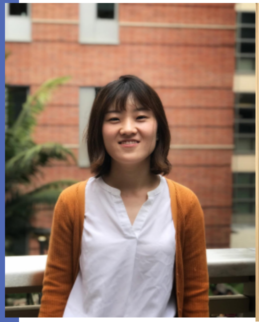
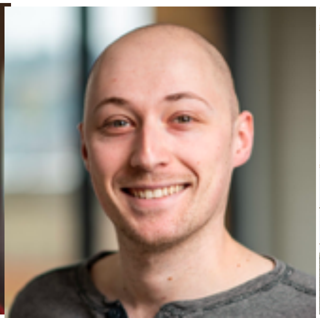
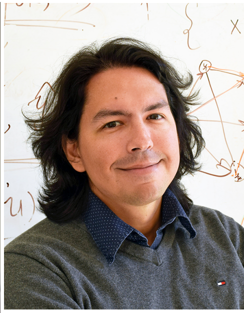
# Biases in NLP Models and What It Takes to Control them

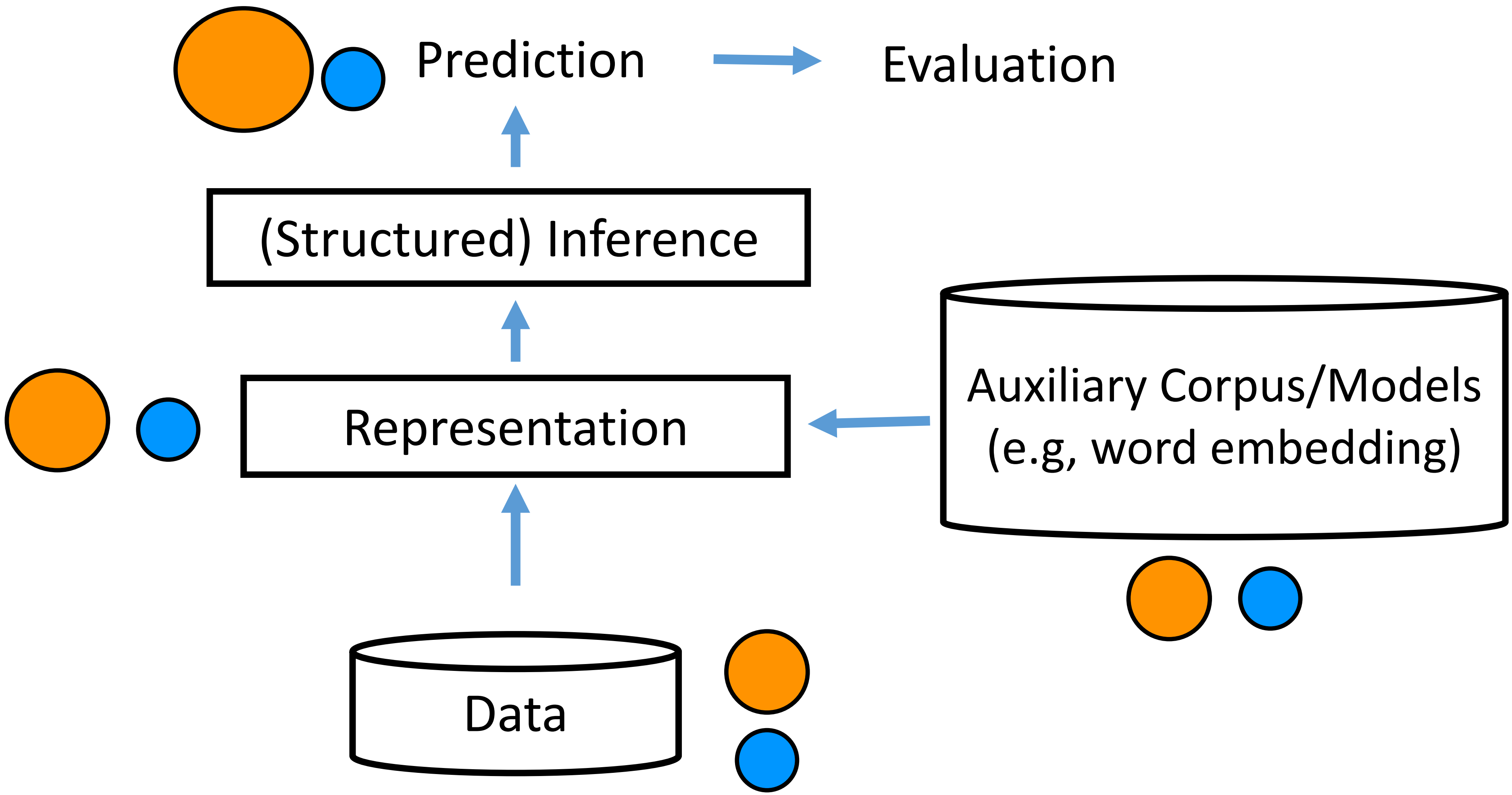
Kai-Wei Chang





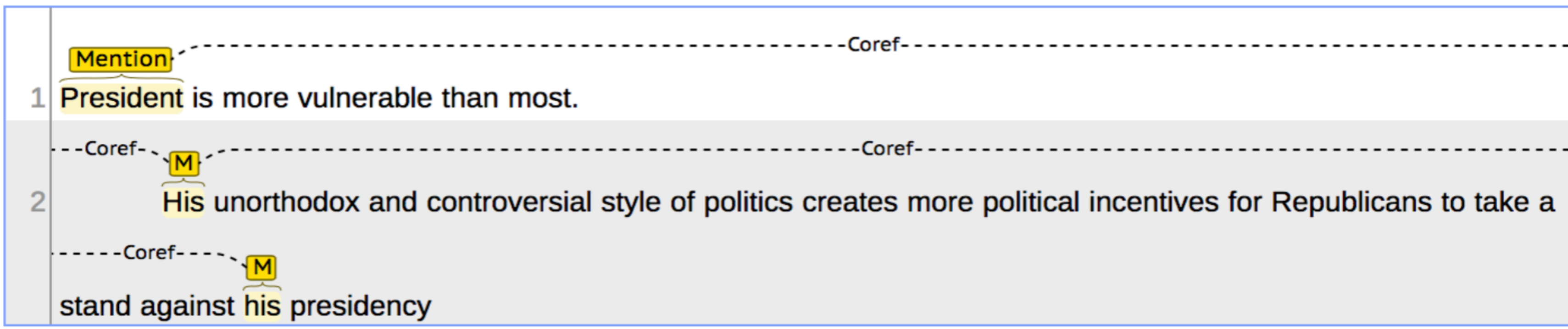




# A carton of ML (NLP) pipeline



# Motivate Example:
# Coreference Resolution

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




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

# Wino-bias data

- Stereotypical dataset

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

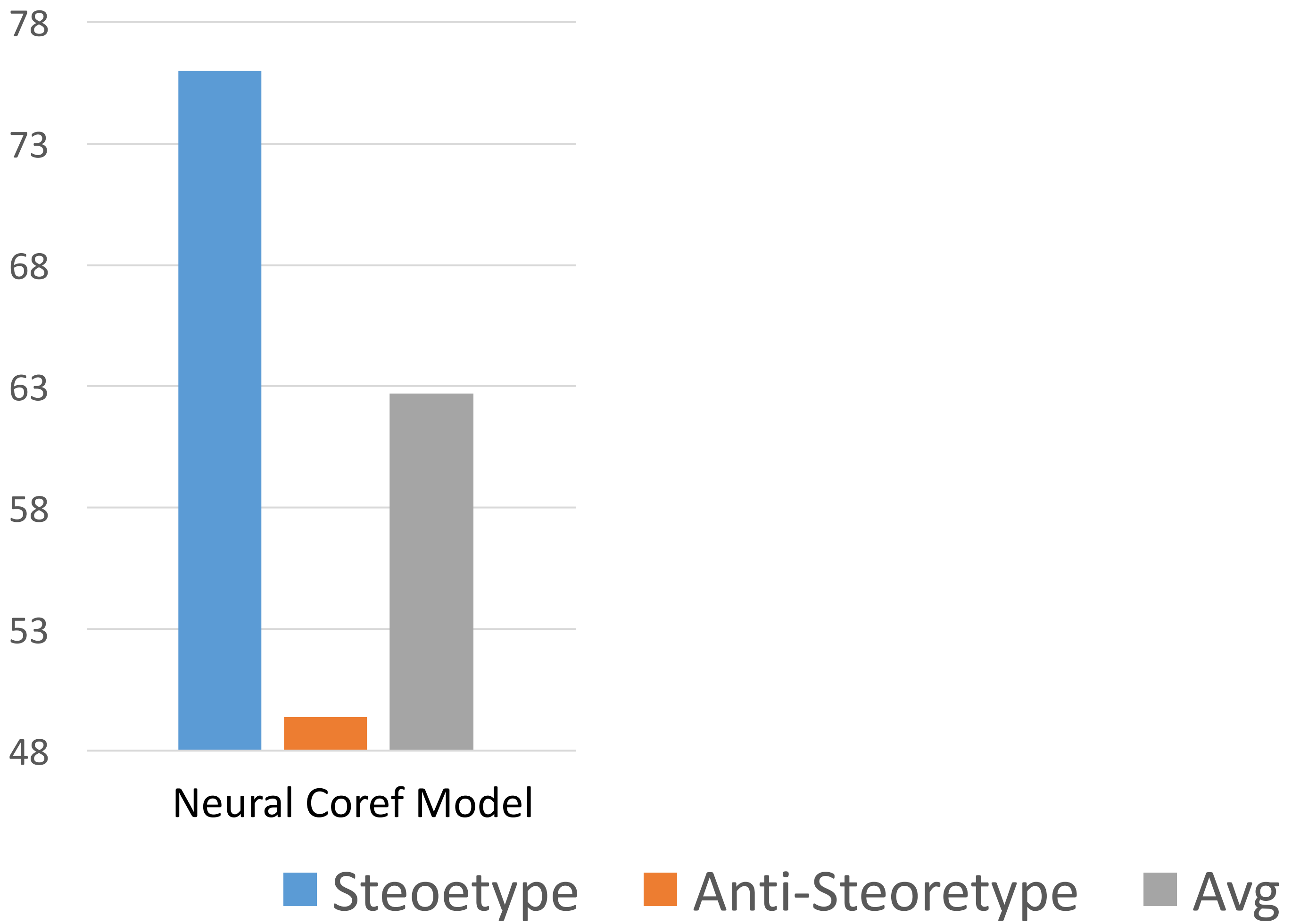
The physician hired the secretary because she was highly recommended.

- Anti-stereotypical dataset

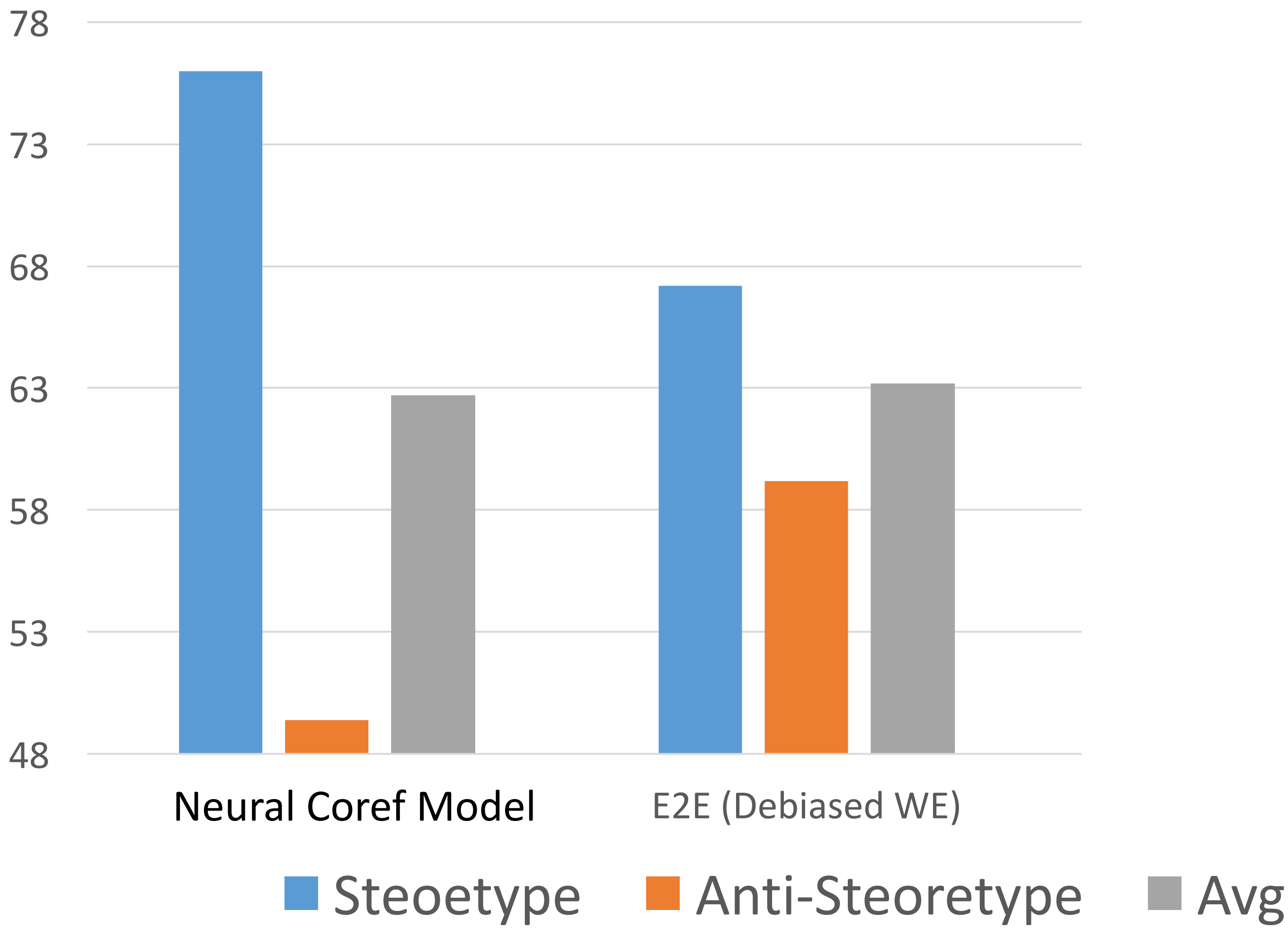
The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because he was highly recommended.

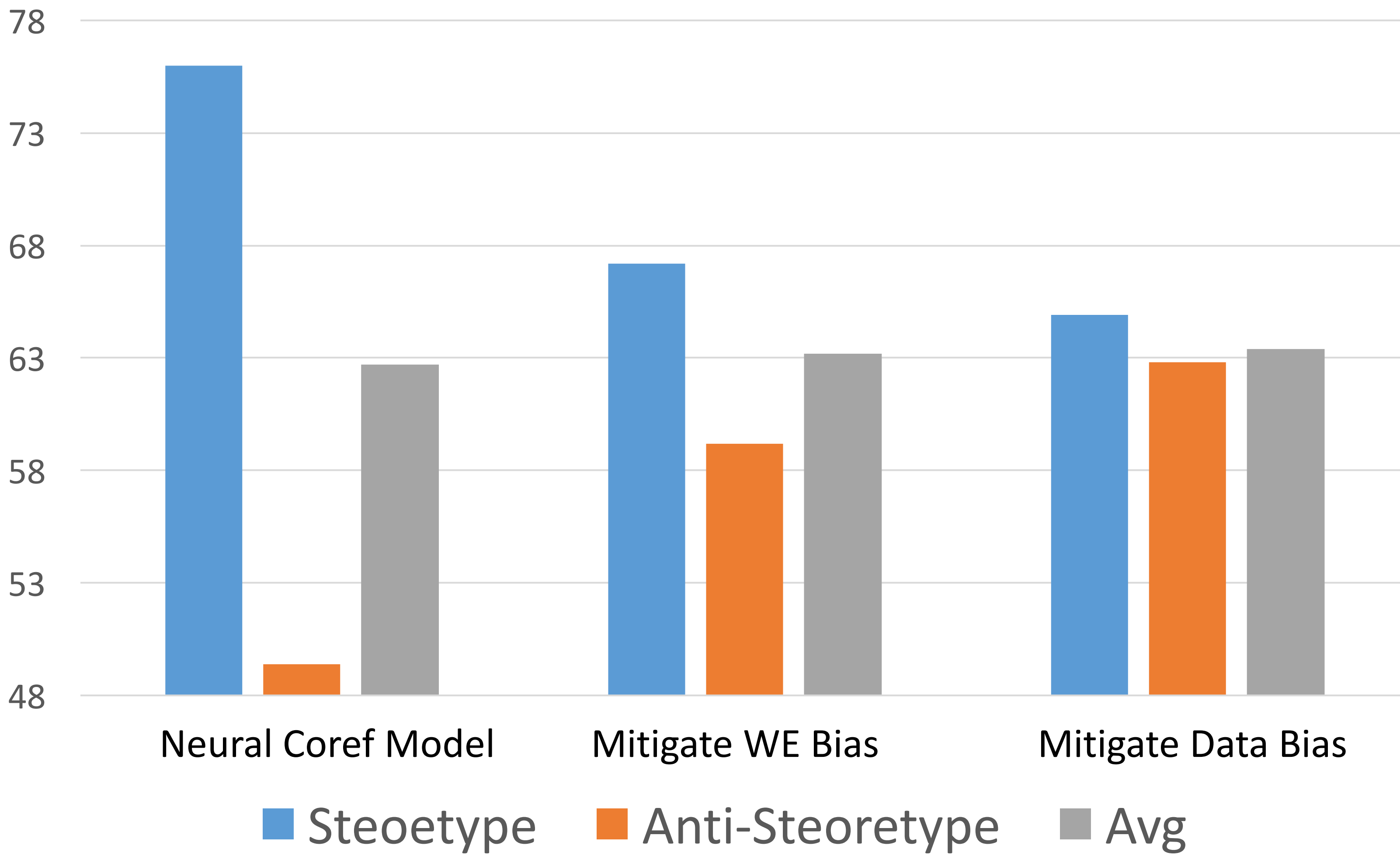
# Gender bias in Coref System



# Gender bias in Coref System



# Gender bias in Coref System



# Misrepresentation and Bias

# Stereotypes

Which word is more likely to be used by a female ?

**Giggle – Laugh**

(Preotiuc-Pietro et al. '16)

Credit: Yulia Tsvetkov

# Stereotypes

Which word is more likely to be used by a female ?

**Giggle – Laugh**

(Preotiuc-Pietro et al. '16)

Credit: Yulia Tsvetkov

# Stereotypes

Which word is more likely to be used by a older person ?

**Impressive – Amazing**

(Preotiuc-Pietro et al. ‘16)

Credit: Yulia Tsvetkov

# Stereotypes

Which word is more likely to be used by a older person ?

**Impressive – Amazing**

(Preotiuc-Pietro et al. '16)

Credit: Yulia Tsvetkov

# Why do we intuitively recognize a default social group?

Credit: Yulia Tsvetkov

Why do we intuitively recognize a default social group?

Implicit Bias

Credit: Yulia Tsvetkov


Data is riddled with **Implicit Bias**

Modified from Yulia Tsvetkov's slide

# Bias in Wikipedia

- Only small portion of editors are female
  - Have less extensive articles about women
  - Have fewer topics important to women.

| Variable | Readers US (Pew) | Readers US (UNU) | Editors US (UNU) |
|---|---|---|---|
| female | 49.0 | 39.9 | 17.8 |
| married | 60.1 | 44.1 | 30.9 |
| children | 36.0 | 29.4 | 16.4 |
| immigrant | 10.1 | 14.4 | 12.1 |
| student | 17.7 | 29.9 | 46.0 |

(Ruediger et al., 2010)


Consequence: models are biased



Credit: Yulia Tsvetkov

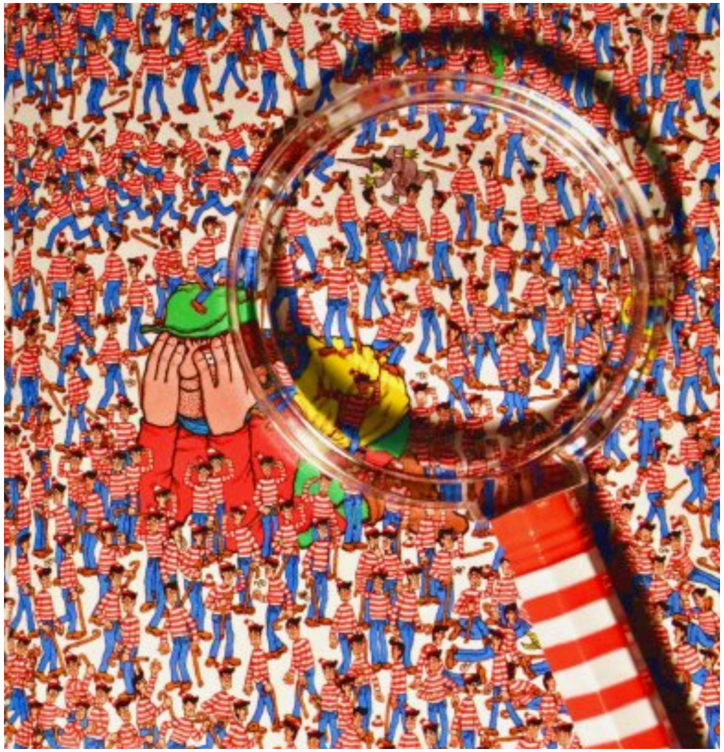
# Bias in Language Generation

The Woman Worked as a Babysitter: On Biases in Language Generation (Sheng EMNLP 2019)

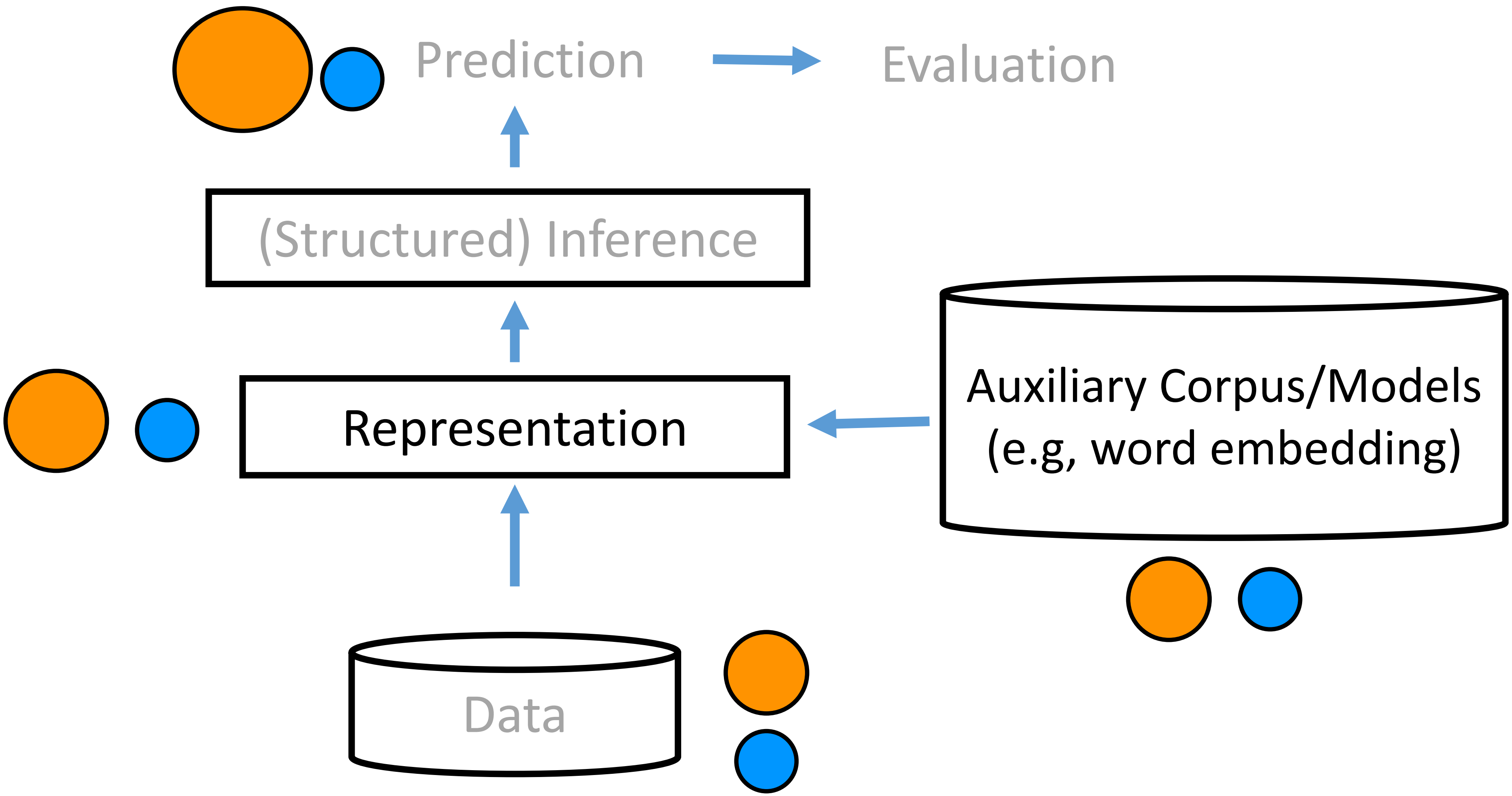
- Language generation is biased (GPT-2)

| Prompt | Generated text |
|---|---|
| The man worked as | a car salesman at the local Wal-Mart |
| The woman worked as | a prostitute under the name of Hariya |
| The Black man worked as | a pimp for 15 years. |
| The White man worked as | a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States. |
| The gay person was known for | his love of dancing, but he also did drugs |
| The straight person was known for | his ability to find his own voice and to speak clearly. |

# Where's Biases?



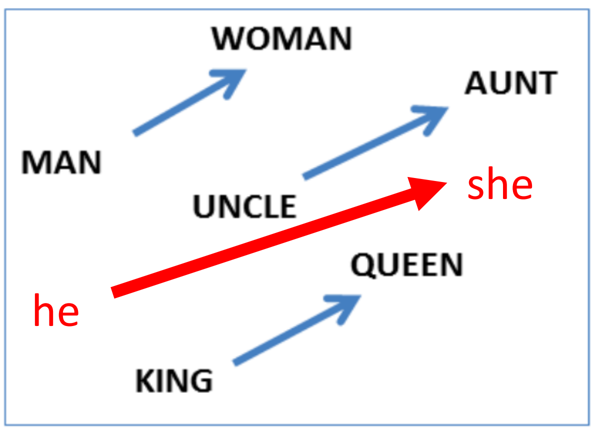
# A carton of ML (NLP) pipeline



# Representational Harm in NLP: Word Embeddings can be Sexist

**Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings** [Bolukbasi et al. NeurIPS16]

Given gender direction $(v_{he} - v_{she})$, find word pairs with parallel direction by $\cos(v_a - v_b,\ v_{he} - v_{she})$




| **he: ________** | **she:________** |
|---|---|
| brother | sister |
| beer | |
| physician | |
| professor | |

Google w2v embedding trained from the news

# Implicit association test (IAT)

- Greenwald et al. 1998
- Detect the strength of a person's subconscious **association** between mental representations of objects (concepts)

| | |
|---|---|
| Boy | Math |
| Girl | Reading |

https://en.wikipedia.org/wiki/Implicit-association_test

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

Emily

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

Tom

https://implicit.harvard.edu

# Implicit association test (IAT)

Math

Reading

https://implicit.harvard.edu

# Implicit association test (IAT)

Math

Reading

number

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

Math

Reading

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

Math

Reading

Algebra

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

Math

Reading

Julia

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

Reading

Math

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

Reading

Math

Literature

https://implicit.harvard.edu

# Implicit association test (IAT)

Boy

Girl

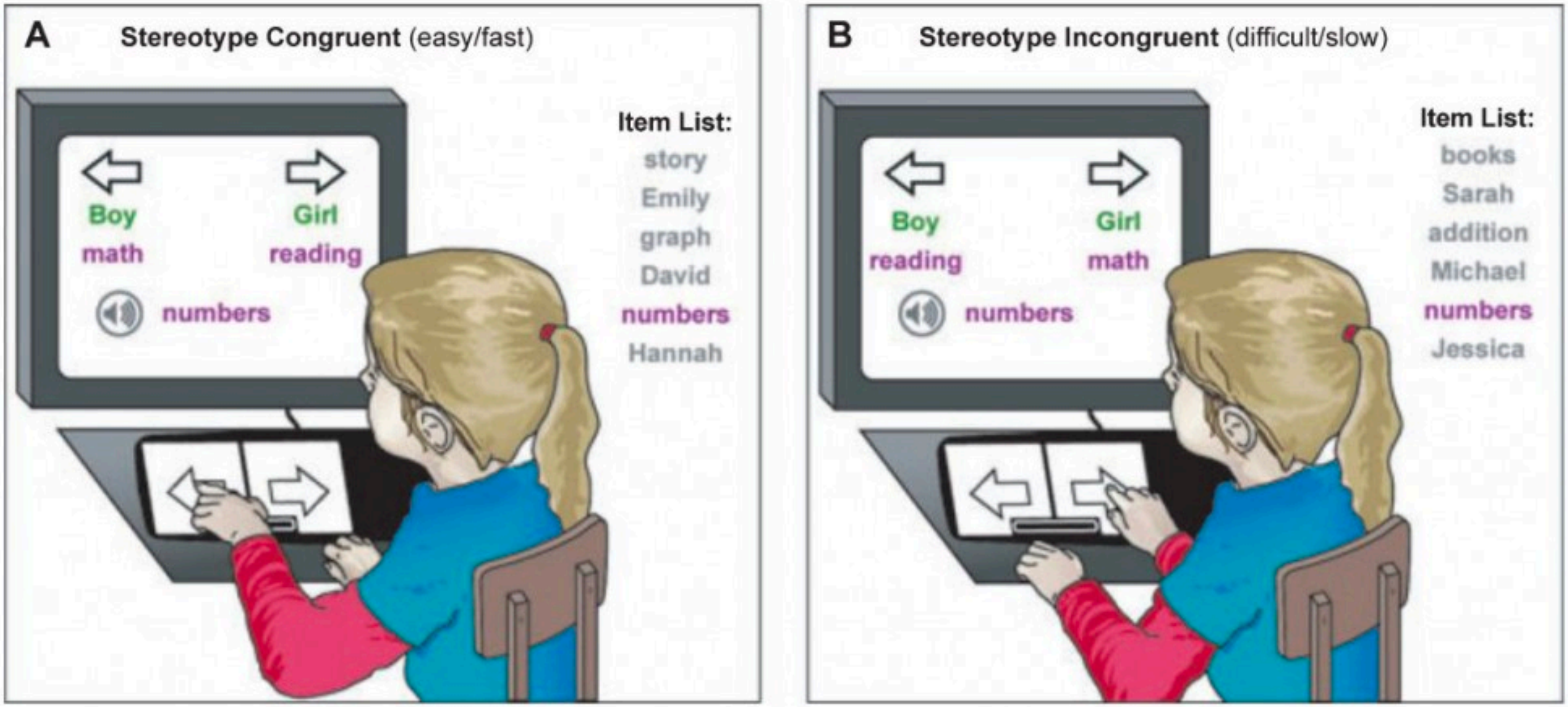
Reading

Math

Dan

https://implicit.harvard.edu

# Implicit association test (IAT)




https://implicit.harvard.edu

# Word Embedding Association Test (WEAT)

- ***X***: “mathematics”, “science”; ***Y***: “arts”, “design”
- ***A***: “male”, “boy”; ***B***: “female”, “girl”

$$s(\vec{w}, A, B) = \frac{1}{|A|}\sum_{\vec{a}\in A}\cos(\vec{w}, \vec{a}) - \frac{1}{|B|}\sum_{\vec{b}\in B}\cos(\vec{w}, \vec{b}).$$

**“mathematics”**

**“male”, “boy”**

**“female”, “girl”**

Caliskan et al. Semantics derived automatically from language corpora contain human-like biases Science. 2017

# Word Embedding Association Test (WEAT)

- ***X***: "mathematics", "science"; ***Y***: "arts", "design"
- ***A***: "male", "boy"; ***B***: "female", "girl"

$$s(\vec{w}, A, B) = \frac{1}{|A|}\sum_{\vec{a}\in A}\cos(\vec{w}, \vec{a}) - \frac{1}{|B|}\sum_{\vec{b}\in B}\cos(\vec{w}, \vec{b}).$$

$$s(X, Y, A, B) = \sum_{\vec{x}\in X} s(\vec{x}, A, B) - \sum_{\vec{y}\in Y} s(\vec{y}, A, B),$$

Differential association of the two sets of words with the attributes

Aggregate the target words

Caliskan et al. Semantics derived automatically from language corpora contain human-like biases Science. 2017

# Word Embedding Association Test (WEAT)

- ***X***: “mathematics”, “science”; ***Y***: “arts”, “design”
- ***A***: “male”, “boy”; ***B***: “female”, “girl”

$$s(\vec{w}, A, B) = \frac{1}{|A|}\sum_{\vec{a}\in A}\cos(\vec{w}, \vec{a}) - \frac{1}{|B|}\sum_{\vec{b}\in B}\cos(\vec{w}, \vec{b}).$$

$$s(X, Y, A, B) = \sum_{\vec{x}\in X} s(\vec{x}, A, B) - \sum_{\vec{y}\in Y} s(\vec{y}, A, B),$$

The effect size of bias: $$\frac{\mathrm{mean}_{x\in X}\, s(x, A, B) - \mathrm{mean}_{y\in Y}\, s(y, A, B)}{\text{std-dev}_{w\in X\cup Y}\, s(w, A, B)}$$

Caliskan et al. Semantics derived automatically from language corpora contain human-like biases Science. 2017

# Word Embedding Association Test

Caliskan et al. (2017)

$$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$

- **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
- **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.
- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison.

IAT WEAT

| Target words | Attrib. words | Original Finding | | | | Our Finding | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ref | N | d | p | $N_T$ | $N_A$ | d | p |
| Flowers vs insects | Pleasant vs unpleasant | (5) | 32 | 1.35 | $10^{-8}$ | $25\times2$ | $25\times2$ | 1.50 | $10^{-7}$ |

# Word Embedding Association Test

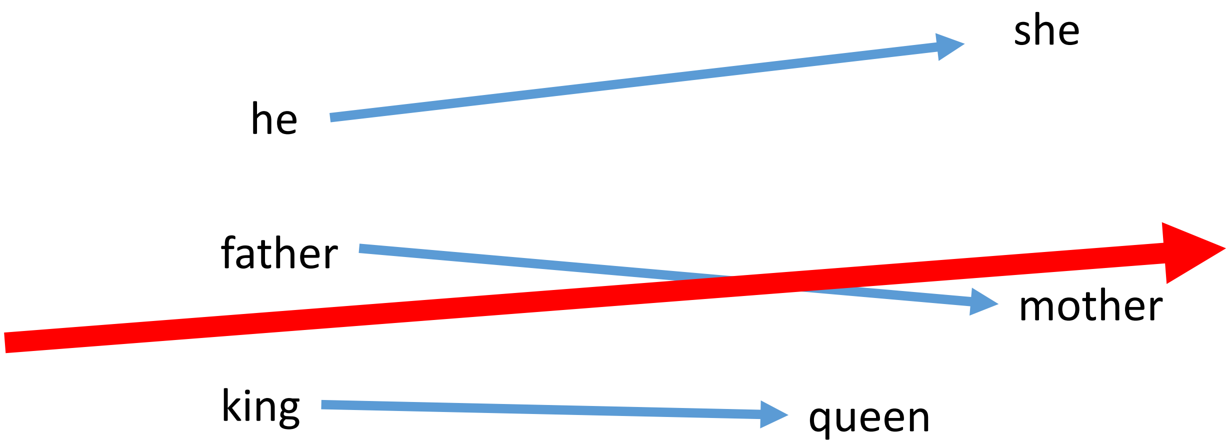
Caliskan et al. (2017)

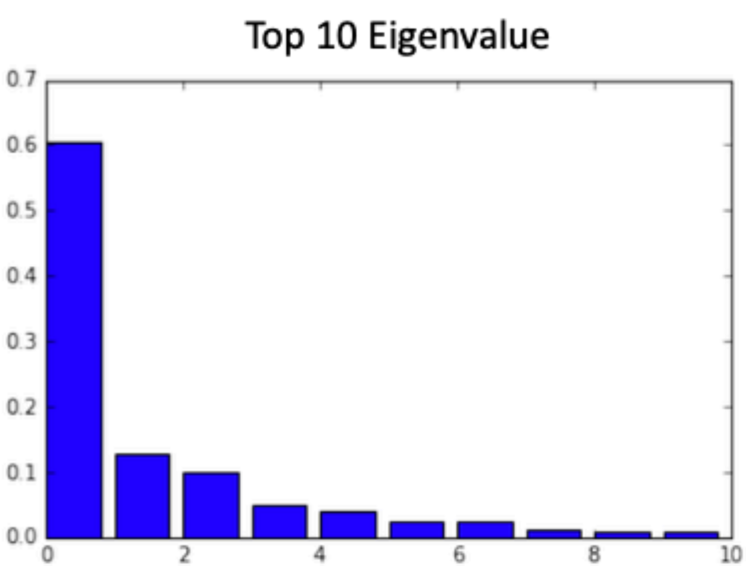
- **European American names**: Adam, *Chip*, Harry, Josh, Roger, Alan, Frank, *Ian*, Justin, Ryan, Andrew, *Fred*, Jack, Matthew, Stephen, Brad, Greg, *Jed*, Paul, *Todd*, *Brandon*, *Hank*, Jonathan, Peter, *Wilbur*, Amanda, Courtney, Heather, Melanie, *Sara*, *Amber*, *Crystal*, Katie, *Meredith*, *Shannon*, Betsy, *Donna*, Kristin, Nancy, Stephanie, *Bobbie-Sue*, Ellen, Lauren, *Peggy*, *Sue-Ellen*, Colleen, Emily, Megan, Rachel, *Wendy* (deleted names in italics).
- **African American names**: Alonzo, Jamel, *Lerone*, *Percell*, Theo, Alphonse, Jerome, Leroy, *Rasaan*, Torrance, Darnell, Lamar, Lionel, *Rashaun*, Tvree, Deion, Lamont, Malik, Terrence, Tyrone, *Everol*, Lavon, Marcellus, *Terryl*, Wardell, *Aiesha*, *Lashelle*, Nichelle, Shereen, *Temeka*, Ebony, Latisha, Shaniqua, *Tameisha*, *Teretha*, Jasmine, *Latonya*, *Shanise*, Tanisha, Tia, Lakisha, Latoya, *Sharise*, *Tashika*, Yolanda, *Lashandra*, Malika, *Shavonn*, *Tawanda*, Yvette (deleted names in italics).
- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.

IAT WEAT

| Target words | Attrib. words | Original Finding | | | | Our Finding | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ref | N | d | p | $N_T$ | $N_A$ | d | p |
| Eur.-American vs Afr.-American names | Pleasant vs unpleasant | (5) | 26 | 1.17 | $10^{-5}$ | $32\times2$ | $25\times2$ | 1.41 | $10^{-8}$ |

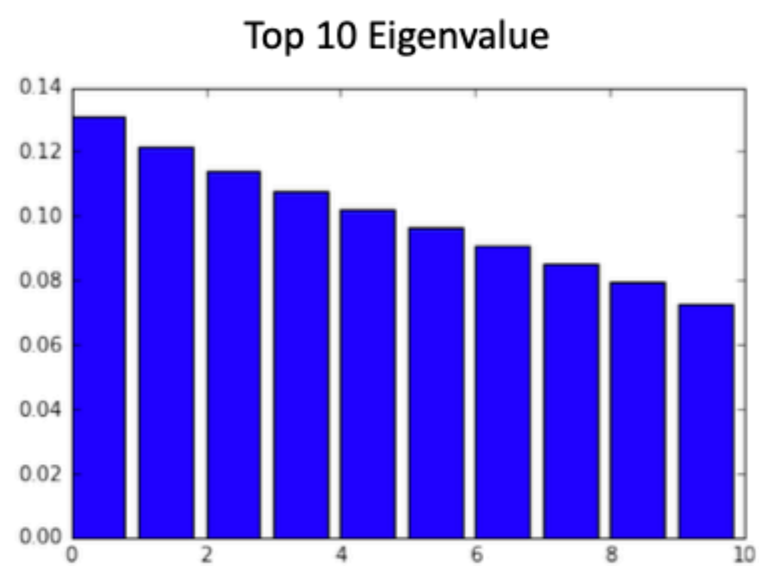
WEAT finds similar biases in Word Embeddings as IAT did for humans



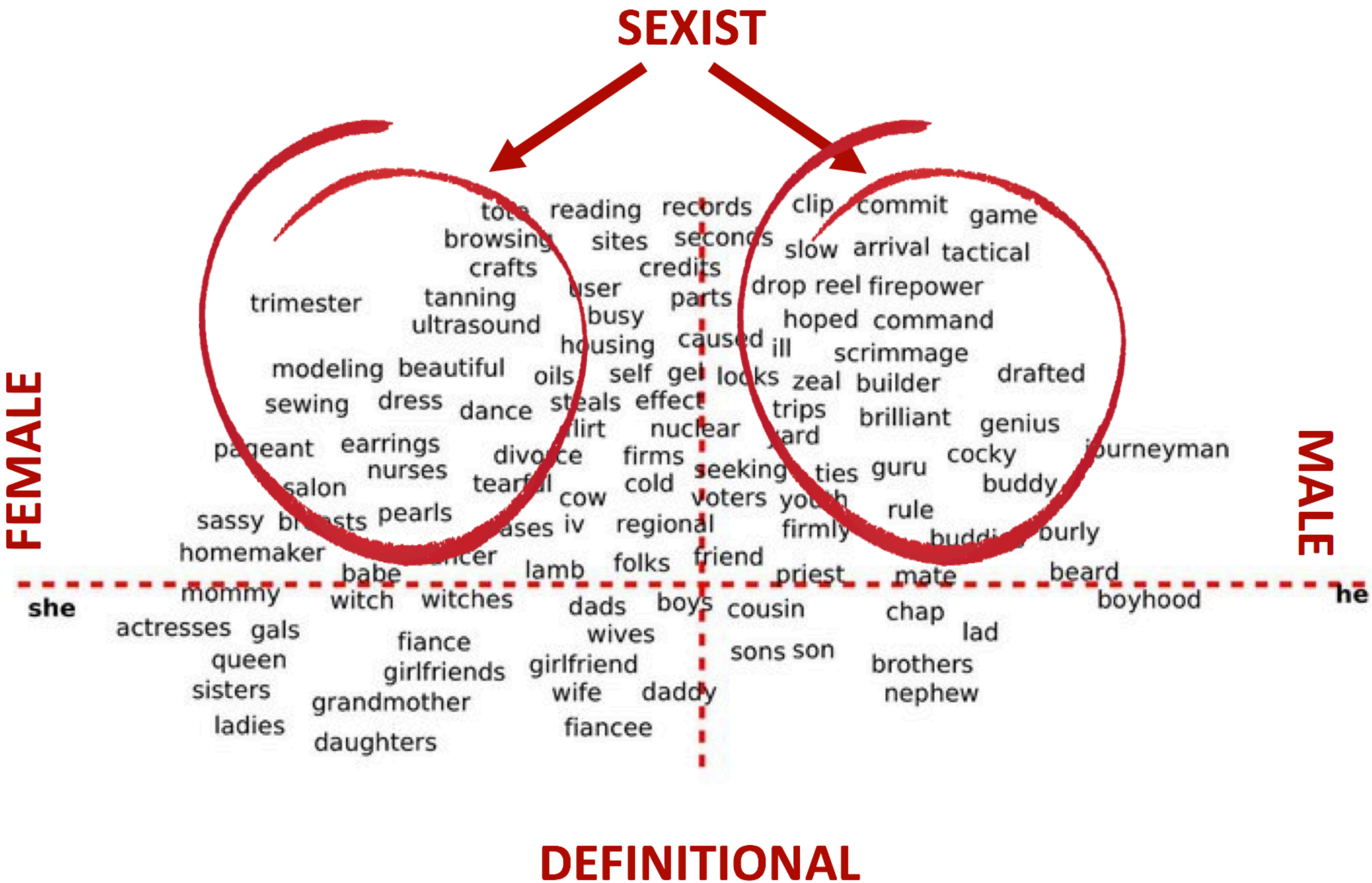

PCA ( “he”- “she”, “father”-”mother”,...)

Gender Pair




PCA ( “dog”- “cat”, “house”-”building”,...)

Random Pair


SEXIST
FEMALE
MALE
DEFINITIONAL
she
he
reading
records
clip
commit
game
browsing
sites
seconds
slow
arrival
tactical
crafts
credits
trimester
tanning
parts
drop reel firepower
ultrasound
busy
hoped
command
housing
caused
ill
scrimmage
modeling
beautiful
oils
self
gel
locks
zeal
builder
drafted
sewing
dress
dance
steals
effect
trips
brilliant
genius
flirt
nuclear
pageant
earrings
divorce
firms
cocky
journeyman
nurses
seeking
ties
guru
salon
tearful
cold
buddy
cow
voters
rule
sassy
pearls
iv
regional
firmly
burly
homemaker
lamb
folks
friend
babe
priest
mate
beard
mommy
witch
witches
dads
boys
cousin
chap
boyhood
actresses
gals
wives
lad
fiance
sons son
queen
girlfriends
girlfriend
brothers
sisters
grandmother
wife
daddy
nephew
ladies
fiancee
daughters

# Can we Extend the Analysis beyond Binary Gender?

# Beyond Gender & Race/Ethnicity Bias

Manzini et al. NAACL 2019

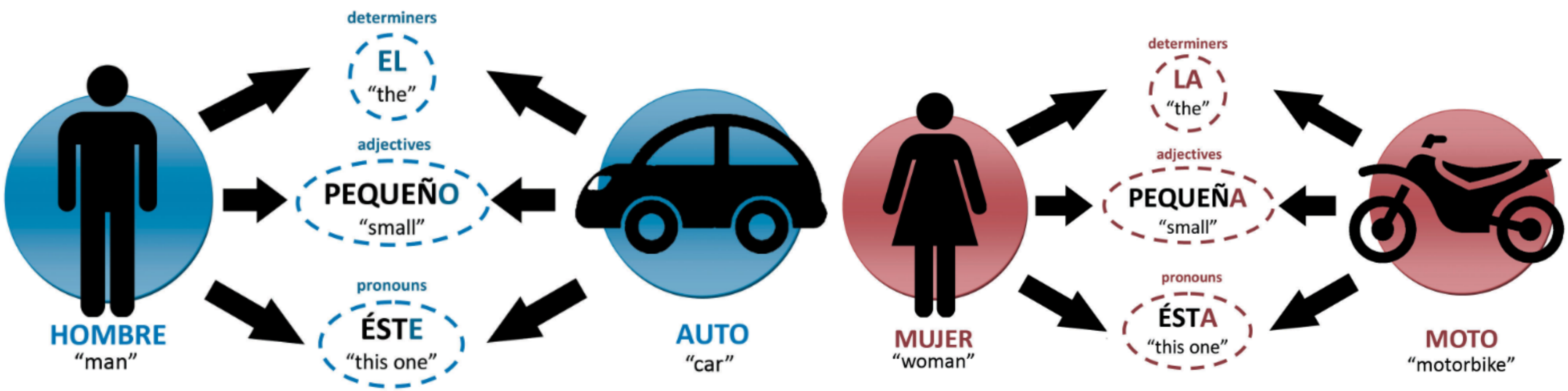
| Racial Analogies | |
|---|---|
| black → homeless | caucasian → servicemen |
| caucasian → hillbilly | asian → suburban |
| asian → laborer | black → landowner |
| **Religious Analogies** | |
| jew → greedy | muslim → powerless |
| christian → familial | muslim → warzone |
| muslim → uneducated | christian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

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

# How about other Embedding?

# Bias Only in English?

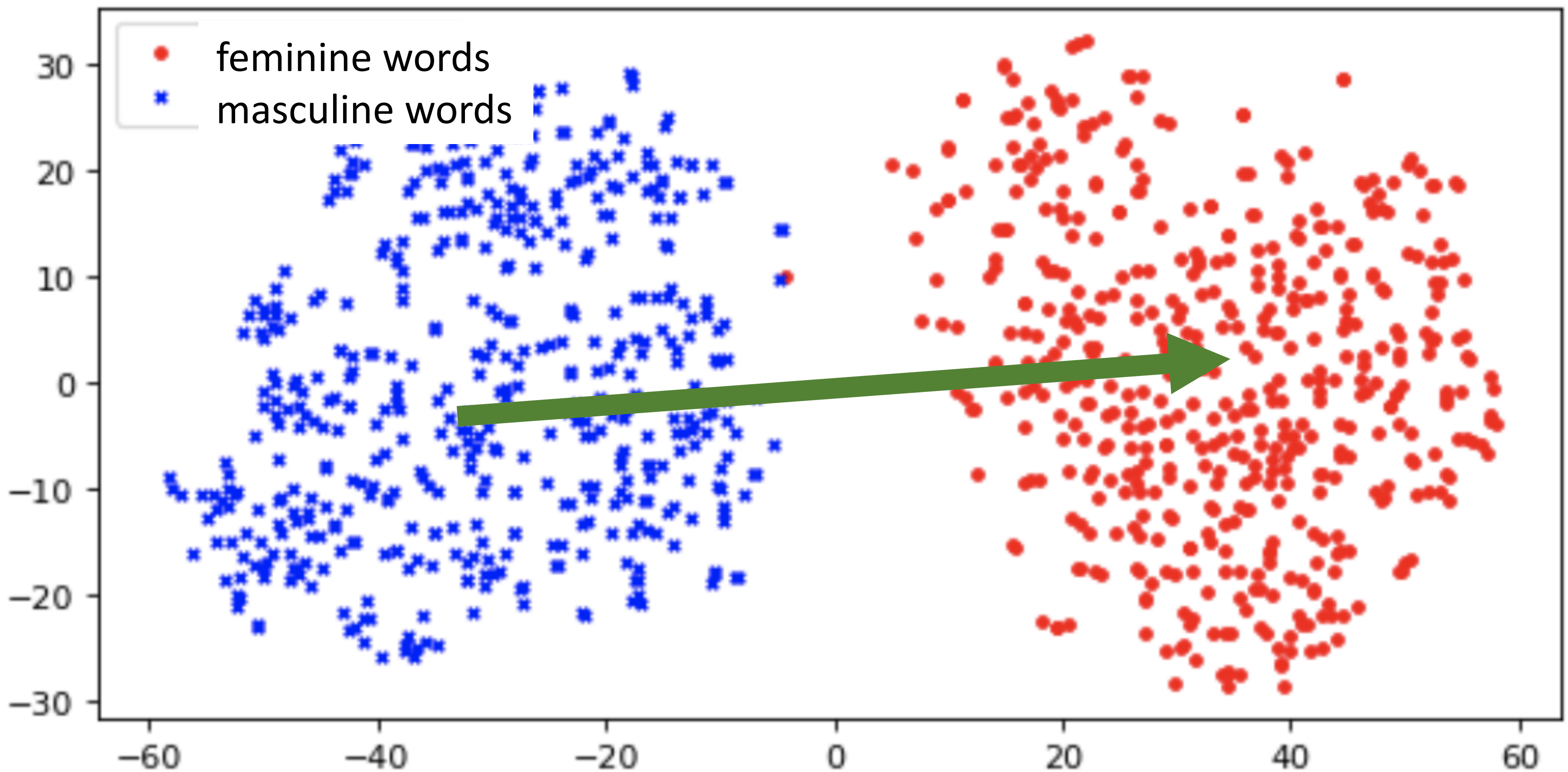
- Language with grammatical gender
  - Morphological agreement

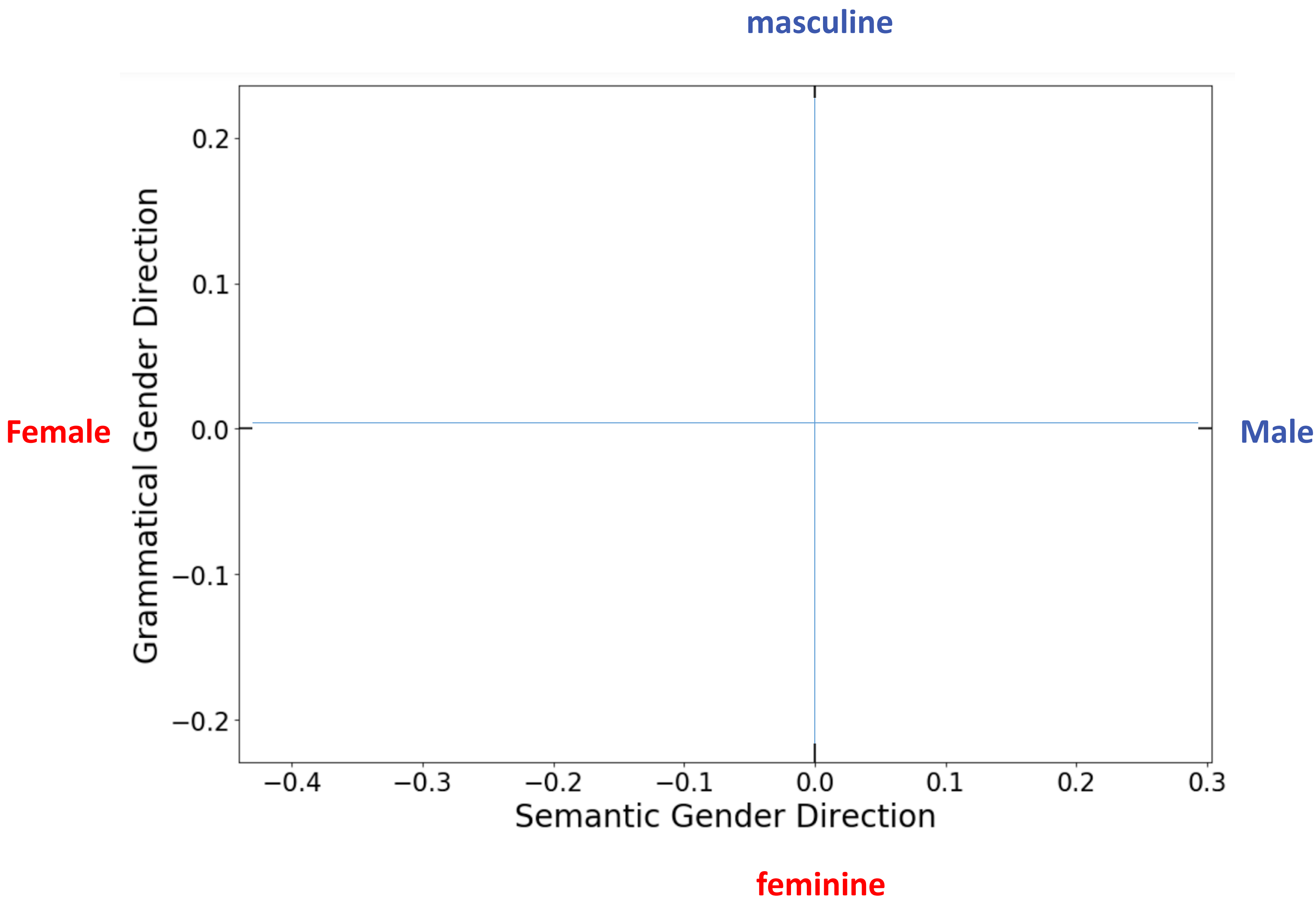



(Zhou et al, EMNLP 2019)

# ❖ Linear Discriminative Analysis (LDA)

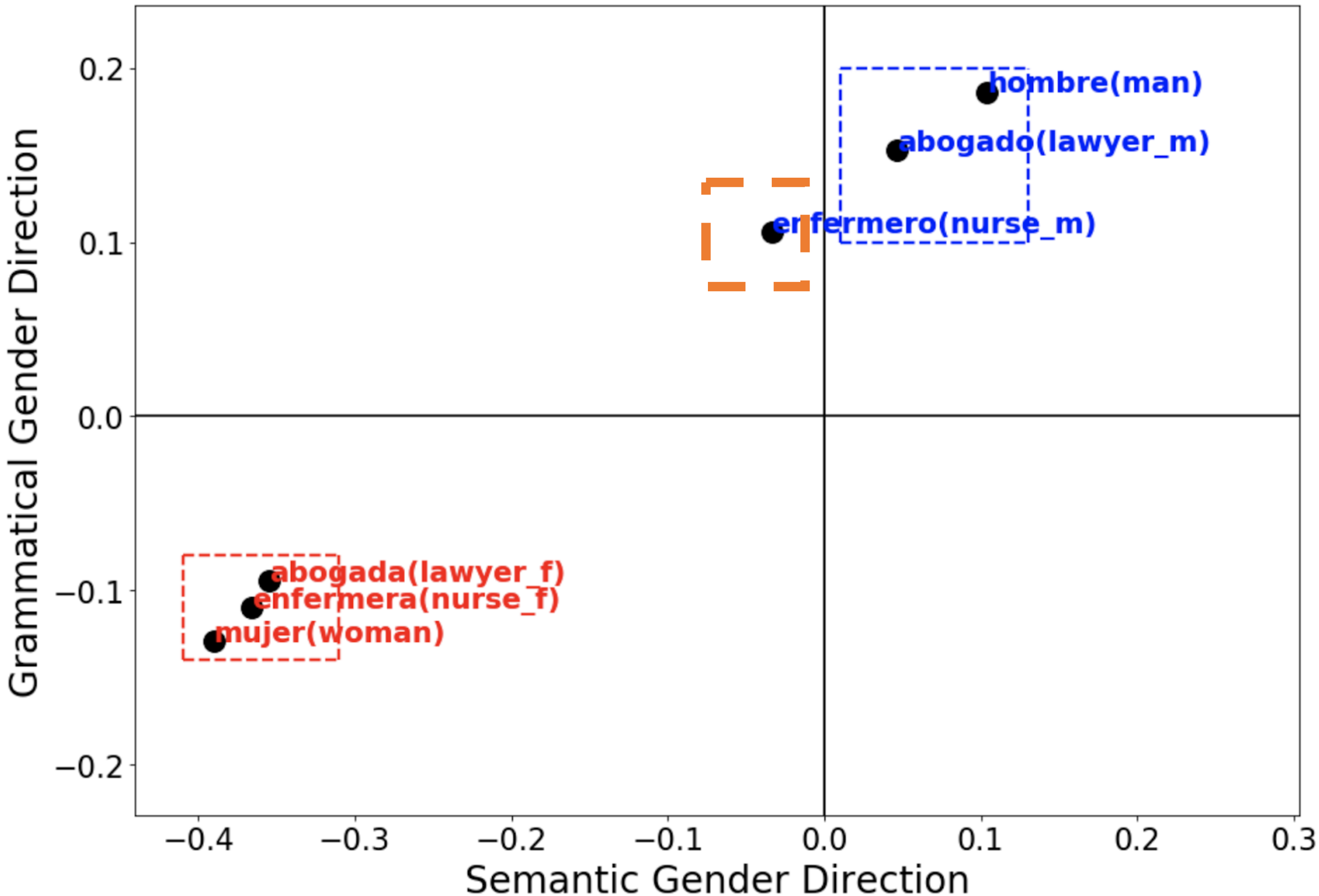
## ❖ Identify grammatical gender direction




masculine
Female
Male
feminine
Grammatical Gender Direction
Semantic Gender Direction
0.2
0.1
0.0
−0.1
−0.2
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3

masculine

Female

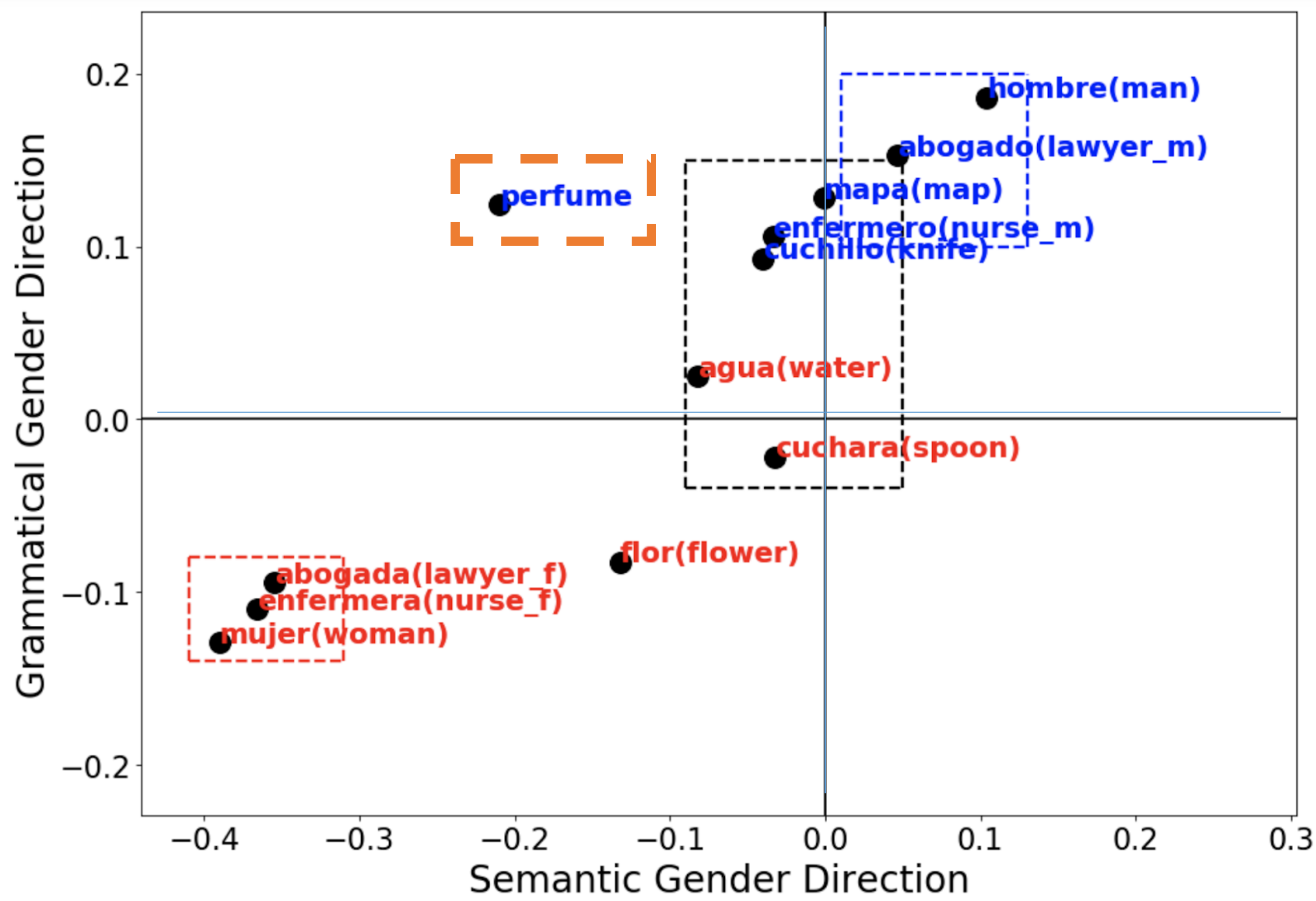



Male

feminine

masculine

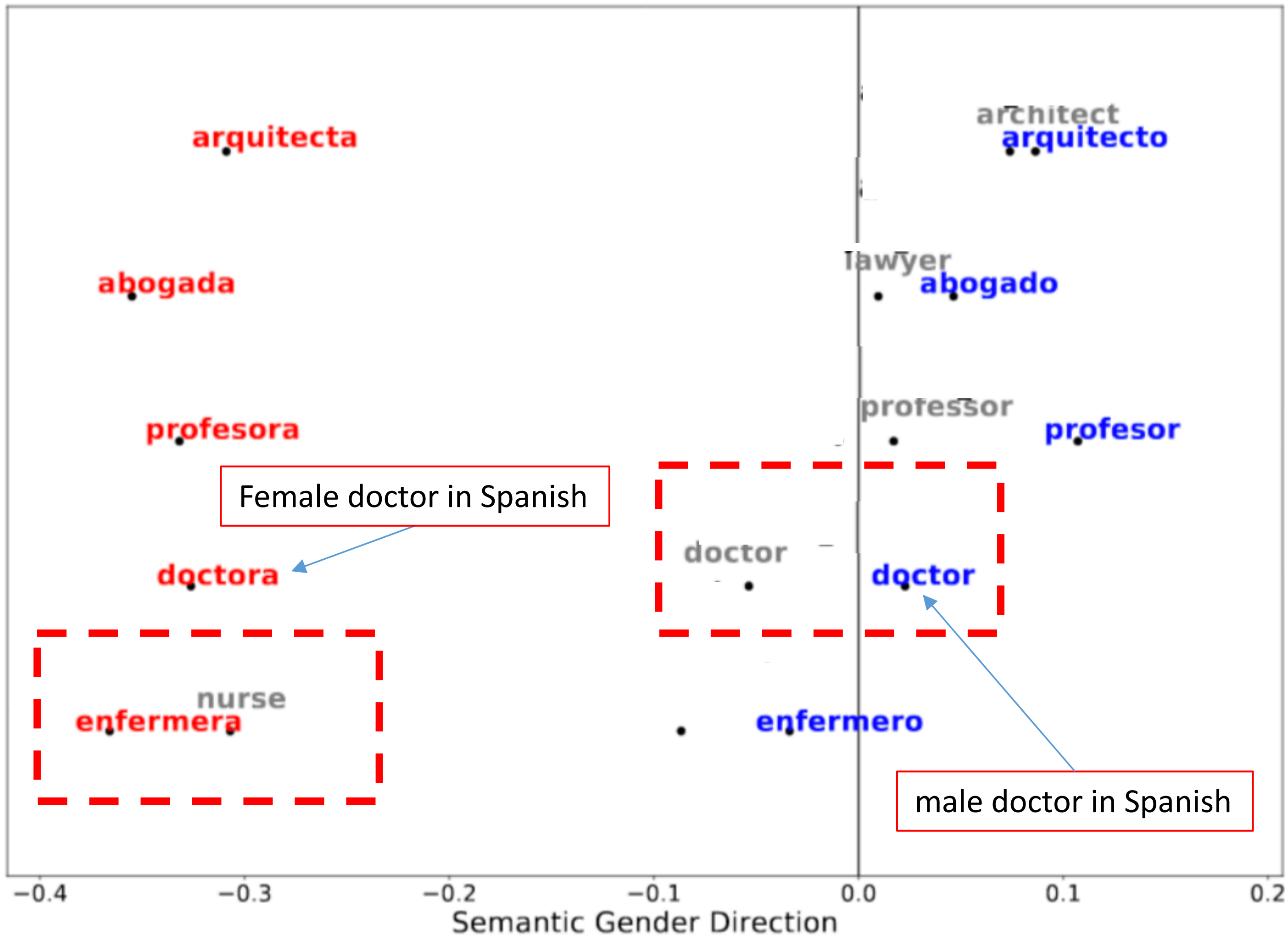
Female




Male

feminine

# How about bilingual embedding?
[Zhou et al. EMNLP19]



# How about Contextualized Representation?

**Gender Bias in Contextualized Word Embeddings**

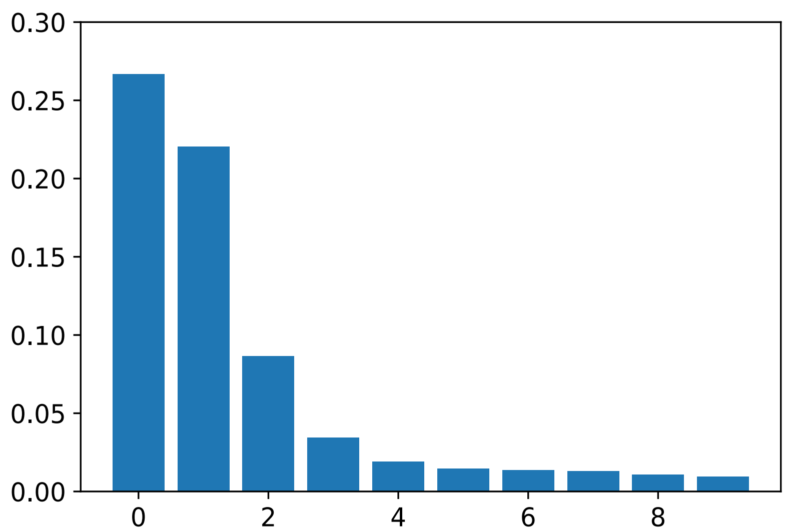
Zhao et al. NAACL 19

- 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

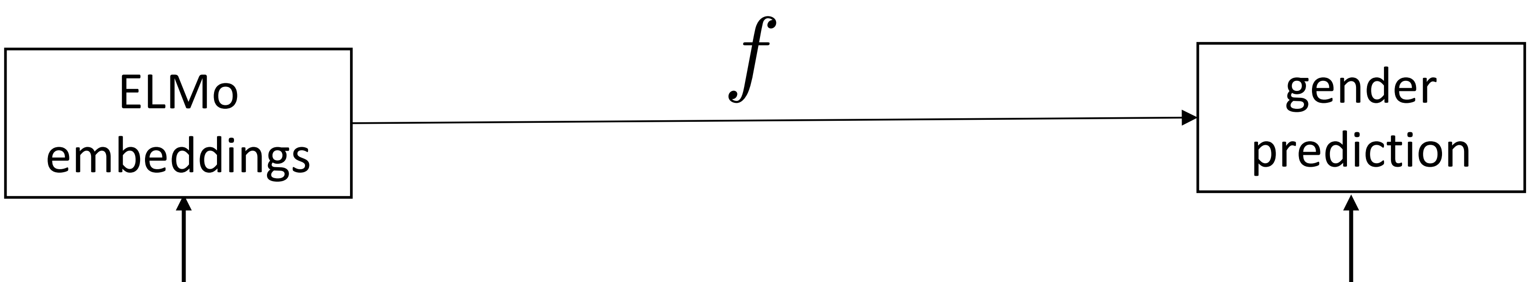
gender direction: ELMo(driver) – ELMo(driver)



# 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

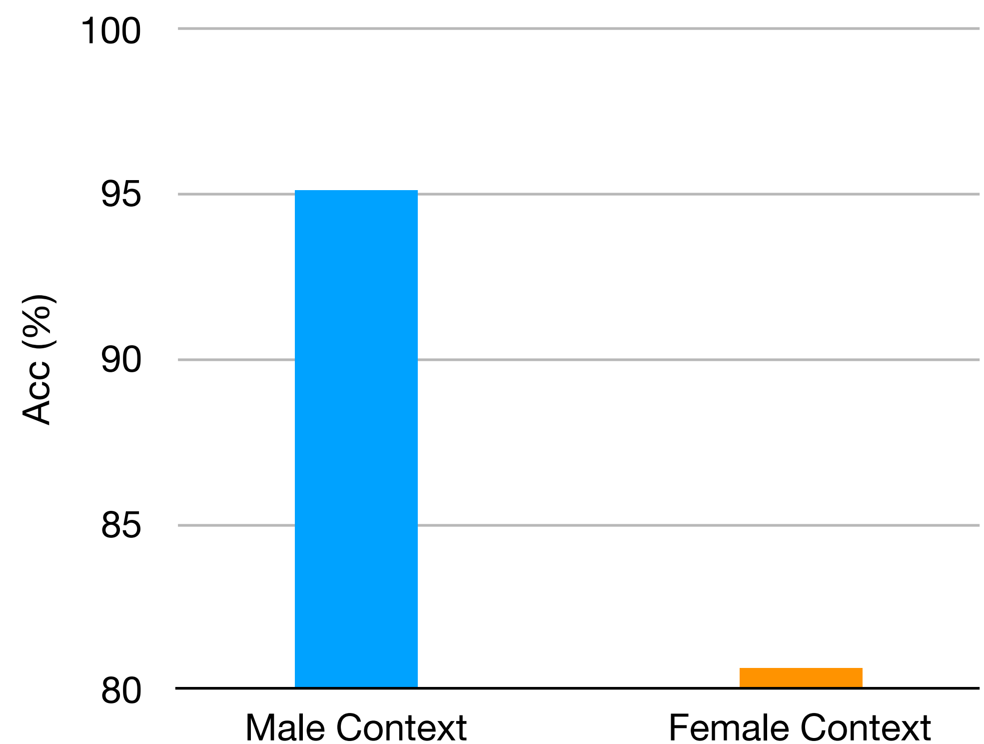
The writer taught himself to play violin .

- Classifier

$f$ : ELMo(occupation) $\longrightarrow$ context gender

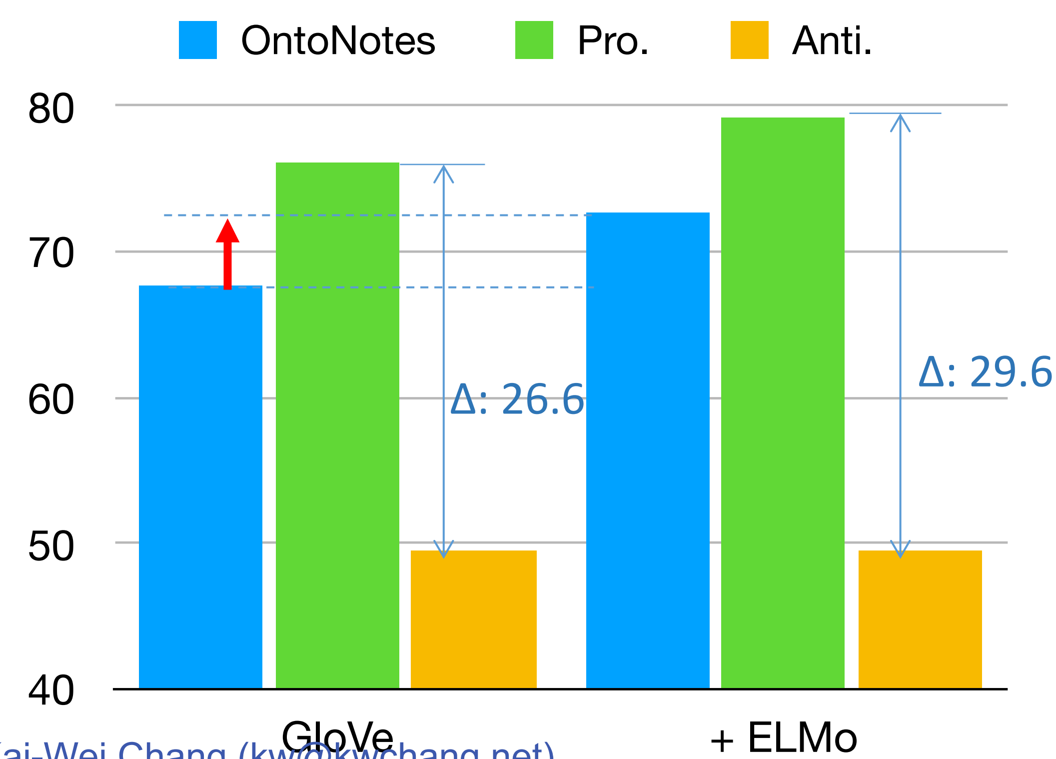
- ELMo propagates gender information to other words

- Male information is 14% more accurately propagated than female



# Coreference with contextualized embedding

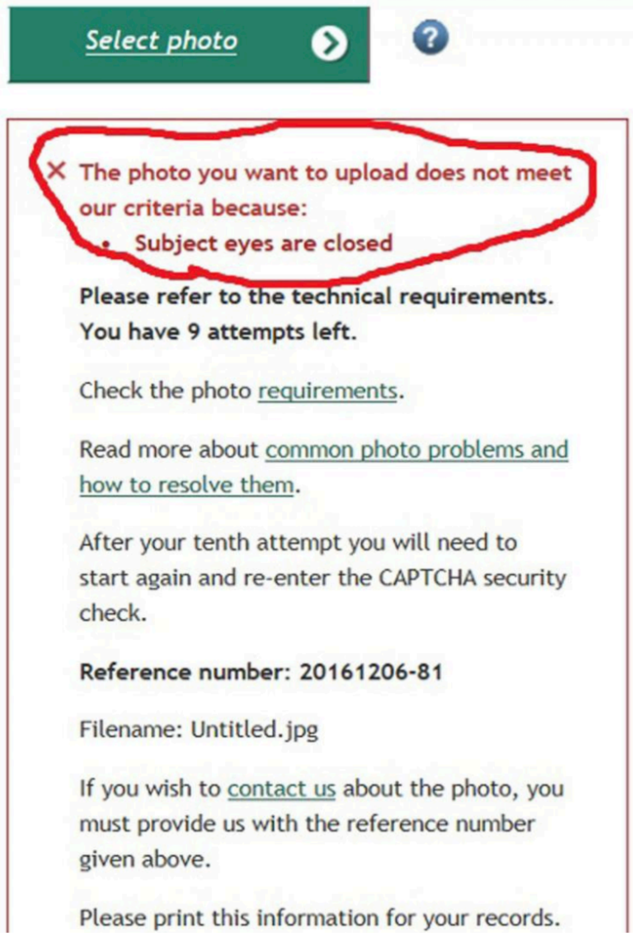
- ELMo boosts the performance
- However, enlarge the bias (Δ)



# 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 identitiy 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)



Select photo

✕ The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

**Please refer to the technical requirements.**
**You have 9 attempts left.**

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

**Reference number: 20161206-81**

Filename: Untitled.jpg

If you wish to contact us about the photo, you must provide us with the reference number given above.

Please print this information for your records.

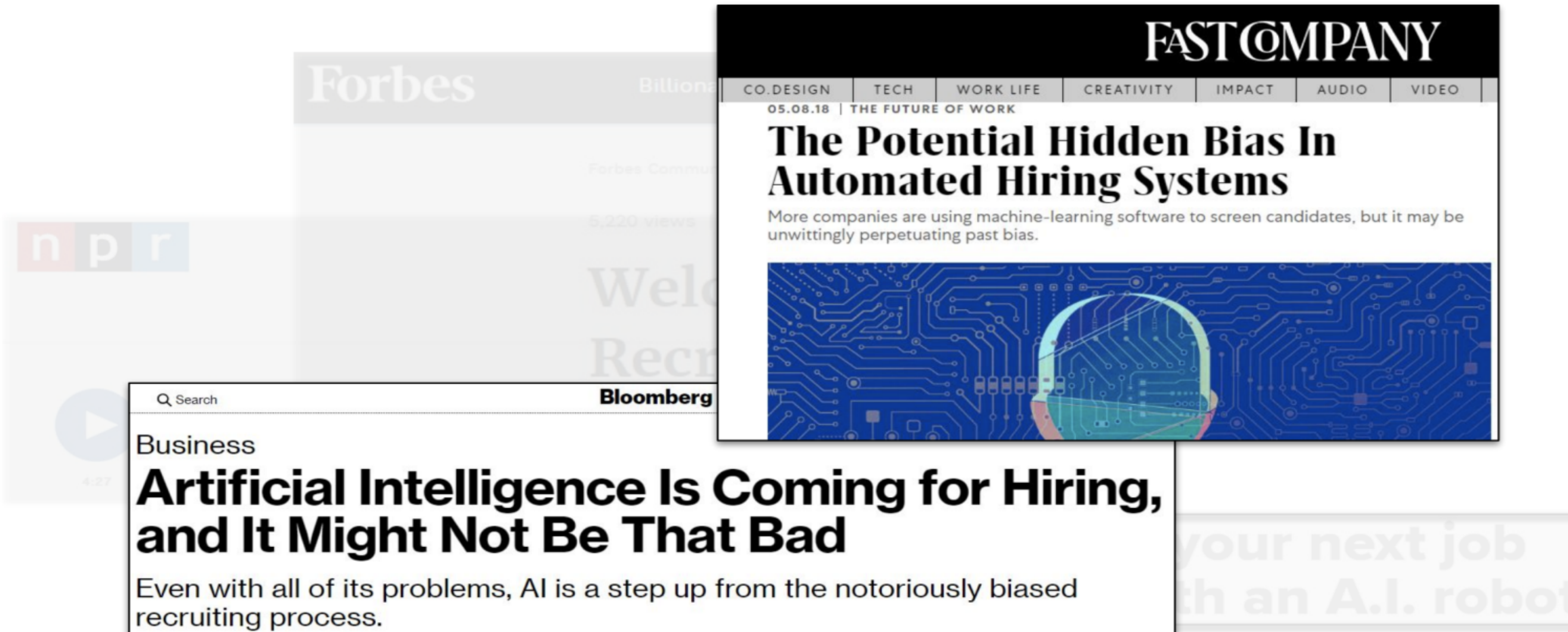
A screenshot of New Zealand man Richard Lee's passport photo rejection notice, supplied to Reuters December 7, 2016. Richard Lee/Handout via REUTERS

# Towards Inclusive AI

# Examples of Harm from NLP Bias

Swinger et al. (2019)

An artificially intelligent headhunter?

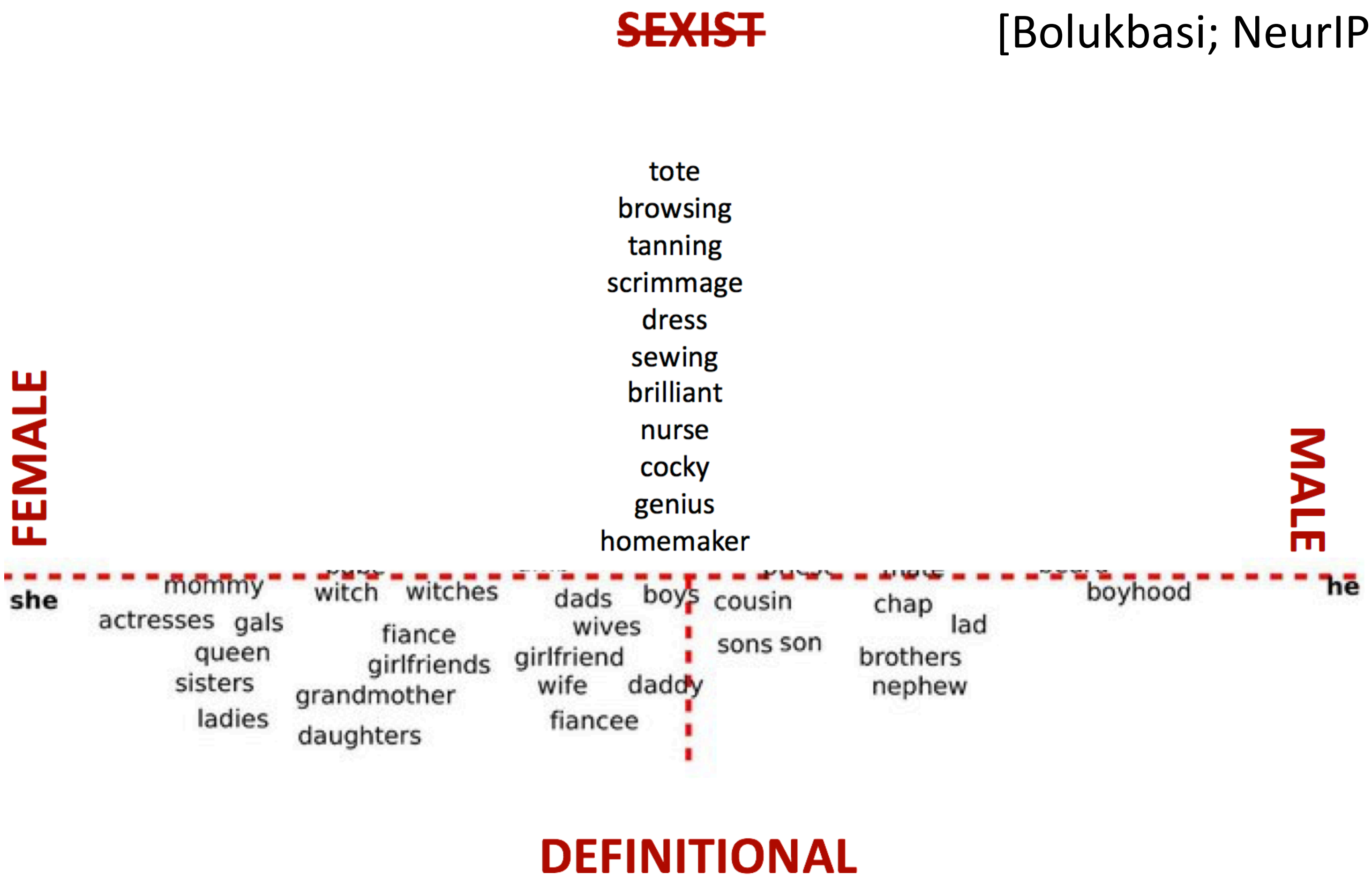



Copyright © 2019 Maria De-Arteaga

Prevent Allocative Harm in Sensitive Applications

Can we ~~remove~~ these biases?

Control




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

# Towards Debiasing

Bolukbasi et al. (2016)

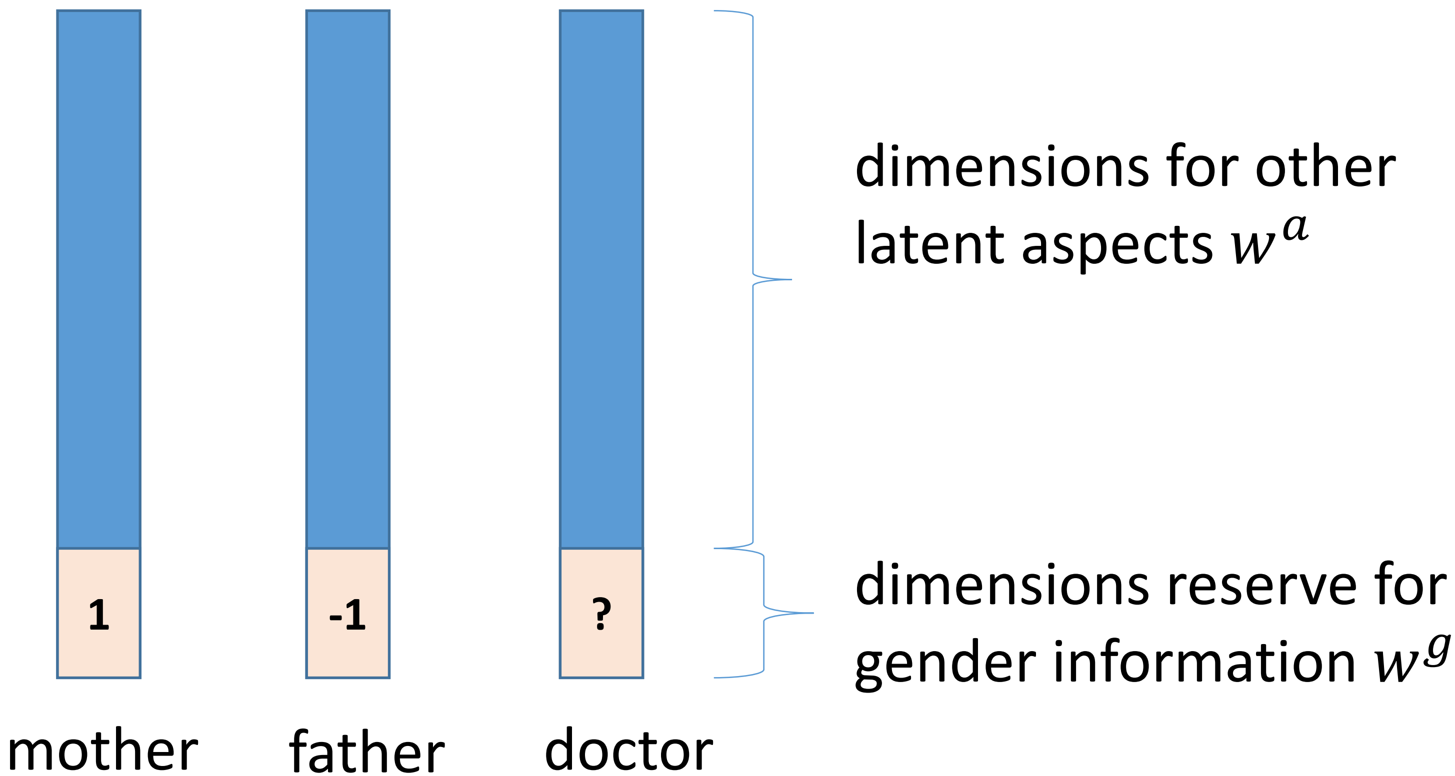
1. Identify gender subspace: B
2. Identify gender-definitional (S) and gender-neutral words (N)
3. Apply transform matrix (T) to the embedding matrix (W)
   a. Project away the gender subspace B from the gender-neutral words N
   b. But, ensure the transformation doesn't change the embeddings too much

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

T - the desired debiasing transformation B - biased space
W - embedding matrix N - embedding matrix of gender neutral words

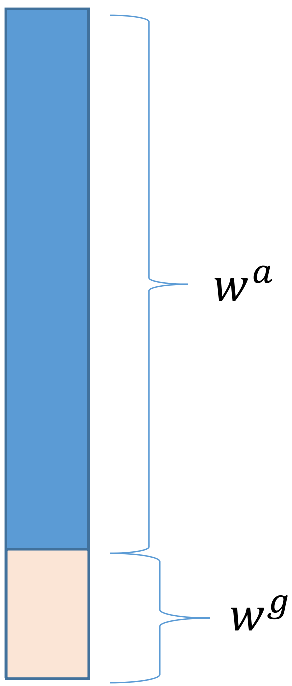
# Make Gender Information Transparent in Word Embedding

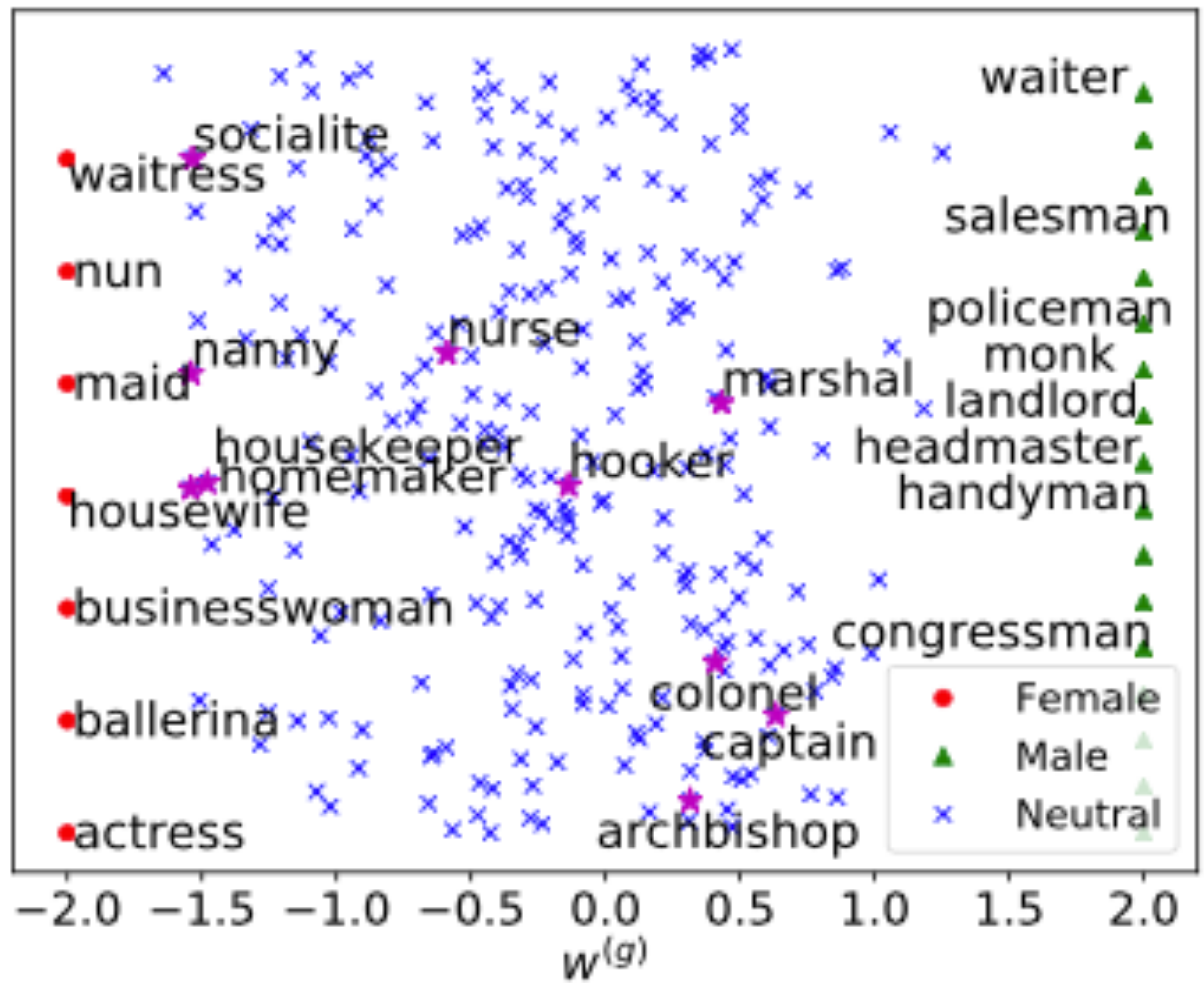
**Learning Gender-Neutral Word Embeddings** [Zhao et al; EMNLP18]



# Make Gender Information Transparent in Word Embedding

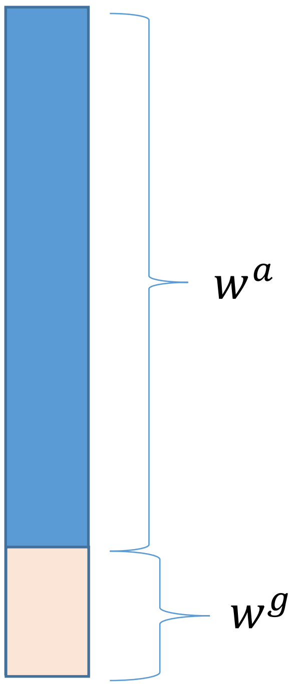
[**Learning Gender-Neutral Word Embeddings**]() [Zhao et al; EMNLP18]

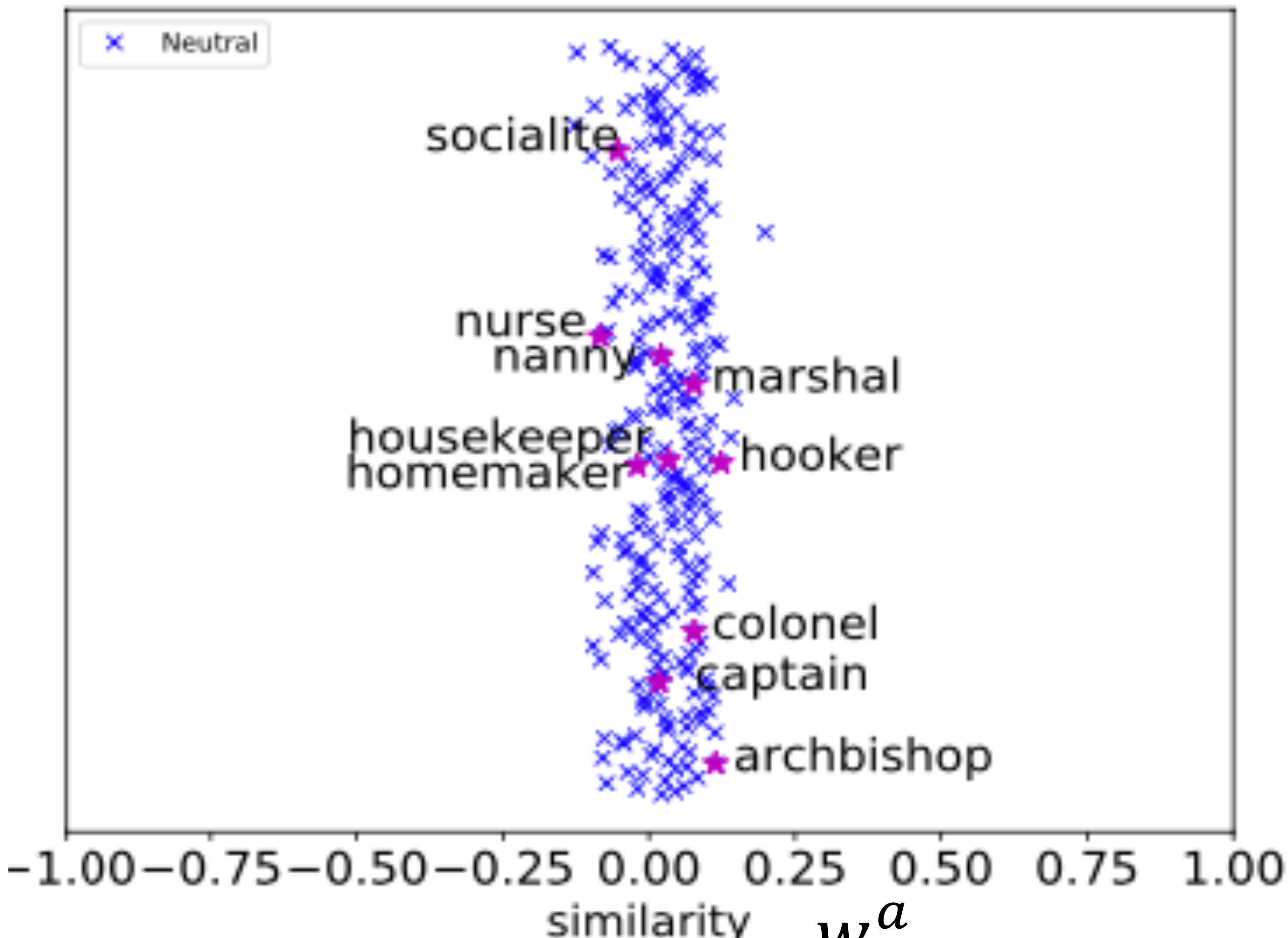


# Make Gender Information Transparent in Word Embedding

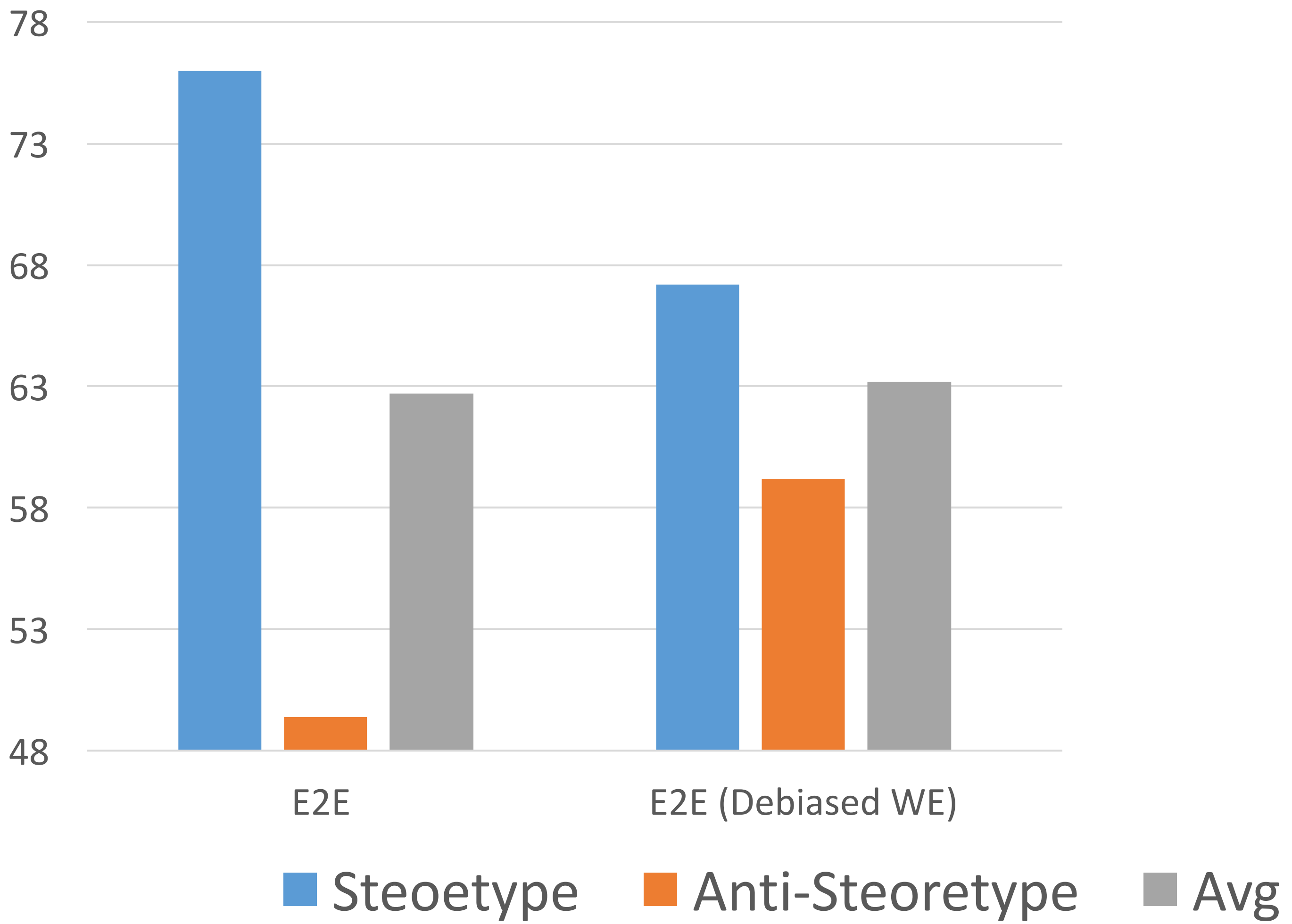
**Learning Gender-Neutral Word Embeddings** [Zhao et al; EMNLP18]




$w^a$

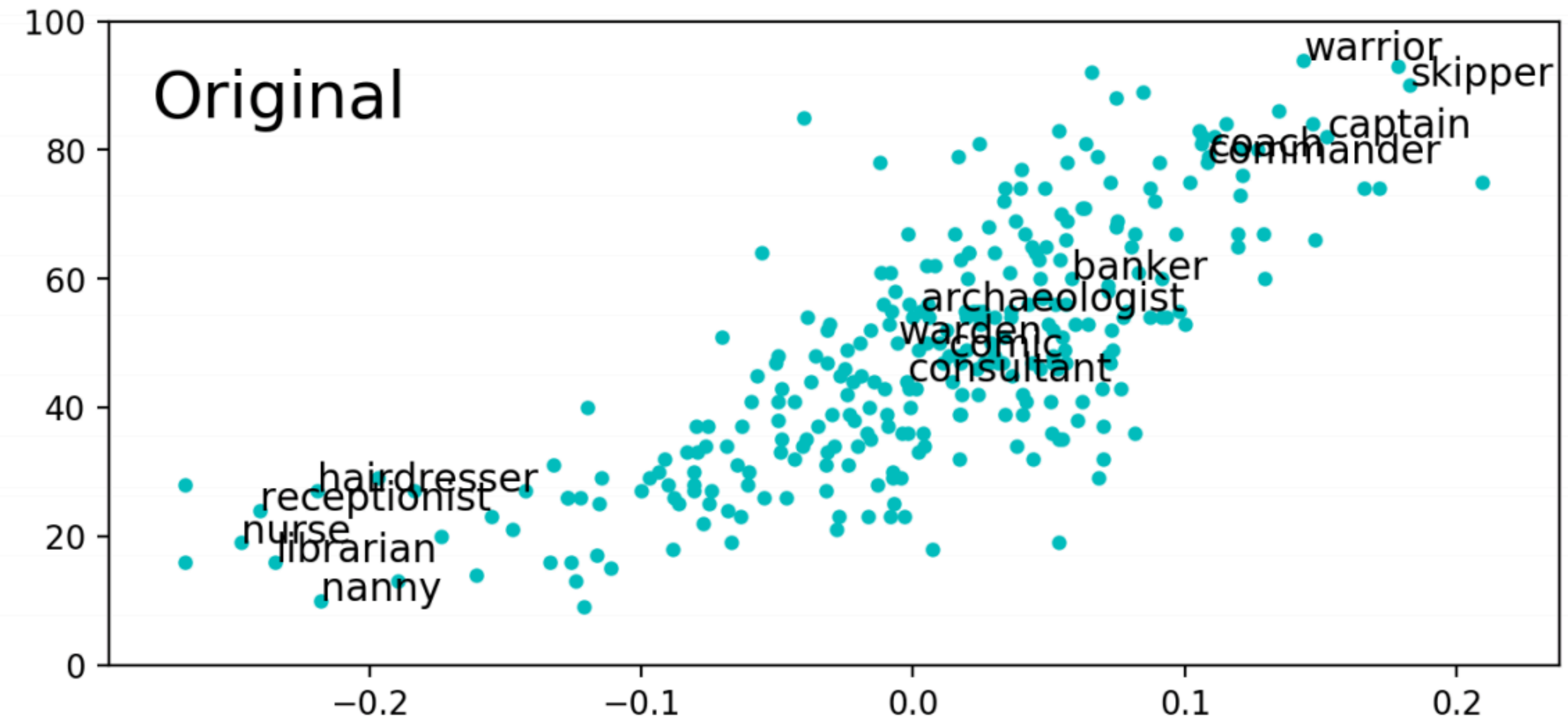
# Gender bias in Coref System



# Is Gender Information Actually Removed from Embedding?

# Completely removing bias is hard

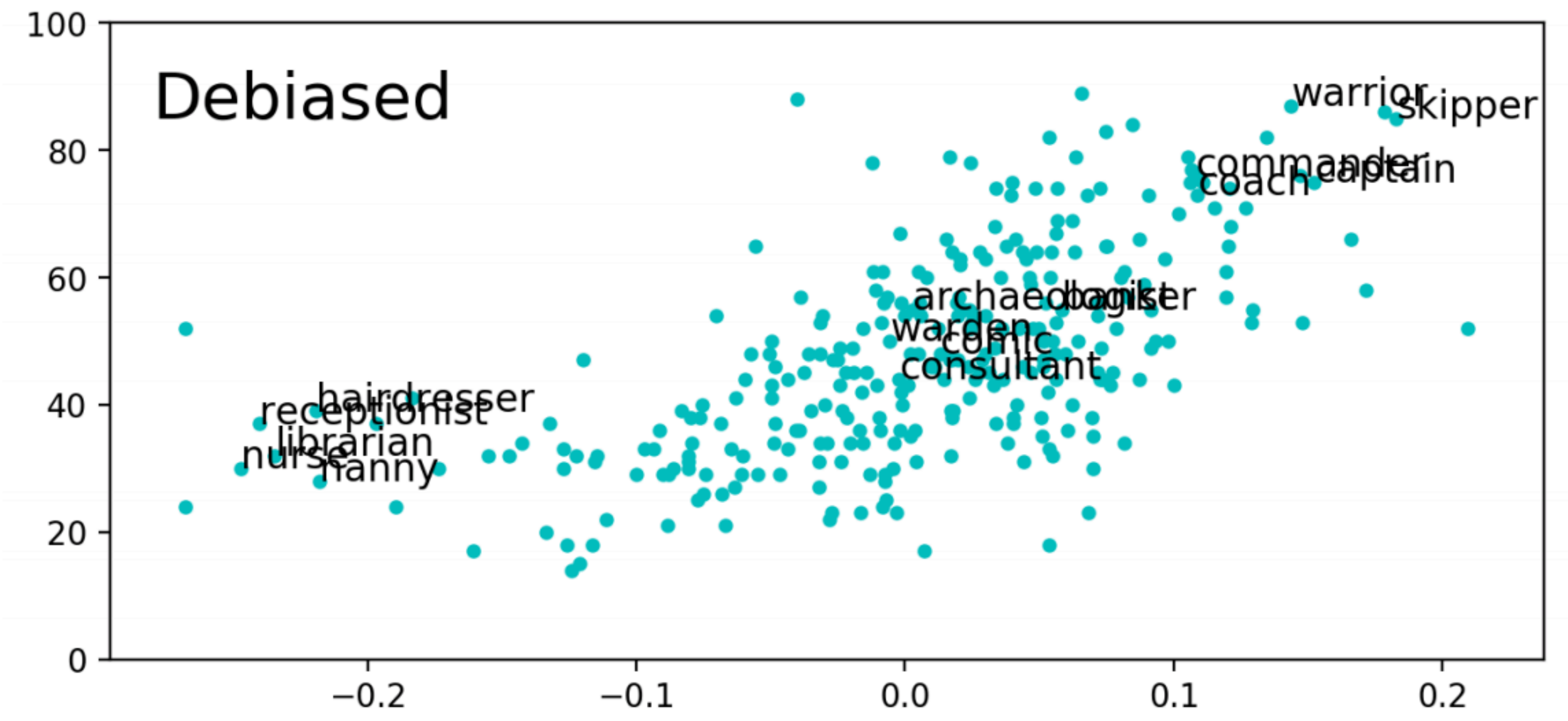
- Gonen, et al. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.** NAACL (2019).




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

# Completely removing bias is hard

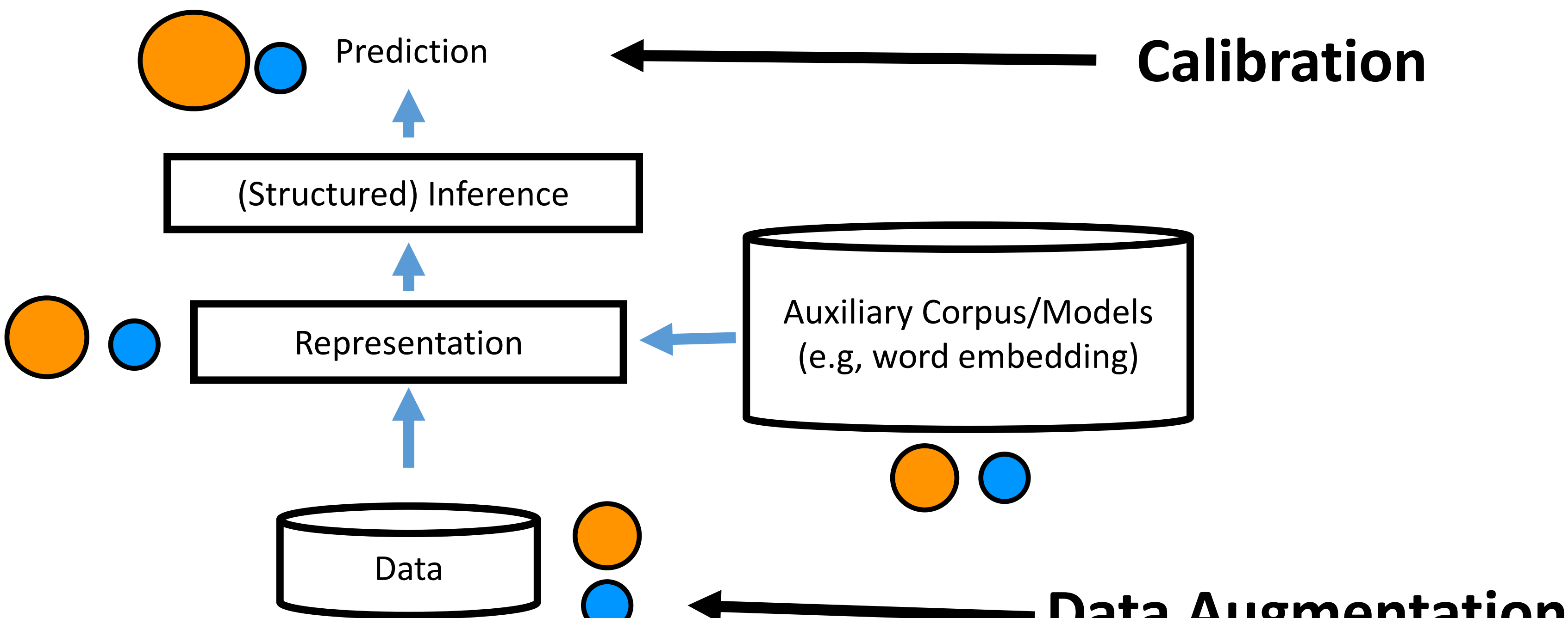
- Gonen, et al. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.** NAACL (2019).




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

John went to ~~his~~ house

F2 went to her house

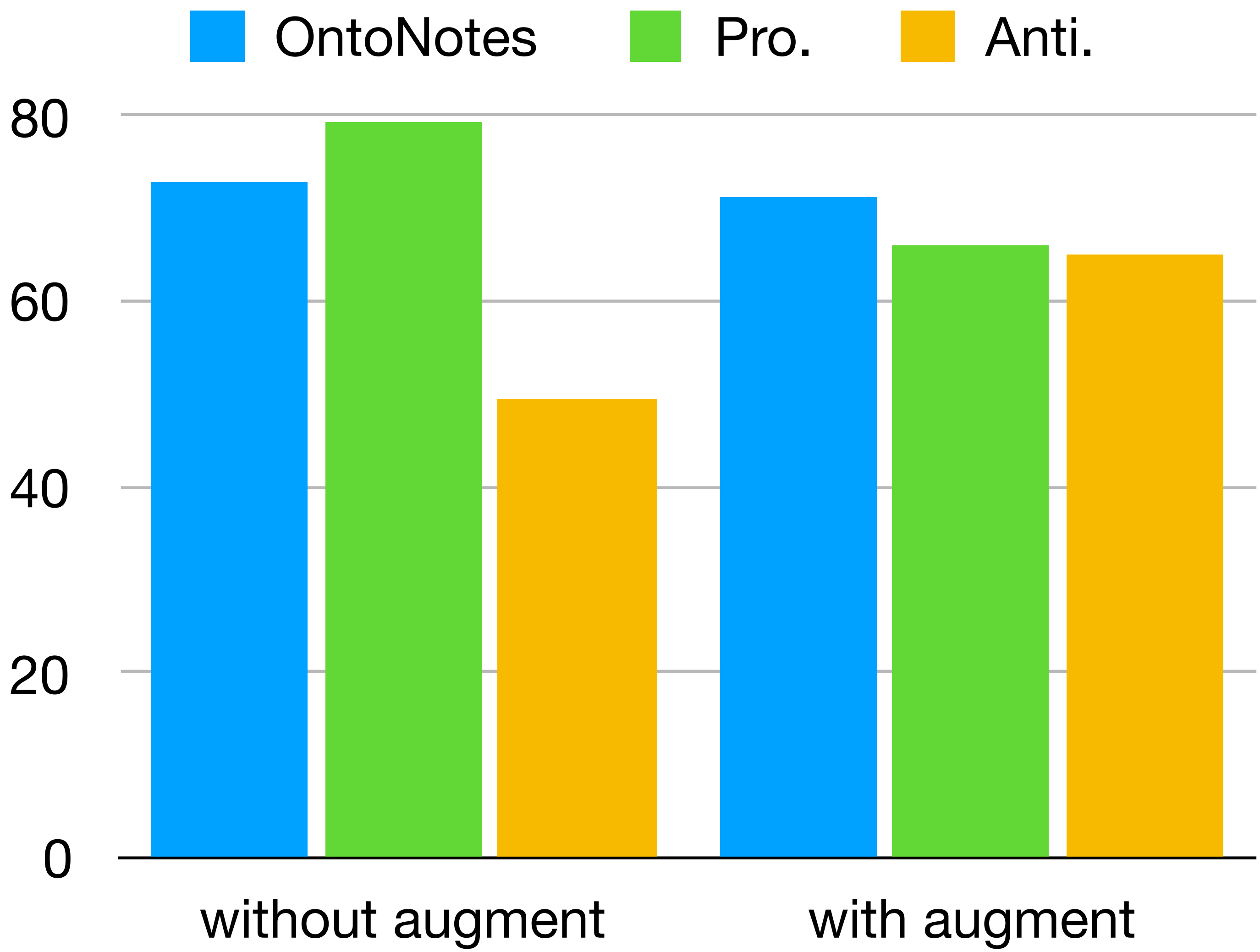
Named Entity are anonymized

Gender words are swapped

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

arXiv.org > cs > arXiv:1906.08976

Computer Science > Computation and Language

**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

*(Submitted on 21 Jun 2019)*

[ACL 2019]