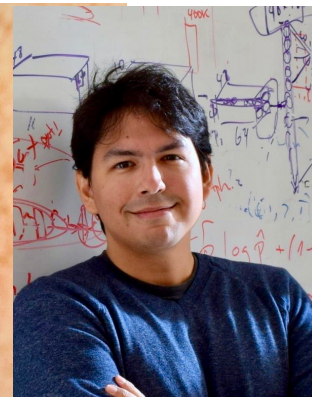
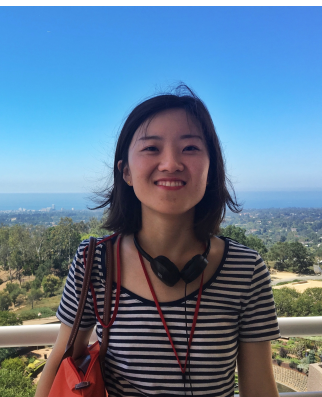


Gender Bias in Contextualized Word Embeddings






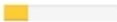





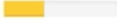






Jieyu Zhao¹, Tianlu Wang², Mark Yatskar³, Ryan Cotterell⁴, Vicente Ordonez², Kai-Wei Chang¹

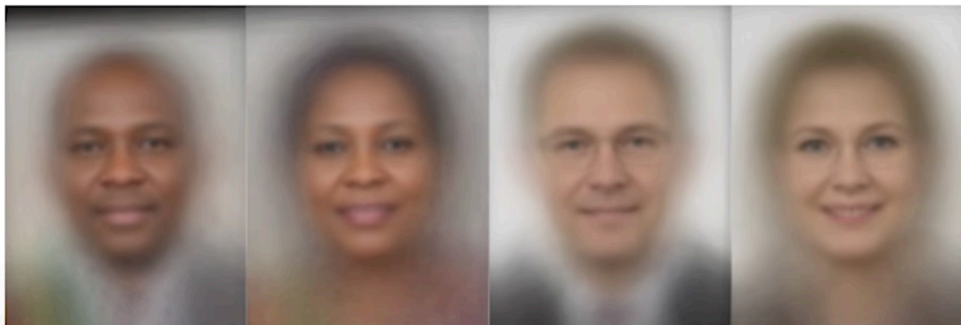
¹UCLA, ²University of Virginia, ³Allen Institute for AI, ⁴University of Cambridge



Two Perspectives of Fairness in ML/NLP

- ML/NLP models should work for everyone

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 



Gender shade: <https://www.youtube.com/watch?v=TWWsW1w-BVo> [Buolamwini& Gebru 18]

Two Perspectives of Fairness in ML/NLP

- ML/NLP models should work for everyone
- ML/NLP models should be aware of potential stereotypes existing in the data/model and avoid affecting downstream tasks

Bias in NLP: Word Embeddings

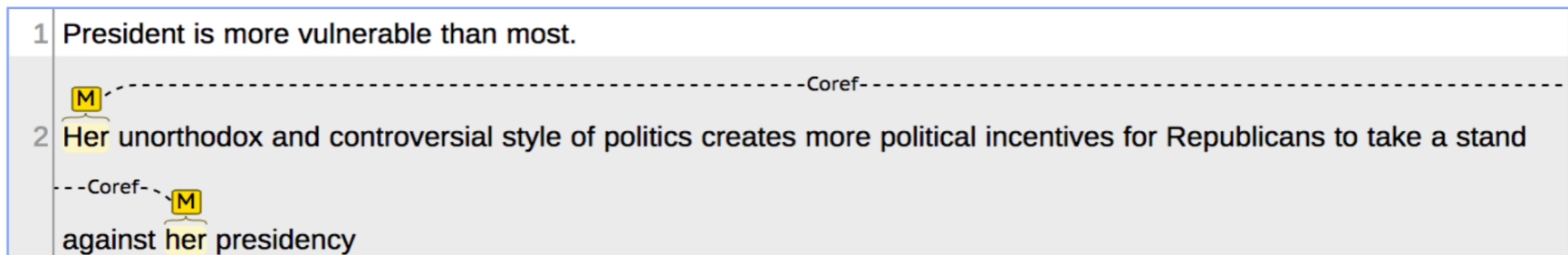
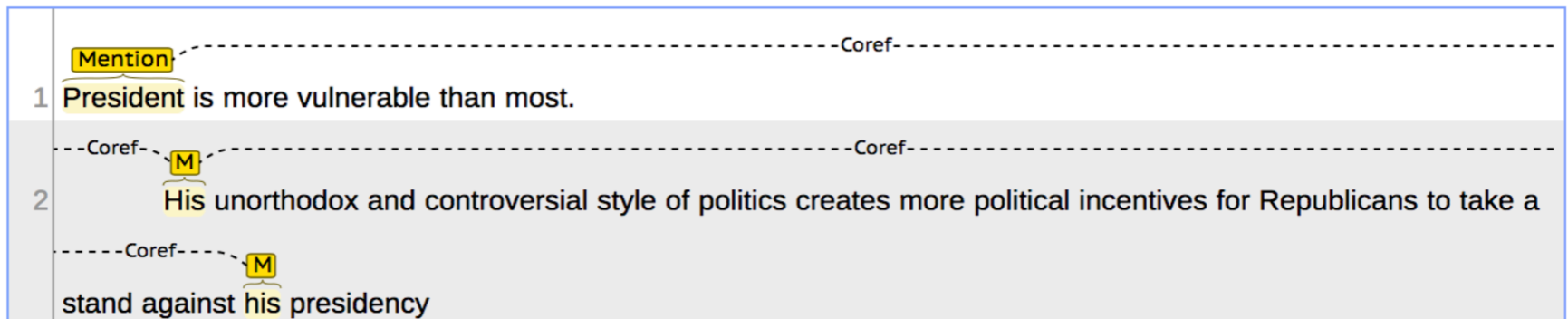
he

she



Bias in NLP: Downstream Task

- Coreference resolution
 - Model fails for “she” when given same context

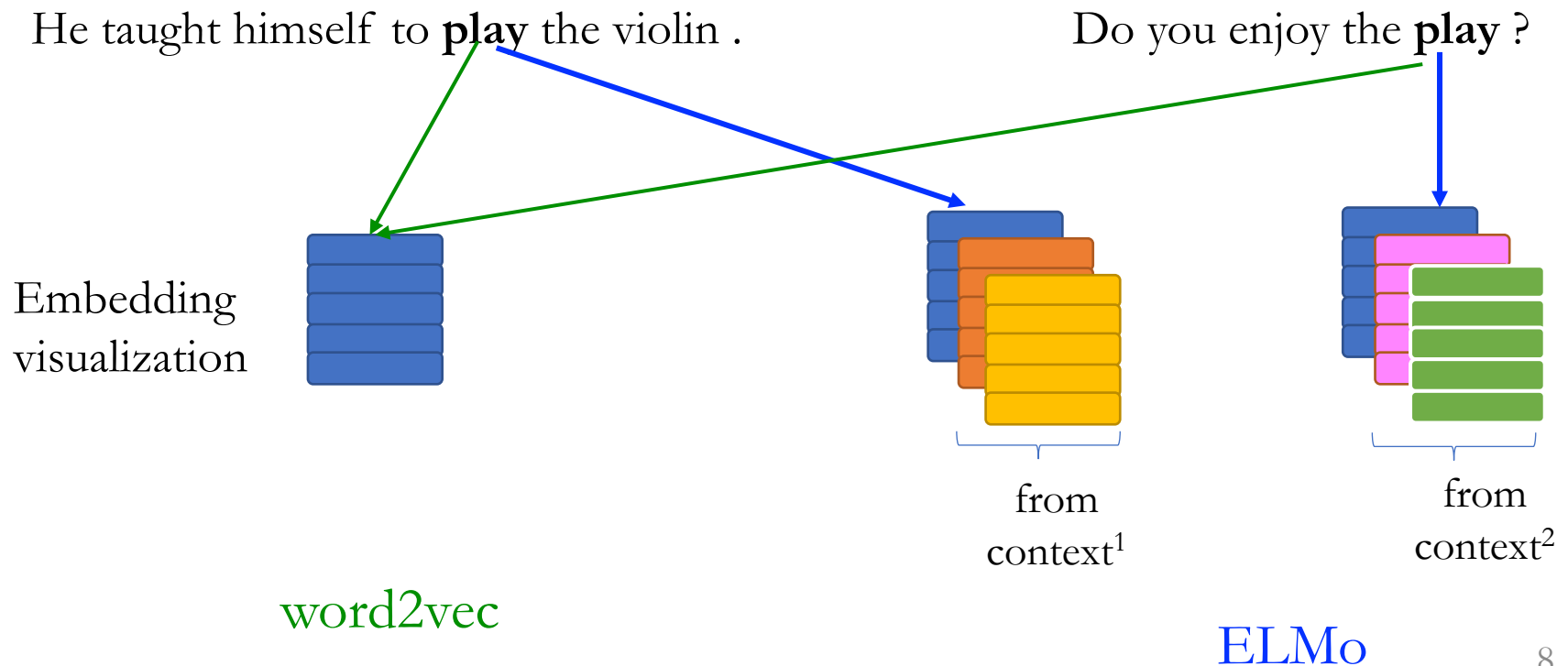


Outline

- Training corpus for ELMo is biased
- Visualize gender geometry in ELMo
- Bias propagates to downstream tasks
- Mitigate the bias

Background: ELMo

- Take LM information
- Assign words with different embeddings based on the surrounding contexts



Bias in ELMo

- Bias Analysis
 - Training Dataset Bias
 - Geometry of the Gender
 - Unequal Treatment of Gender in ELMo
 - Downstream task – Coreference resolution

Bias in ELMo

Training Dataset Bias

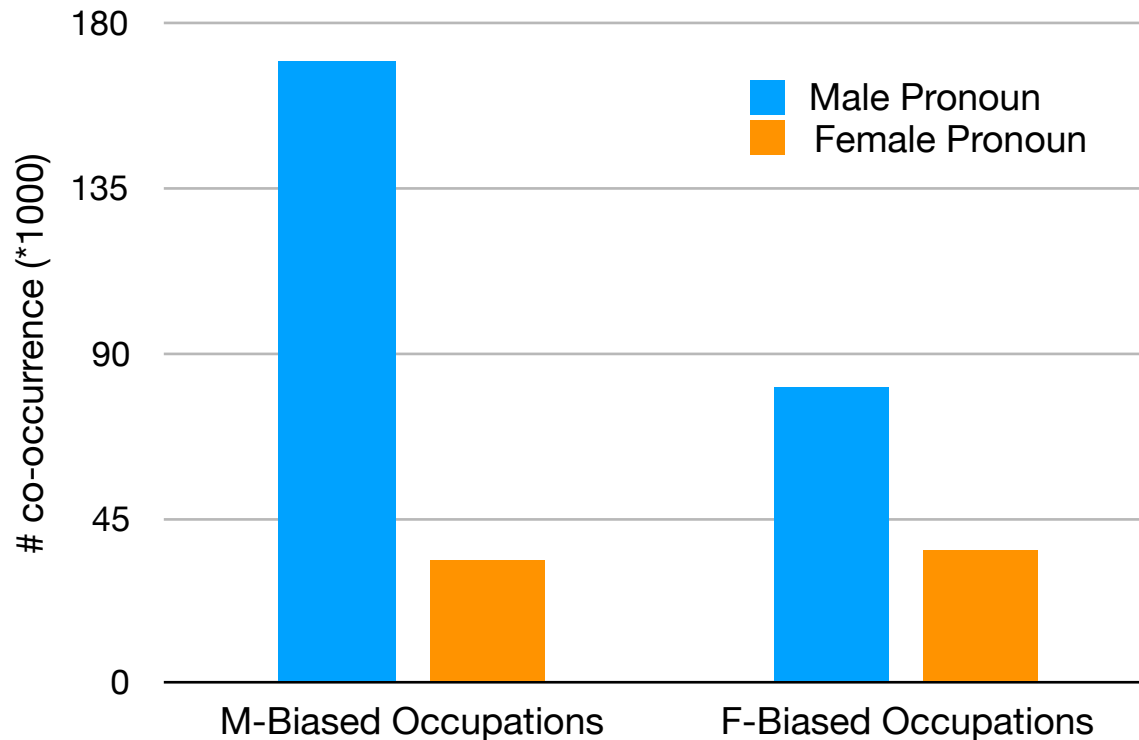
- Dataset is biased towards **male**

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)

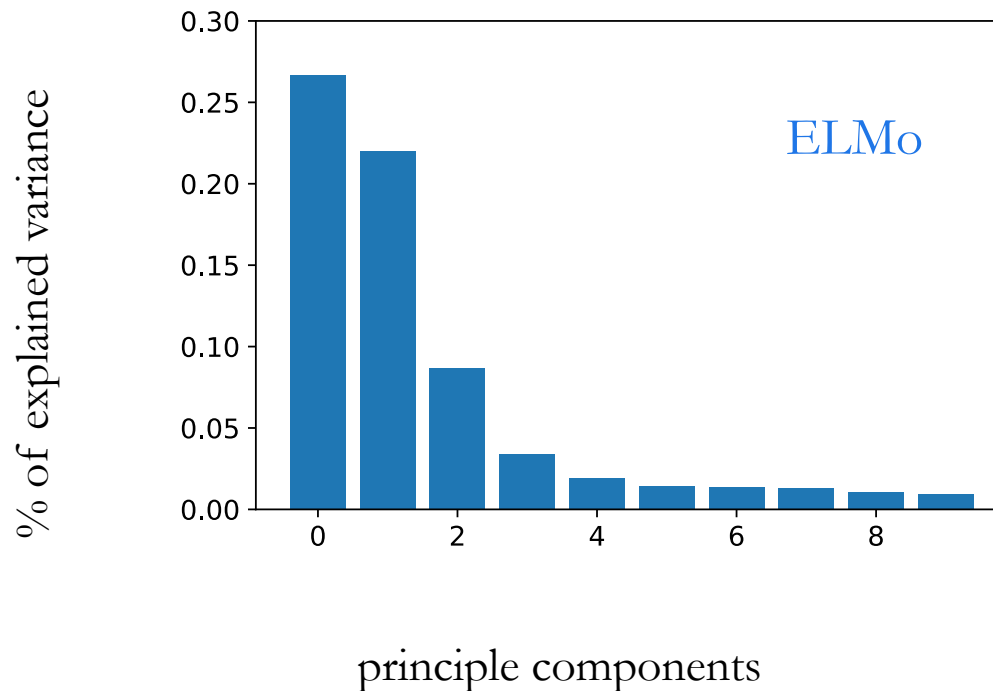
- Male pronouns co-occur more frequently with occupation words¹



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

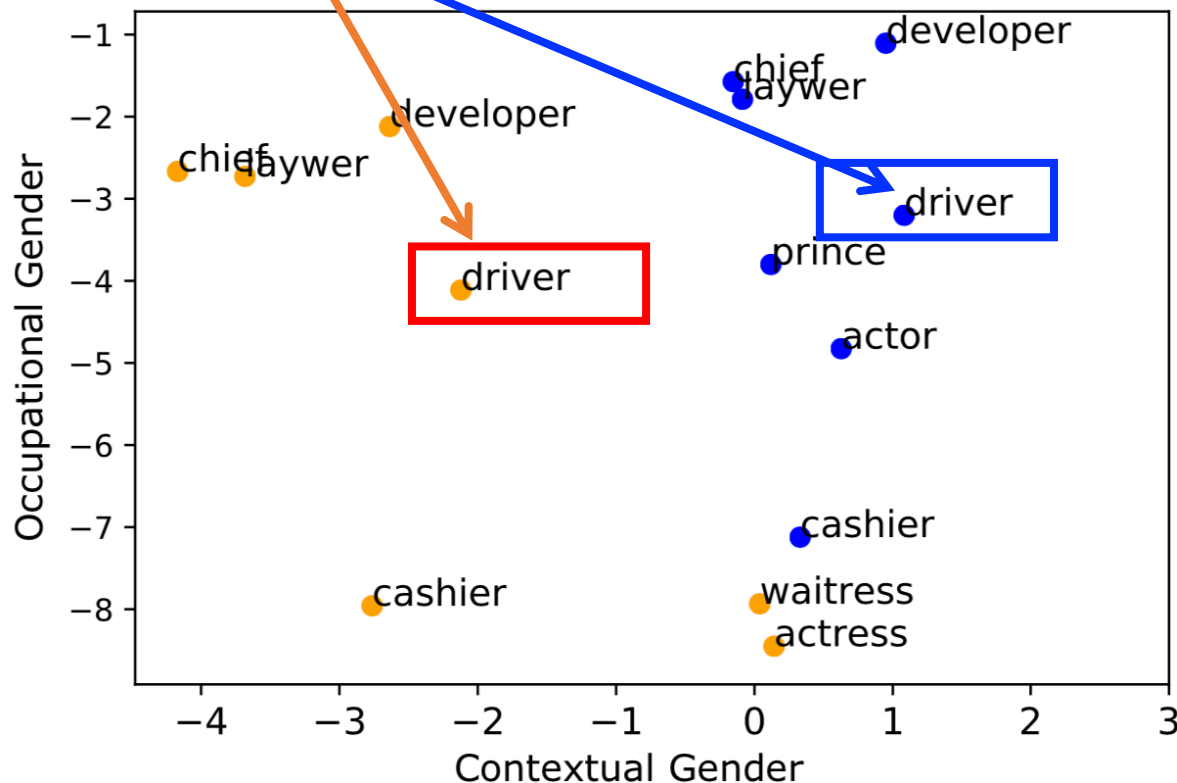
Geometry of Gender in ELMo

- ELMo has two principle components



Geometry of Gender in ELMo

- The **driver** transported the counselor to the hospital because **she** was paid
- The **driver** transported the counselor to the hospital because **he** was paid



● Female context

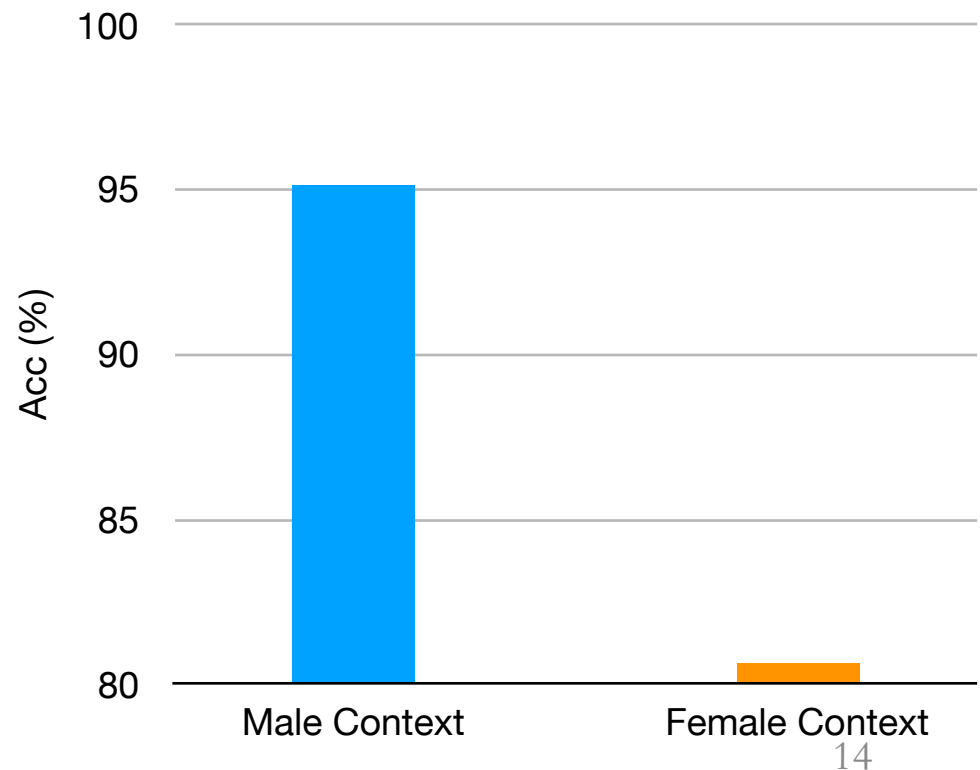
● Male context

Unequal Treatment of Gender

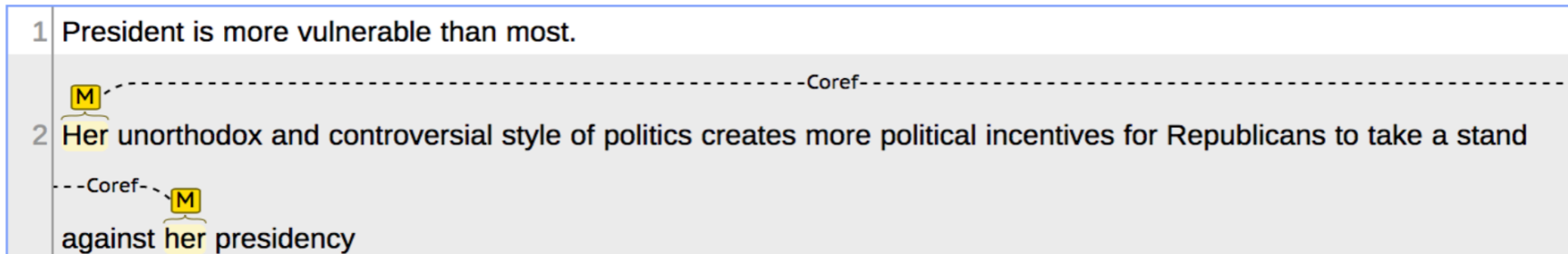
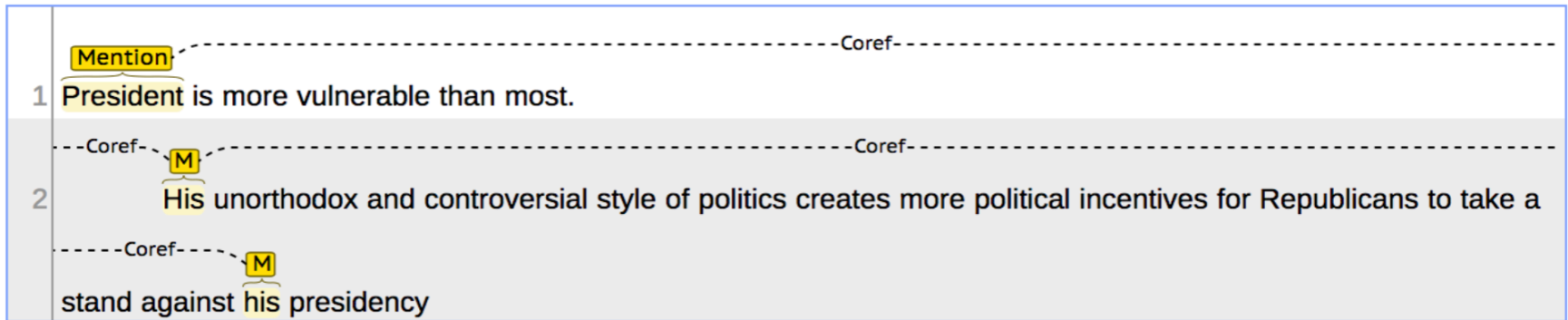
- Classifier

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

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



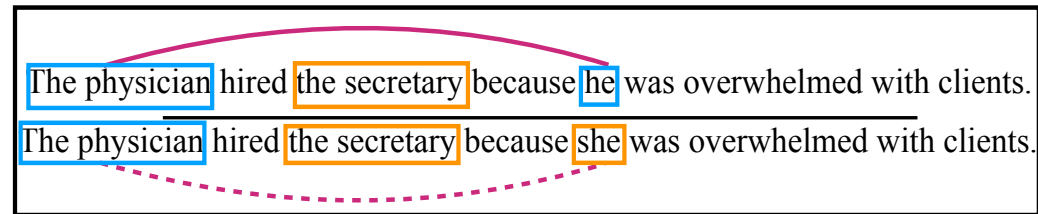
Bias in Downstream Task -- Coreference Resolution



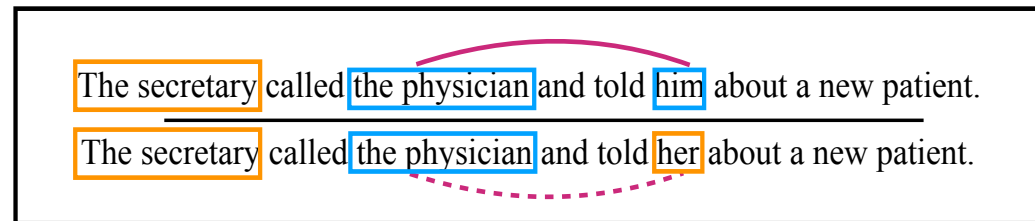
Bias in Downstream Task: Coreference Resolution

- WinoBias dataset¹
 - Pro-Stereotypical and Anti-Stereotypical dataset
- Bias: different performance between Pro. and Anti. dataset.

Semantics Only



w/ Syntactic Cues



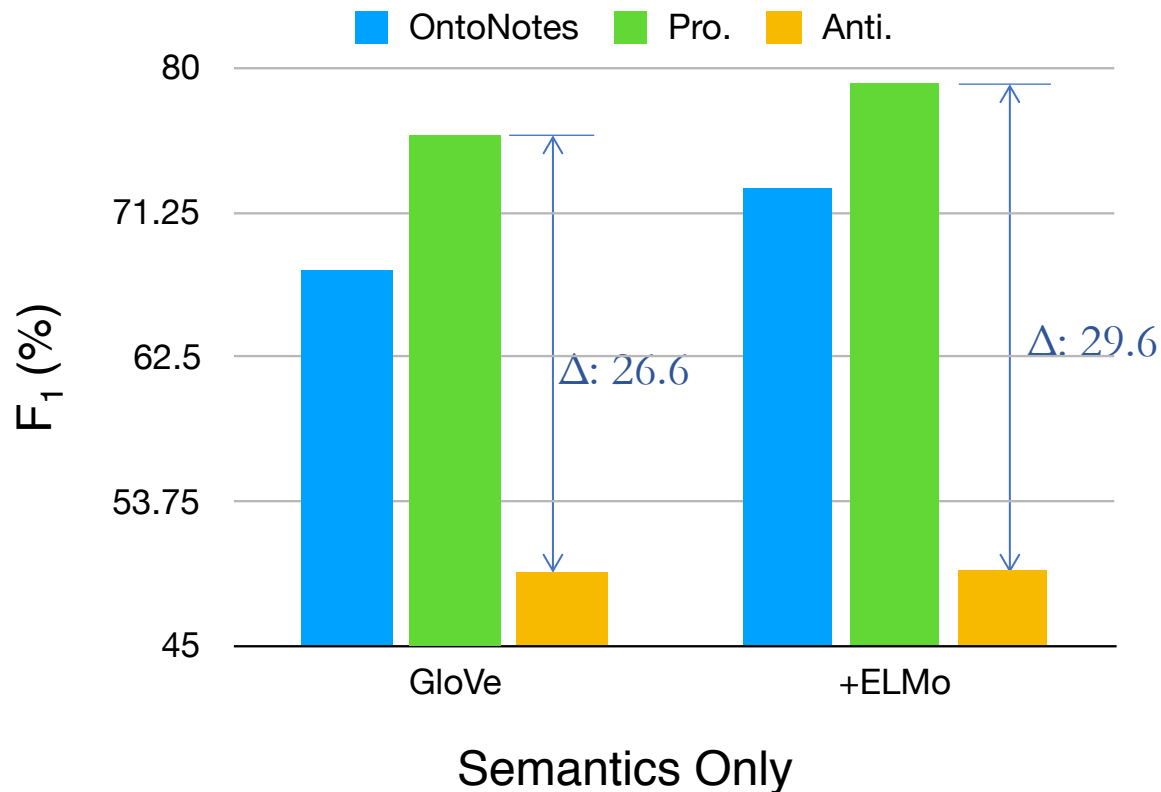
— Pro.

- - - Anti.

¹<https://uclanlp.github.io/corefBias>

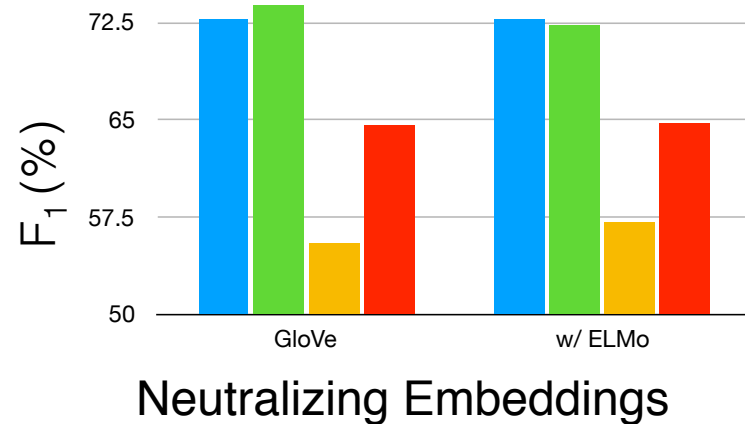
Bias in Coreference

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



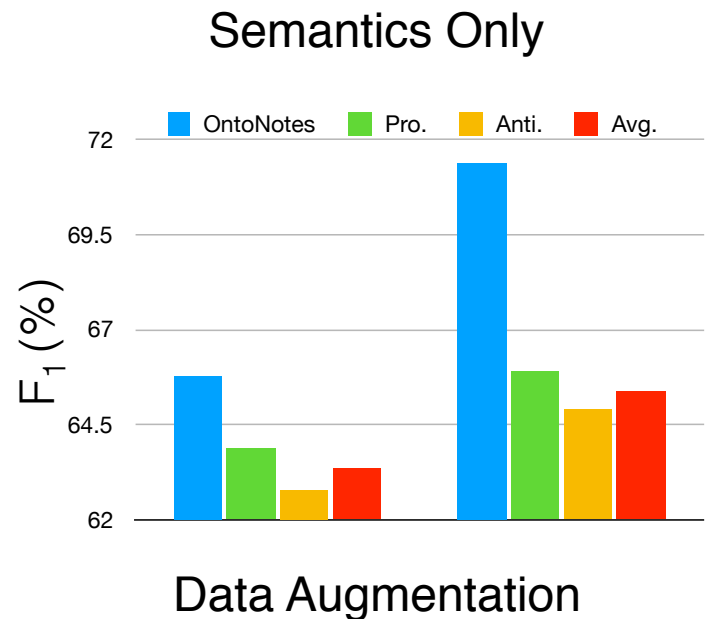
Mitigate Bias (Method 1)

- Neutralizing ELMo Embeddings
 - Generating gender swapped test variants
 - Average the ELMo embeddings for test dataset
 - Do not need retrain; keeps the performance
 - lightweight

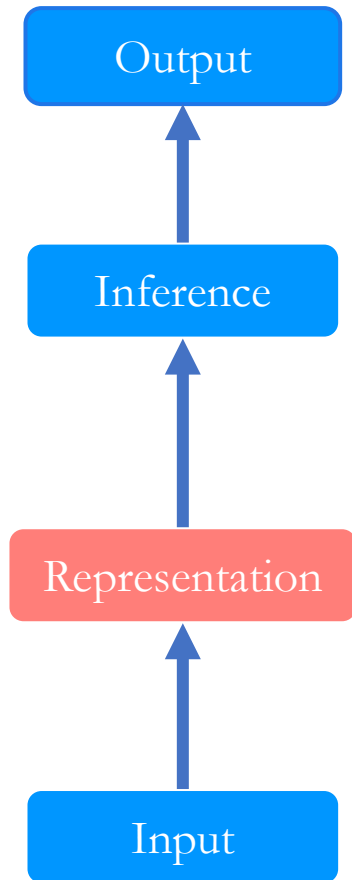


Mitigate Bias (Method 2)

- Data Augmentation
 - Generate gender swapped training variants
 - Better mitigation; need retrain



Bias in NLP/ML



- **Zhao** et al. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints
- Bolukbasi et al. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
- **Zhao** et al. Learning Gender-Neutral Word Embeddings
- Elazar & Goldberg. Adversarial Removal of Demographic Attributes from Text Data
- Wang et al. Adversarial Removal of Gender from Deep Image Representations
- Xie et al. Controllable Invariance through Adversarial Feature Learning
- **Zhao** et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods
- Park et al. Reducing Gender Bias in Abusive Language Detection

Thank you!