

Bias In NLP Tasks

Caleb Kaiji Lu

Spring 18739

Agenda

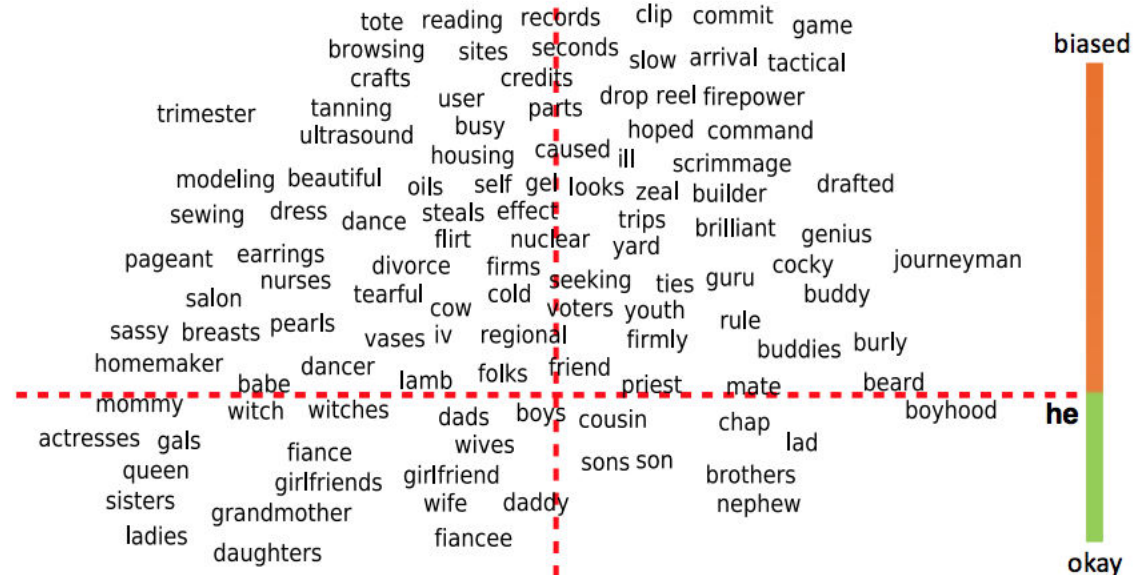
- Bias in Word Embeddings
- Bias in Neural Coreference Resolution
- Debiasing Method
- Future Work

Agenda

- Bias in Word Embeddings
- Bias in Neural Coreference Resolution
- Debiasing Method
- Future Work

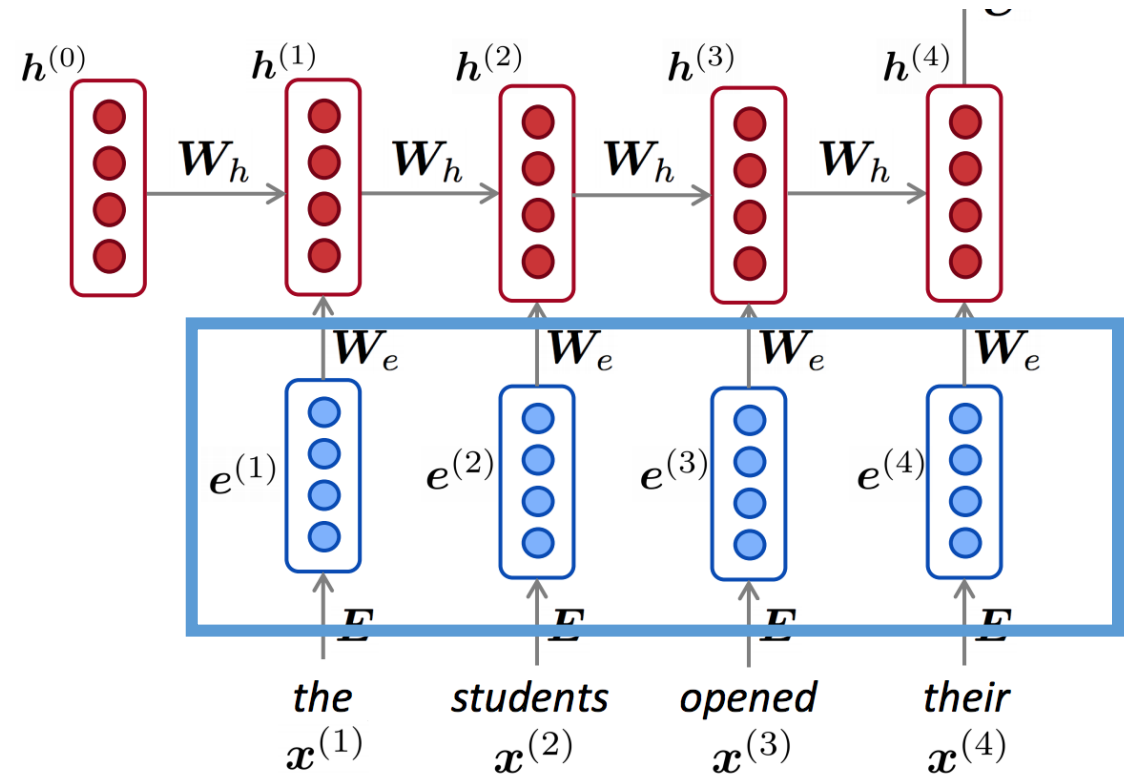
Bias In Word Embeddings

- Man is to Programmer - Woman is to (Homemaker)
- Debiasing Method
 - Remove Gender Component from the vector space



The Usage of Word Embeddings

- The first step in most NLP tasks
- Language Modeling

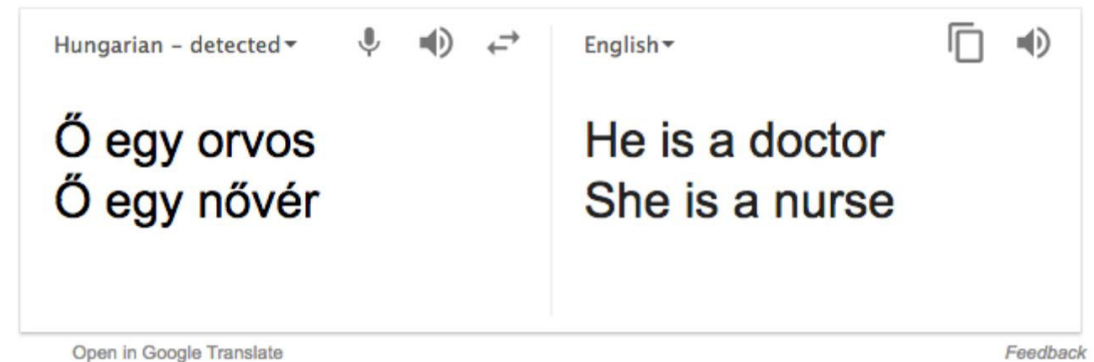
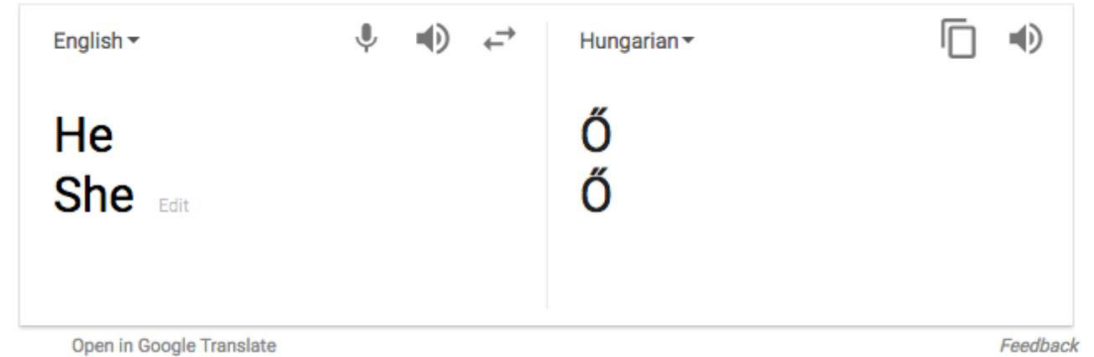


Word Embeddings: Trainable or Fixed?

- Word Embedding can be used to replace words as inputs to the model
 - Efficient
 - Will cover most vocabularies if the dataset is small
- Word Representations can be trained as part of a neural net work to solve a particular task
 - Learn Useful representations specific to the task
 - POS: Cat & Human would be similar
 - Expensive
 - Dataset might be too small to learn useful representations
 - Dataset might not cover all the vocabularies
 - Usually word embedding layer is initialized with pretrained word embeddings(word2vec, glove)

Does Bias Exist in Downstream Tasks?

- If so, How does bias in word embedding play a role in this?
- Does word embedding play all the role in this?
 - If not, where does the bias come from?
 - How do we fix it?



Word Embeddings: Trainable or Fixed?

- Word Embedding can be used to replace words as inputs to the model
 - Efficient
 - Will cover most vocabularies if the dataset is small

Debiasing Word embeddings maybe Helpful?

- Word Representations can be trained as part of a neural net work to solve a particular task
 - Learn Useful representations specific to the task
 - POS: Cat & Human would be similar
 - Expensive
 - Dataset might be too small to learn useful representations
 - Dataset might not cover all the vocabularies
 - Usually word embedding layer is initialized with pretrained word embeddings(word2vec, glove)

The bias would be relearned?

Debias the weights after training?

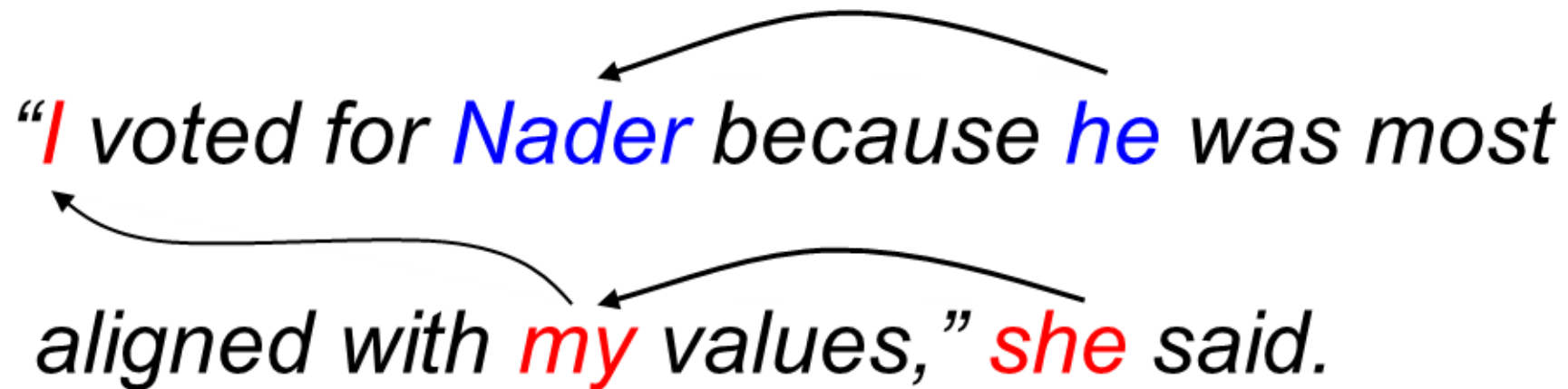
Agenda

- Bias in Word Embeddings
- **Bias in Neural Coreference Resolution**
- Debiasing Method
- Future Work

Coreference Resolution

- The task of finding words and expressions(mentions) that refer to the same entity in the world.

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

The diagram illustrates coreference resolution in the sentence "I voted for Nader because he was most aligned with my values," she said. Three curved arrows indicate the relationships: one from "Nader" to "he", one from "my" to "I", and one from "she" to the start of the sentence.

Nader is the antecedent of **he**

Coreference Resolution

- Noun phrases refer to entities in the world, many pairs of noun phrases co-refer, some nested inside others

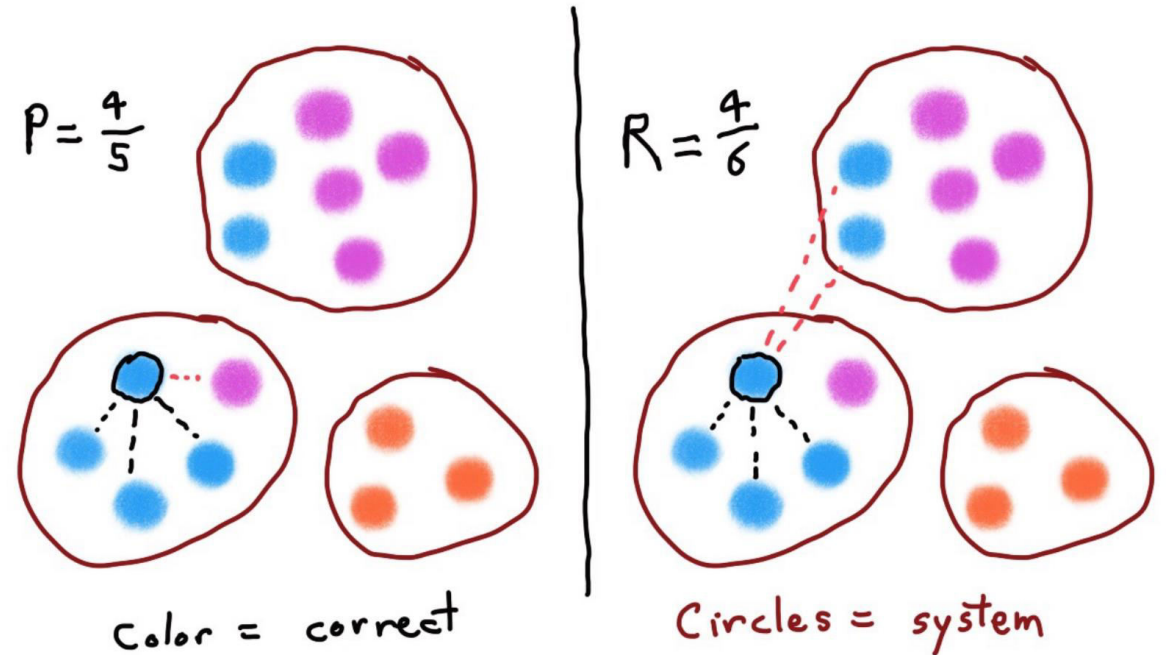
John Smith, CFO of Prime Corp. since 1986,
saw his pay jump 20% to \$1.3 million
as the 57-year-old also became
the financial services co.'s president.

Applications of Coreference Resolution

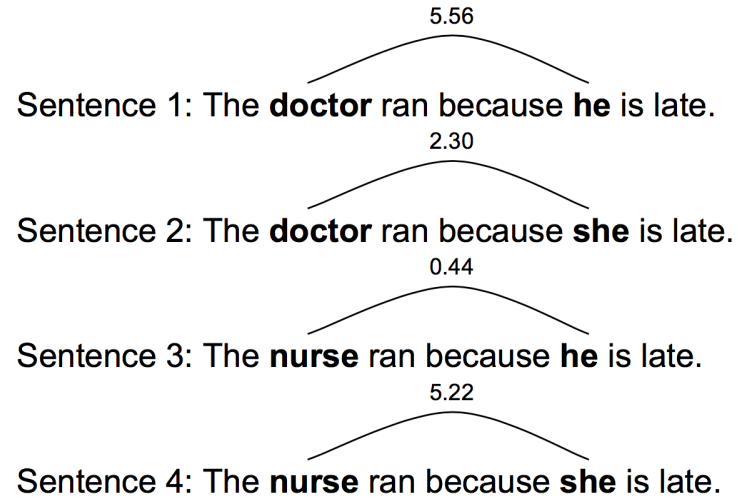
- Full text understanding:
 - Understanding an extended discourse
 - Machine translation (if languages have different features of gender, number, etc.)
 - Text summarization, including things like web snippets
 - Tasks like information extraction and question answering
 - Correctly answering often involves resolving anaphora
 - He married Claudia Ross in 1971.

Evaluation of Coreference Resolution

- B Cube Metric:
- Other Metrics:
 - MUC
 - CEAF
- All of them are sort of evaluating getting coreference links/clusters right and wrong

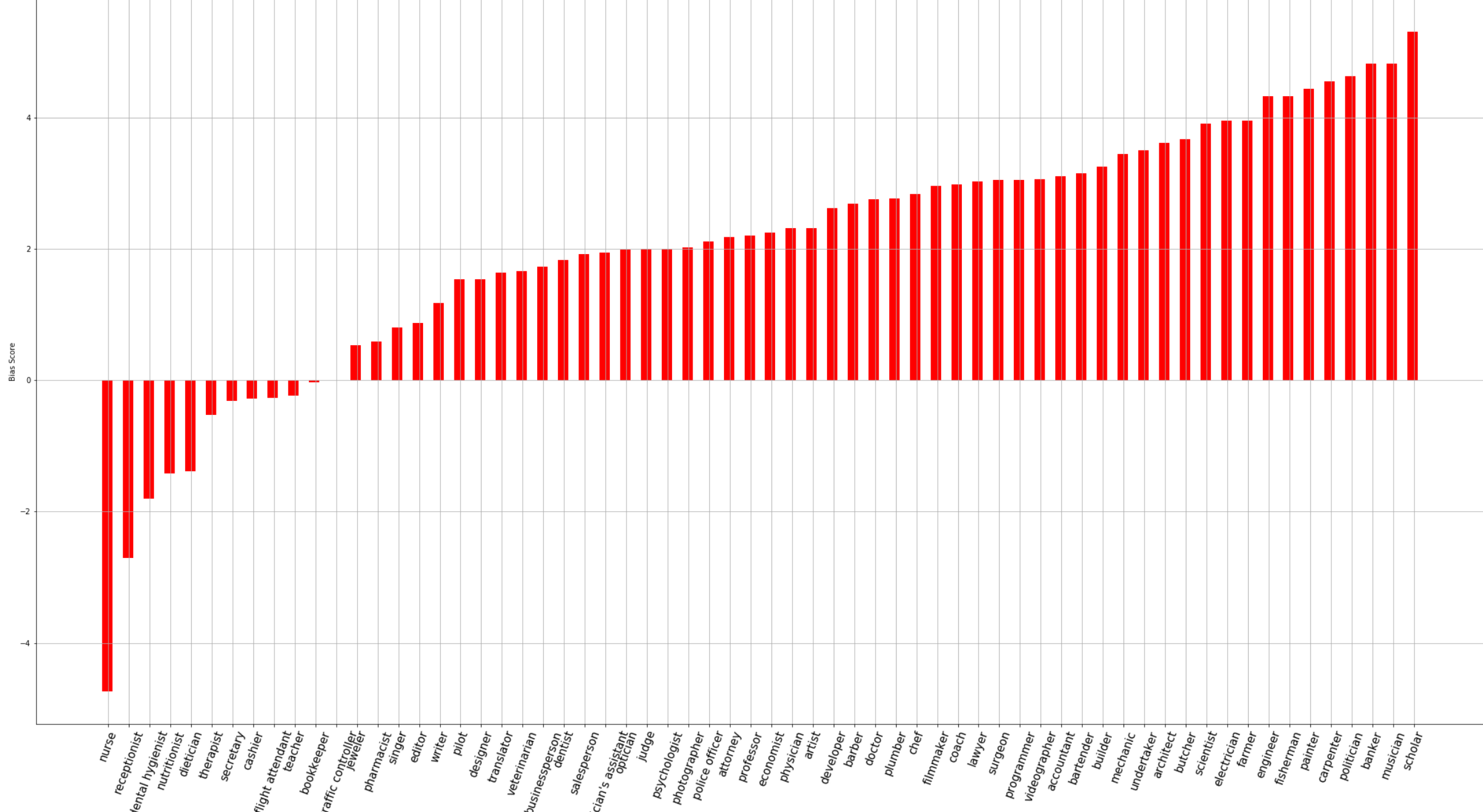


Bias in Coreference Resolution



$$\text{Bias}(\text{doctor}) = 5.56 - 2.30 = 2.26$$

$$\text{Bias}(\text{Nurse}) = 0.44 - 5.22 = -4.78$$



Coreference Resolution

- Traditional Rule-Based Model (What is being used right now)
 - 1. Begin at the NP immediately dominating the pronoun
 - 2. Go up tree to first NP or S. Call this X, and the path p.
 - 3. Traverse all branches below X to the left of p, left-to-right, breadth-first. Propose as antecedent any NP that has a NP or S between it and X
 - 4. If X is the highest S in the sentence, traverse the parse trees of the previous sentences in the order of recency. Traverse each tree left-to-right, breadth first. When an NP is encountered, propose as antecedent. If X not the highest node, go to step 5.
 - 5. From node X, go up the tree to the first NP or S. Call it X, and the path p.
 - 6. If X is an NP and the path p to X came from a non-head phrase of X (a specifier or adjunct, such as a possessive, PP, apposition, or relative clause), propose X as antecedent (The original said “did not pass through the N’ that X immediately dominates”, but the Penn Treebank grammar lacks N’ nodes....)
 - 7. Traverse all branches below X to the left of the path, in a left-to-right, breadth first manner. Propose any NP encountered as the antecedent
 - 8. If X is an S node, traverse all branches of X to the right of the path but do not go below any NP or S encountered. Propose any NP as the antecedent.
 - 9. Go to step 4

Rules don't always work

- [The city council] refused [the women] a permit because they feared violence.
- [The city council] refused [the women] a permit because they advocated violence.
 - Winograd (1972)

Neural Coreference Resolution Models

- A Mention Ranking Model
 - Given all mentions are labeled
 - Explicitly rank all candidate antecedents for a mention

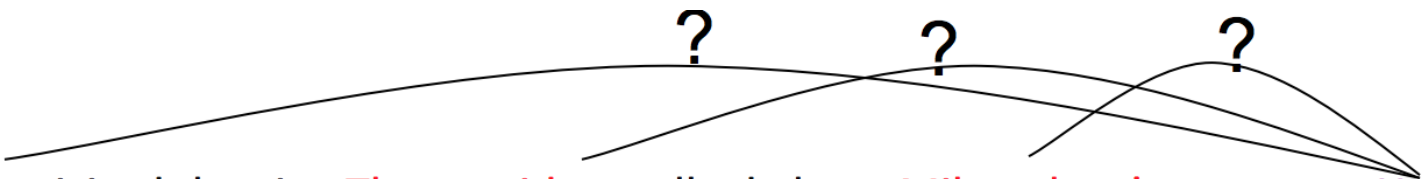
“I voted for Nader because he was most aligned with my values,” she said.

The diagram illustrates coreference resolution in the sentence: "I voted for Nader because he was most aligned with my values," she said. Arrows indicate the following relationships: an arrow from "I" to "Nader", an arrow from "Nader" to "he", an arrow from "my" to "I", and an arrow from "she" to "I".

Coreference Score

- For each candidate antecedent, a score is given by the model.

Mr. Obama visited the city. The president talked about Milwaukee's economy. He mentioned new jobs.

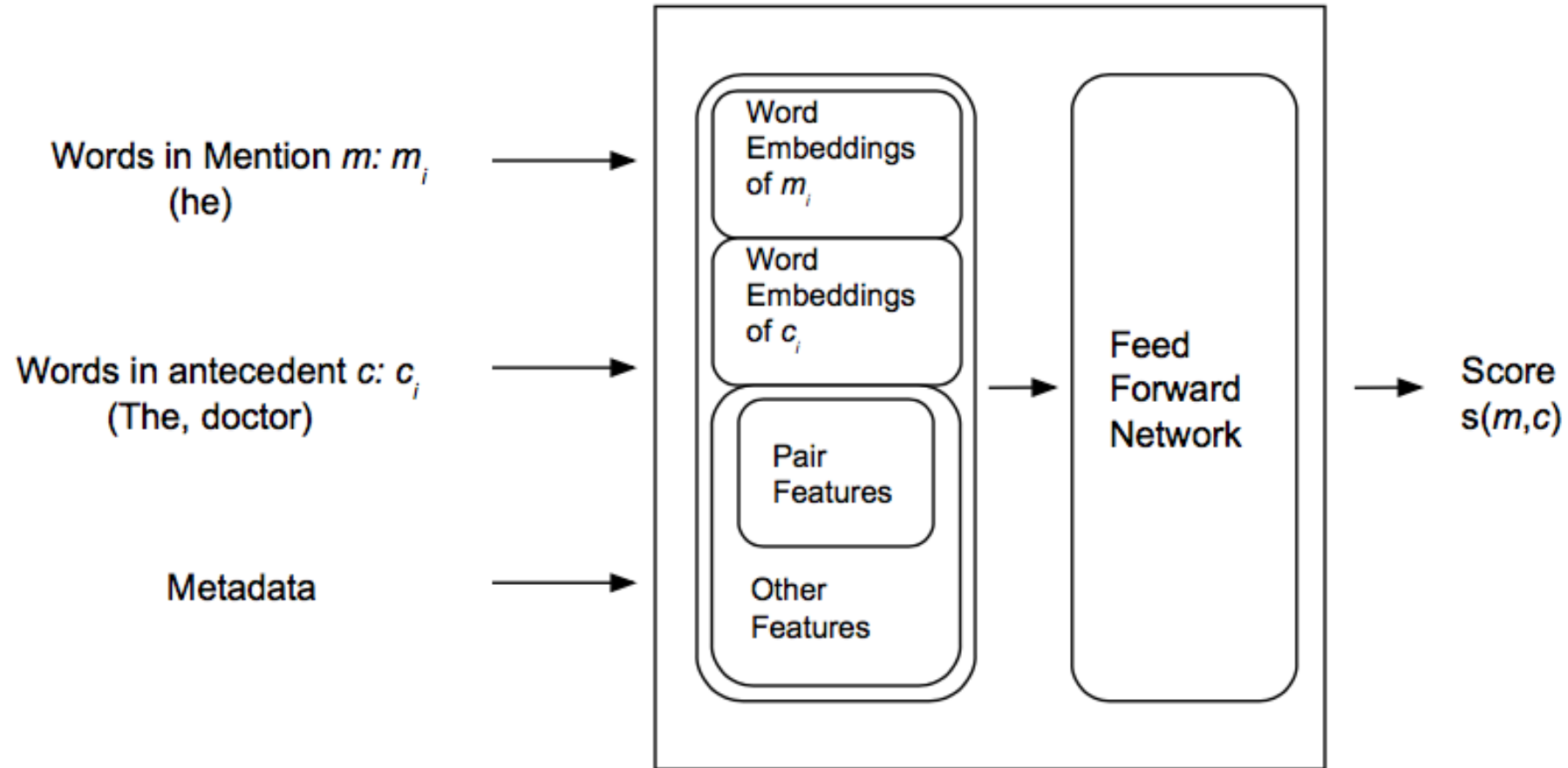


The diagram illustrates the coreference resolution process. Three candidate antecedents are shown above the text: 'Mr. Obama', 'The president', and 'Milwaukee's'. Each candidate is connected to the pronoun 'He' by a curved line. A question mark is placed above each of these three lines, indicating that the model is evaluating the score for each potential antecedent to determine the most likely coreferent.

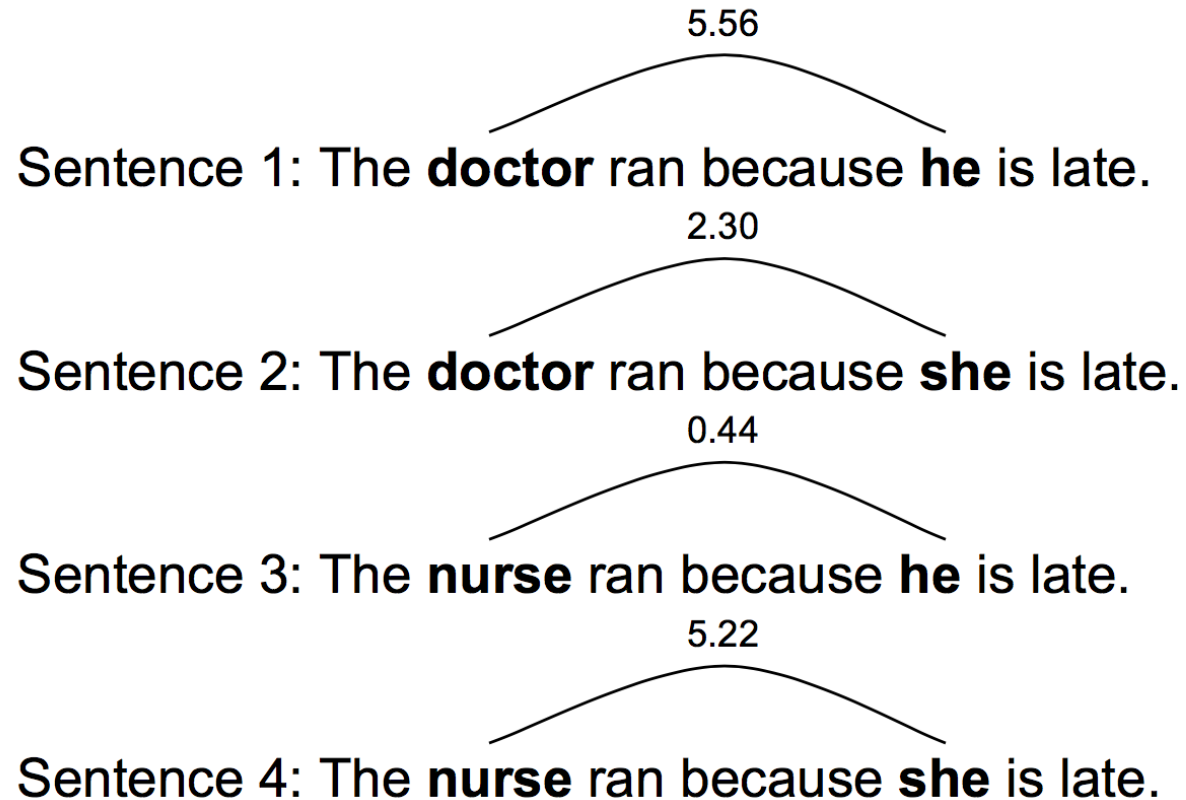
- Assign each mention its highest scoring candidate antecedent according to the model

Neural Coreference Resolution Models

Ex: The Doctor ran because he is late .



Bias in Coreference Resolution



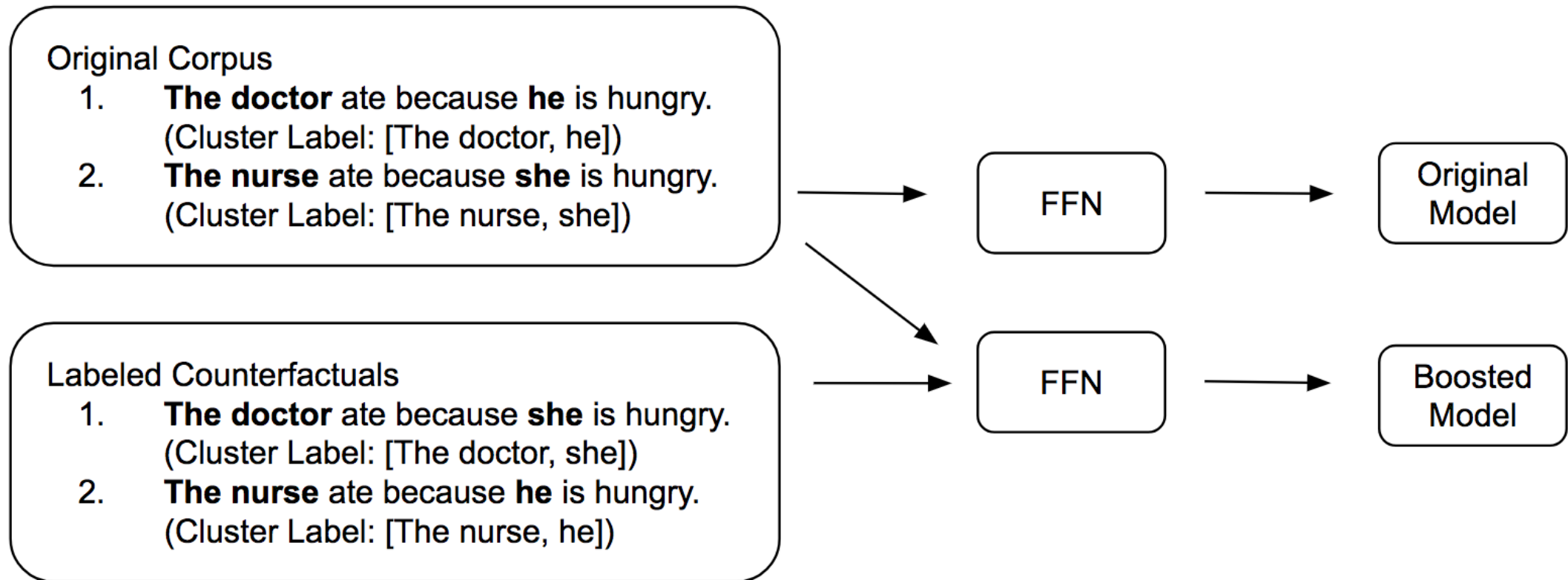
Agenda

- Bias in Word Embeddings
- Bias in Neural Coreference Resolution
- **Debiasing Method**
- Future Work

The bias comes from the data

- Simplest solution: Collect better data!
 - Not realizable
- Fix the model
 - Make the model learn less bias!
 - Well, we don't know how the model works...
 - (Need Explanations!)
 - Invasive, could hurt performance
- Synthesize Better data?
 - Create counterfactuals!
 - Data Augmentation
 - Out-of-distribution from real-world data
 - Need Labeling

Counterfactual Retraining in NCR



Counterfactual Retraining in NCR

- Step 1: Create a list of gendered word pairs
 - (He,she),(man,woman),(actor,actress),(monk,nun).....(actors,actresses)
- Step 2: Identify the correct situations to flip
 - Bill Clinton....He is a great President
 - Can't flip! ~~Bill Clinton....She is a great President.~~
 - Rule: If the gendered word is in the same cluster with a proper noun, should not flip.
- Step 3: Handle other corner cases
 - Ex: Her (his/him)

Experiment on Real NCR Modles

- Model I (Lee et al 2017)
 - EMNLP 2017
 - Fixed Embedding layer
 - Glove, 300 dimension
 - Some other model variations
- Model II (Clark et al 2016)
 - EMNLP 2016
 - Trainable embedding layer
 - Initialized with word2vec
 - 50 dimensions

Model I (Lee et al, 2017)

Table 1: Comparison of 4 models for Lee et al. [2017]

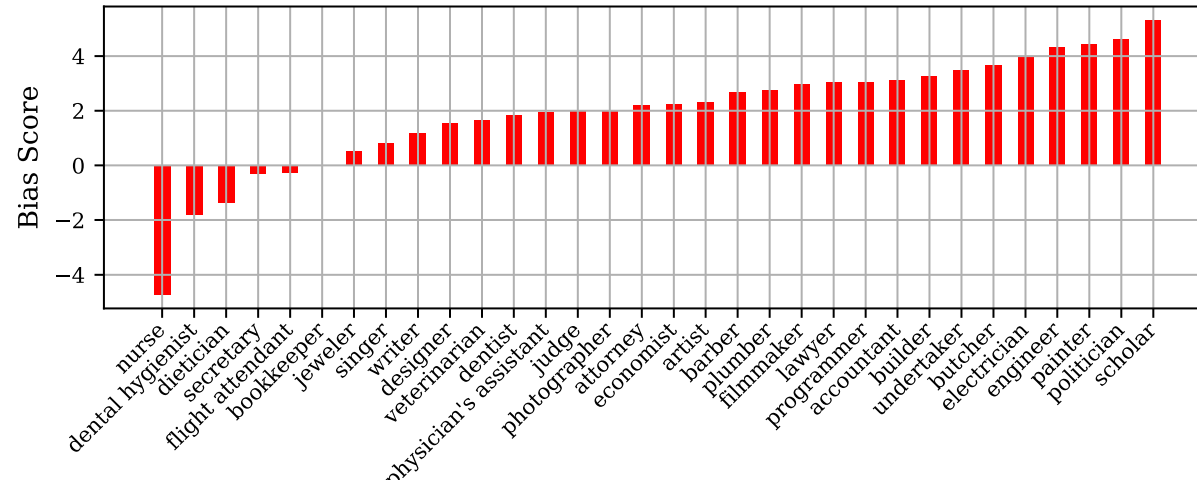
Index	Model	Test Acc. (F1)	Δ Test Acc.	L1 Bias	Δ Bias%
1.1	Original Model	67.24 ¹	-	2.46	-
1.2	Boosted Model	67.04	-0.20	0.91	-63%
1.3	Original Model w/ Debiased Embeddings	66.70	-0.54	1.35	-45%
1.4	Boosted Model w/ Debiased Embeddings	66.82	-0.42	0.51	-79%

For N sentence templates, the gender bias towards occupation word o is defined as

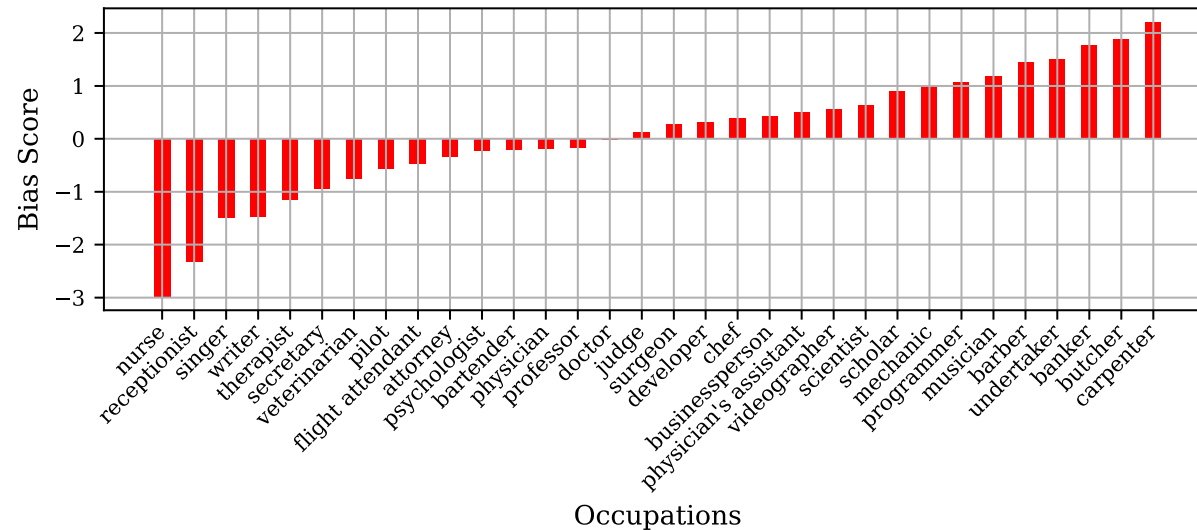
$$Bias_o = \frac{1}{N} \sum_{i=1}^N |s_i(she, o) - s_i(he, o)|$$

Model I (Lee.et al,2017)

- Original Model

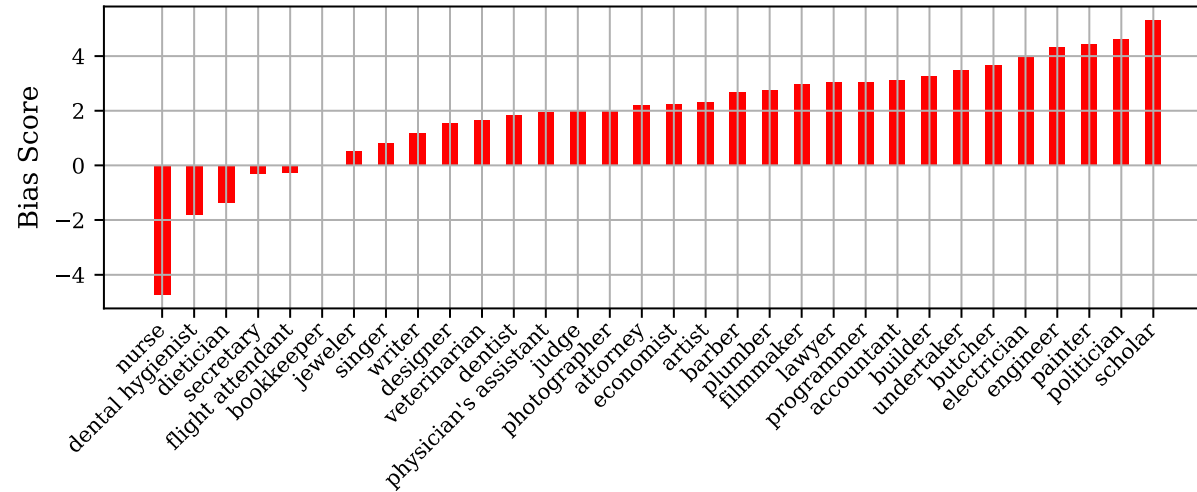


- Boosted Model

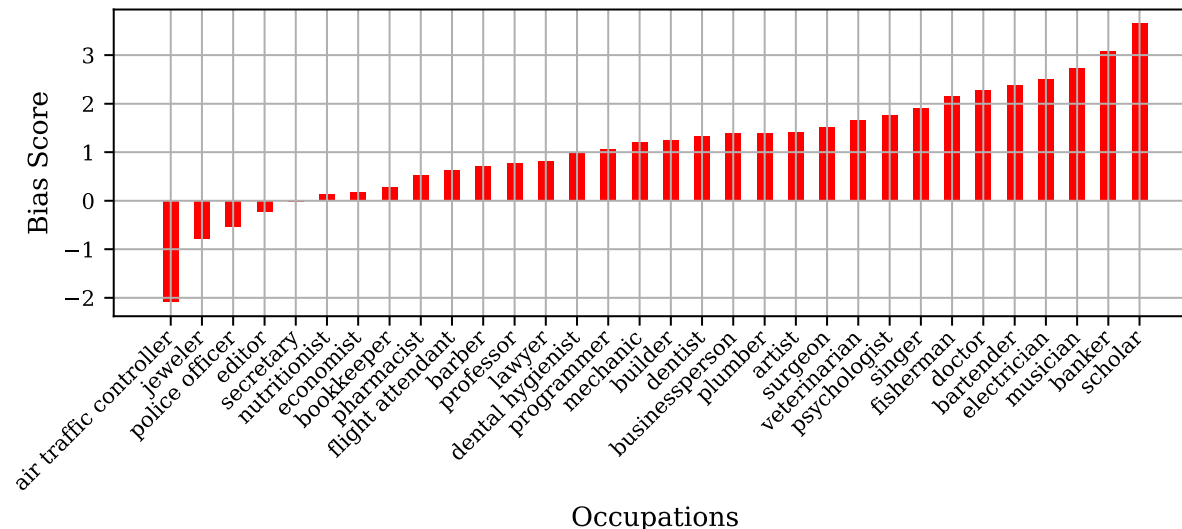


Model I (Lee.et al,2017)

- Original Model

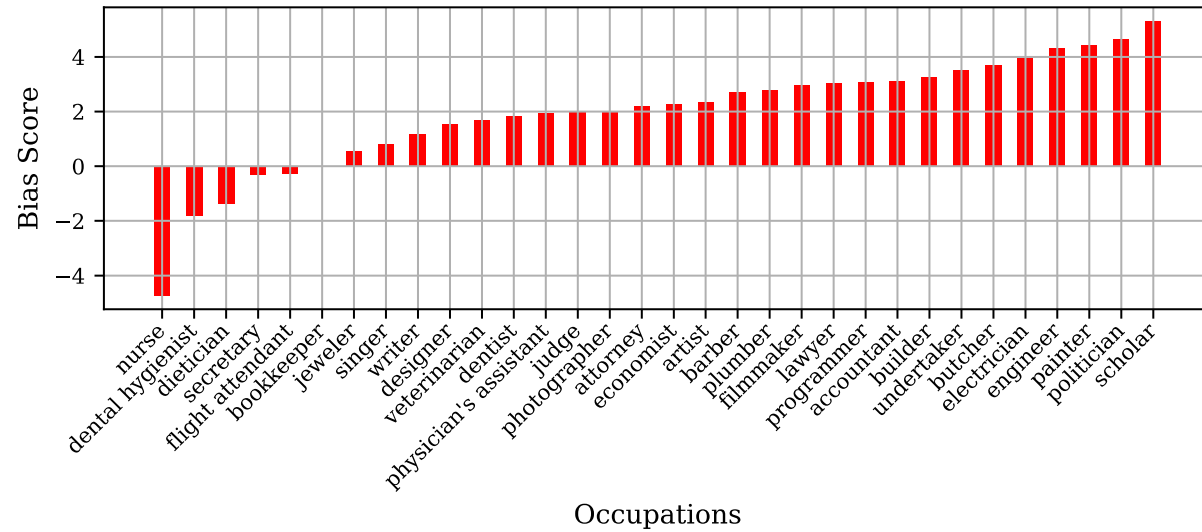


- Original Model w/
debiased
Embedding

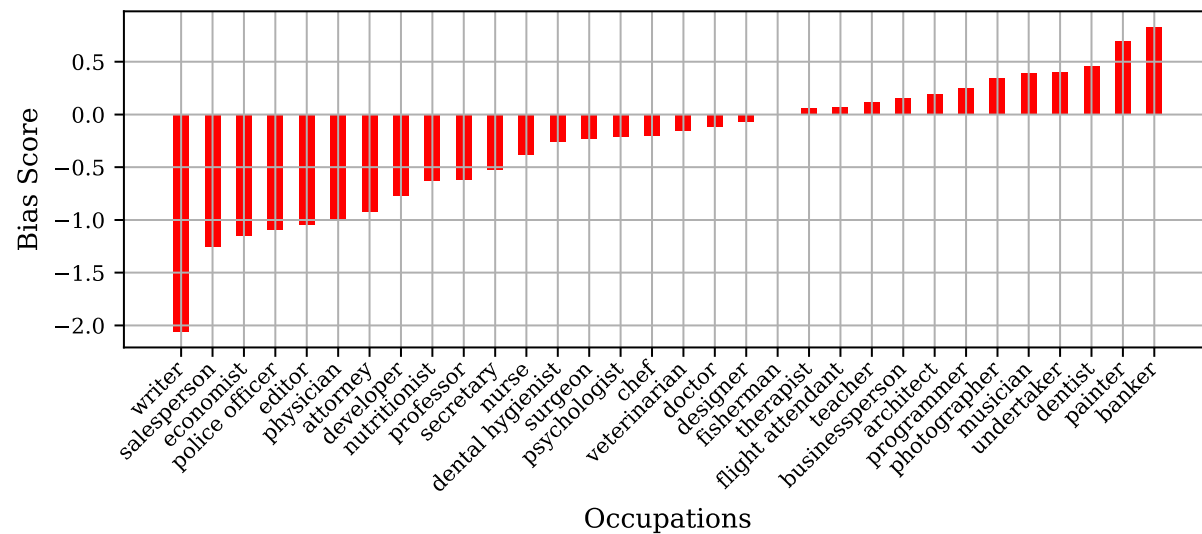


Model I (Lee.et al,2017)

- Original Model



- Boosted Model + Debiased WE



Model II(Clark & Manning, 2016)

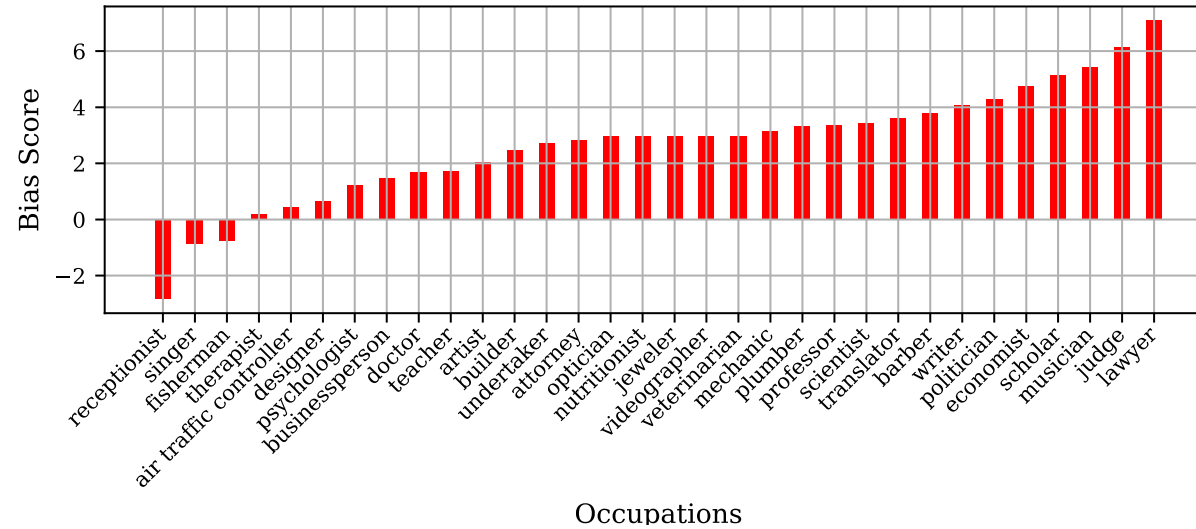
Table 3: Comparison of 5 models for Clark and Manning [2016b]

Model	Test Acc. (F1)	Δ Test Acc.	$L1$ Bias	Δ Bias%
Orig. Model	69.10	-	2.95	-
Orig. Model with debiased initialization	68.82	-0.28	2.50	-15%
Orig. Model with debiased weights	66.04	-3.06	0.9	-69%
Boosted Model	69.02	-0.08	0.93	-68%
Boosted Model with debiased initialization	68.5	-0.60	0.72	-75%

- The Usage of Word Embeddings
 - Only for initialization
 - Learnable Embedding layer
- Debiasing of WE
 - Debiasing Initialization
 - Debiasing the weights after the model is trained

Model II(Clark & Manning, 2016)

- Original Model

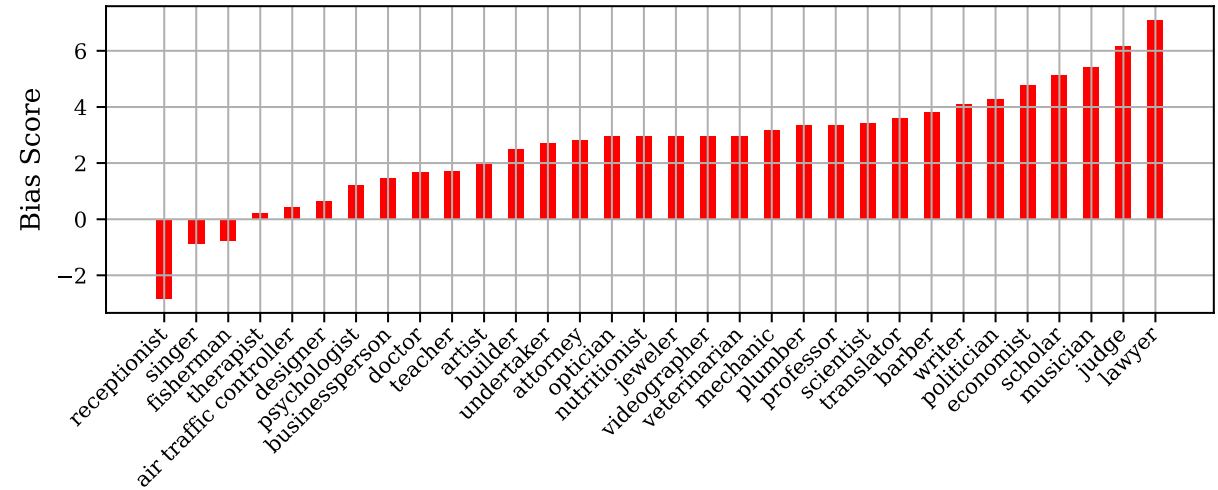


- Boosted Model

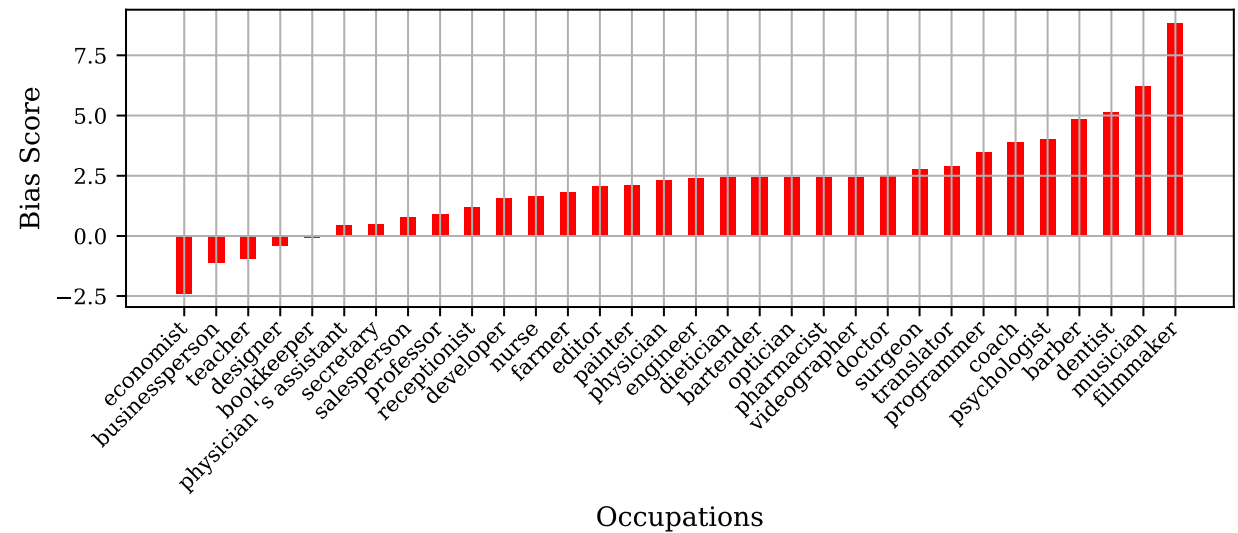


Model II(Clark & Manning, 2016)

- Original Model



- Original Model with debiased initialization



Agenda

- Bias in Word Embeddings
- Bias in Neural Coreference Resolution
- Debiasing Method
- **Future Work**

Future Work

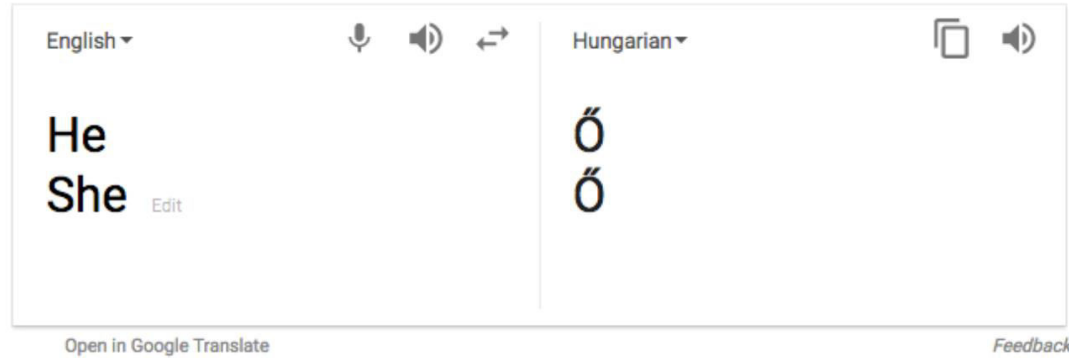
- Bias in Other NLP tasks
 - Language Modeling
 - Machine Translation

Bias in Language Modeling

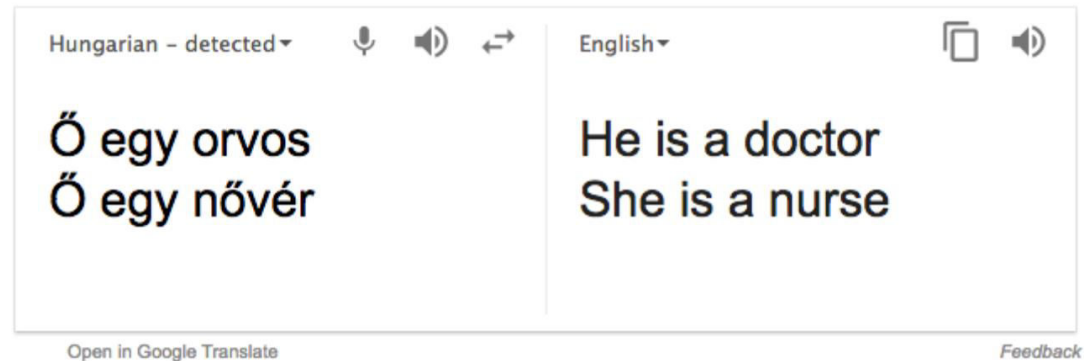
Sentence	Perplexity of individual sentence
The doctor is a man .	26.17
The doctor is a woman .	28.85
The professor is a man .	36.16
The professor is a woman .	37.51
The nurse is a man .	52.93
The nurse is a woman .	34.42

Bias in Machine Translation

- English-Hungarian



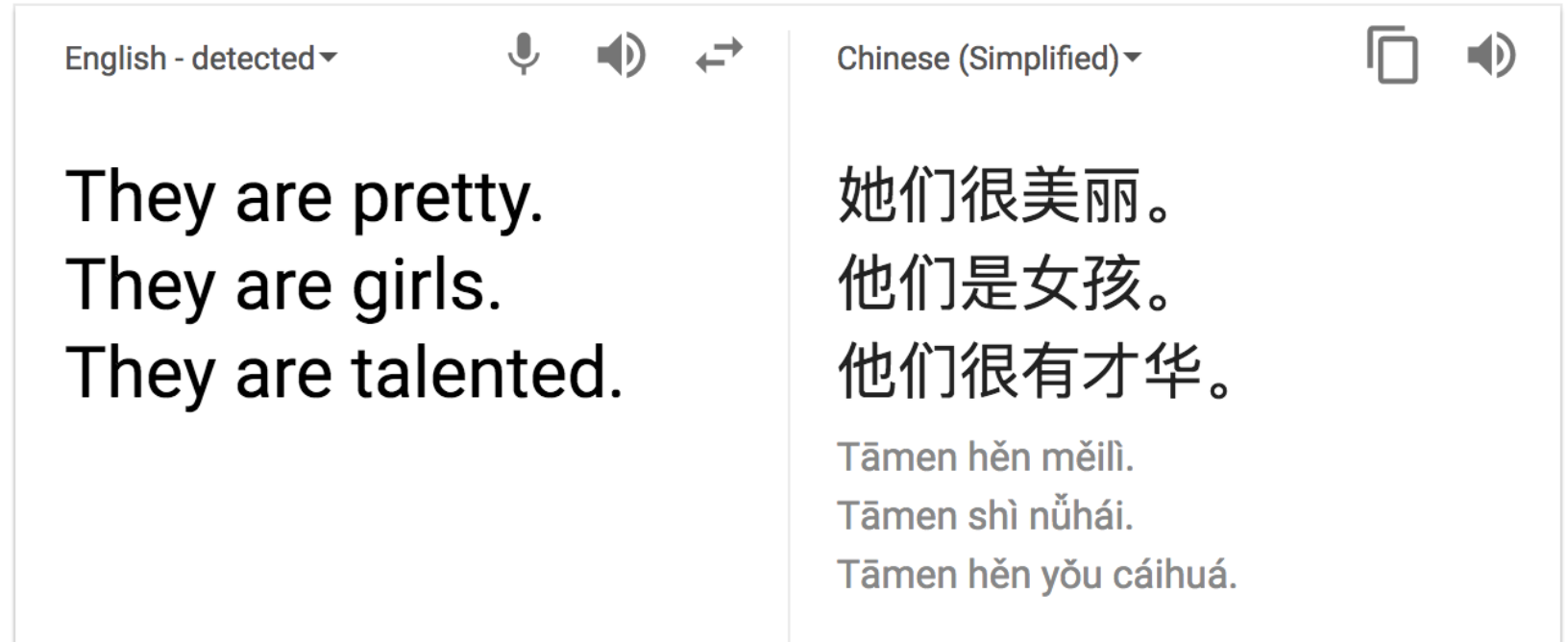
A screenshot of the Google Translate interface. The left panel is labeled 'English' and contains the text 'He' and 'She' with an 'Edit' link next to 'She'. The right panel is labeled 'Hungarian' and contains the output 'Ő' and 'Ő'. At the bottom, there are links for 'Open in Google Translate' and 'Feedback'.



A screenshot of the Google Translate interface. The left panel is labeled 'Hungarian - detected' and contains the text 'Ő egy orvos' and 'Ő egy nővér'. The right panel is labeled 'English' and contains the output 'He is a doctor' and 'She is a nurse'. At the bottom, there are links for 'Open in Google Translate' and 'Feedback'.

Bias in Machine Translation

- English-Chinese



The screenshot shows a machine translation interface with two columns. The left column is labeled 'English - detected' and contains three lines of text: 'They are pretty.', 'They are girls.', and 'They are talented.'. The right column is labeled 'Chinese (Simplified)' and contains three lines of Chinese text: '她们很美丽。', '他们是女孩。', and '他们很有才华。'. Below the Chinese text are three lines of pinyin: 'Tāmen hěn měilì.', 'Tāmen shì nǚhái.', and 'Tāmen hěn yǒu cáihuá.'. The interface also includes icons for microphone, speaker, and a double-headed arrow, as well as a copy icon and a speaker icon in the top right corner.

English - detected

Chinese (Simplified)






They are pretty.
They are girls.
They are talented.

她们很美丽。
他们是女孩。
他们很有才华。

Tāmen hěn měilì.
Tāmen shì nǚhái.
Tāmen hěn yǒu cáihuá.

Bias in Machine Translation

- English-French

<p>English - detected ▾   </p> <p>This nurse is talented. This male nurse is talented. Edit</p>	<p>French ▾  </p> <p>Cette infirmière a du talent. Cet infirmier a du talent.</p>
--	---