HW5 Part II + Bias in NLP

Spring 2019

Caleb Kaiji Lu

HW5 Part II: Debiasing Word Embeddings
[Bolukabasi 2016]

- Man: King :: Woman:Queen
- Paris: France :: Tokyo:Japan

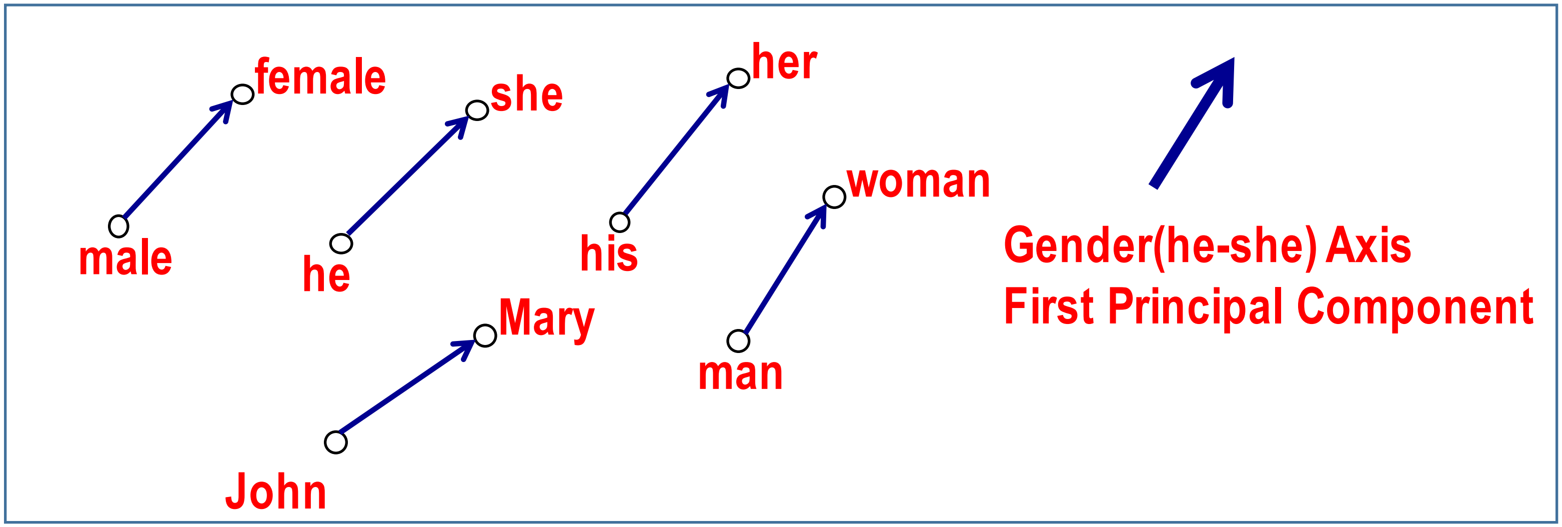
- He:Brother :: She: Sister
- He:Blue :: She:Pink
- He:Doctor :: She:Nurse
- He:Realist :: She:Feminist
- She:Pregnancy :: He:Kidney Stone
- She:Baking::He:Roasting
- She:Blonde::He:Blond

HW5 Part II: Debiasing Word Embeddings

[Bolukabasi 2016]

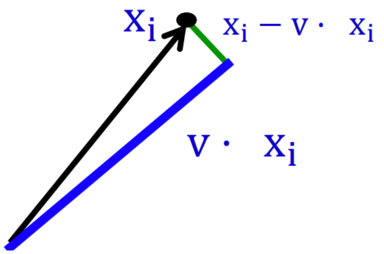
- To be released today
- Three steps
  - Identify gender subspace (PCA using SVD)
  - Neutralize
  - Equalize
- Evaluation
  - Analogy completion for he—she
  - Analogy completion for a WE evaluation dataset
- Three word files:
  - Gender-definitional words (for identifying the gender subspace)
  - Gender-specific words(for identifying words to neutralize)
  - Equalized pairs (words to equalize)

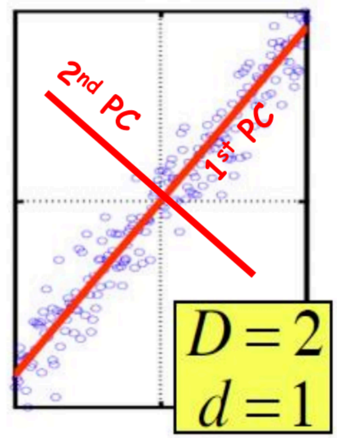
The Geometry of Gender



Principal Component Analysis

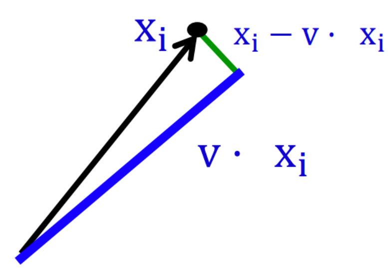
- Principal Components (PC) are orthogonal directions that capture most of the variance in the data.
  - 1^{st} PC – direction of greatest variability in data
  - 2^{nd} PC – Next orthogonal (uncorrelated) direction of greatest variability (remove all variability in first direction, then find next direction of greatest variability)
  - And so on…





Principal Component Analysis (PCA)

- Let v1, v2, …, vd denote the d principal components.
  - V is orthonormal
- Let X = [x1, x2, … , xn] (columns are the datapoints)
  - Data points are centered
- Find vector that maximizes sample variance of projected data
  - Find vector that minimizes the average reconstruction error



Principal Component Analysis (PCA)

- Blackboard

Identify Gender Subspace

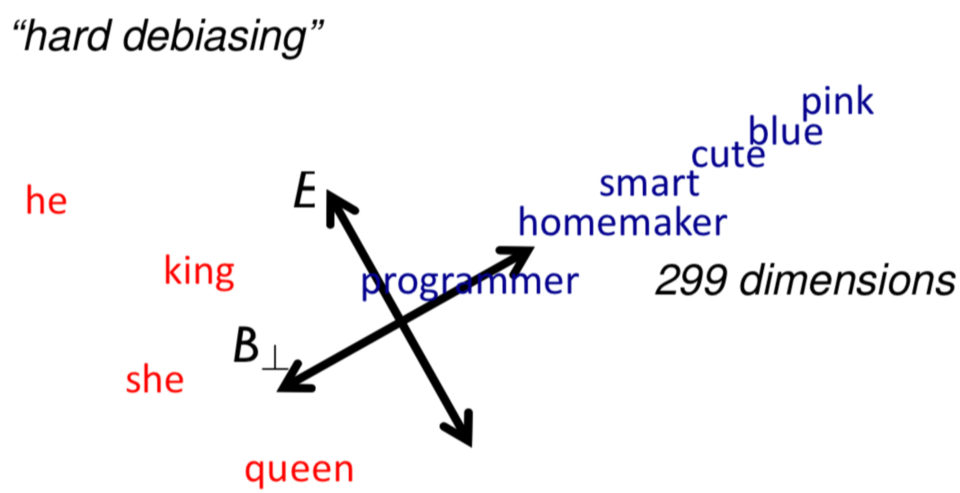
Step 1: Identify gender subspace. Inputs: word sets W, defining sets $D_1, D_2, \ldots, D_n \subset W$ as well as embedding $\left\{\vec{w} \in \mathbb{R}^d\right\}_{w \in W}$ and integer parameter $k \geq 1$. Let

$$\mu_i := \sum_{w \in D_i} \vec{w}/|D_i|$$

be the means of the defining sets. Let the bias subspace B be the first k rows of SVD($\mathbf{C}$) where

$$\mathbf{C} := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \mu_i)^T (\vec{w} - \mu_i) \big/ |D_i|.$$

Neutralize and equalize



Neutralize and Equalize

Step 2a: Hard de-biasing (neutralize and equalize). Additional inputs: words to neutralize $N \subseteq W$, family of equality sets $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ where each $E_i \subseteq W$. For each word $w \in N$, let $\vec{w}$ be re-embedded to

$$\vec{w} := (\vec{w} - \vec{w}_B)/\|\vec{w} - \vec{w}_B\|.$$

For each set $E \in \mathcal{E}$, let

$$\begin{aligned} \mu &:= \sum_{w \in E} w/|E| \\ \nu &:= \mu - \mu_B \\ \text{For each } w \in E, \quad \vec{w} &:= \nu + \sqrt{1 - \|\nu\|^2}\,\frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|} \end{aligned}$$

- B: gender subspace
- w_B: projection of w on B
- BlackBoard

Agenda

- Introduction
- Gender Bias in NLP tasks
- Counterfactual Data-Augmentation
- Gender Bias in RNN Language Models
- Neural Coreference Resolution Basics
- Gender Bias in Coreference Resolution

Natural Questions

- Does bias exist downstream tasks?
- Does mitigating bias in word embeddings **also mitigate bias in the downstream tasks?**
- Does mitigating bias in word embeddings **impact the performance of the downstream tasks?**

Bias in NLP tasks

- Bias in language modeling

	A		B	$\ln \Pr[B \mid A]$
$1_{\square}$:	**He** is a	|	**doctor**.	-9.72
$1_{\bigcirc}$:	**She** is a	|	**doctor**.	-9.77
$2_{\square}$:	**He** is a	|	**nurse**.	-8.99
$2_{\bigcirc}$:	**She** is a	|	**nurse**.	-8.97

- Bias in Coreference resolution

$1_{\square}$: The **doctor** $\overbrace{\text{ran because}}^{5.08}$ **he** is late.

$1_{\bigcirc}$: The **doctor** $\overbrace{\text{ran because}}^{1.99}$ **she** is late.

$2_{\square}$: The **nurse** $\overbrace{\text{ran because}}^{-0.44}$ **he** is late.

$2_{\bigcirc}$: The **nurse** $\overbrace{\text{ran because}}^{5.34}$ **she** is late.

Bias in NLP tasks[Lu,18]

- Definition of bias:
  - Causal Testing
  - **Define** Matched pairs of individuals (instances) that differ in only a targeted concept (gender)
  - **Calculate** difference in outcomes (conditional log-likelihood)
  - Causal influence of the concept in the scrutinized model

	A	B	$\ln \Pr[B \mid A]$
$1_{\square}$:	**He** is a	**doctor**.	-9.72
$1_{\bigcirc}$:	**She** is a	**doctor**.	-9.77
$2_{\square}$:	**He** is a	**nurse**.	-8.99
$2_{\bigcirc}$:	**She** is a	**nurse**.	-8.97

Figure: Two matched Pairs

Bias in NLP tasks [Lu,18]

- Matched Pairs
  - Templates: He/She is a/an | [OCCUPATION]
  - Aggregate templates
  - Aggregate occupation words (crosslisted from US labor data and language model vocabulary)

	A		B	$\ln \Pr[B \mid A]$
$1_{\square}$:	**He** is a	$\mid$	**doctor**.	-9.72
$1_{\bigcirc}$:	**She** is a	$\mid$	**doctor**.	-9.77
$2_{\square}$:	**He** is a	$\mid$	**nurse**.	-8.99
$2_{\bigcirc}$:	**She** is a	$\mid$	**nurse**.	-8.97

Figure: Two matched Pairs

How to Eliminate the Bias

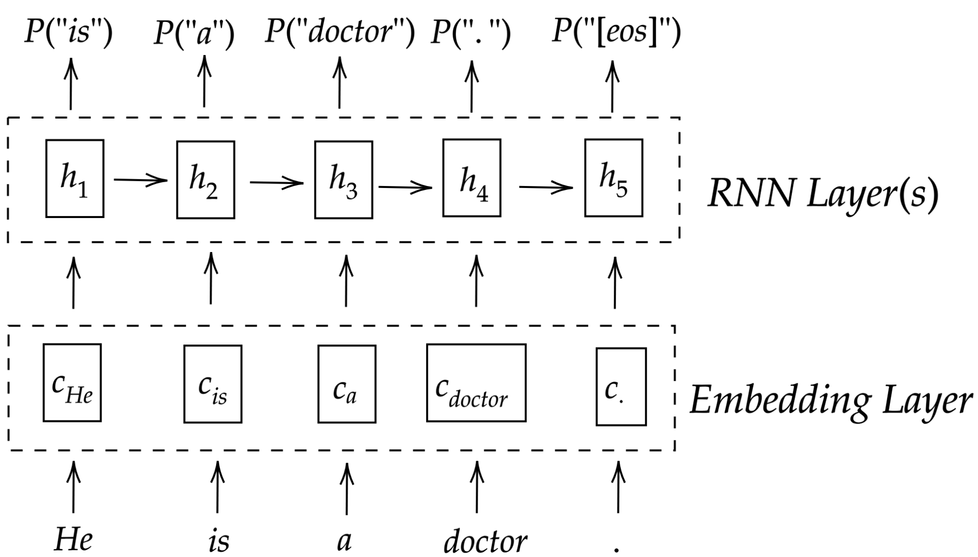
- Simplest solution: Collect unbiased data
  - Not realizable
- Change the model parameters/ Change the objective function

Previously: Debiasing by changing training objective [Zhang, 2018]

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A$$

- For each analogy in the dataset, we let x = (x1, x2, x3)
  - x1 = he; x2 = doctor; x3 = doctor; x4 = ?
- Original Model(Lp)
  - Ground truth for the fourth word $v = x_2 + x_3 - x_1$
  - Estimate for the fourth word: $\hat{y} = v - ww^T v$
- Adversarial Model(LA)
  - Estimate for Adversarial network: $\hat{z} = w_2^T \hat{y}$
  - Ground truth for Adversarial Network: $z = proj_g y$

Previously: Debiasing by changing the model parameters



- Debiasing the embedding layer?

Word Embeddings: Trainable or Fixed?

Word Embedding can be used to replace words as inputs to the model

- Pros
  - Efficient
  - Handle OOV cases if the training dataset is small
- Cons:
  - Cannot tailor to the task
- Debiasing word embedding maybe helpful

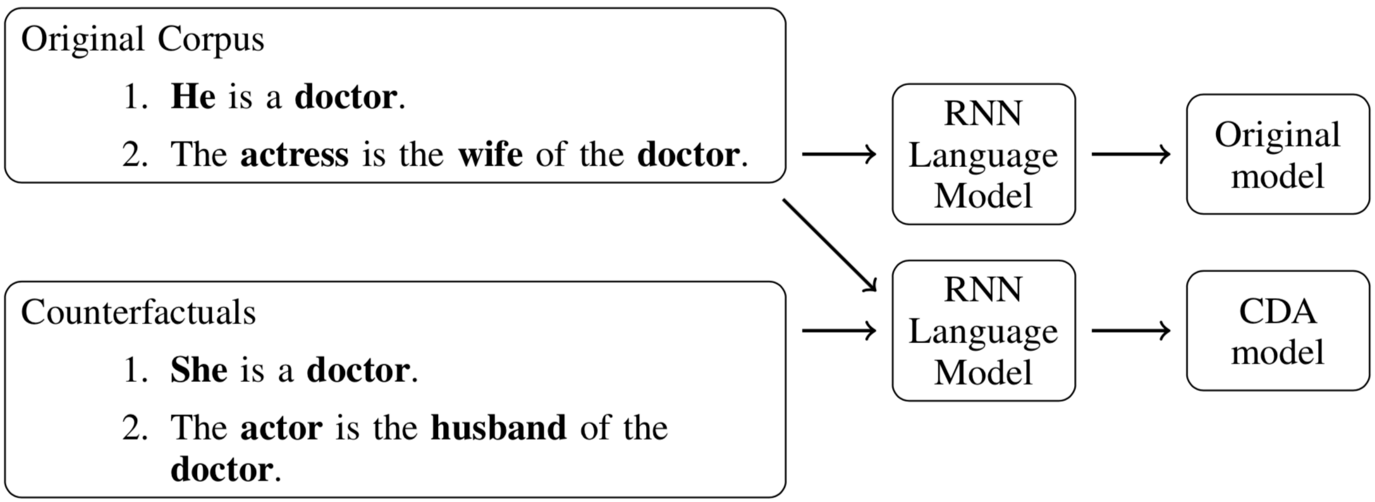
Word Embedding can be trained as part of the model

- Pros:
  - Learn Useful representations specific to the task
- Cons:
  - Expensive
  - Dataset might be too small to learn useful representations
  - Dataset might not cover all the vocabularies
- Debiasing Word embedding may not be helpful
  - Destroy the model
  - Bias is relearned

How to Eliminate the Bias

- Simplest solution: Collect unbiased data
  - Not realizable
- Fix the model / Change the objective function
  - Invasive, could hurt performance
  - Model-dependent
- **<u>Synthesize Unbiased data</u>**
  - Model-agnostic
  - Counterfactual Data Augmentation

Debiasing by Synthesizing data: Counterfactual Data Augmentation



- Generate a new sentence by flipping gender-specific words to their counterparts of opposite gender
- Add the new sentences to the training data
- Train a new model

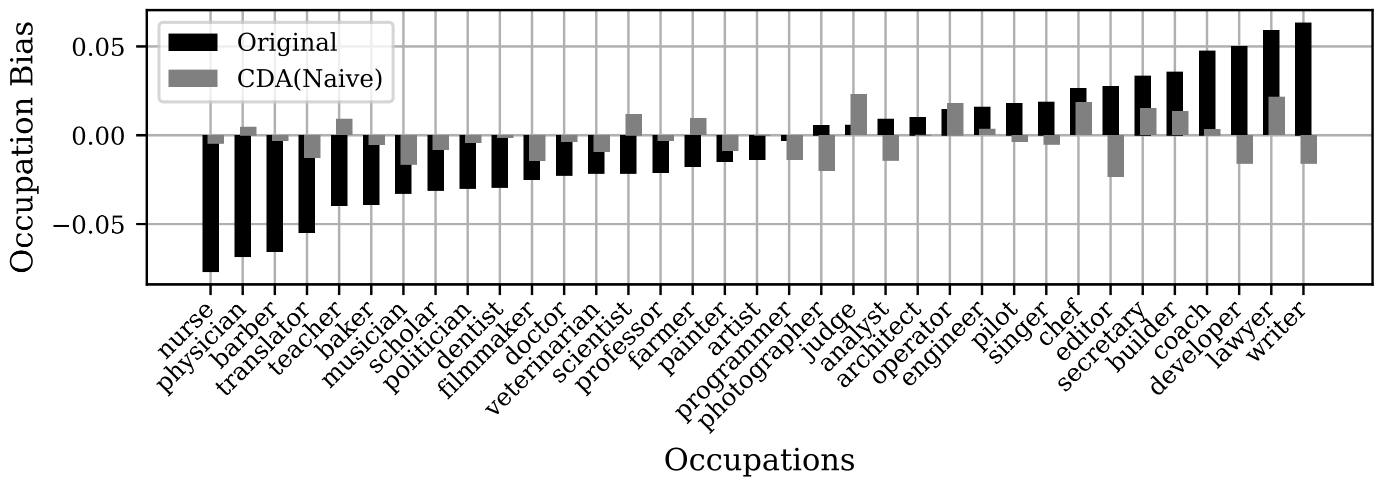
Counterfactual Data Augmentation

- Identify the list of gendered word pairs
  - (he,she), (man,woman), (actor,actress), (monk,nun), (actors,actresses),…..
- Make sure that the flipped sentences are grammatically correct
  - “Bill Clinton’s wife is Hillary."
    - Can’t flip! ~~Bill Clinton’s husband is Hillary.~~
    - Rule: If the gendered word refers to the same person/entity with a proper noun, we shall not flip.
  - Handle other corner cases
    - Ex: her (his/him)
- Could be applied to other NLP tasks

Experiment 1: Language Modeling

- Models:
  - A benchmark LSTM
    - Embedding size: 1500
    - LSTM cell size: 1500
  - Debiasing :
    - Debias the trained embedding [baseline]($\overrightarrow{\mathrm{WED}}$)
    - CDA(naïve): Flip every gender-specific words without any grammatical constraints
    - CDA(grammar): CDA(naïve) + grammatical constraint
    - Initialize the embedding layer from baseline and train on augmented dataset ($\overleftarrow{\mathrm{WED}}$)
- Data:
  - Wiki-text2 dataset
  - 36718 sentences, at least 7579 sentences with one gender-specific word

Results



- Occupation Bias
  - Negative occupation bias: biased towards female; Positive occupation bias: biased towards male
  - The bias in the original model roughly aligns with expectations on gender-occupation stereotypes in the real world
- Applying CDA consistently mitigate bias for almost all occupations.

Results

Config	Test Perp.	ΔTest Perp.	AOB	ΔAOB%
No debias	83.39	-	0.030	-
$\overrightarrow{\text{WED}}$	1128.15	+1044.76	0.0024	-92%
$\overleftarrow{\text{WED}}$	85.16	+1.77	0.013	-57%
CDA (g_{grammar})	84.03	+0.64	0.021	-30%
CDA (g_{naive})	83.63	+0.24	0.010	-67%

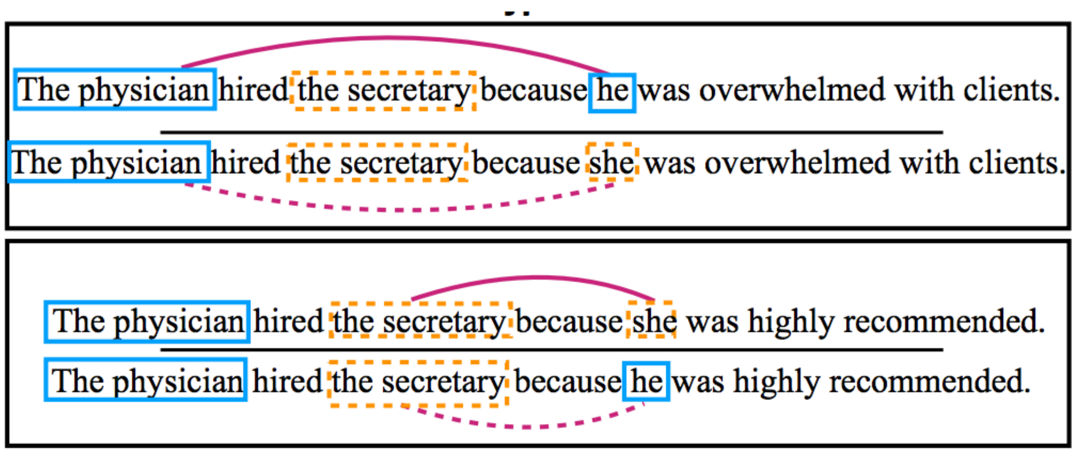
- AOB: Aggregate Occupation Bias; Test Perp: Test Perplexity
- Both CDA mitigate bias while preserving the performance
  - CDA(naïve) has surprisingly better performance

Results

Config	Test Perp.	ΔTest Perp.	AOB	ΔAOB%
No debias	83.39	-	0.030	-
$\overrightarrow{\mathrm{WED}}$	1128.15	+1044.76	0.0024	-92%
$\overleftarrow{\mathrm{WED}}$	85.16	+1.77	0.013	-57%
CDA (g_{grammar})	84.03	+0.64	0.021	-30%
CDA (g_{naive})	83.63	+0.24	0.010	-67%

- Apply word embedding debiasing after the model is trained $\overrightarrow{\mathrm{WED}}$)greatly reduces bias, but also destroys the model performance
  - Reason for low bias: low variance of the output score distribution
- Apply word embedding debiasing ($\overleftarrow{\mathrm{WED}}$)and continue training on the augmented dataset:
  - Reintroduce bias back

Bias in Coreference Resolution



(a) Coreference resolution

Coreference Resolution Basics

- Identify all mentions that refer to the same real world entity
  - Mentions: words/phrases that refers to a real entity in the world
  - Antecedent of a mention: other mention/mentions that precedes said mention, which refers to the same entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Coreference Resolution in Two Steps

- Detect the mentions (easy)

  "[I] voted for [Nader] because [he] was most aligned with [[my] values]," [she] said

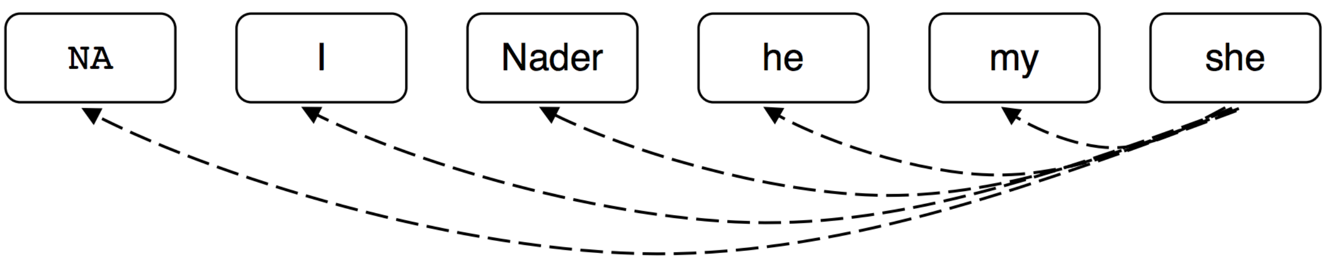
- Cluster the mentions (hard)

  "[I] voted for [Nader] because [he] was most aligned with [[my] values]," [she] said

A Mention Ranking System [clark & Manning, 2016]

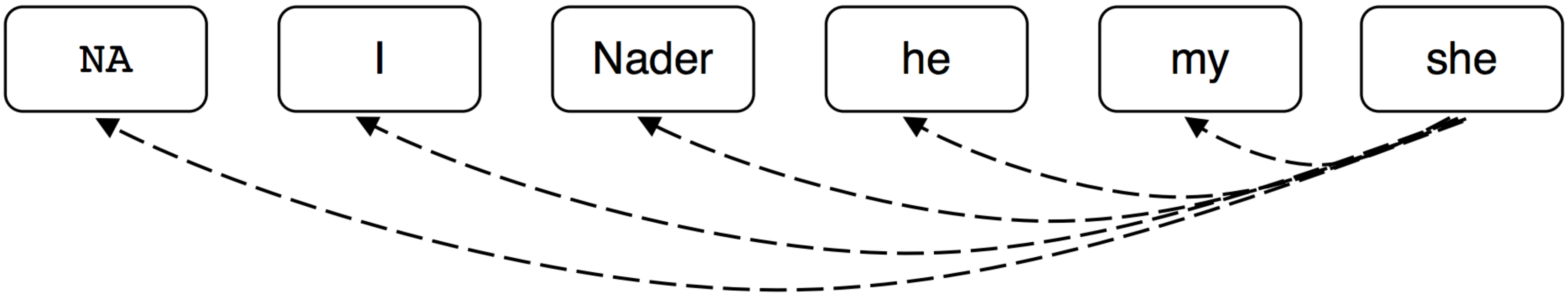
"[I] voted for [Nader] because [he] was most aligned with [[my] values]," [she] said

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything ("singleton" or "first" mention)



best antecedent for *she*?

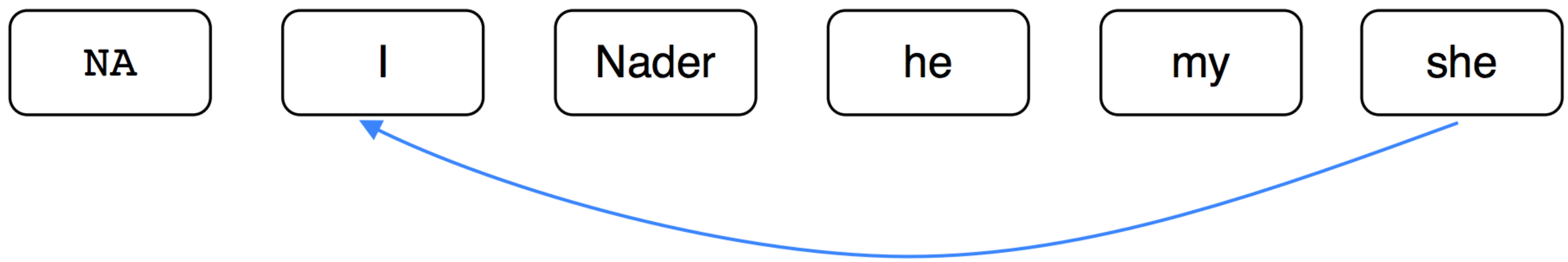
A Mention Ranking System [clark & Manning, 2016]



p(NA, she) = 0.1
p(I, she) = 0.5
p(Nader, she) = 0.1
p(he, she) = 0.1
p(my, she) = 0.2

Apply a softmax over the scores for candidate antecedents so probabilities sum to 1

A Mention Ranking System [clark & Manning, 2016]



only add highest scoring coreference link

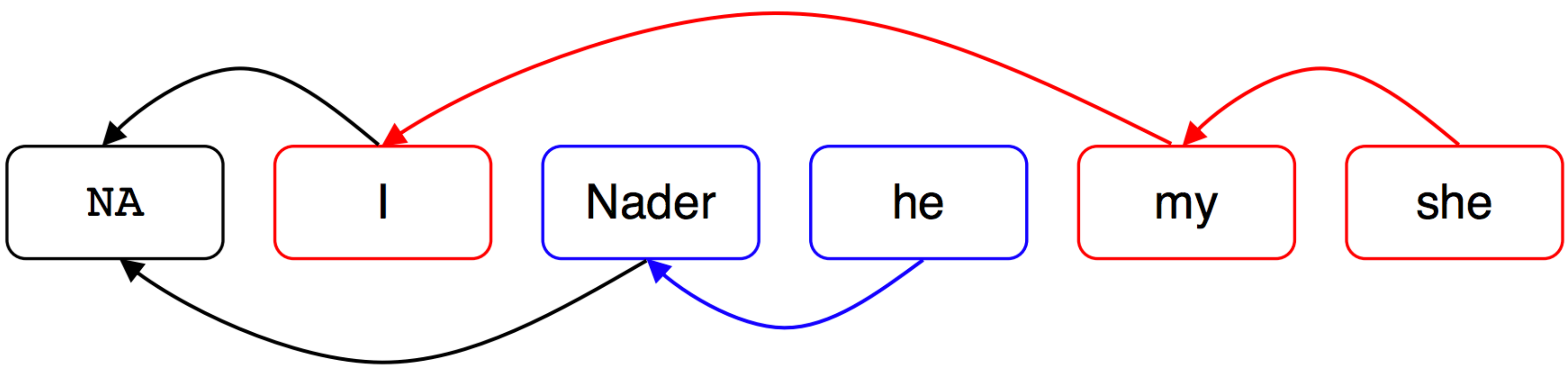
p(NA, she) = 0.1
p(I, she) = 0.5
p(Nader, she) = 0.1
p(he, she) = 0.1
p(my, she) = 0.2

Apply a softmax over the scores for candidate antecedents so probabilities sum to 1

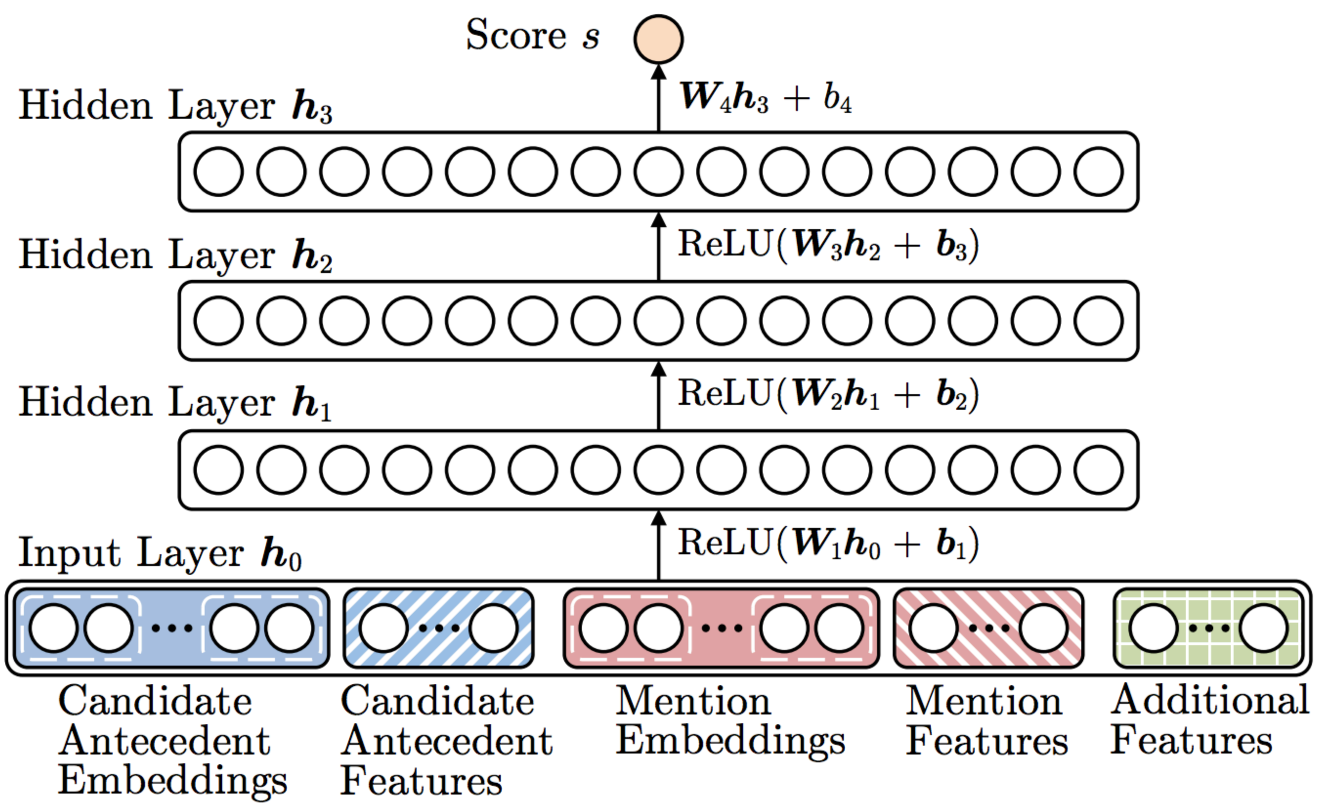
A Mention Ranking System [clark & Manning, 2016]

- Test Time:
  - Cluster the pairs

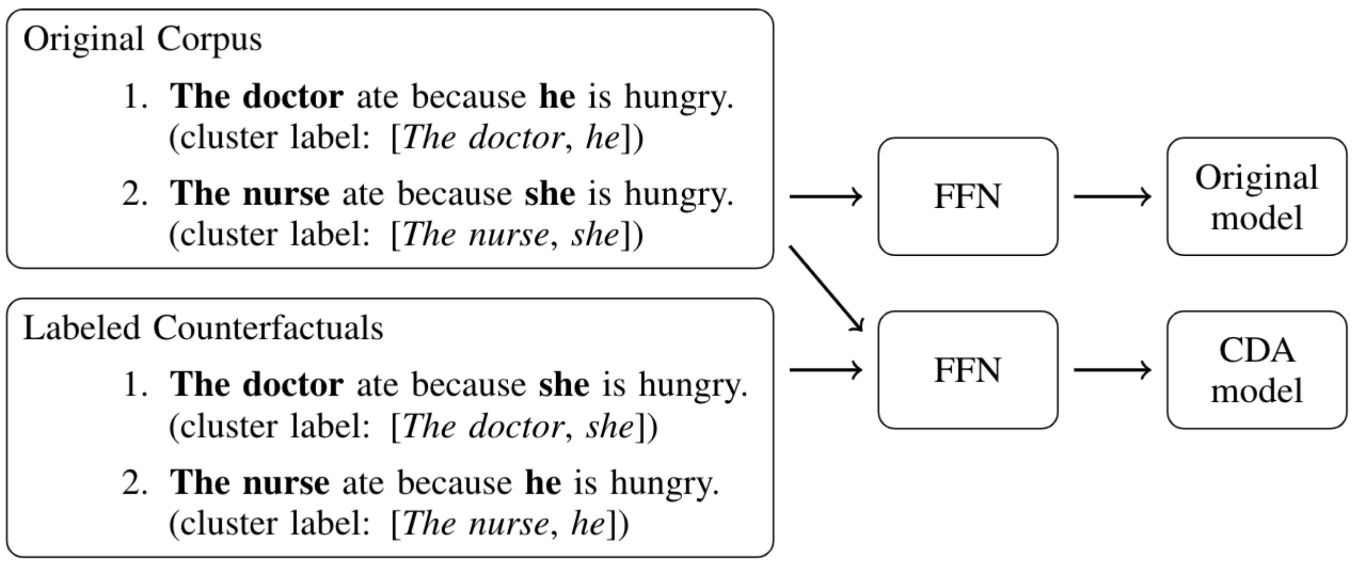
“[I] voted for [Nader] because [he] was most aligned with [[my] values],” [she] said



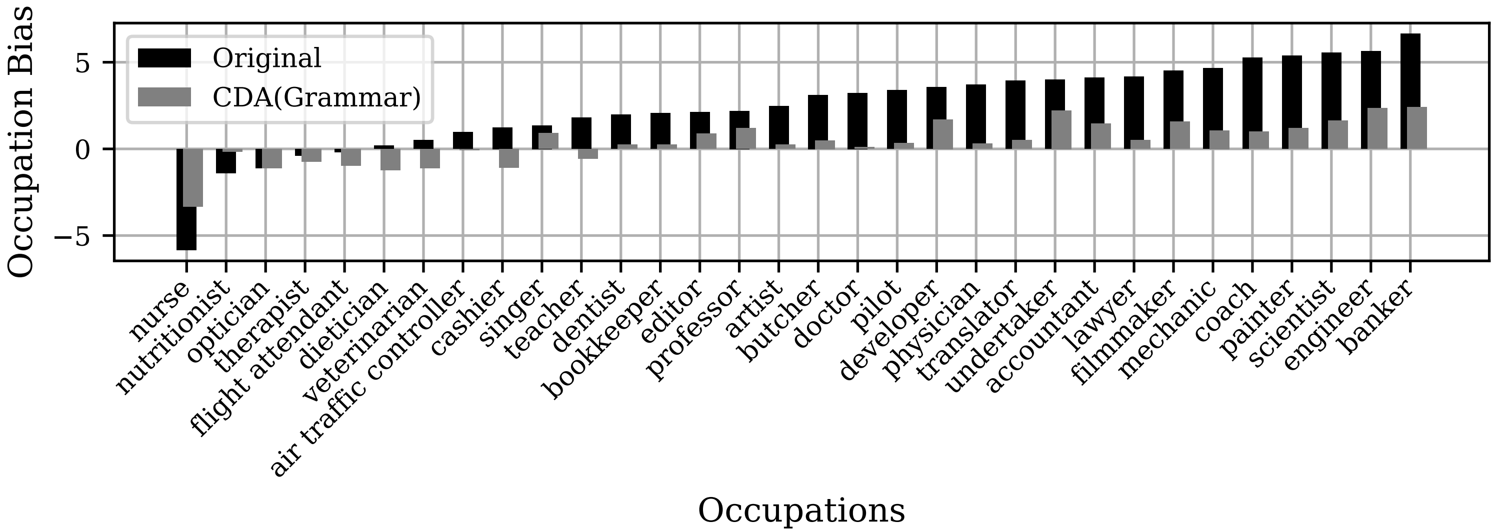
Neural Coref Model [clark & Manning, 2016]



CDA for Neural Coref Resolution



Bias in Coreference Resolution



- Occupation Bias
  - Negative occupation bias: biased towards female; Positive occupation bias: biased towards male
  - The bias in the original model roughly aligns with expectations on gender-occupation stereotypes in the real world
- Applying CDA consistently mitigate bias for almost all occupations.

Bias in Coreference Resolution

Index	Debiasing Configuration	Test Acc. (F1)	ΔTest Acc.	AOB	ΔAOB%
1.1	None	67.20[4]	-	3.00	-
1.2	CDA (g_{grammar})	67.40	+0.20	1.03	-66%
1.3	WED	67.10	-0.10	2.03	-32%
1.4	CDA (g_{grammar}) w/ WED	67.10	-0.10	0.51	-83%

Table 2: Comparison of 4 debiasing configurations for NCR model of Lee et al. [2017].

- Additive Effect of :
  - Fixing the embeddings using debiasing
  - Fixing other parameters using counterfactual data augmentation

Bias in Coreference Resolution

Index	Debiasing Configuration	Test Acc. (F1)	ΔTest Acc.	AOB	$\pm$AOB	ΔAOB%
2.1	None	69.10	-	2.95	2.74	-
2.2	$\overleftarrow{\text{WED}}$	68.82	-0.28	2.50	2.24	-15%
2.3	$\overrightarrow{\text{WED}}$	66.04	-3.06	0.9	0.14	-69%
2.4	$\overleftarrow{\text{WED}}$ and $\overrightarrow{\text{WED}}$	66.54	-2.56	1.38	-0.54	-53%
2.5	CDA (g_{grammar})	69.02	-0.08	0.93	0.07	-68%
2.6	CDA (g_{grammar}) w/ $\overleftarrow{\text{WED}}$	68.5	-0.60	0.72	0.39	-75%
2.7	CDA (g_{grammar}) w/ $\overrightarrow{\text{WED}}$	66.12	-2.98	2.03	-2.03	-31%
2.8	CDA (g_{grammar}) w/ $\overleftarrow{\text{WED}}$, $\overrightarrow{\text{WED}}$	65.88	-3.22	2.89	-2.89	-2%

Table 3: Comparison of 8 debiasing configurations for NCR model of Clark and Manning [2016b]. The $\pm$AOB column is aggregate occupation bias with preserved signs in aggregation.

Summary

- Gender bias exists in downstream tasks
  - Language Models
  - Coreference Resolution
- Can effectively reduce bias by training on augmented dataset
- Previous methods of addressing bias in word embeddings
  - Hurts performance if done after a model is trained
  - Reintroduces the bias back if initialized before a model is trained
  - Additive effect if the embedding is pretrained

Questions?

- References:

  - T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.
  - K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender bias in neural natural language processing," *arXiv preprint arXiv:1807.11714*, 2018.
  - Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking corefer- ence models. *arXiv preprint arXiv:1609.08667*, 2016b.
  - Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*, 2017.