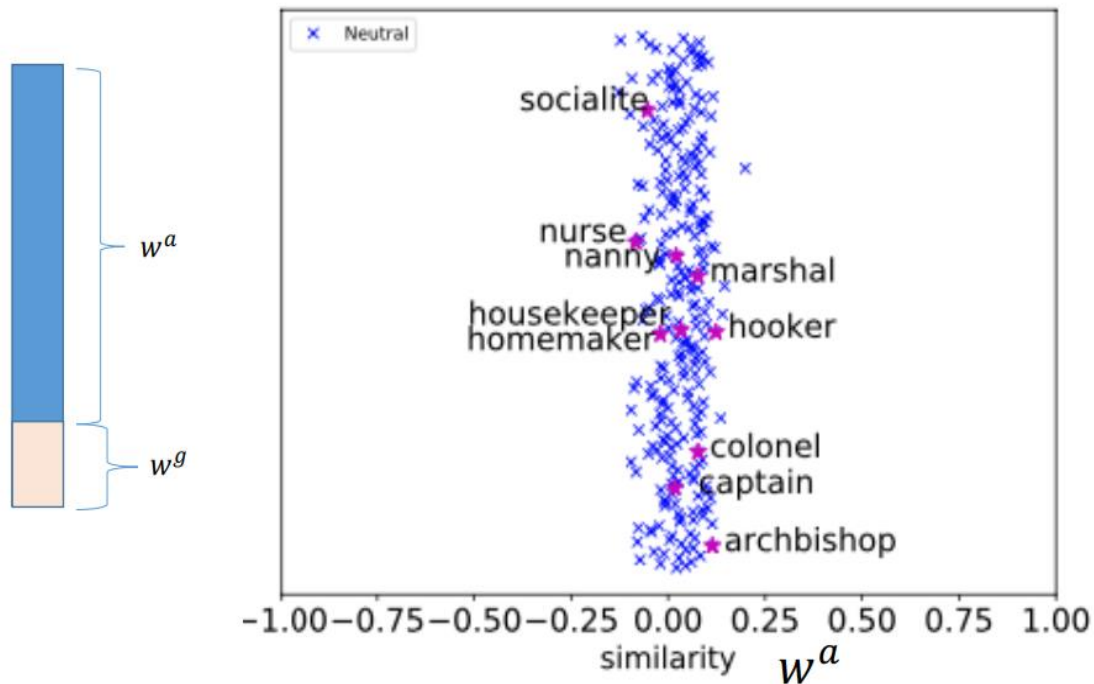
Make Gender Information Transparent in Word Embedding (Cont.)

- Learning Gender-Neutral Word Embeddings [Zhao et al; EMNLP18]



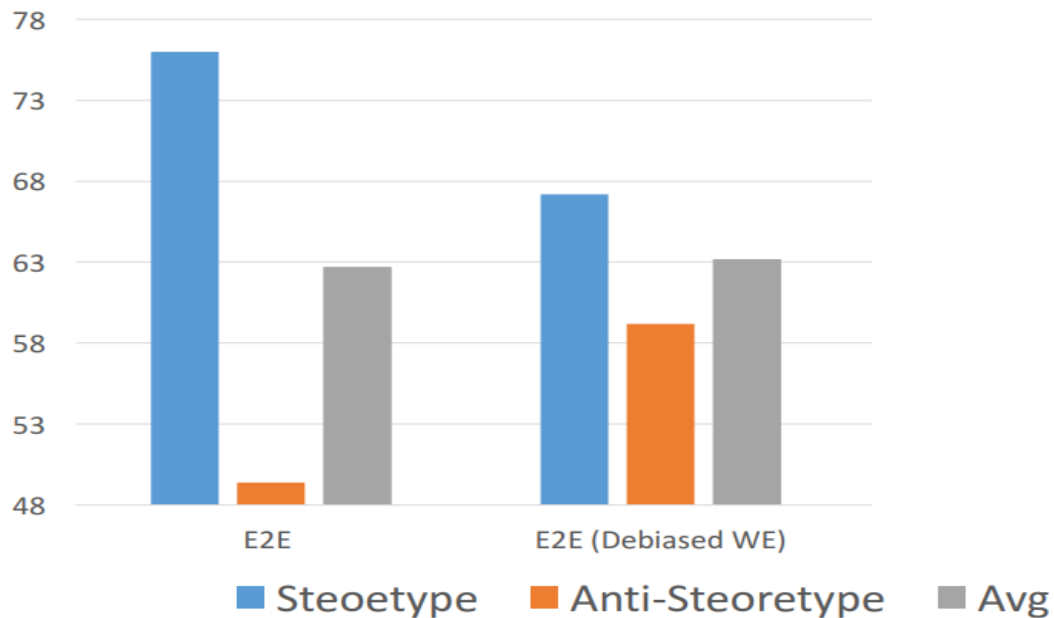
Make Gender Information Transparent in Word Embedding (Cont.)

- Learning Gender-Neutral Word Embeddings [Zhao et al: EMNLP18]



Gender bias in Coref System

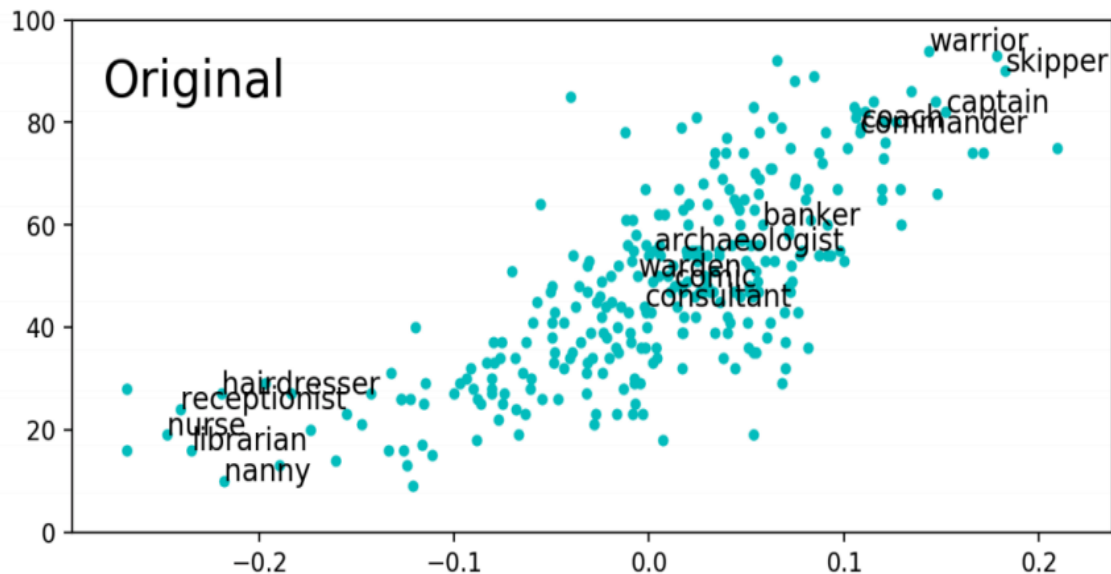
Gender bias in Coref System



Is Gender Information Actually Removed from
Embedding?

Completely removing bias is hard

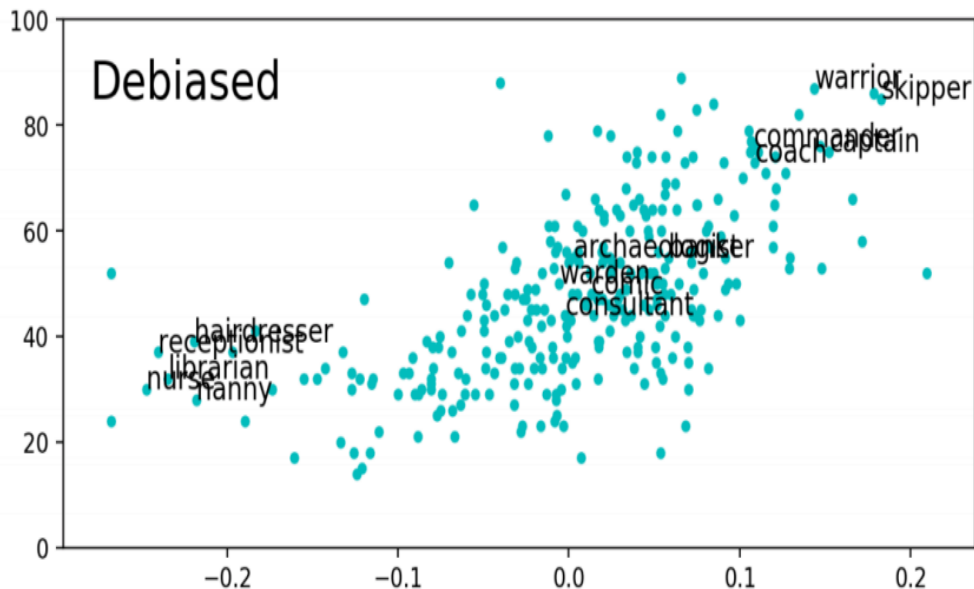
- Gonen, et al. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.** NAACL (2019).



Number of male neighbors for each occupation x-axis: original bias

Completely removing bias is hard (Cont.)

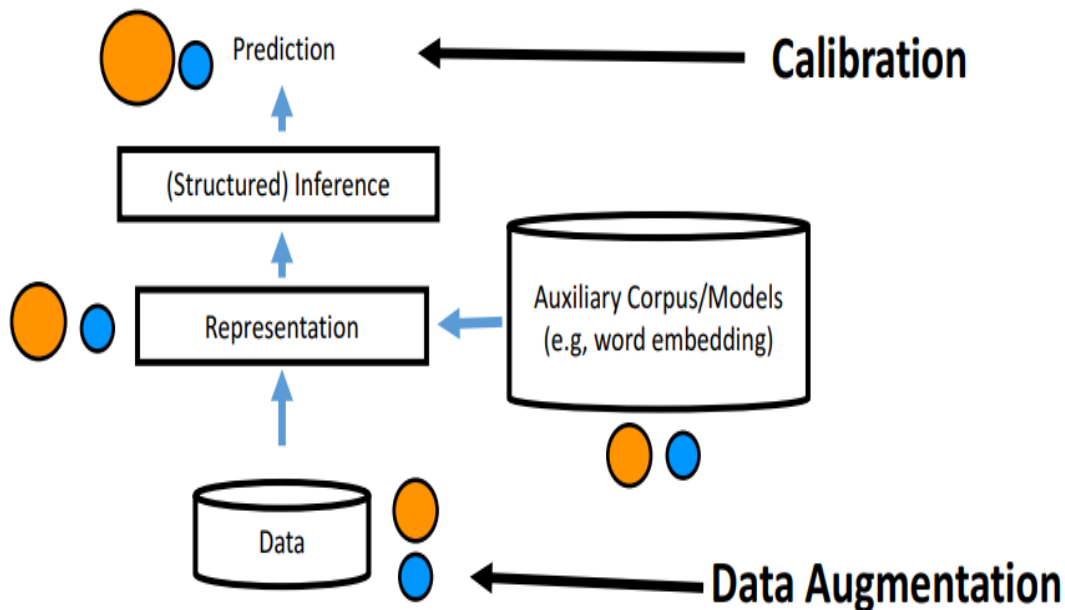
- Gonen, et al. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.** NAACL (2019).



Number of male neighbors for each occupation x-axis: original bias

Should We Debias Word Embedding?

- Awareness is better than blindness (Caliskan et. al. 17)



Wino-Bias Data

– Stereotypical Dataset:

The physician hired the secretary because he was overwhelmed with clients.




The physician hired the secretary because she was highly recommended.



– Anti-Stereotypical Dataset:

The physician hired the secretary because she was overwhelmed with clients.

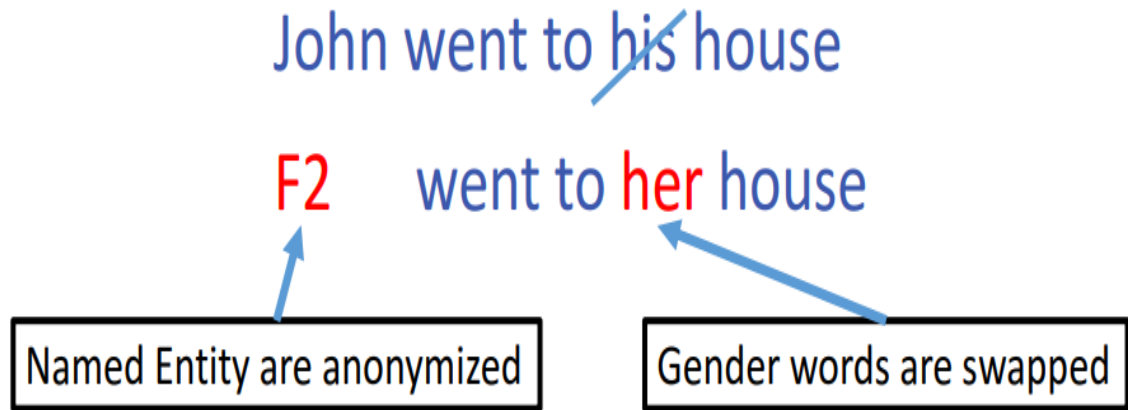


The physician hired the secretary because he was highly recommended.



Data Augmentation-- Balance the data

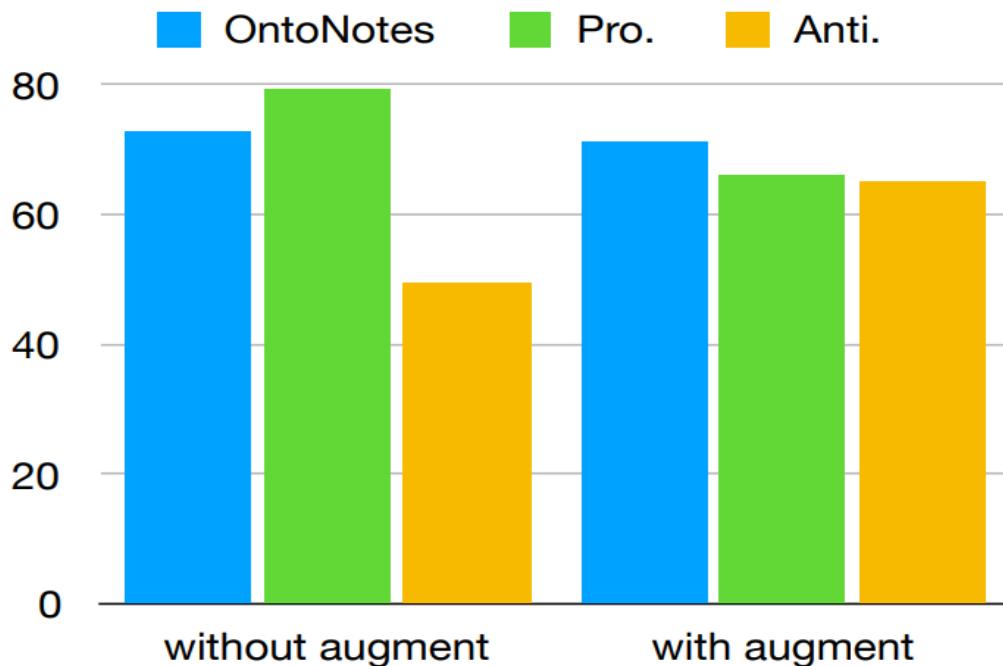
- Gender Swapping -- simulate sentence in opposite gender



Better than down/up sampling

This idea has been used in computer vision as well

Reduce Bias via Data Augmentation in Coreference Resolution



Various Biases are embedded in NLP models

Controlling Biases is still an open problem

Further Readings

- Subramonian, A. (2021, June). Fairness and Bias Mitigation: A practical guide into the AllenNLP Fairness module (<https://guide.allennlp.org/fairness>)
- Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., ... & Van Der Wal, O. (2022, March). You Reap What You Sow:
- On the Challenges of Bias Evaluation Under Multilingual Settings. (<https://aclanthology.org/2022.bigscience-1.3.pdf>)
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.W. and Wang, W.Y., 2019. Mitigating gender bias in natural language processing: Literature review. arXiv preprint arXiv:1906.08976.
- Linguistics 575: Societal Impacts of NLP (https://faculty.washington.edu/ebender/2021_575/)
- Blodgett, S.L., Barocas, S., Daumé, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. ACL.

Further Readings

- Chang, K.W., Ordonez, V., Mitchell, M., Prabhakaran, V (2019). Tutorial: Bias and Fairness in Natural Language Processing. EMNLP 2019.
- Rathore, A., Dev, S., Phillips, J.M., Srikumar, V., Zheng, Y., Yeh, C., Wang, J., Zhang, W., & Wang, B. (2021). VERB: Visualizing and Interpreting Bias Mitigation Techniques for Word Representations. ArXiv, abs/2104.02797.
- Dev, S., Sheng, E., Zhao, J., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Peng, N., & Chang, K. (2021). What do Bias Measures? ArXiv, abs/2108.03362.
- Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J.M., & Chang, K. (2021). Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. ArXiv, abs/2108.12084.