UCLA NLP

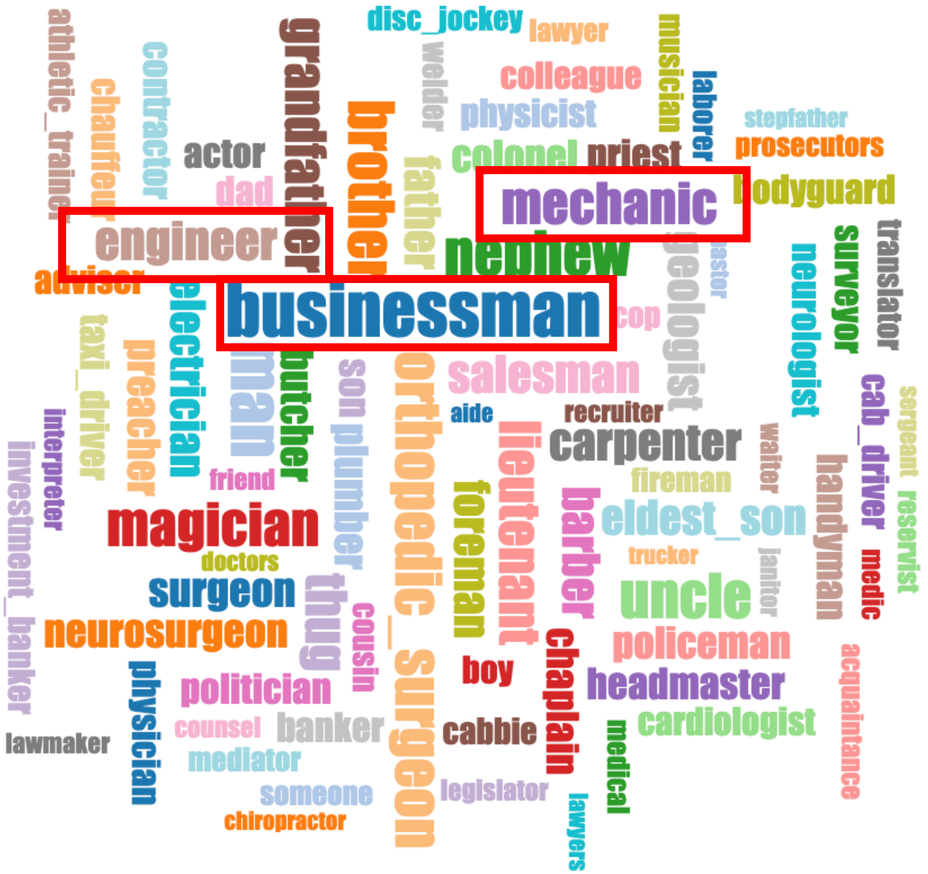
Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[1], Tianlu Wang[2], Mark Yatskar[3], Ryan Cotterell[4], Vicente Ordonez[2], Kai-Wei Chang[1]

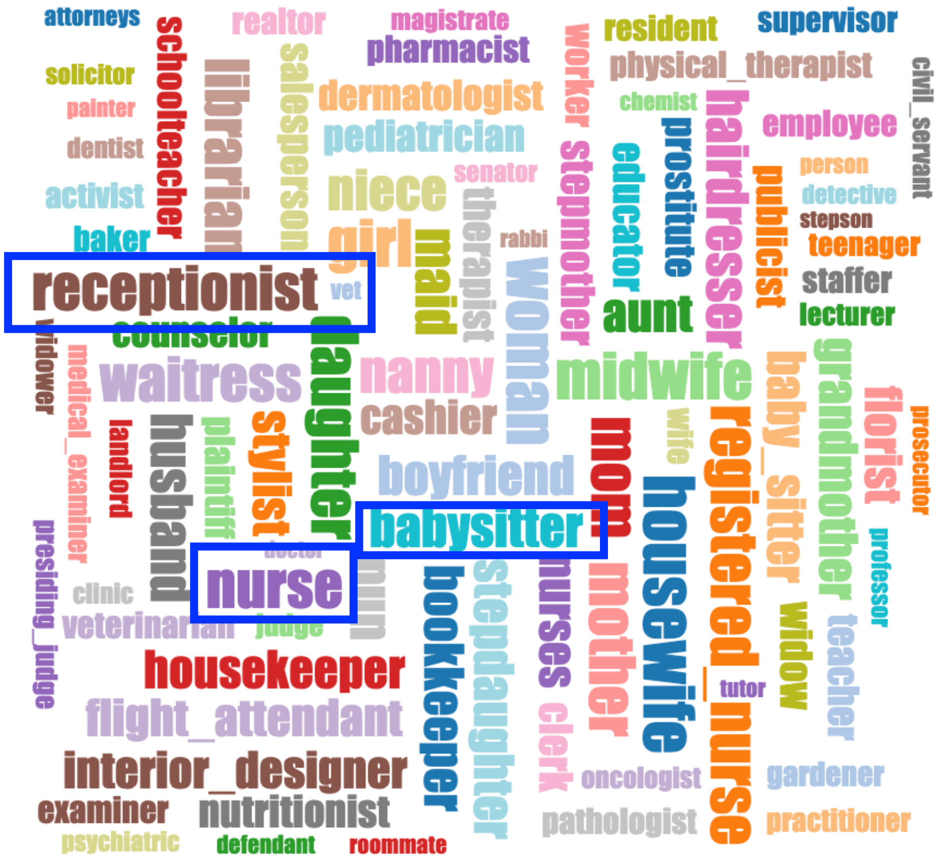
[1]UCLA, [2]University of Virginia, [3]Allen Institute for AI, [4]University of Cambridge

Bias in NLP: Word Embeddings

he



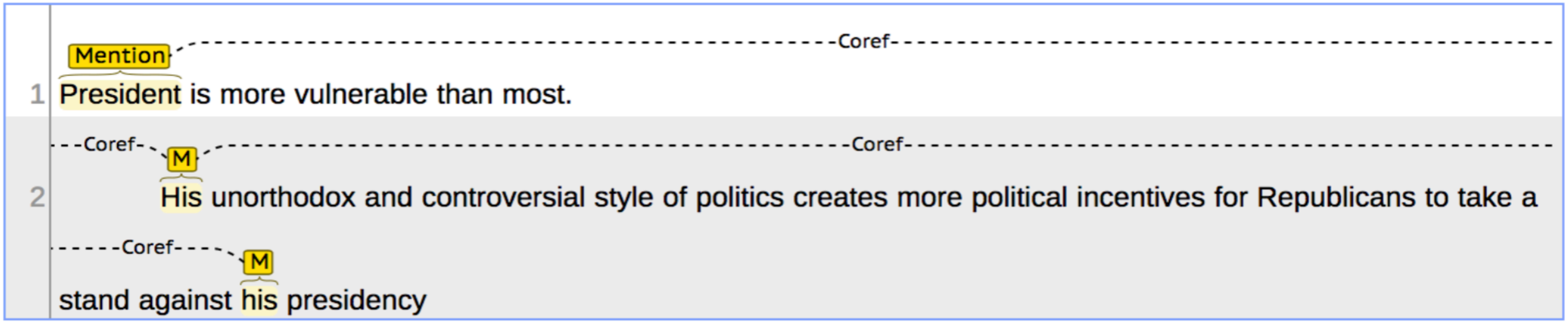
she



http://wordbias.umiacs.umd.edu/

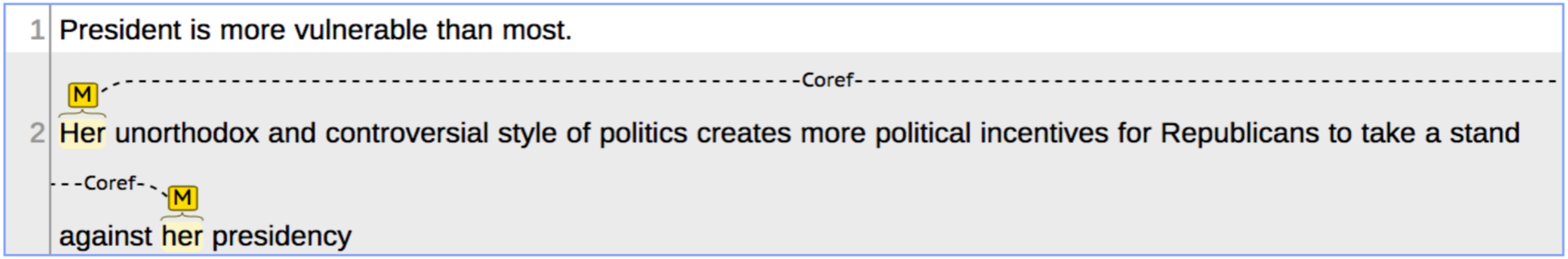
Bias in NLP: Downstream Task

- Coreference resolution is biased[1,2]
  - Model fails for "she" when given same context



1 President is more vulnerable than most.

2 His unorthodox and controversial style of politics creates more political incentives for Republicans to take a stand against his presidency



1 President is more vulnerable than most.

2 Her unorthodox and controversial style of politics creates more political incentives for Republicans to take a stand against her presidency

[1] Zhao et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL 2018

[2] Rudinger et al. Gender Bias in Coreference Resolution. NAACL 2018

Contextualized Word Embeddings

CoVe[1]

ELMo[2]

BERT[2]

Great performance improvement! 😎

Bias? 😕

[1] https://www.vecteezy.com

[2] https://www.pinclipart.com

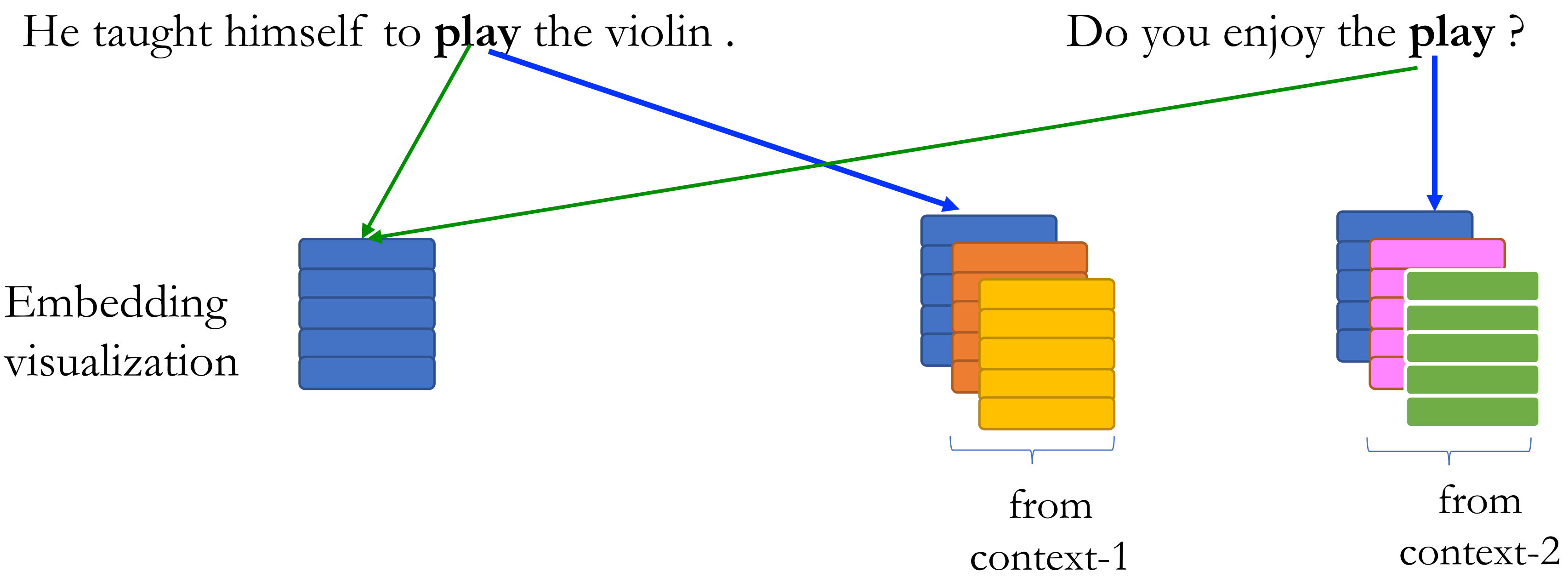
Outline - 1

- ELMo is sensitive to gender
  - Training corpus is biased
  - ELMo treats genders unequally
  - Bias propagates to downstream tasks

In this work, the analysis is in English.

Background: ELMo

- Make use of a pretrained language model
- Embed corresponding context into the representations



He taught himself to **play** the violin .

Do you enjoy the **play** ?

Embedding visualization

from context-1

from context-2

word2vec

ELMo

Bias in ELMo

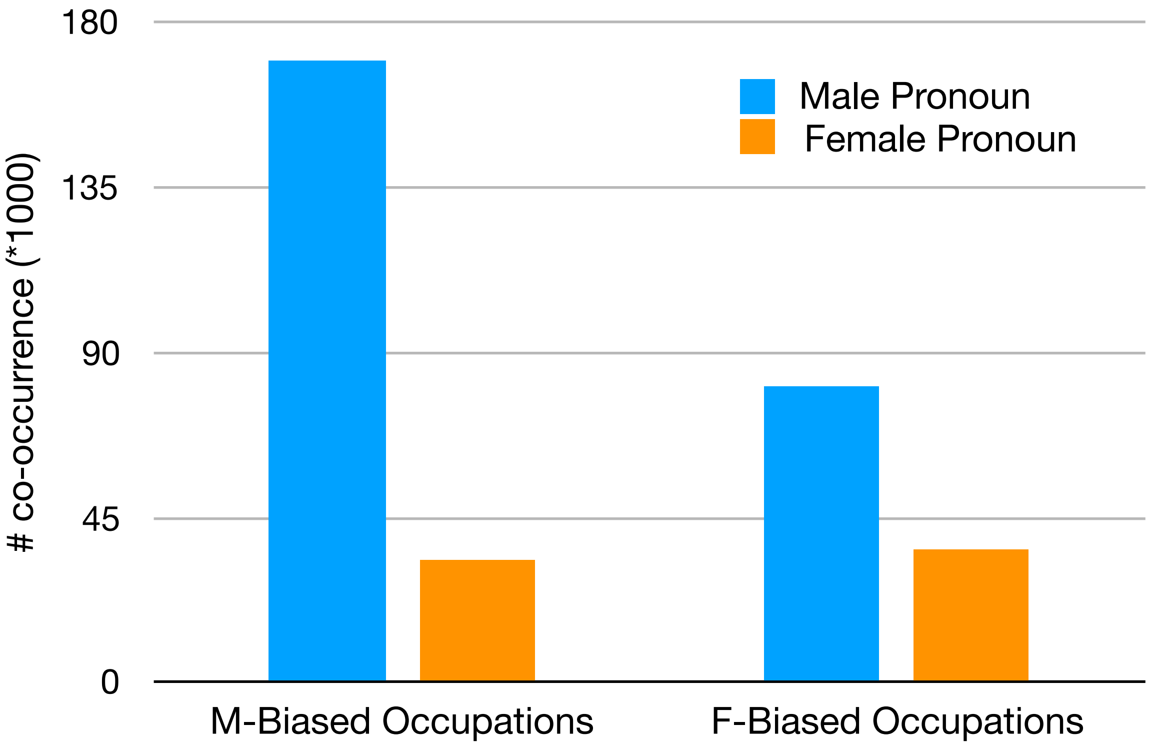
- Training Dataset Bias
  - Dataset is biased towards **man**

Gender	**Male Pronouns**	**Female Pronouns**
Occurrence (*1000)	5,300	1,600

- Male pronouns (he, him, his) occur 3 times more often than females' (she, her)

Bias in ELMo (continued)

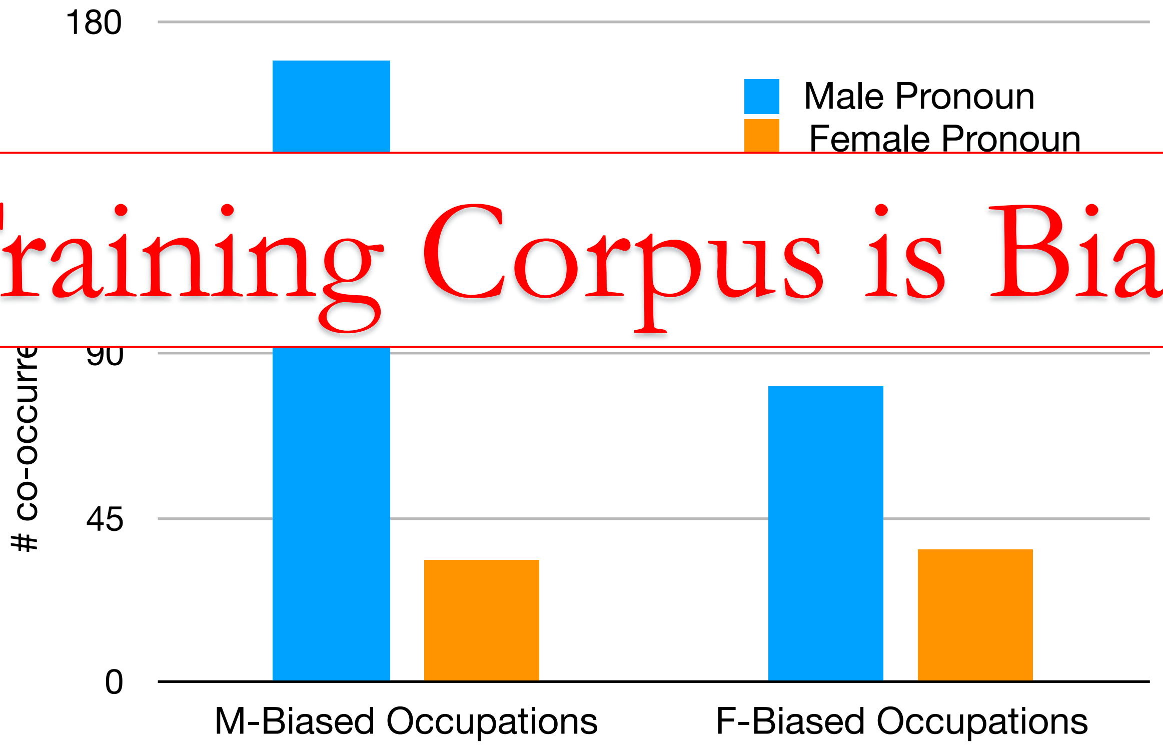
- Male pronouns co-occur more frequently with occupation words[1]



[1] Zhao et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL 2018

Bias in ELMo (continued)

- Male pronouns co-occur more frequently with occupation words[1]



Training Corpus is Biased

[1] Zhao et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL 2018

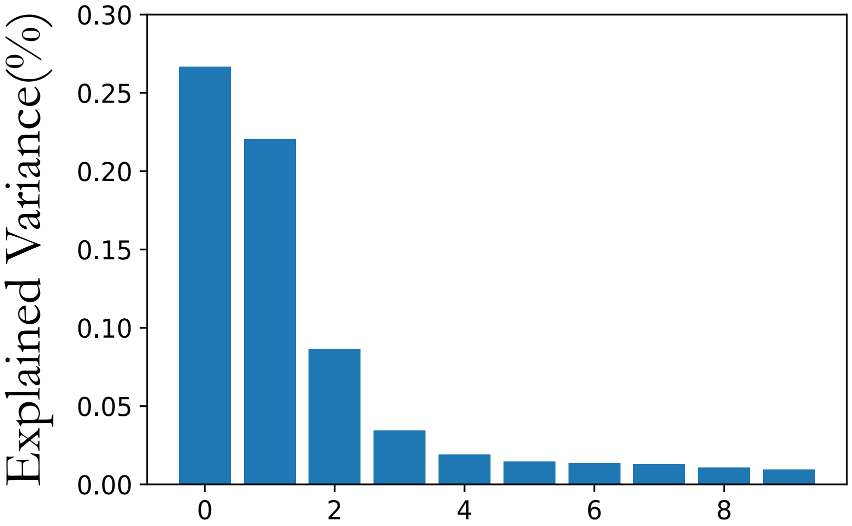
Gender Geometry in ELMo

- First two components explain more variance than others

(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

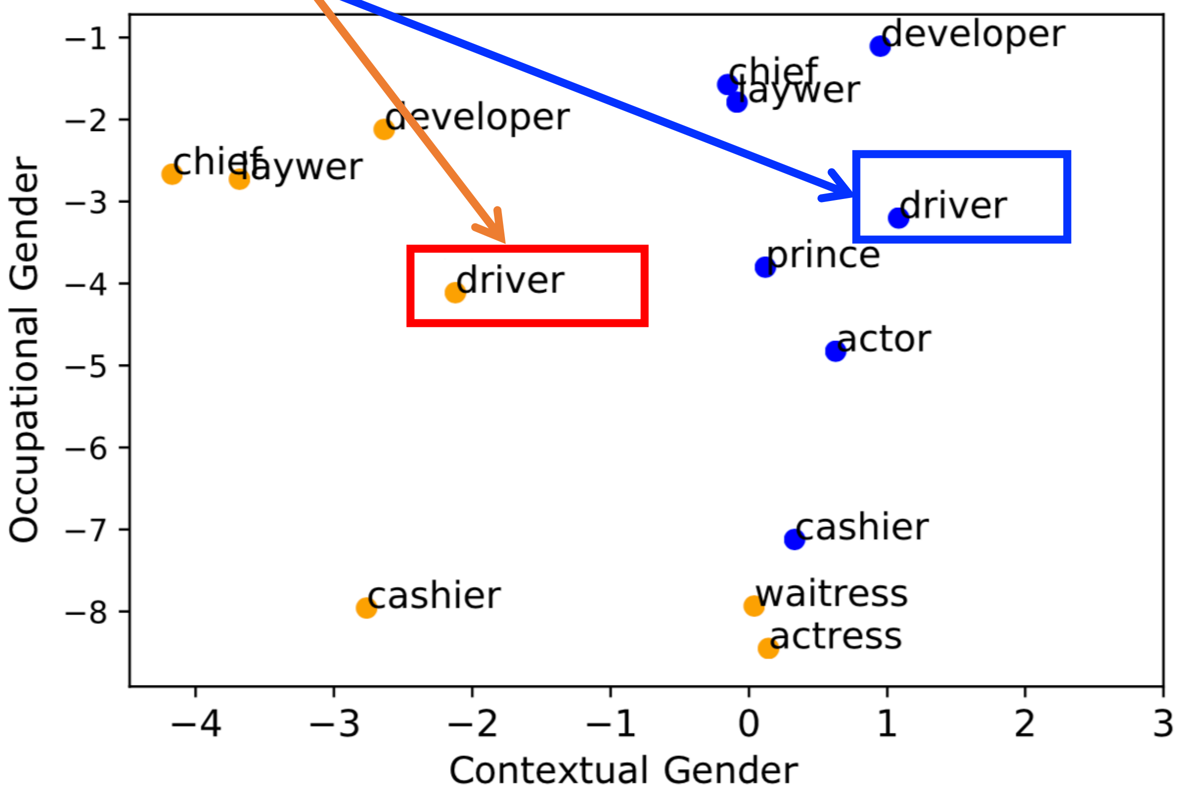
gender direction: ELMo(driver) – ELMo(driver)



Gender Geometry in ELMo

● The driver stopped the car at the hospital because **she** was paid to do so

● The driver stopped the car at the hospital because **he** was paid to do so

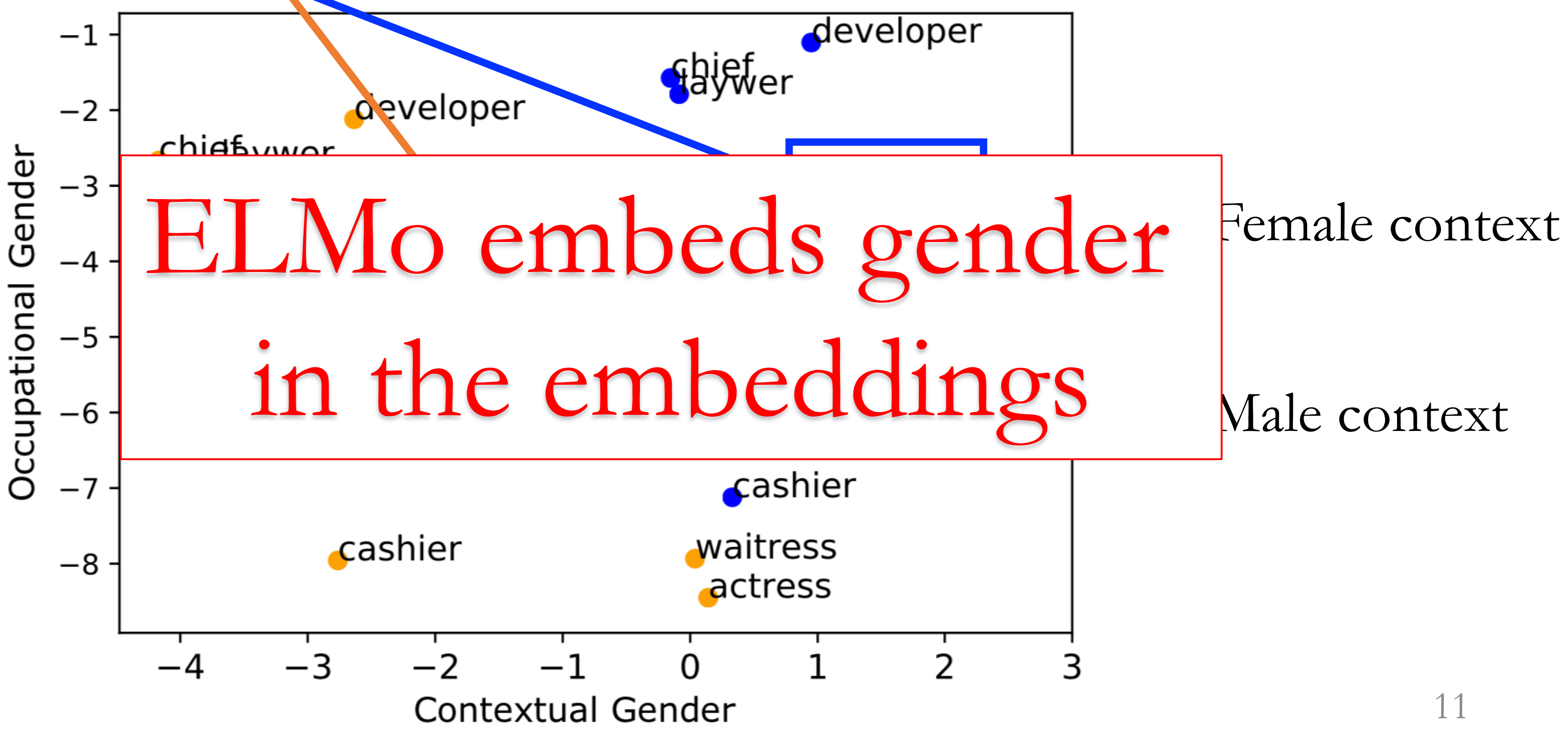


● Female context

● Male context

Gender Geometry in ELMo

- ● The driver stopped the car at the hospital because **she** was paid to do so
- ● The driver stopped the car at the hospital because **he** was paid to do so

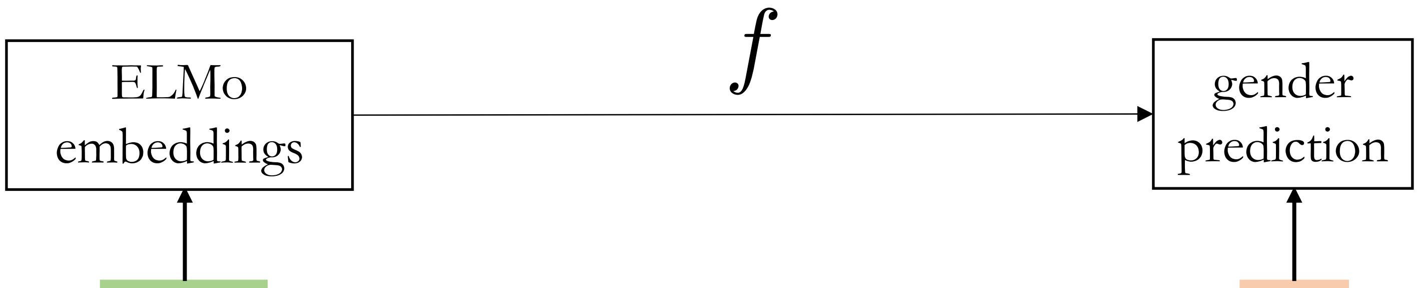


ELMo embeds gender in the embeddings

Unequal Treatment of Gender

- Classifier

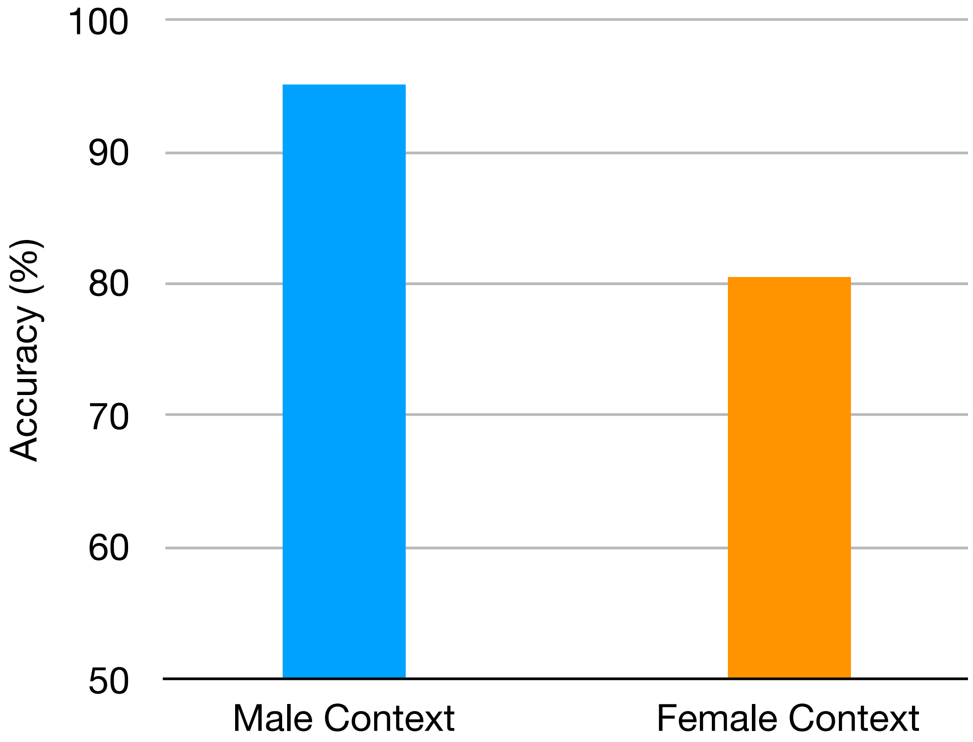
$$f : \text{ELMo(occupation)} \longrightarrow \text{context gender}$$ ● ●



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

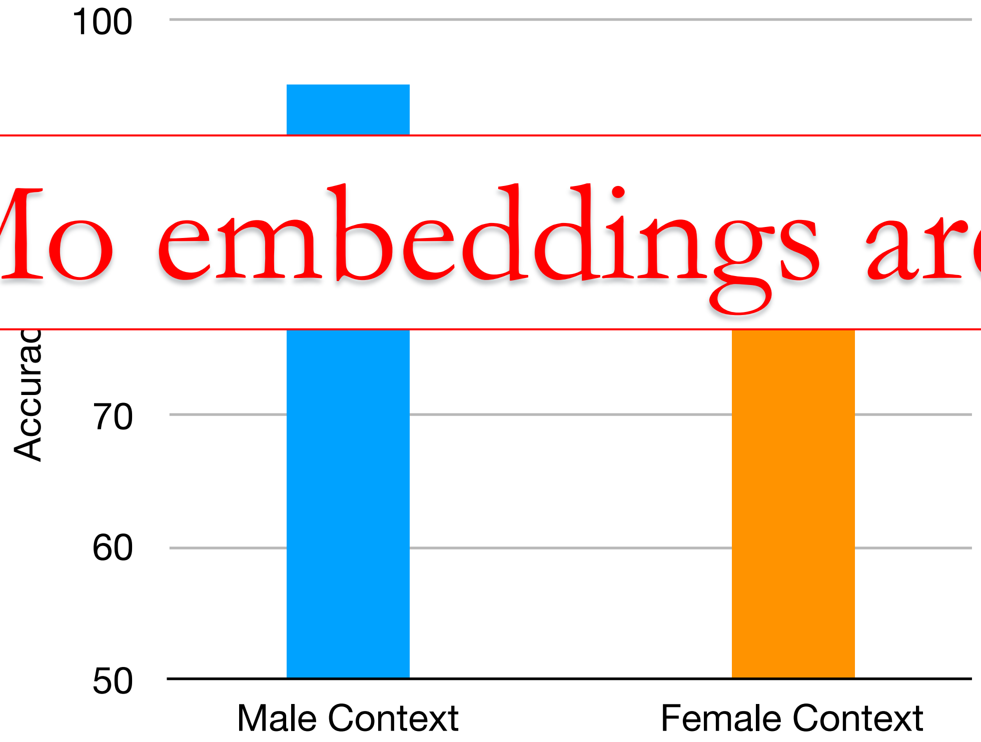
Unequal Treatment of Gender (continued)

- ELMo propagates gender information from the context
- Male information is 14% more accurately propagated than female



Unequal Treatment of Gender (continued)

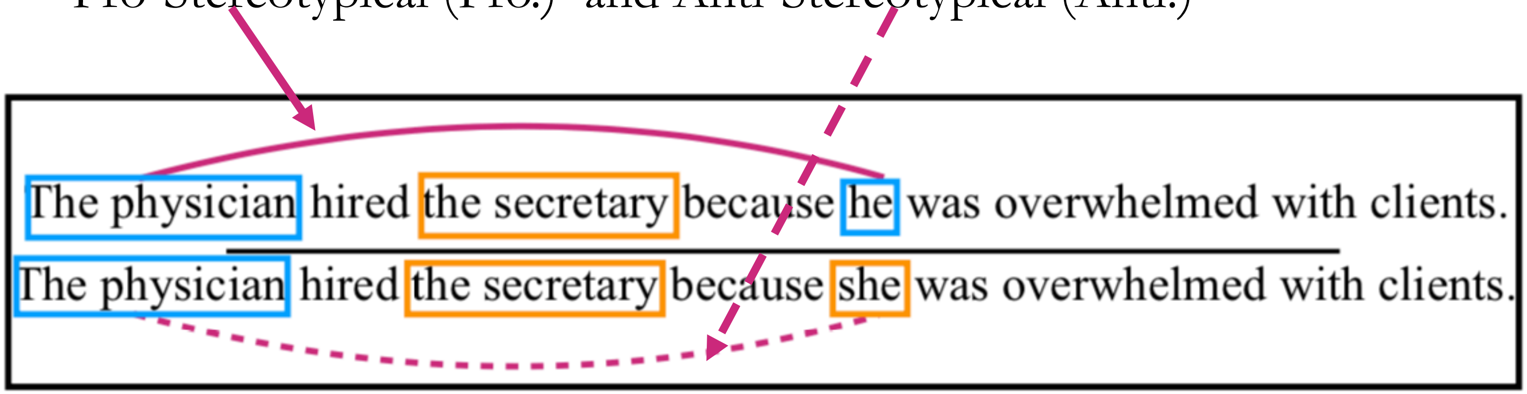
- ELMo propagates gender information from the context
- Male information is 14% more accurately propagated than female



ELMo embeddings are biased

Bias in Downstream Task: Coreference Resolution in English

- WinoBias dataset[1]
  - Pro-Stereotypical (Pro.) and Anti-Stereotypical (Anti.)



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

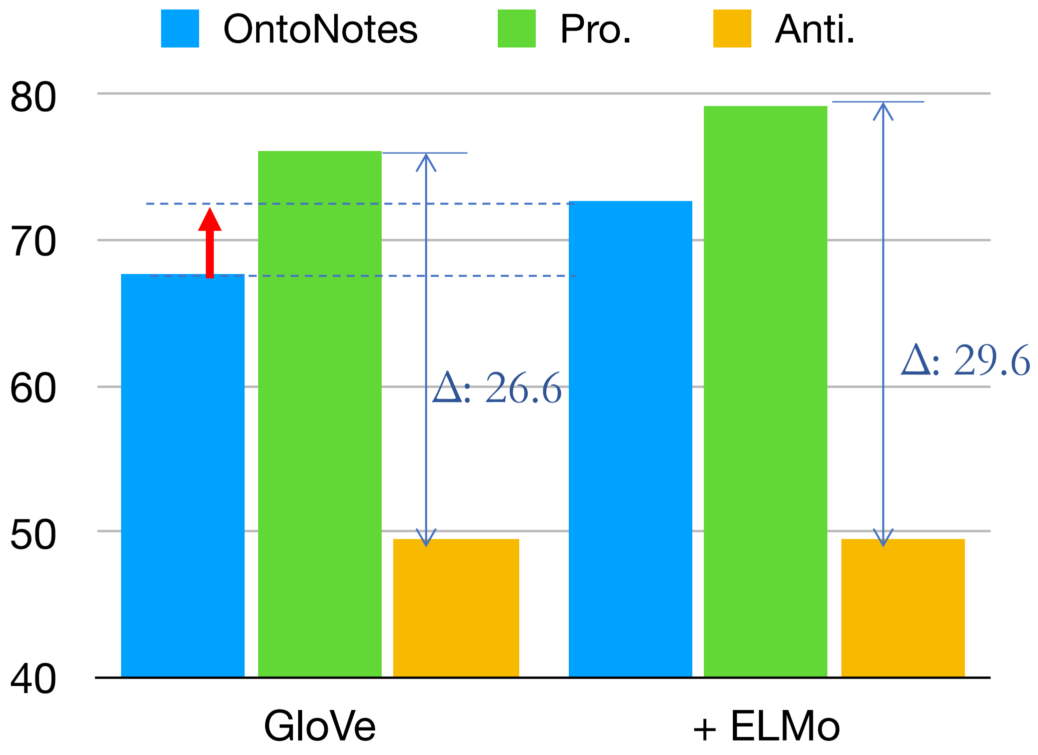
The physician hired the secretary because she was overwhelmed with clients.

- **Bias**: performance difference between Pro. and Anti. dataset.

[1] https://uclanlp.github.io/corefBias

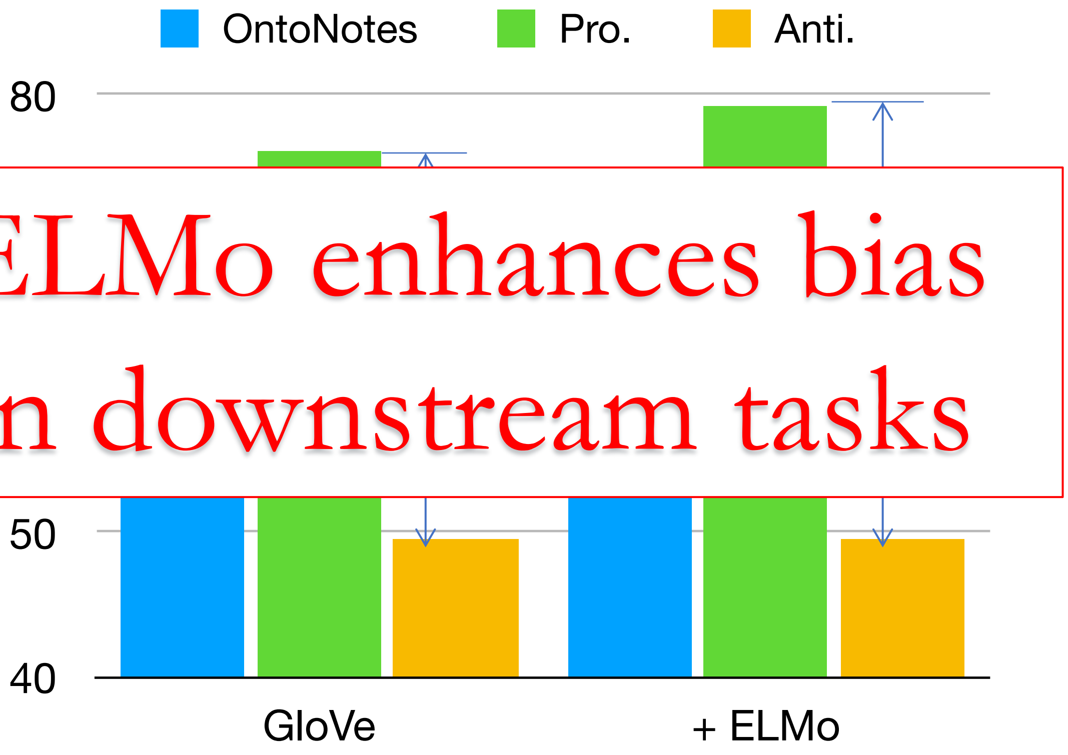
Bias in Coreference

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



Bias in Coreference

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



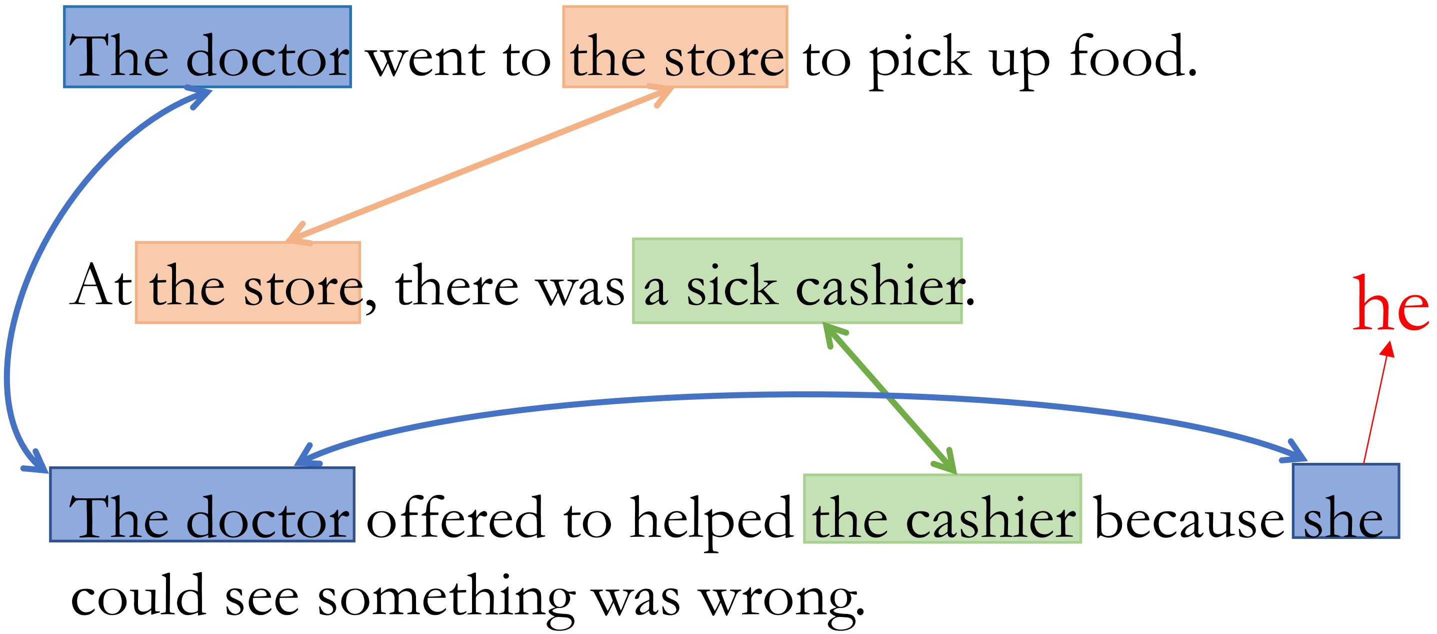
ELMo enhances bias in downstream tasks

Outline - 2

- Mitigation Bias
  - Gender swapping
  - Data augmentation
  - Neutralizing ELMo

Mitigate Bias

- Gender Swapping[1]



The doctor went to the store to pick up food.

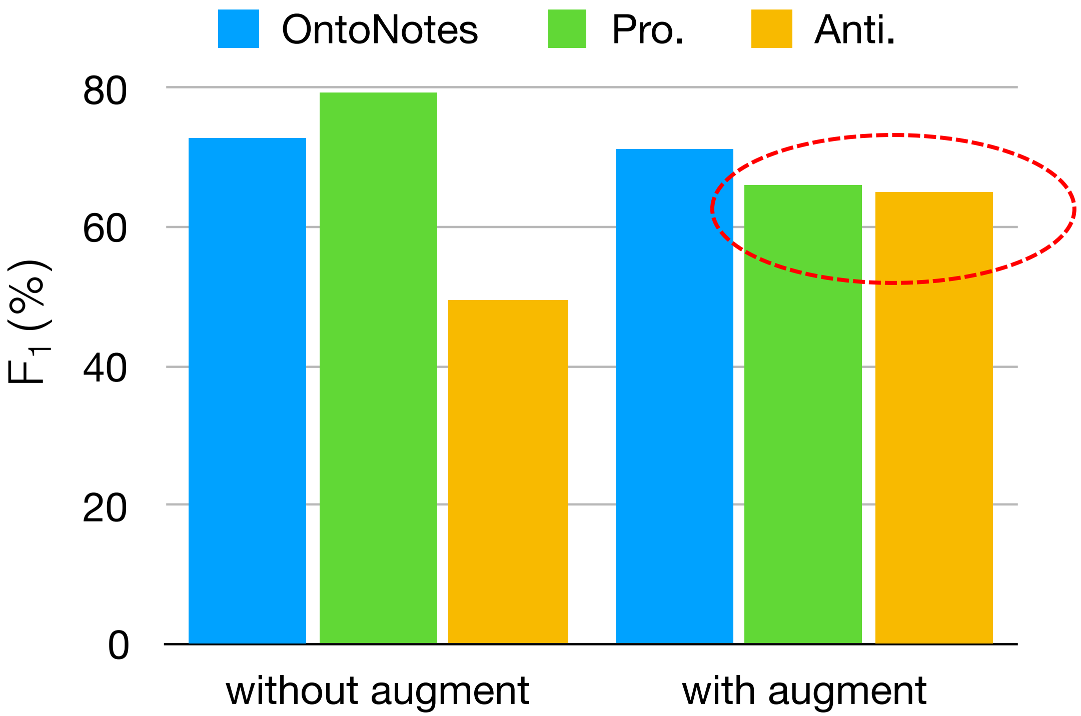
At the store, there was a sick cashier.

The doctor offered to helped the cashier because she could see something was wrong.

[1] Zhao et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL 2018

Mitigate Bias (Method 1)

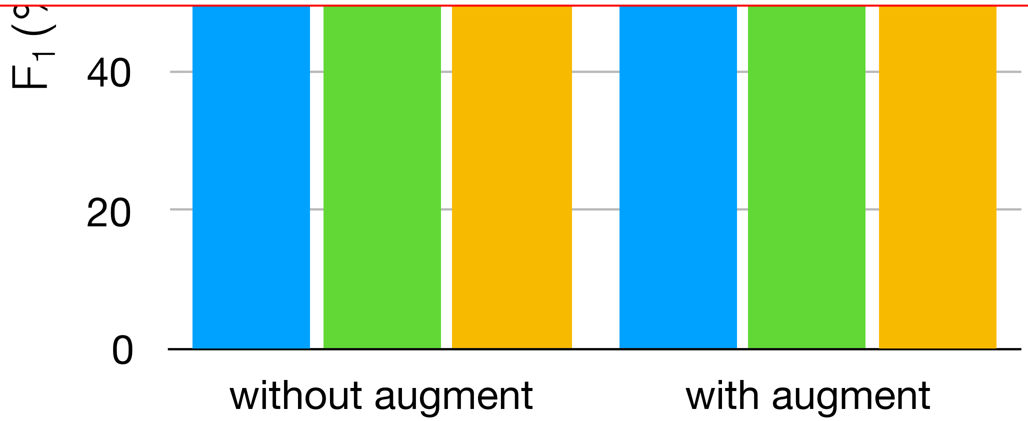
- Data Augmentation
  - Generate gender swapped training variants
  - Re-train on the union dataset
  - Almost mitigate all the bias shown in WinoBias



Mitigate Bias (Method 1)

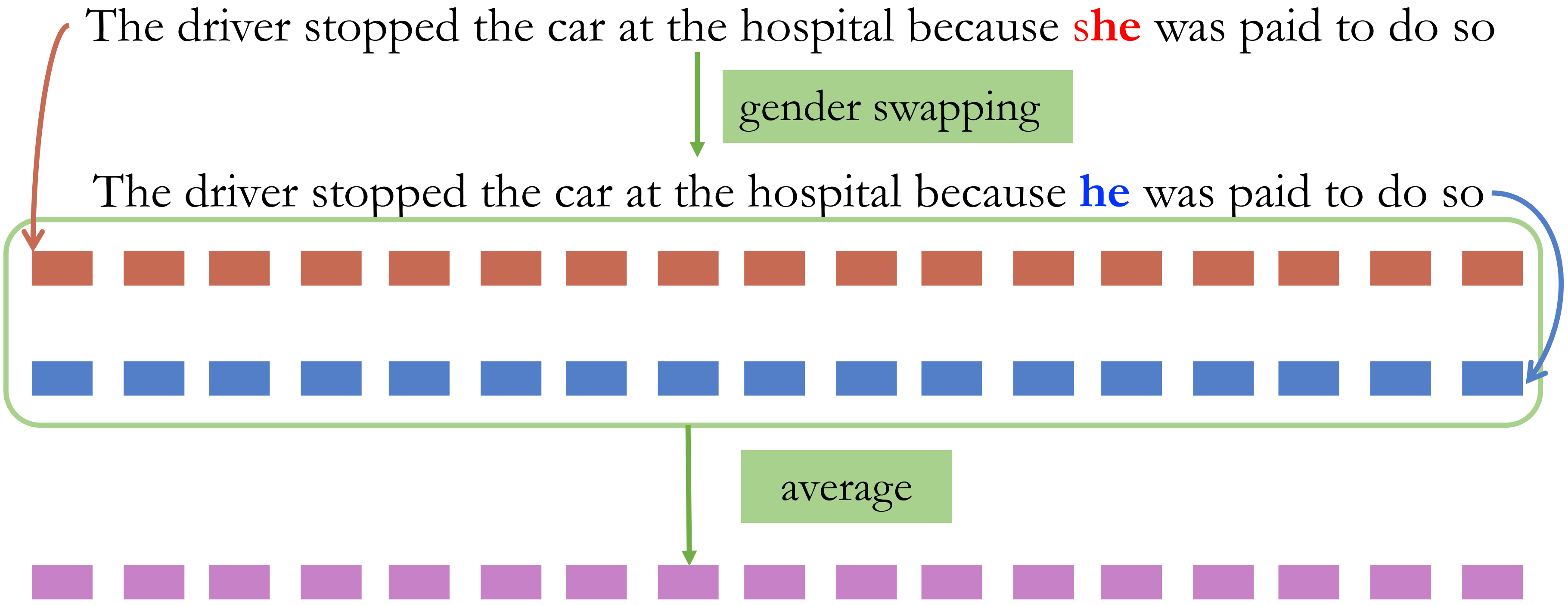
- Data Augmentation
  - Generate gender swapped training variants
  - Re-train on the union dataset
  - Almost mitigate all the bias shown in WinoBias

Data augmentation is effective.
What if we don't want to retrain?



Mitigate Bias (Method 2)

- Neutralize ELMo Embeddings
  - Average the ELMo embeddings for test dataset



The driver stopped the car at the hospital because **she** was paid to do so

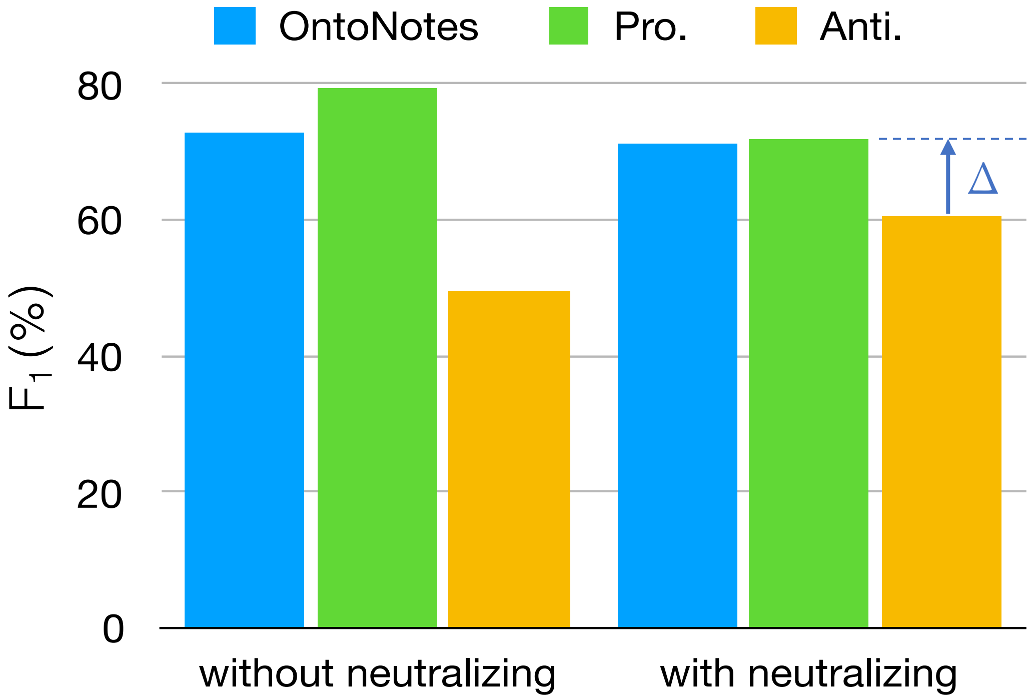
gender swapping

The driver stopped the car at the hospital because **he** was paid to do so

average

Mitigate Bias (Method 2)

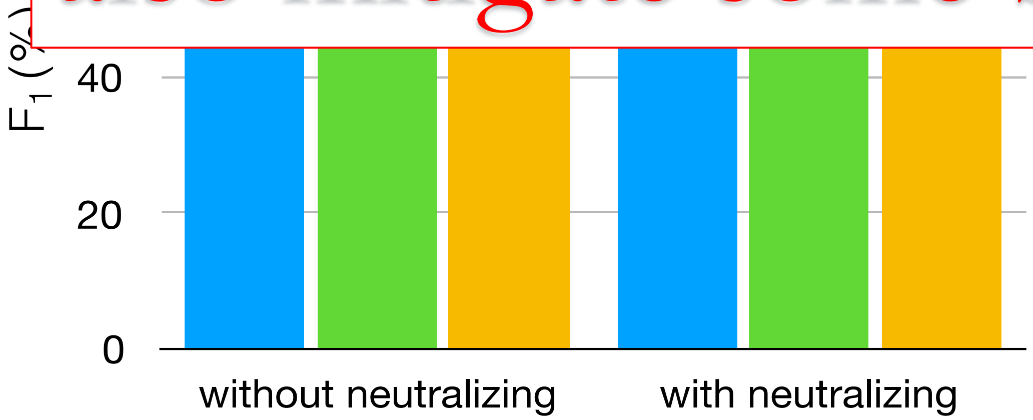
- Neutralize ELMo Embeddings
  - Lightweight; keeps the performance
  - Mitigate some of the bias



Mitigate Bias (Method 2)

- Neutralize ELMo Embeddings
  - Lightweight; keeps the performance
  - Mitigate some of the bias

Lightweight but can also mitigate some bias



Conclusion

- ELMo is sensitive to gender
  - Training corpus is biased to man
  - ELMo treats genders unequally
  - Bias propagates to downstream tasks

- Mitigation Bias
  - Data augmentation
  - Neutralizing ELMo

Thank you!