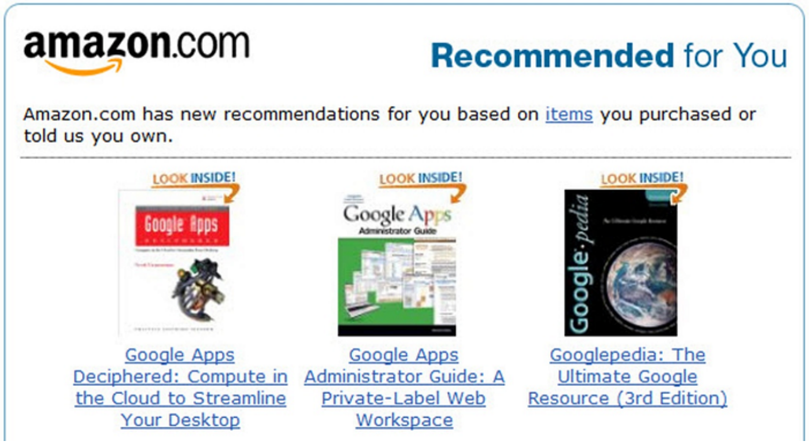
UCLA NLP

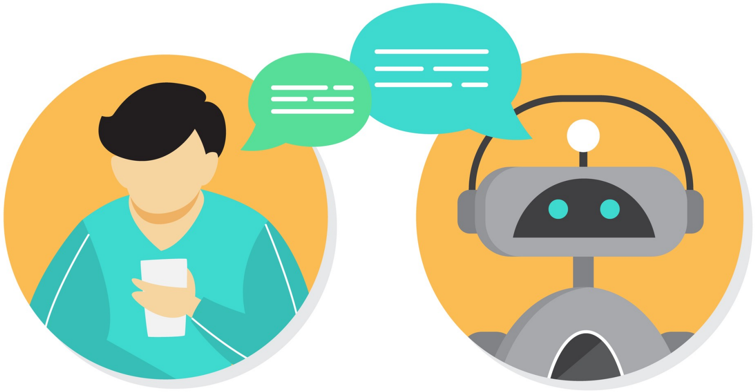
Detecting and Mitigating Bias in NLP

Jieyu Zhao

https://jyzhao.net

NLP models are prevalent





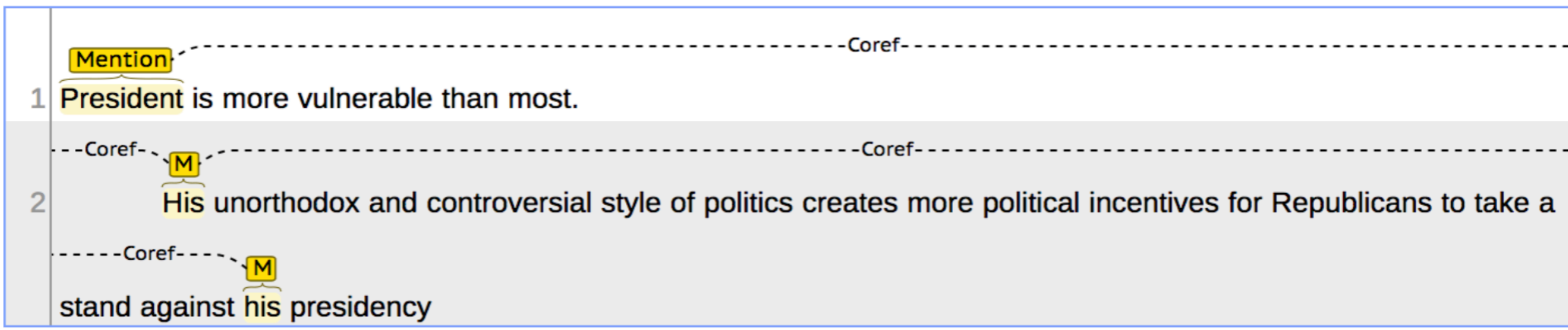
Bias in Visual Semantic Role Labeling (vSRL)

Cooking	
Role	Noun
agent	
food	vegetable
container	bowl
tool	knife
place	kitchen

- Jieyu Zhao, et al. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." EMNLP (2017). Best Long Paper Award

Bias in Coreference Resolution

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

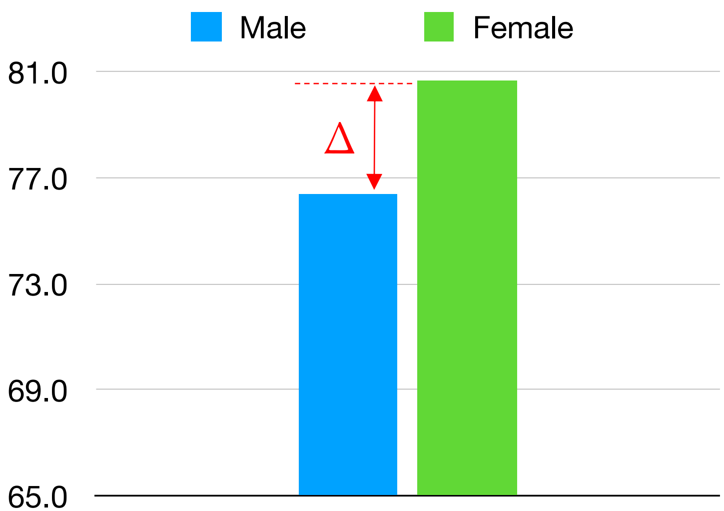


Change his → her ?

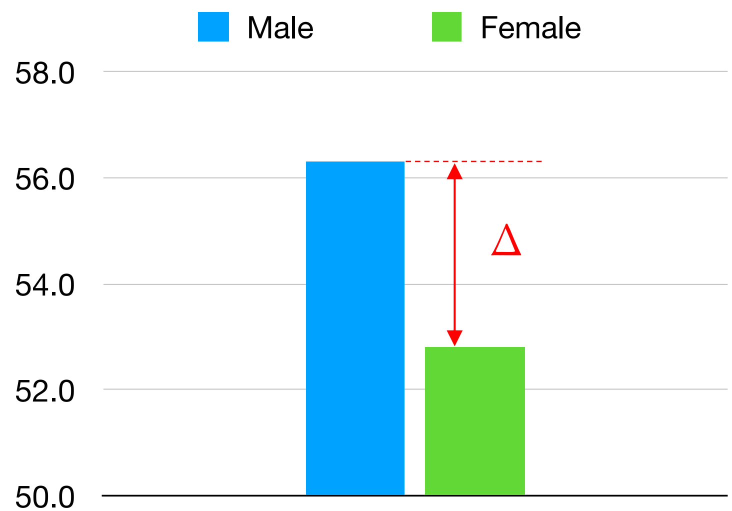
[1]Zhao et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL 2018
[2]Rudinger et al. Gender Bias in Coreference Resolution. NAACL 2018

Bias in Transfer Learning: Bio Prediction

- Edmund J. Bourne, PhD, is a psychologist in northern California, … He is author of several books, …
- Dr. Constance Milbrath is a developmental psychologist, … Her interests at HELP are in the ethno-cultural determinants …



EN dataset



Cross-lingual transfer

Jieyu Zhao, et al. Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer. ACL 2020.

Bias in NLP

Mitigating Gender Bias in Natural Language Processing: Literature Review

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, William Yang Wang

Language (Technology) is Power: A Critical Survey of "Bias" in NLP

Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach

- R. Rudinger et al. Social Bias in Elicited Natural Language Inferences. ENLP 2017
- L. Dixon et al. Measuring and Mitigating Unintended Bias in Text Classification. AAAI 2017
- S. Kiritchenko et al. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. SEM 2018
- J. Park et al. Reducing Gender Bias in Abusive Language Detection. EMNLP 2018
- N. Schluter. The Glass Ceiling in NLP. EMNLP 2018
- K. Webster et al. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. TACL 2018
- G. Stanovsky et al. Evaluating Gender Bias in Machine Translation. ACL 2019
- T. Manzini et al. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multi-class Bias in Word Embedding. NAACL 2019
- E. Sheng, et al. The Woman Worked as a Babysitter: On Biases in Language Generation. EMNLP 2019
- M. De-Arteage et al. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. FAT 2019
- K. Chang et al. Tutorial: Bias and Fairness in Natural Language Processing. EMNLP 2019
- …

Outline

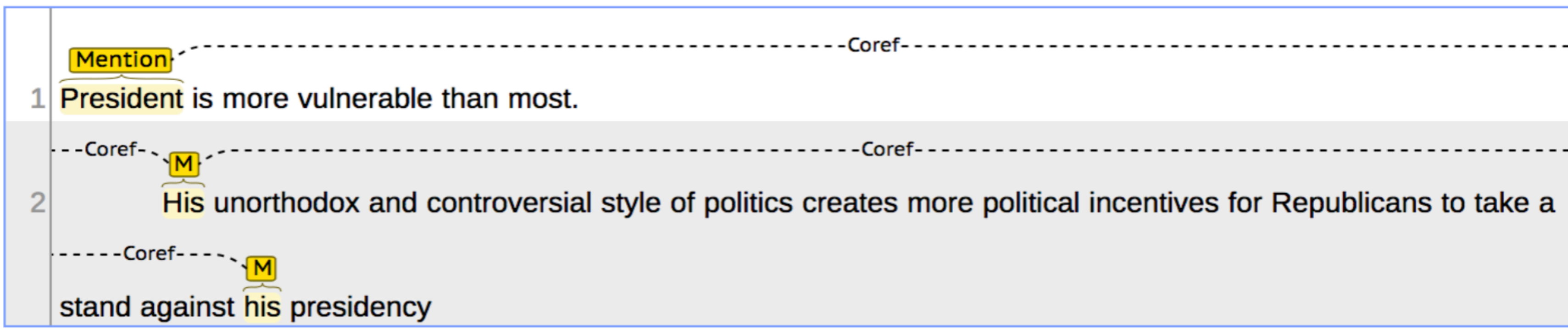
Bias in NLP Modules

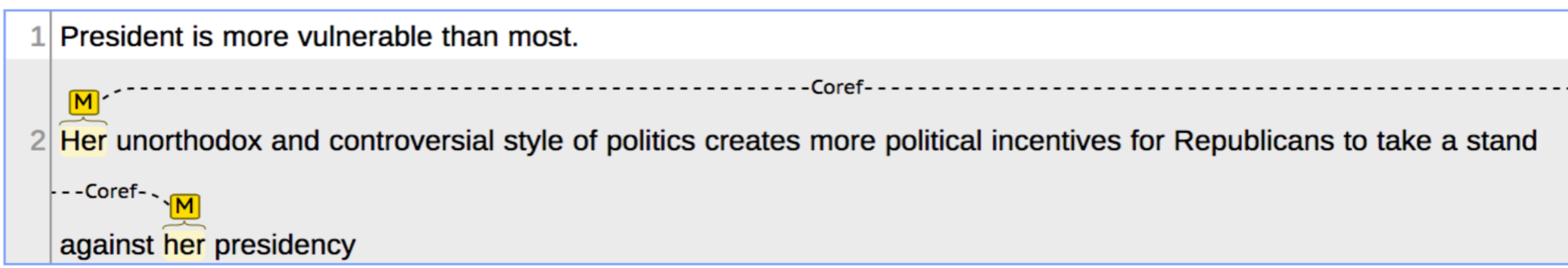
Bias in Language Representations

Bias Amplification

Bias in NLP: Coreference Resolution

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



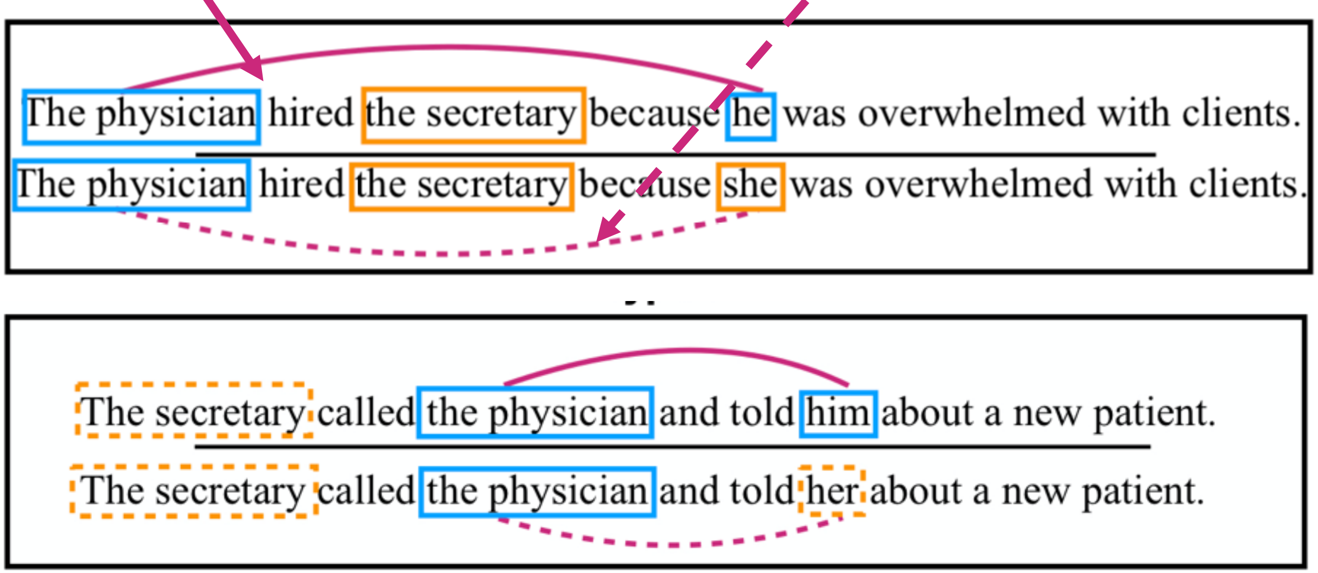


[1]Zhao et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL 2018

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

Evaluate Bias

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

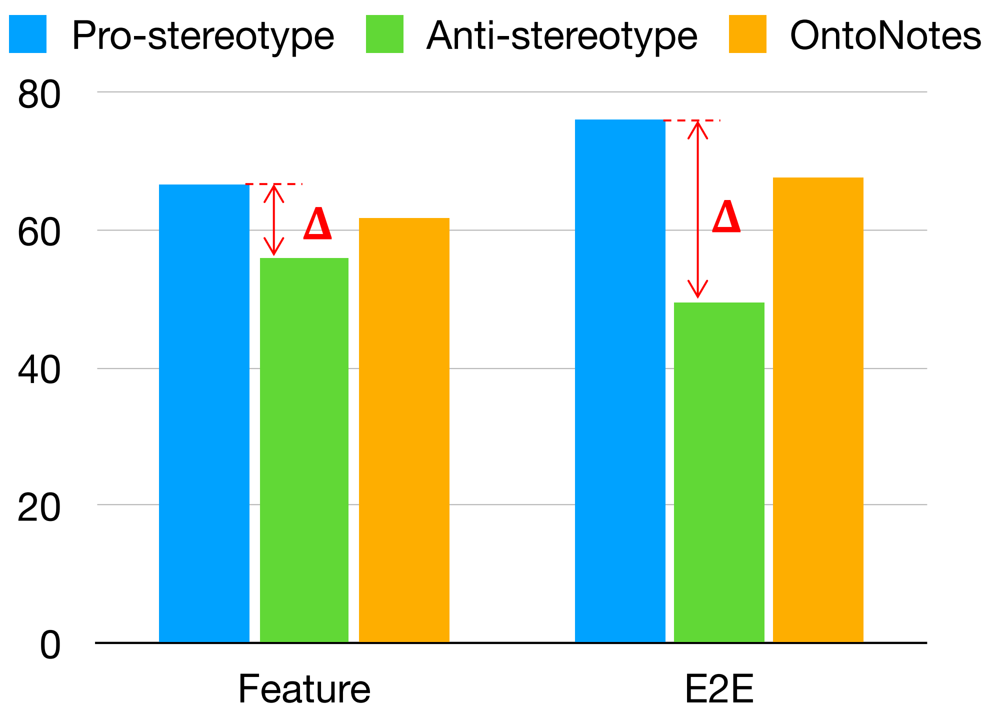


- **Bias**: performance difference between Pro. and Anti. dataset.

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

Bias in Coreference Resolution

- Bias exists in different coreference systems



Source of Bias

- **Training Dataset Bias**

I. 80% of entities headed by gendered pronouns are male.

II. Male gendered mentions are more likely to contain a job.

- **Resource Bias**

I. Word embeddings[1] : "man" is closer to "programmer" than "woman".

II. Gender lists: corpus-based gender statistics

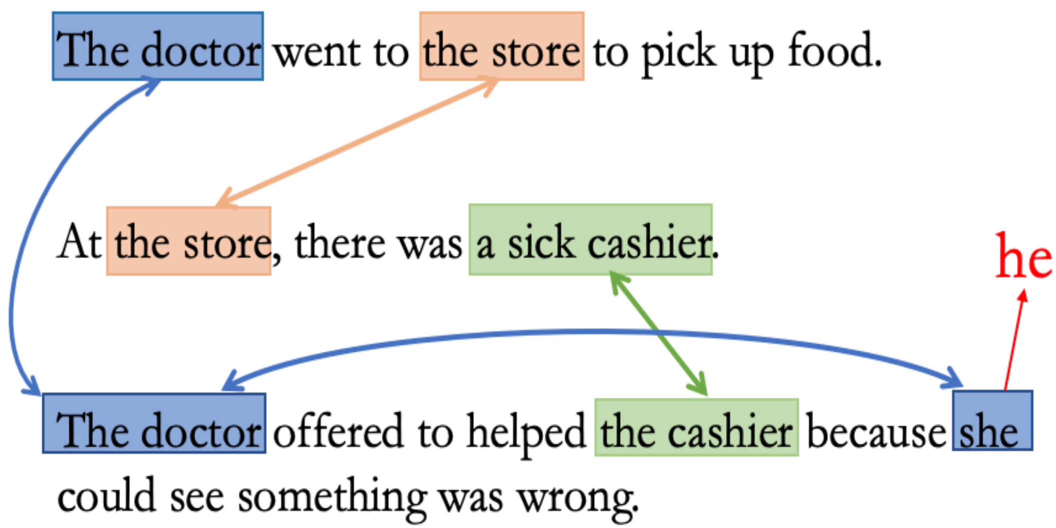
[1]Bolukbasi et al. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016.

Mitigate Bias

- **Gender Swapping**

I. Build a dictionary of gendered terms

II. An additional training corpus where all male entities are swapped for female entities and vice-versa (*Aug.*)

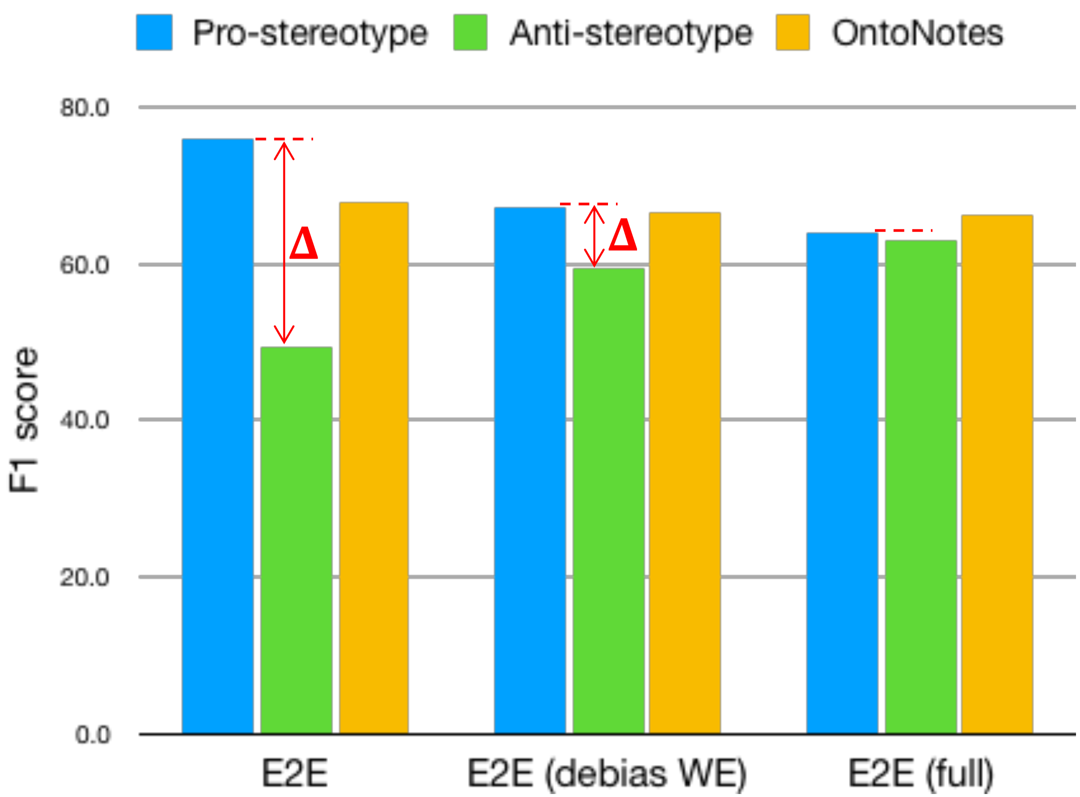


- **Modify the resource**

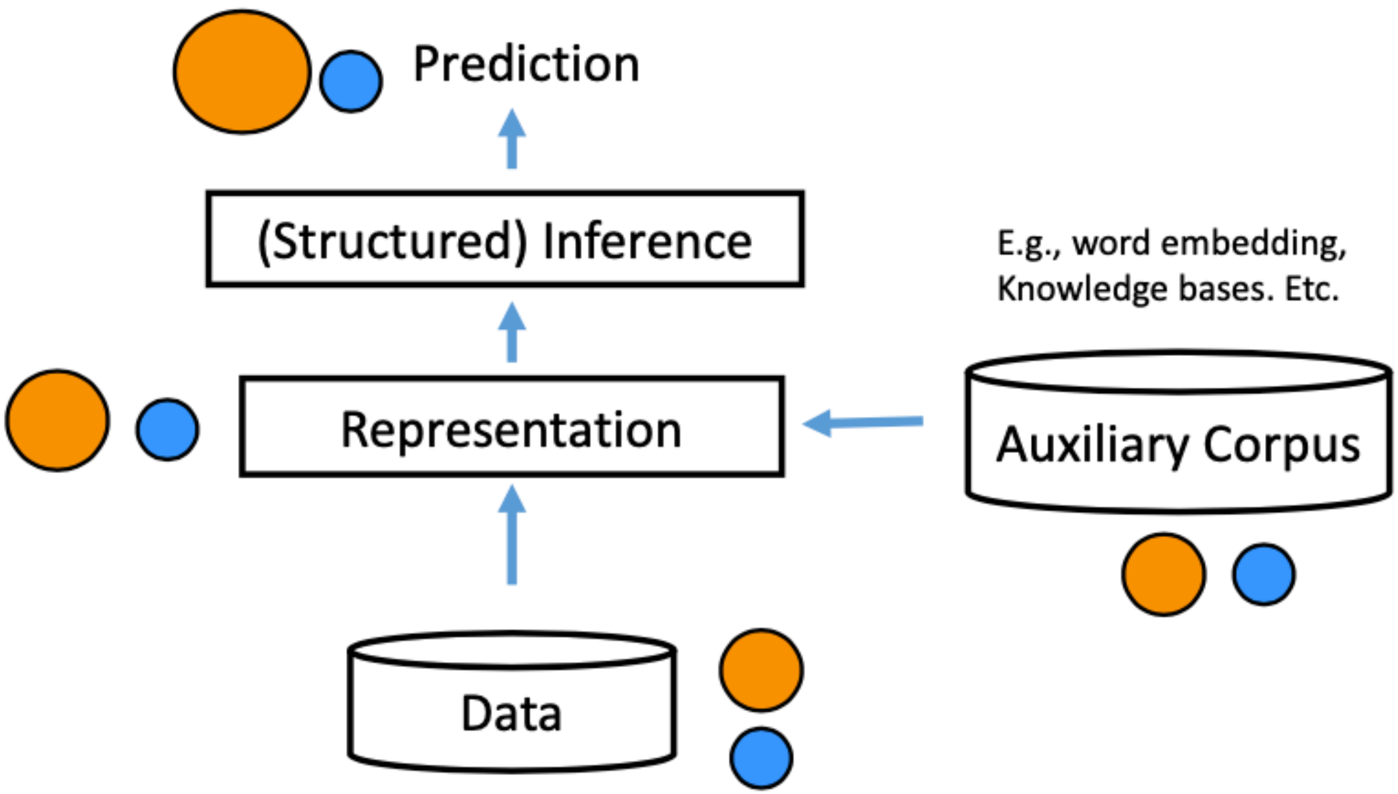
I. Debiased word embeddings or balanced gender list

Bias Mitigation

- Data Augmentation
- Using Debiased Word Embeddings



A Carton of NLP Pipeline



credit: Kai-Wei Chang

UCLA NLP

Outline

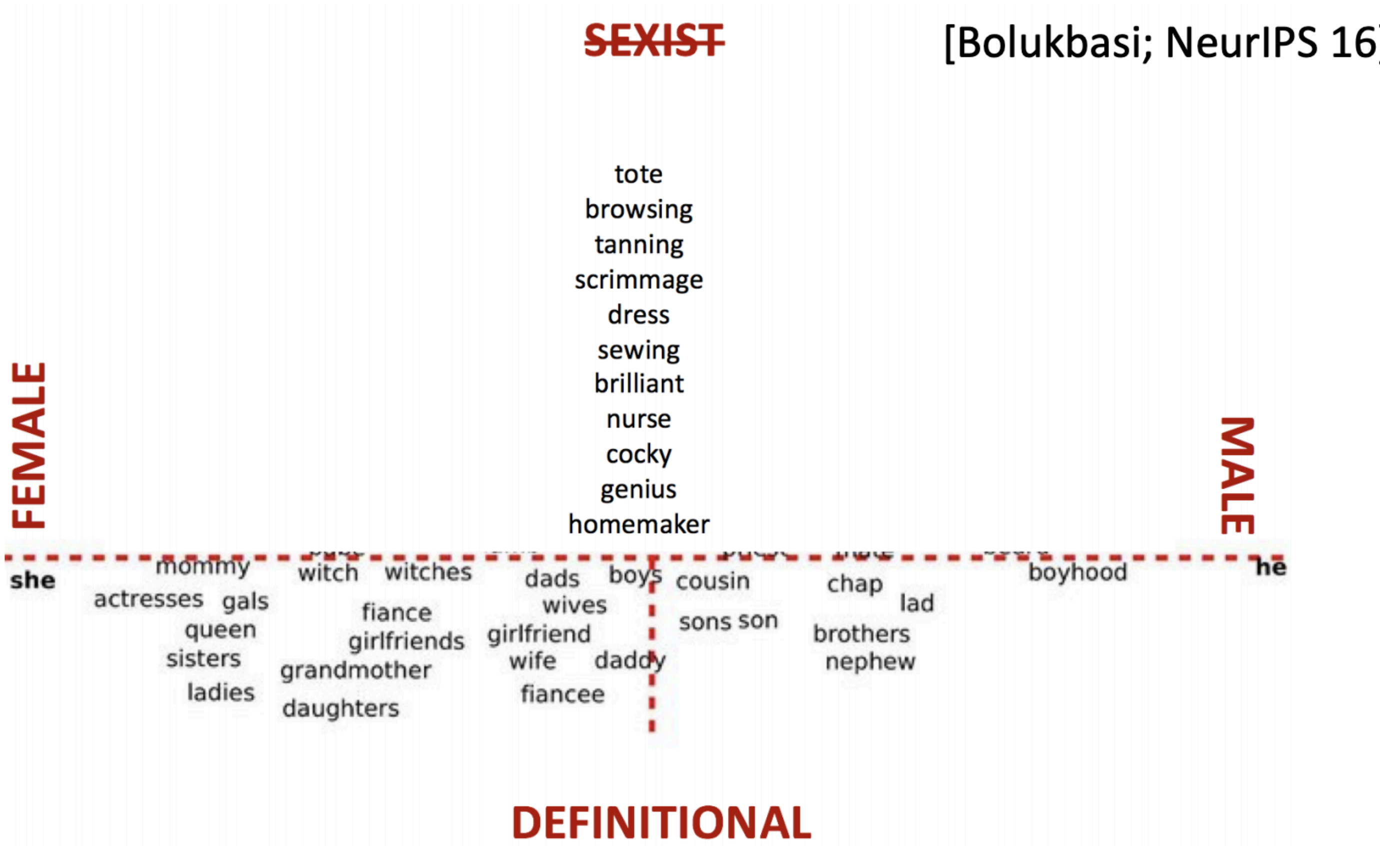
Bias in NLP Modules

Bias in Language Representations

Bias Amplification

UCLA NLP

Bias in Word Embeddings

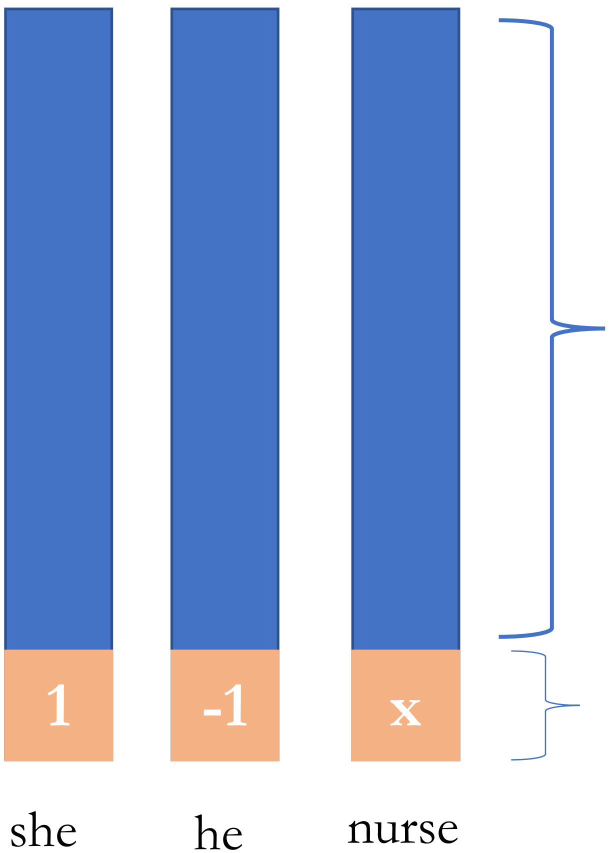


This can be done by projecting gender direction out from gender neutral words using linear operations.

credit: Kai-Wei Chang

Learning Gender-Neutral Word Embeddings

- Goal: To learn an embeddings "without" gender information encoded
- GN-GloVe: To retain gender info in certain dimensions



$$w = [w^{(a)}; w^{(g)}]$$

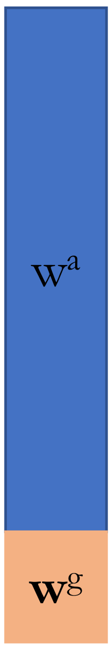
Dimensions for other latent aspects w^a

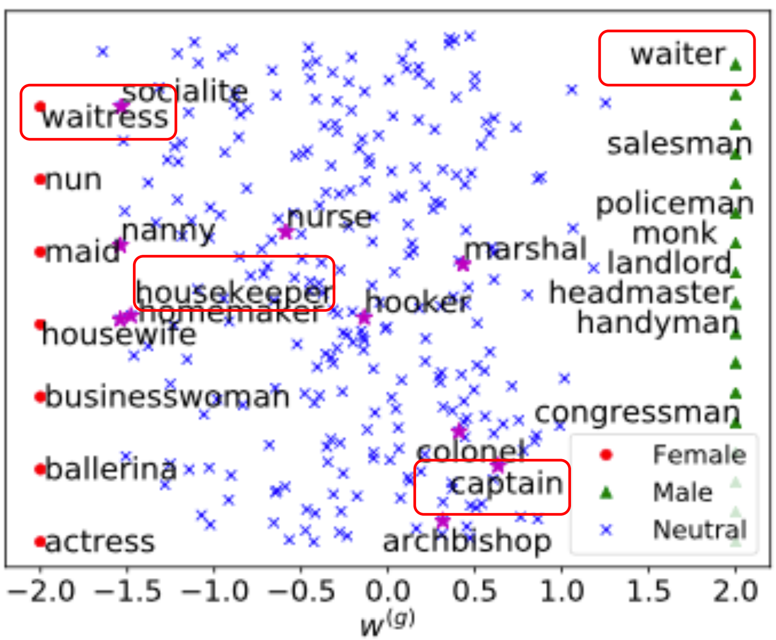
Dimensions reserved for gender info only w^g

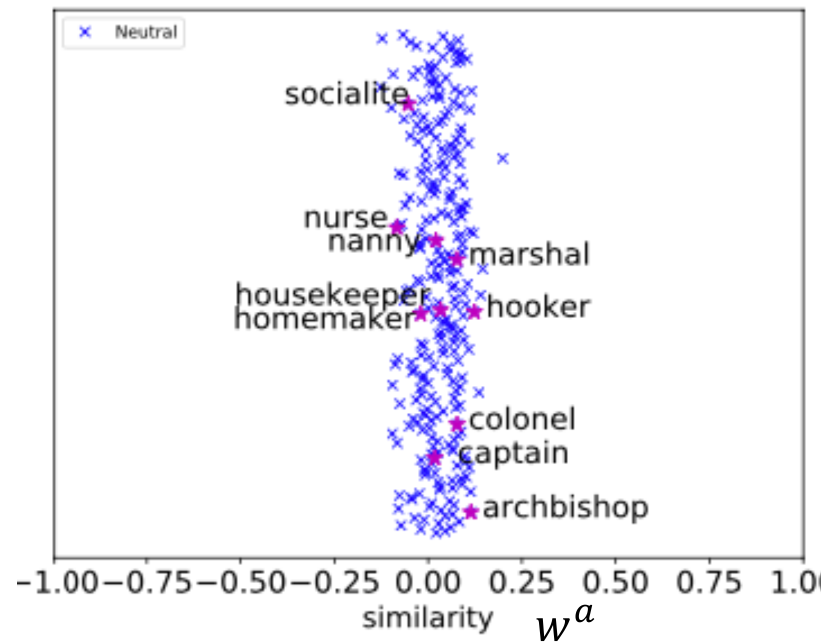
[Jieyu Zhao, et al. EMNLP 2018]

Learning Gender-Neutral Word Embeddings

- GN-GloVe separates the gender info with other aspects



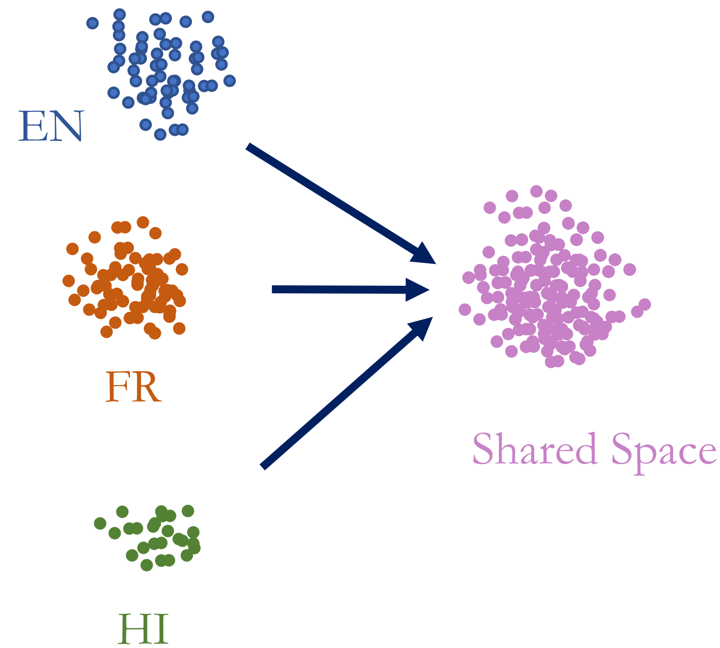




⚖️ H. Gonen, et al. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. NAACL (2019)

Gender Bias in Multilingual Embeddings and Cross-Lingual Transfer Learning

- 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 targets affect the bias
  - Existing mitigation method helps but cannot completely remove the bias.



[Jieyu Zhao, et al. ACL 2020]

UCLA NLP

Gender Bias in Contextualized Embeddings

CoVe

ELMo

BERT

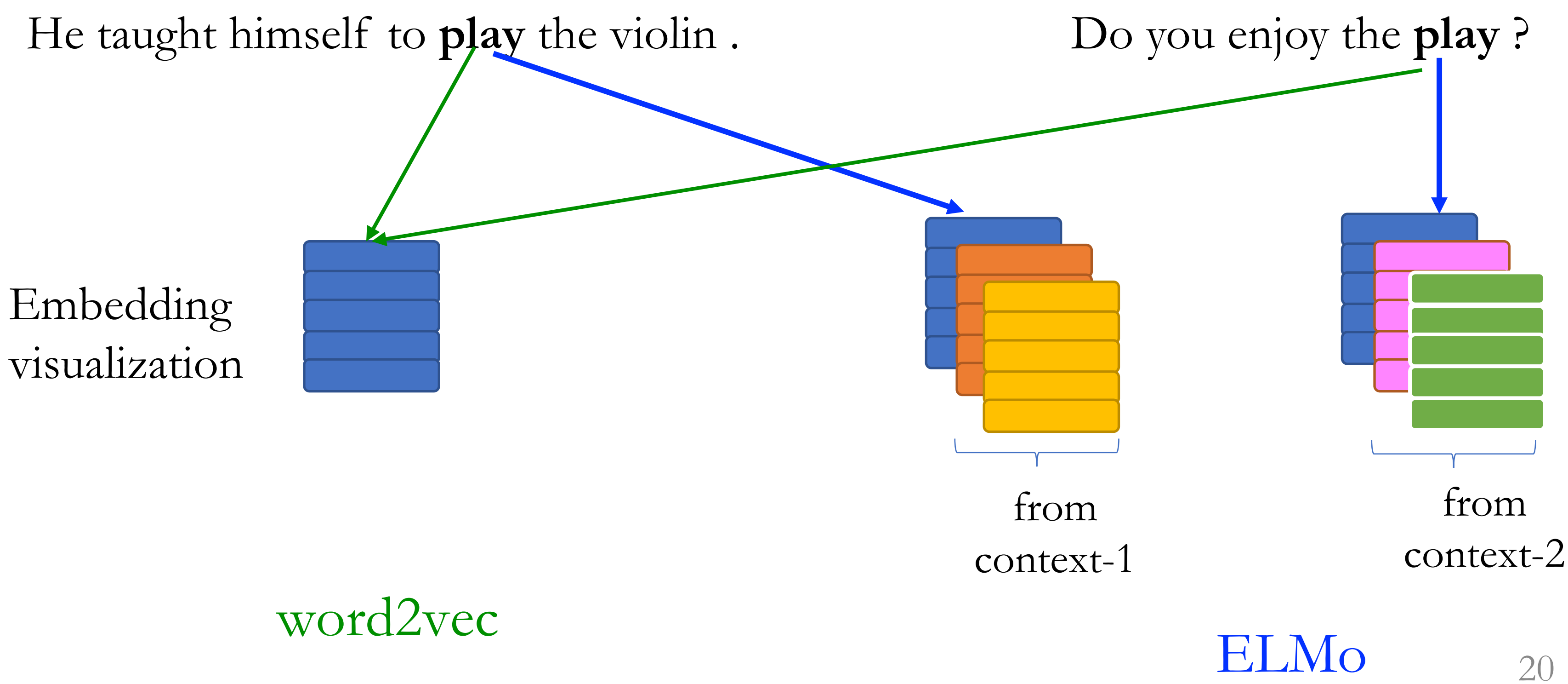
Great performance improvement!

Bias?

In this work, we do the analysis in EN.

Background: ELMo

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



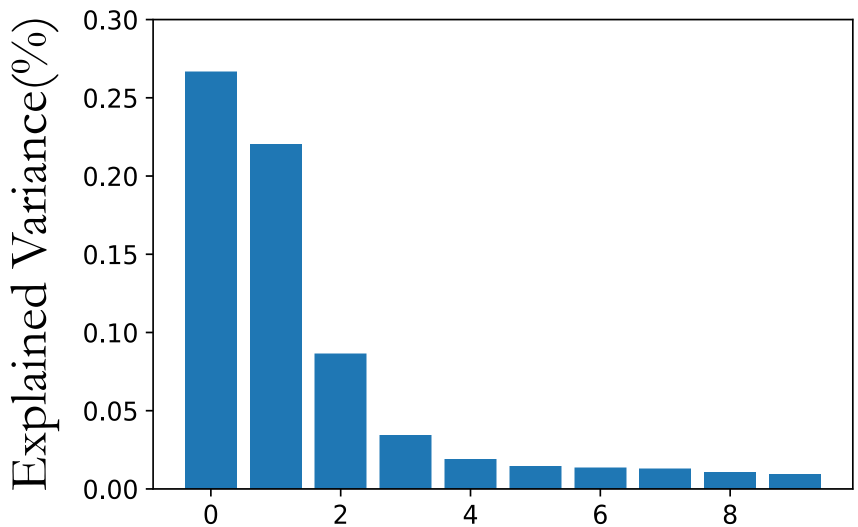
Gender Geometry in ELMo

- First two components explain more variance than others

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

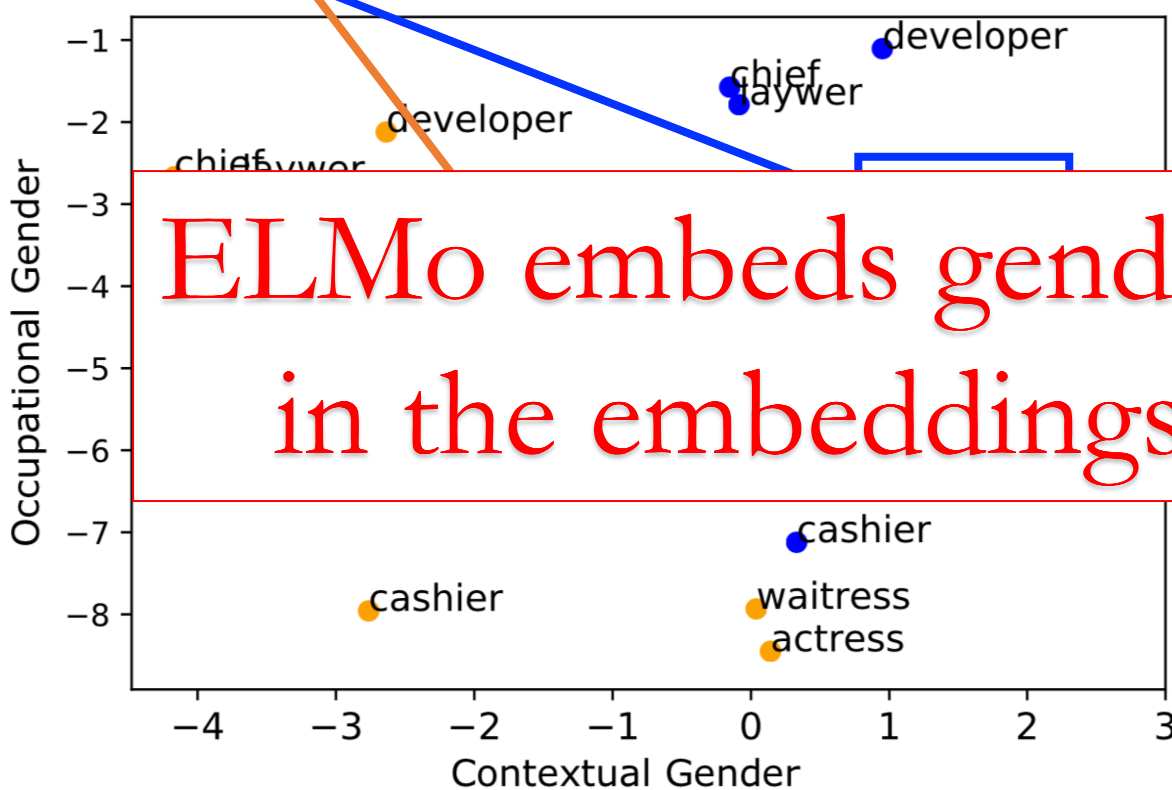
(Masculine) The driver stopped the car at the hospital because **he** was paid to do so

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

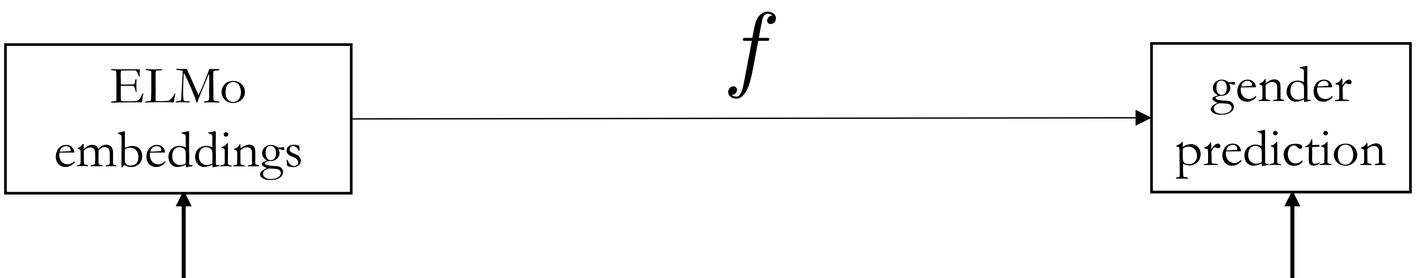
ELMo embeds gender in the embeddings

UCLA NLP

Unequal Treatment of Gender

- Classifier

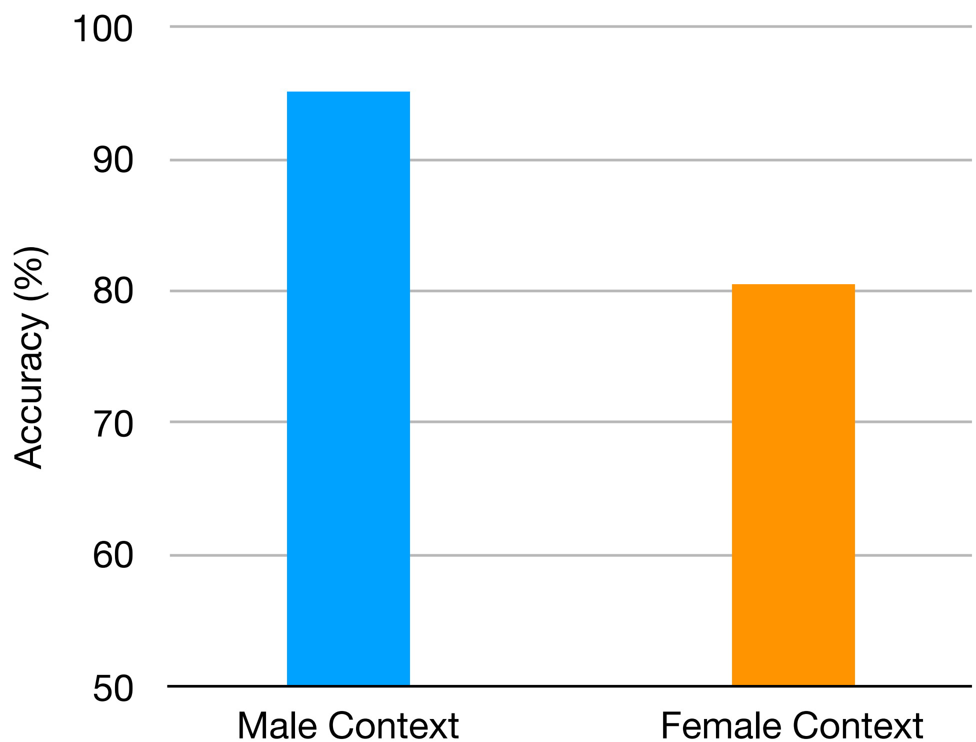
$$f: \text{ELMo(occupation)} \longrightarrow \text{context gender}$$



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

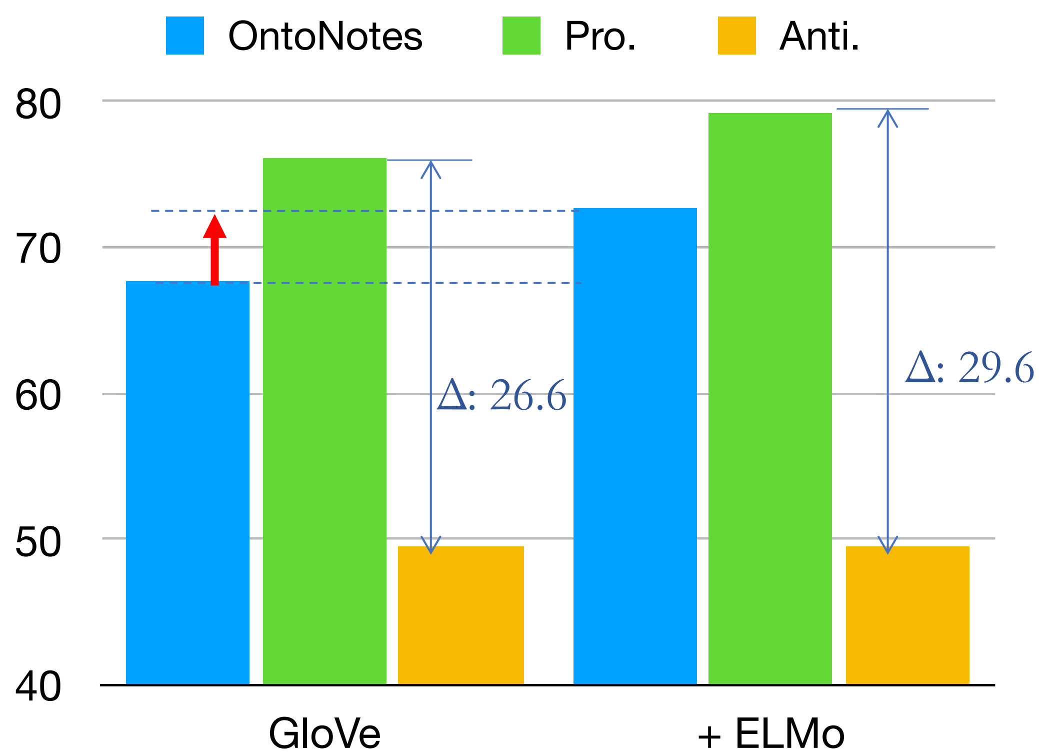
Unequal Treatment of Gender (continued)

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



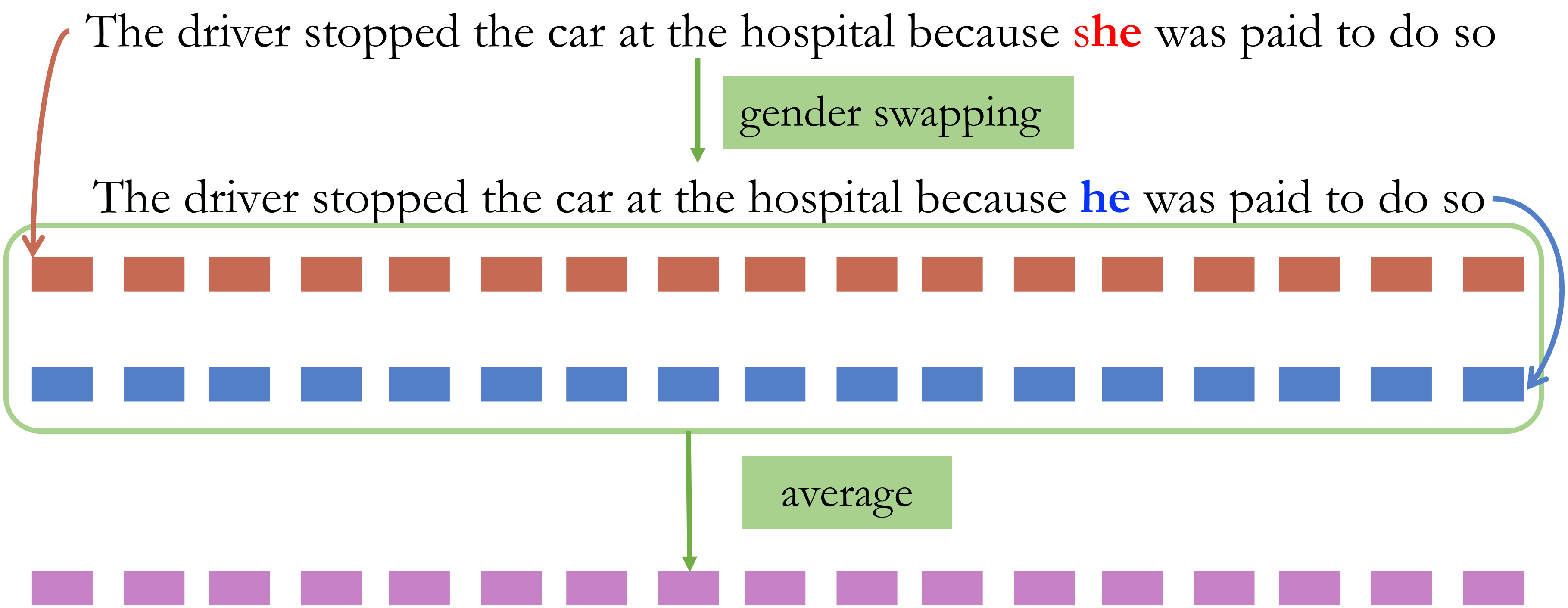
Bias in Coreference Resolution

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



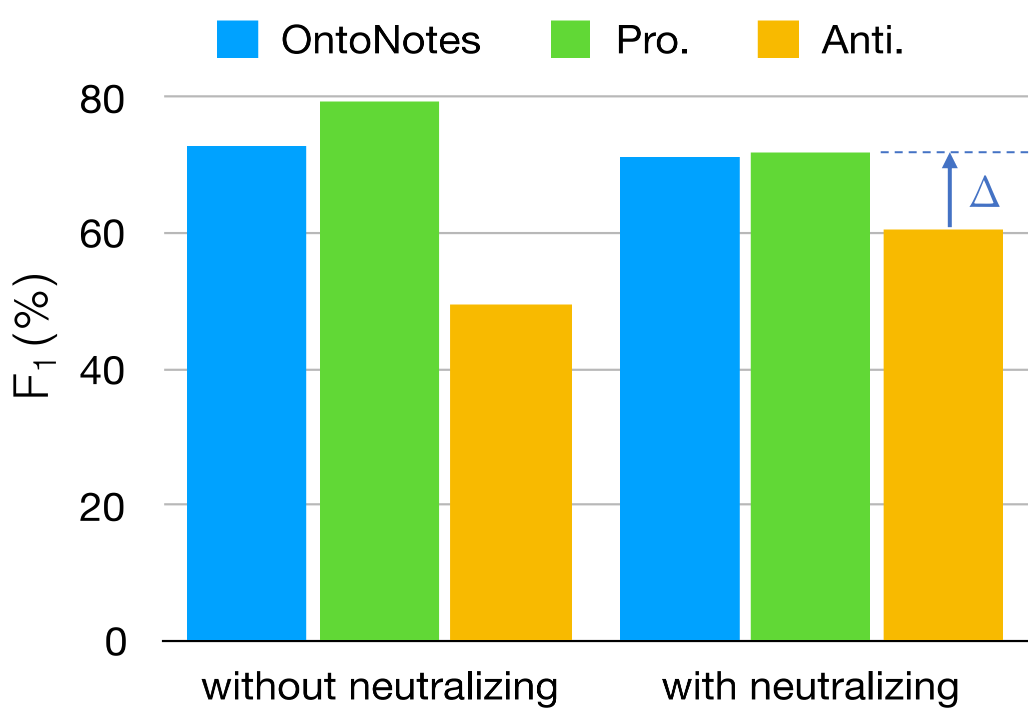
Mitigate Bias

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



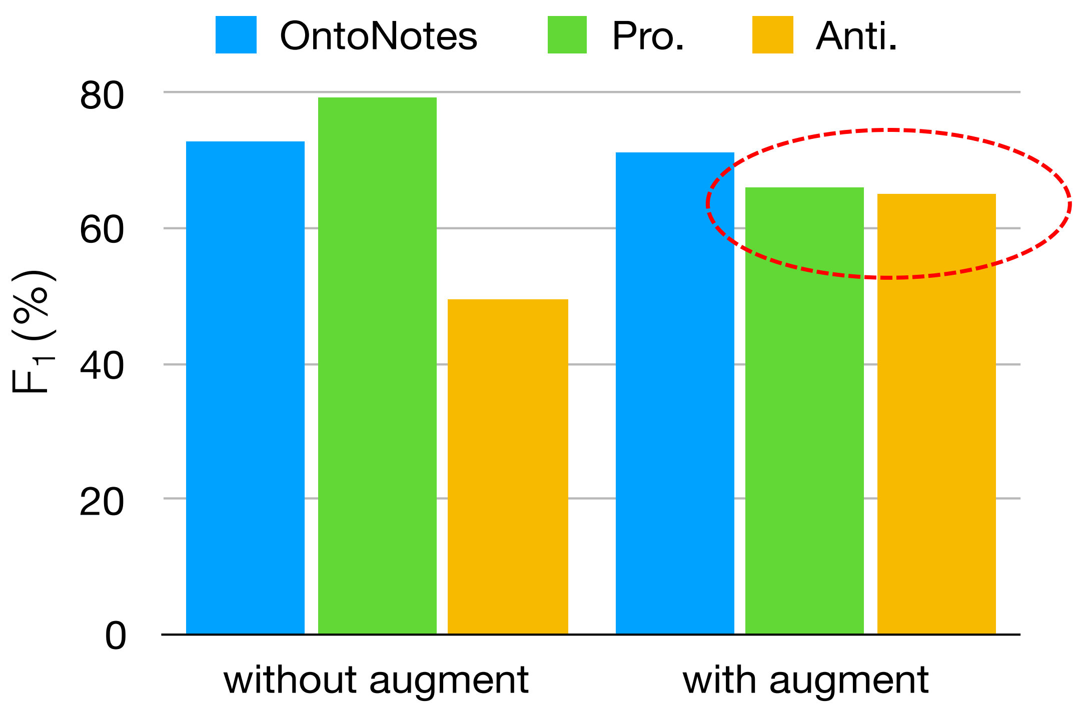
Mitigate Bias

- Neutralize ELMo Embeddings
  - Lightweight; keeps the performance
  - Mitigate some of the bias (in WinoBias)



Mitigate Bias

- Data Augmentation
  - Retrain the model
  - Mitigate almost all the biases (in WinoBias)



Outline

Bias in NLP Modules

Bias in Language Representations

Bias Amplification

What's the agent for this image?

Cooking	
Role	Noun
agent	
food	vegetable
container	bowl
tool	knife
place	kitchen

- Jieyu Zhao, et al. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." EMNLP (2017). Best Long Paper Award

Dataset Gender Bias

33%

Male

66%

Female

imsitu.org

credit to: Mark Yatskar

Gender Bias Amplification

16%

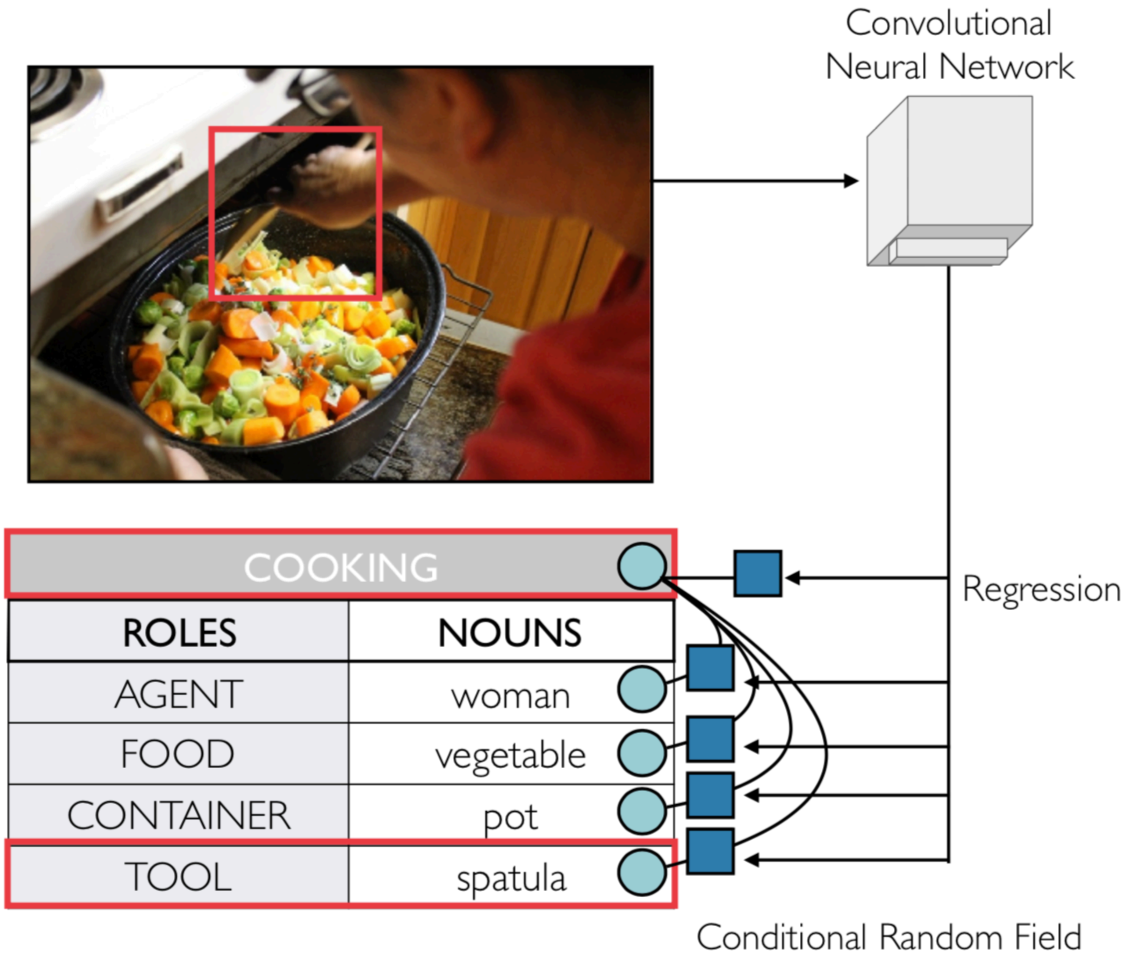
Male

84%

Female

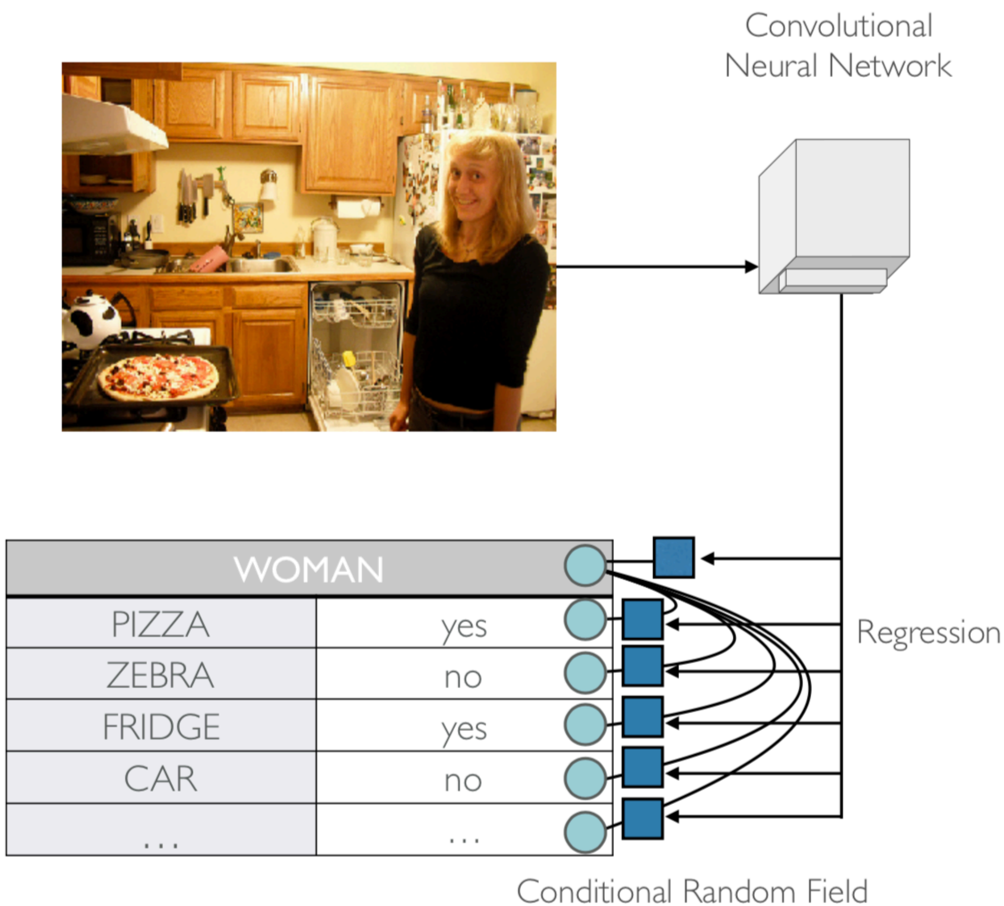
imsitu.org

imSitu: visual Semantic Role Labeling (Activity/Verb)



UCLA NLP

MLC: COCO Multi-Label Classification (Object/Noun)



Visualize and Quantify the Bias

Training Gender Ratio (▲ noun)

Training Set

▲ snowboard
● woman
● man

● MAN	
▲ snowboard	yes
refrigerator	no
bowl	no

● WOMAN	
▲ snowboard	yes
refrigerator	no
bowl	no

$$\frac{\#(\blacktriangle\text{snowboard}, \bullet\text{man})}{\#(\blacktriangle\text{snowboard}, \bullet\text{man}) + \#(\blacktriangle\text{snowboard}, \bullet\text{woman})} = 2/3$$

Model Bias Amplification

Predicted Gender Ratio (verb)

Development Set

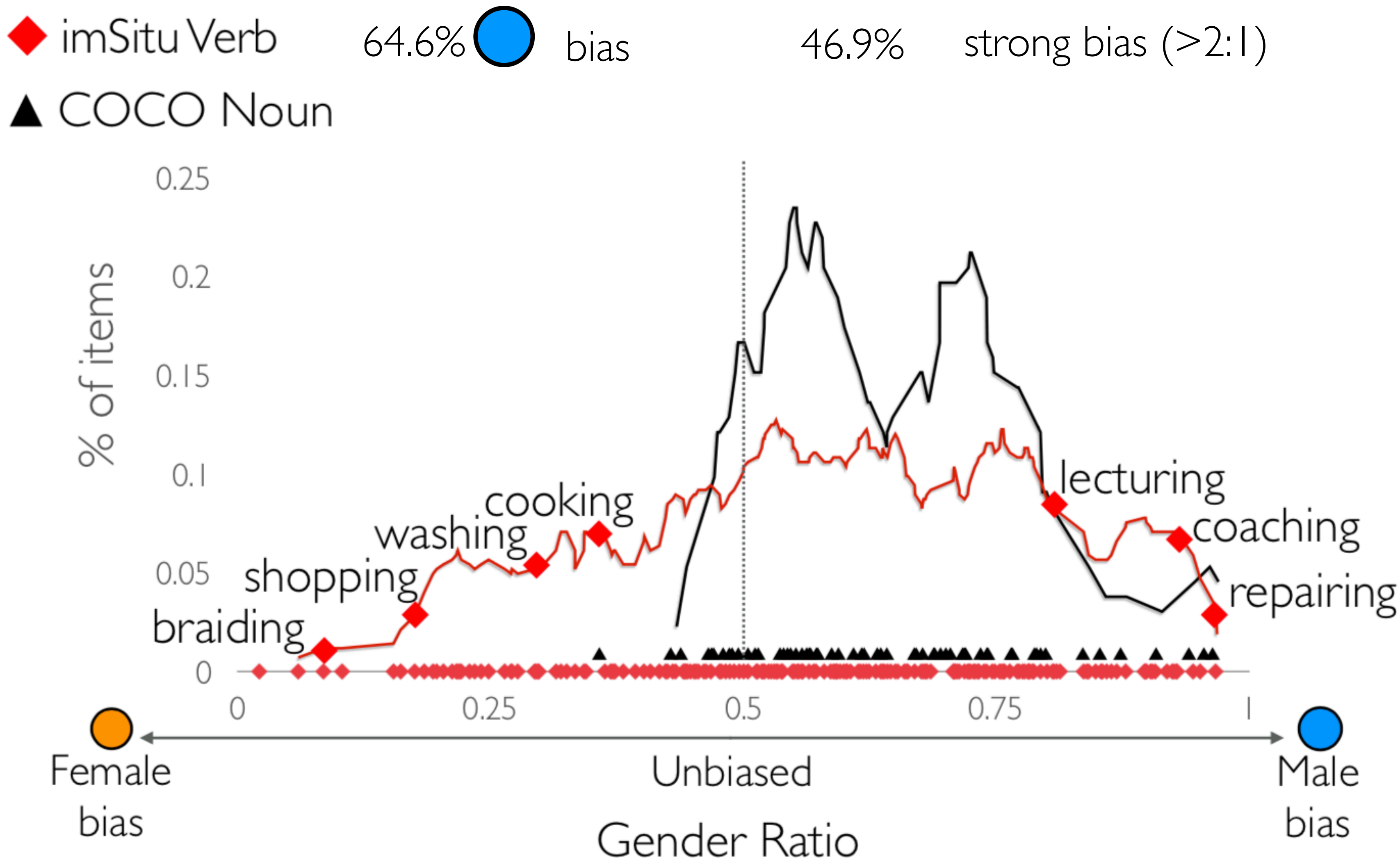
- cooking
- woman
- man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

$$\frac{\#(\text{cooking}, \text{man})}{\#(\text{cooking}, \text{man}) + \#(\text{cooking}, \text{woman})} = 1/6$$

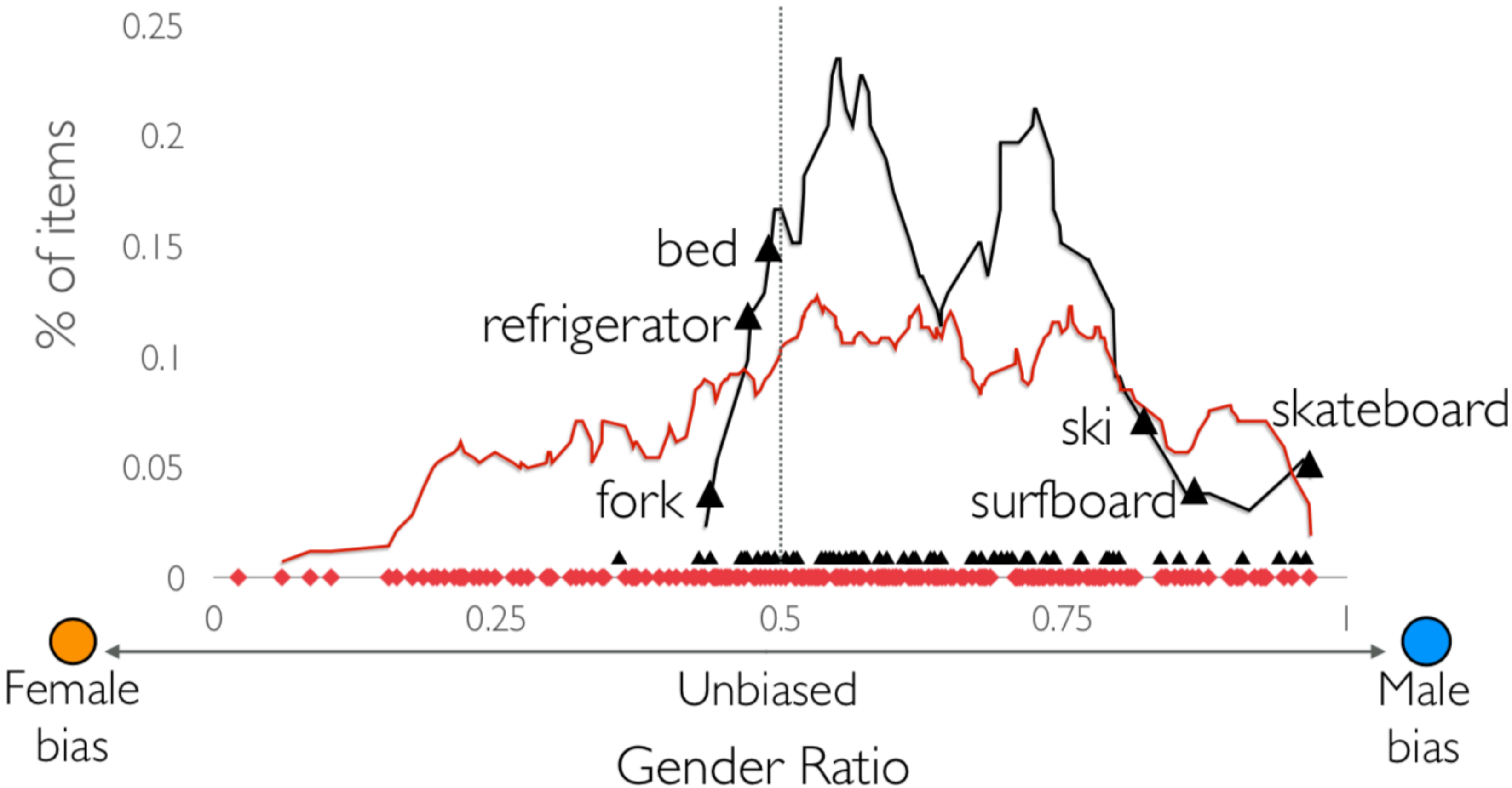
Gender Dataset Bias



Gender Dataset Bias

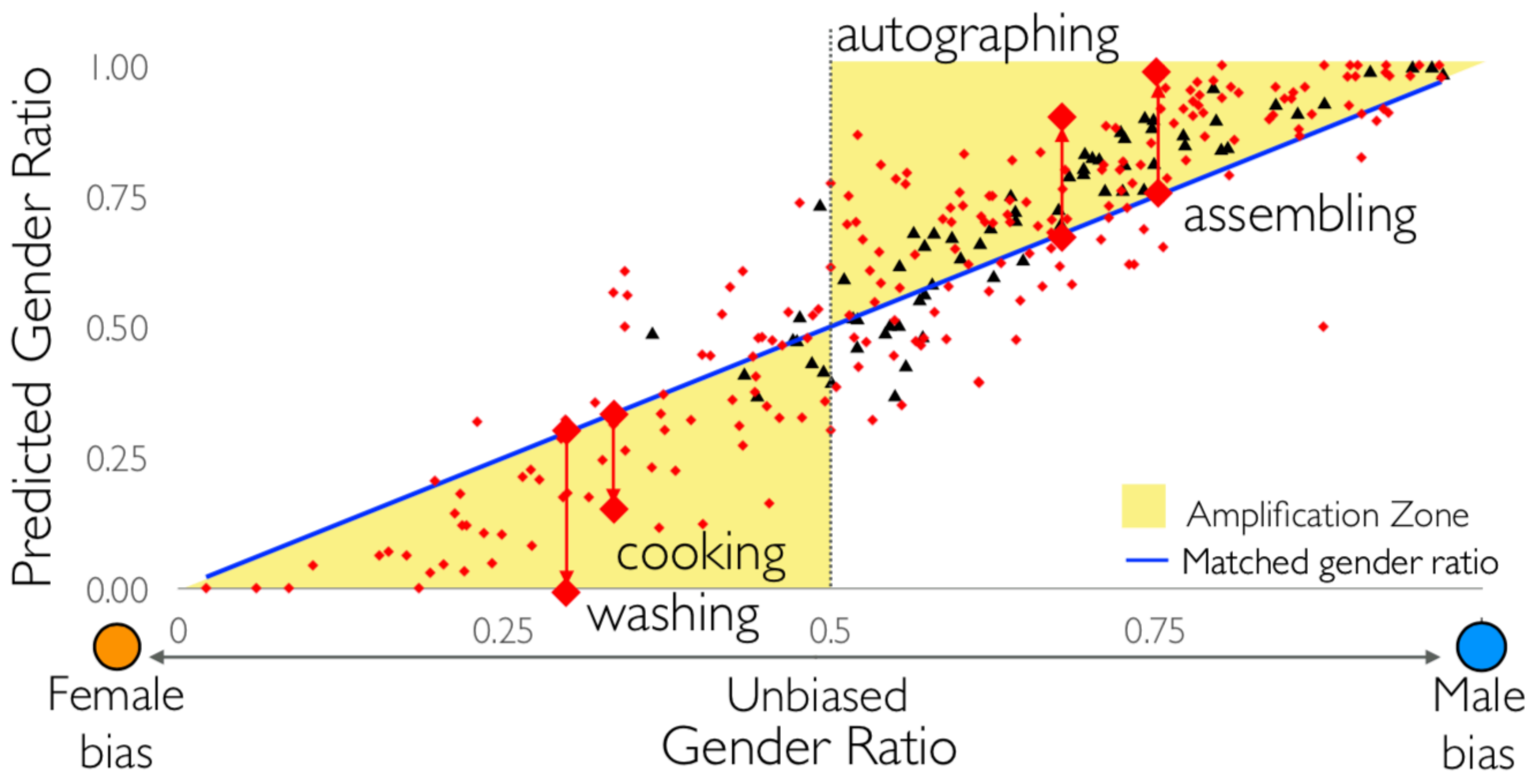
◆ imSitu Verb 64.6% bias 46.9% strong bias (>2:1)

▲ COCO Noun 86.6% bias 37.9% strong bias (>2:1)

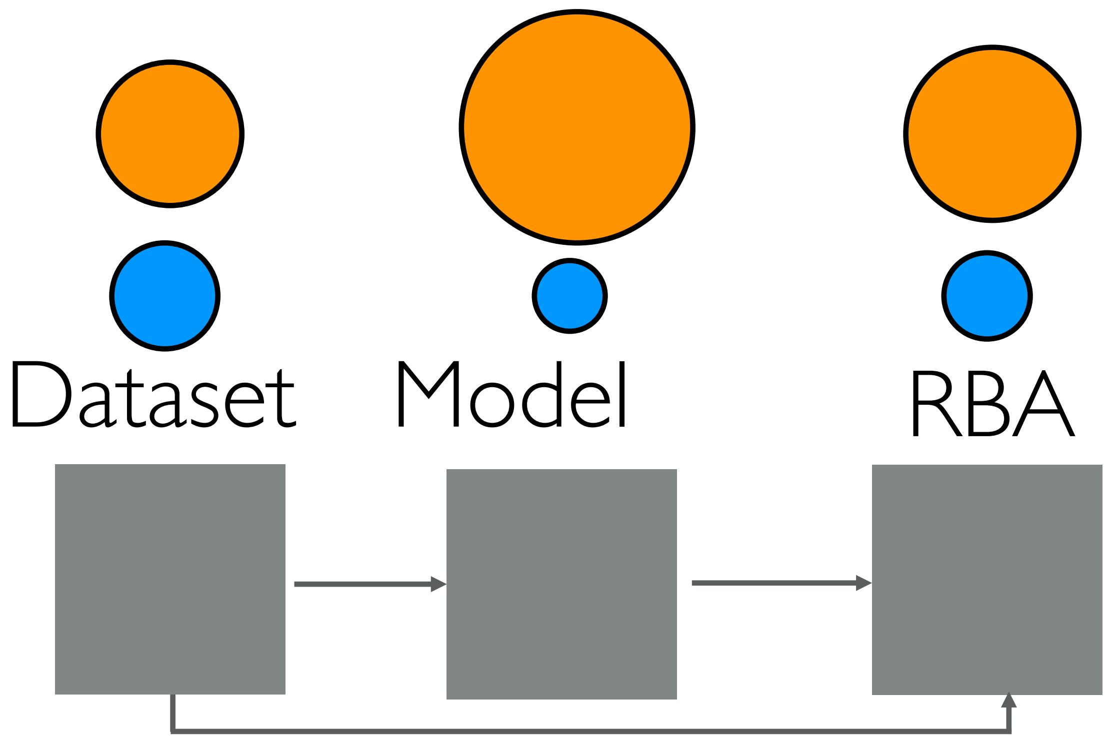


Model Bias Amplification

◆ imSitu Verb	69%	bias$\uparrow$	.05	$\lvert$bias$\uparrow\rvert$	strong bias(> 2:1) :	.07 $\lvert$bias$\uparrow\rvert$
▲ COCO Noun	73%	bias$\uparrow$	.04	$\lvert$bias$\uparrow\rvert$	strong bias(> 2:1) :	.08 $\lvert$bias$\uparrow\rvert$



Reducing Bias Amplification (RBA)



- Make the model avoid making biased decisions

Reducing Bias Amplification (RBA)

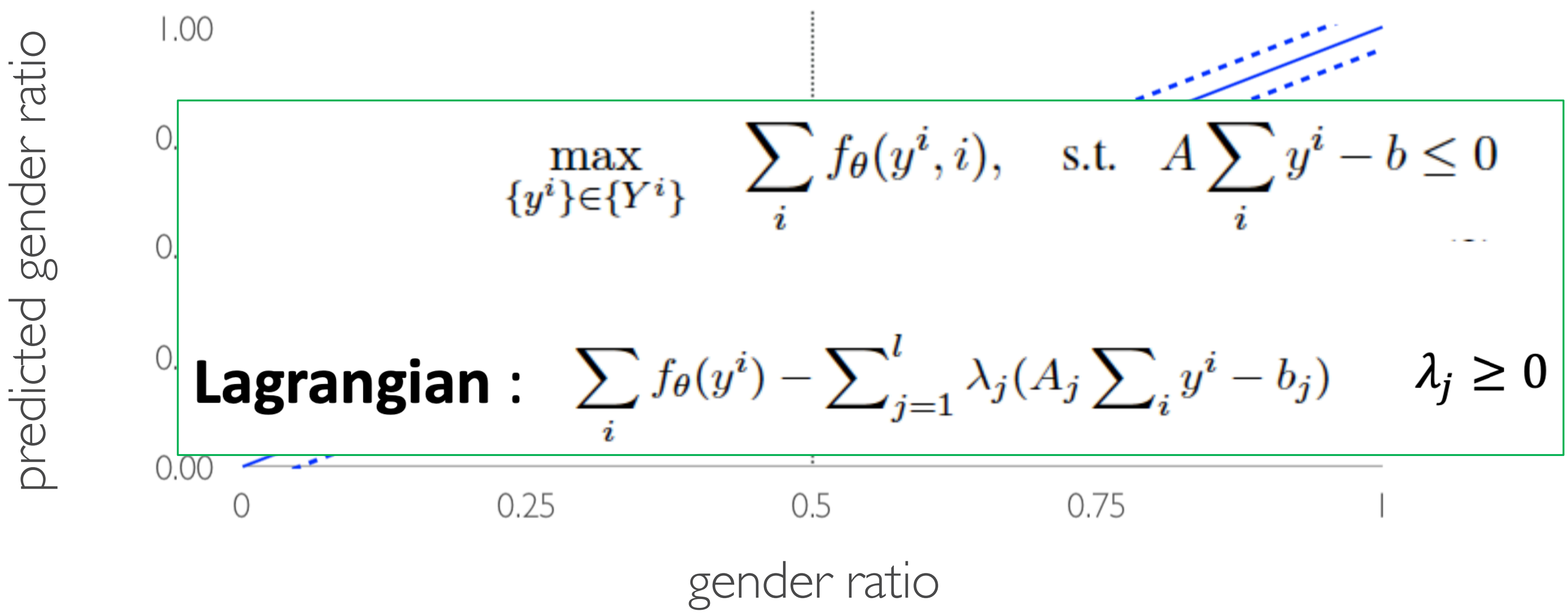
Integer Linear Program

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

Goal of the original model

$$\forall \text{ points} \quad \left| \text{Training Ratio} - \underset{f(y_1 \ldots y_n)}{\text{Predicted Ratio}} \right| <= \text{margin}$$

Our control for calibration



$$\max_{\{y^i\}\in\{Y^i\}} \sum_i f_\theta(y^i, i), \quad \text{s.t.} \quad A\sum_i y^i - b \leq 0$$

Lagrangian : $\sum_i f_\theta(y^i) - \sum_{j=1}^{l} \lambda_j (A_j \sum_i y^i - b_j) \qquad \lambda_j \geq 0$

Bias De-amplification

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\boxed{\sum_i \max_{y_i} \ s(y_i, \text{image})}$$

$$\left|\underset{(1/2)}{\text{Training Ratio}} - \text{Predicted Ratio}\right| <= \text{margin}$$

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Bias De-amplification

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} \; s(y_i, \text{image})$$

$$\left| \underset{(1/2)}{\text{Training Ratio}} - \text{Predicted Ratio} \right| <= \text{margin}$$

- Lagrange Multiplier (λ) Per Constraint

inference

update λ

update potentials

Bias De-amplification

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left|\text{Training Ratio} - \text{Predicted Ratio}\right| <= \text{margin}$$

(1/2)

- Lagrange Multiplier (λ) Per Constraint

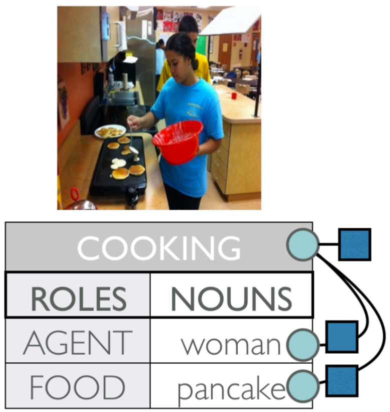
inference

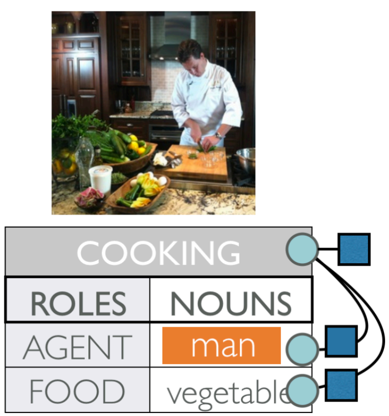
update λ

update potentials

Bias De-amplification

Lagrangian Relaxation



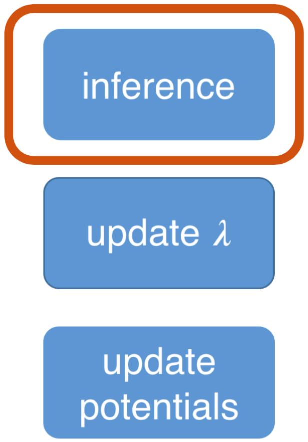


$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\left| \text{Training Ratio} - \text{Predicted Ratio} \right| <= \text{margin}$$

(1/2)

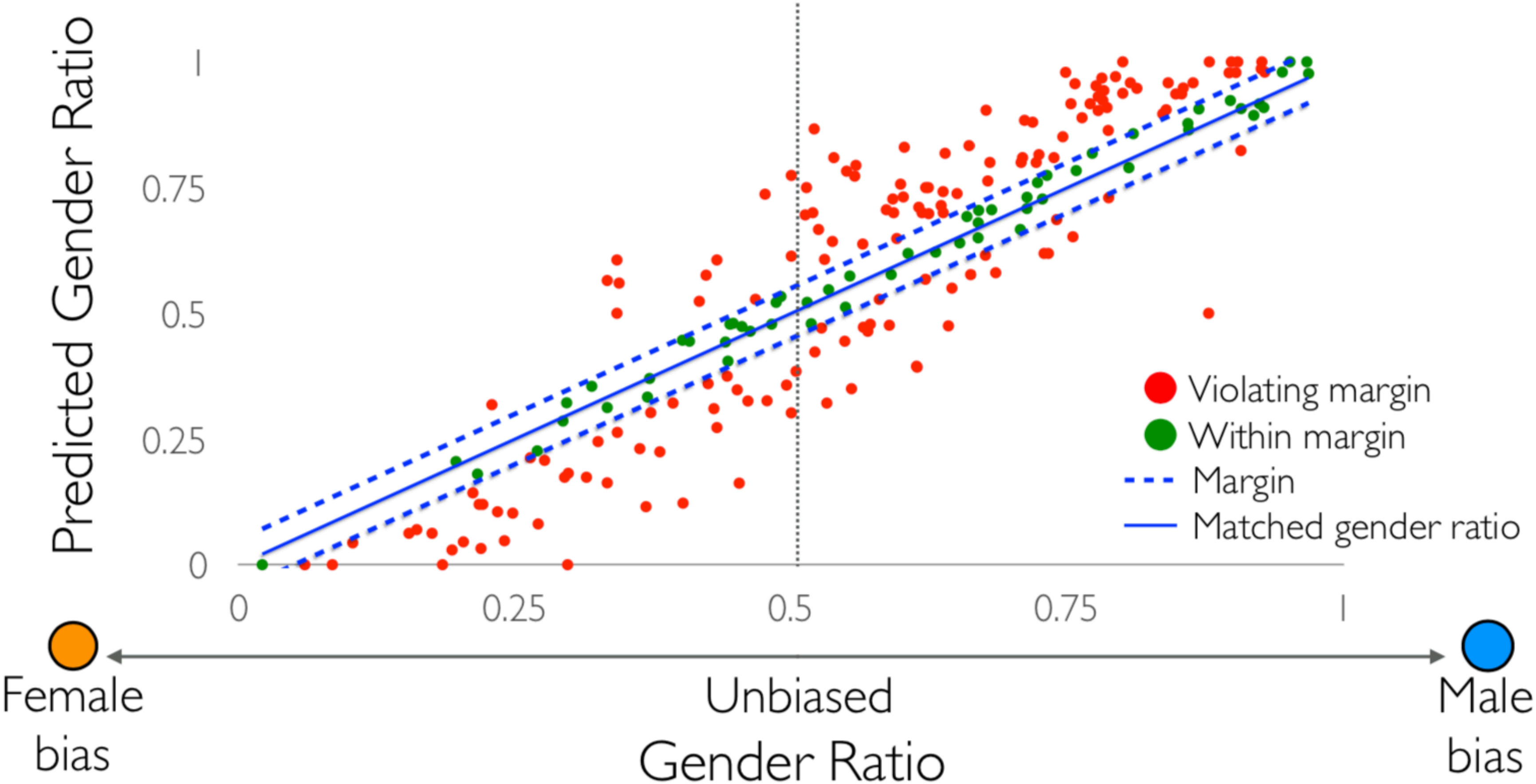
- Lagrange Multiplier (λ) Per Constraint



UCLA NLP

Gender Bias De-amplification in imSitu

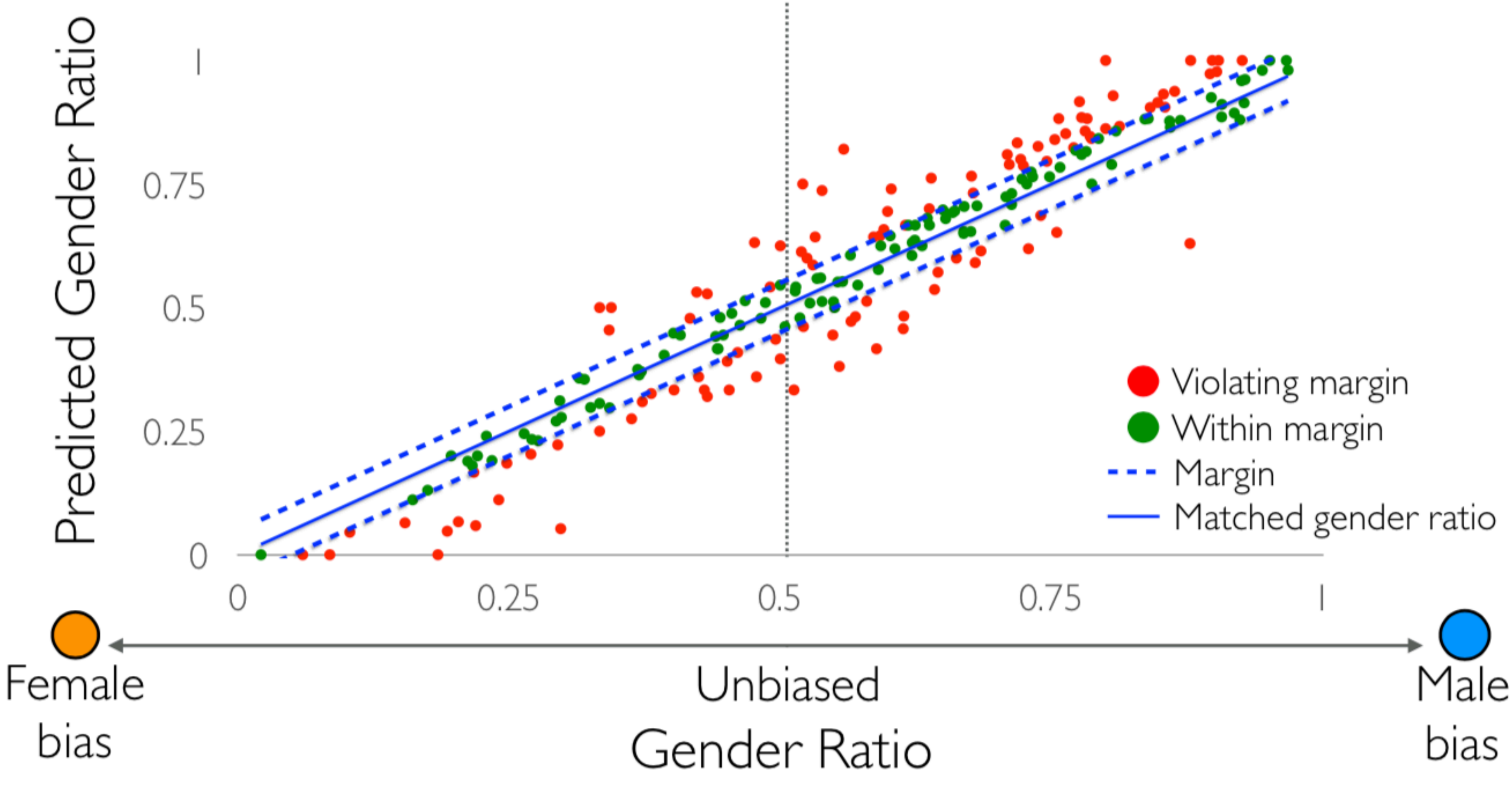
imSitu Verb   Violation: 72.6%   .050 |bias↑| 24.07 acc.



UCLA NLP

Gender Bias De-amplification in imSitu

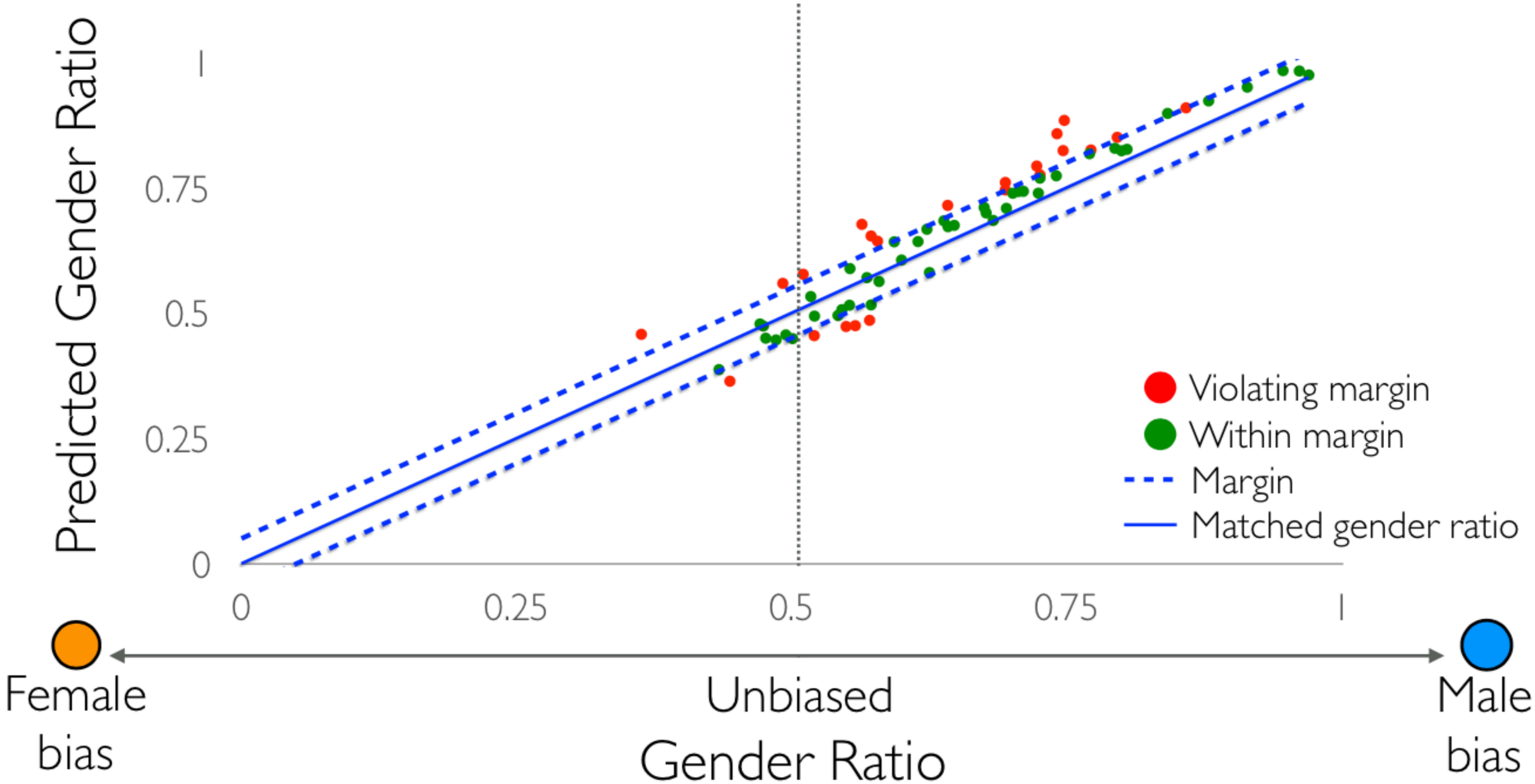
imSitu Verb	Violation: 72.6%	.050 |bias↑|	24.07 acc.
w/ RBA	Violation: 50.5%	.024 |bias↑|	23.97 acc.



UCLA NLP

Gender Bias De-amplification in COCO

COCO Noun	Violation: 60.6%	.032 |bias↑|	45.27	mAP
w/ RBA	Violation: 36.4%	.022 |bias↑|	45.19	mAP



UCLA NLP

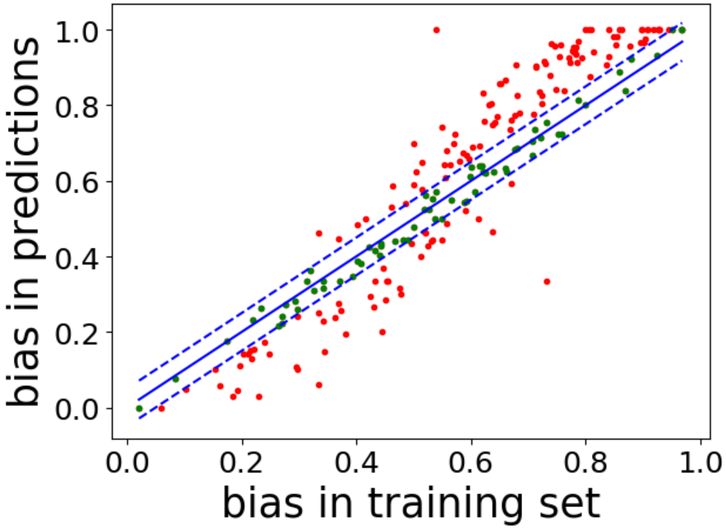
Mitigating Gender Bias Amplification in Distribution by Posterior Regularization[1]

- Top Prediction vs. Distribution Prediction
  - Top prediction (Zhao et. al. 17):
    - Model is forced to make one decision
      - Even similar probabilities for "female" and "male" predictions
    - Potentially amplify the bias
  - Distribution of predictions
    - A better view of understanding bias amplification
    - Model is trained using regularized maximum likelihood objective

[1] S. Jia*, T. Meng*, **J. Zhao** and K. Chang. ACL 2020

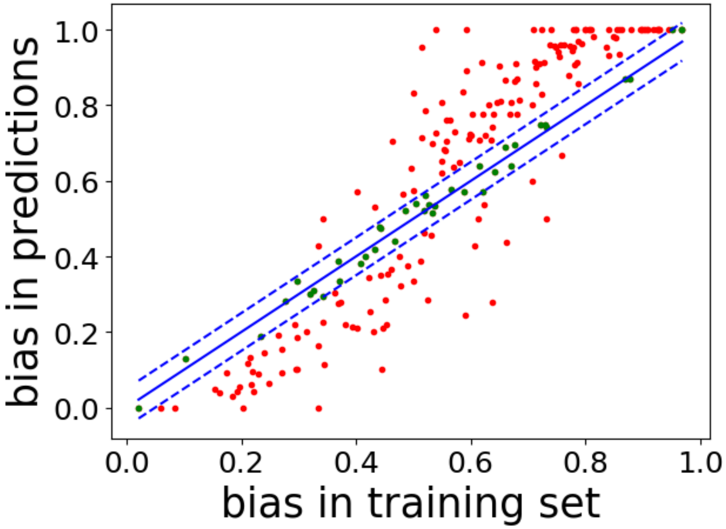
Bias Amplification in Distribution

51.4% violations



Posterior Distribution

81.6% violations

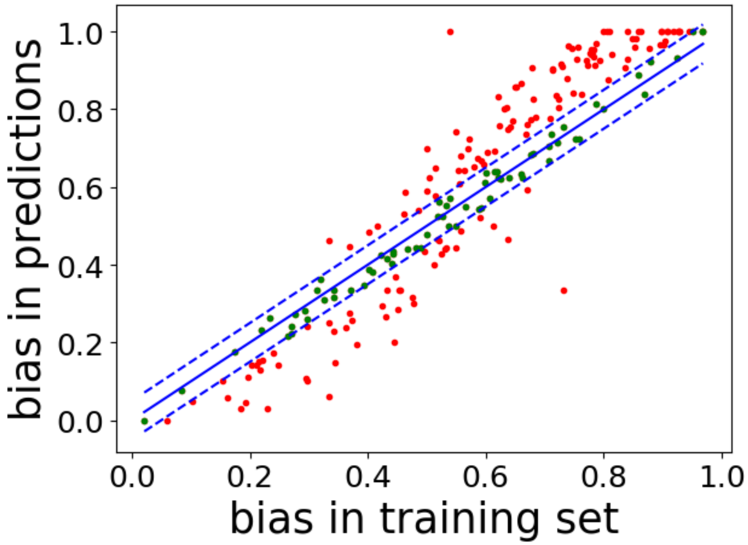


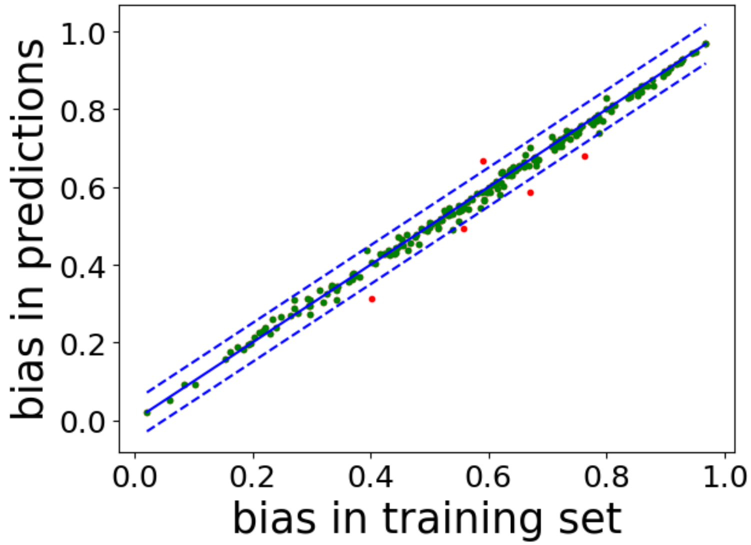
Top prediction (EMNLP'17)

S. Jia*, T. Meng*, **J. Zhao** and K. Chang. ACL 2020

Bias Mitigation Using Posterior Regularization

vSRL	Violation: 51.4%	Amplification: 0.032	Acc.: 23.2%
w/ PR	Violation: 2%	Amplification: -0.005	Acc.: 23.1%

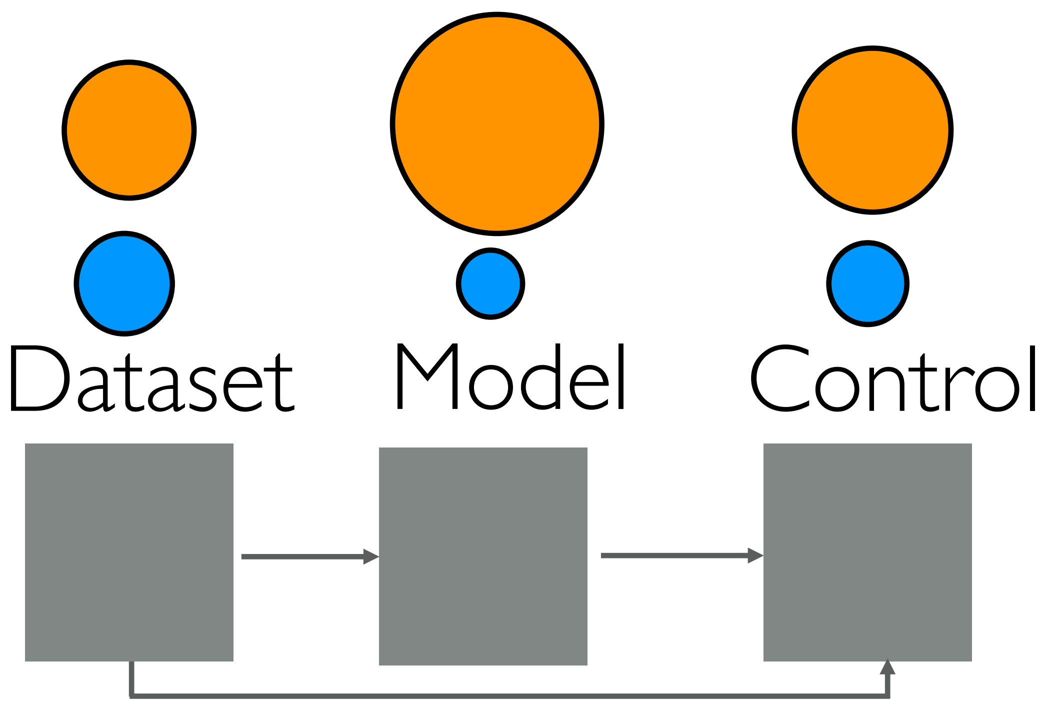




Bias Amplification

Top Prediction
Distribution Prediction

Lagrange Regularization
Posterior Regularization



UCLA NLP

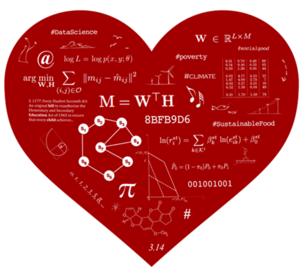
Conclusion

- ❖ Biases are embedded in NLP models
- ❖ Controlling Biases is still an open problem

Our Group Page:
http://web.cs.ucla.edu/~kwchang/members/

K G - B I A S

#Data4Good: Machine Learning in Social Good Applications
June 24, New York City, @ICML



AI for Social Good