Bias in Word Embeddings

Caleb Kaiji Lu

Class 18739 Spring 2019

Agenda

- Recap on word embeddings
- Bias in word embeddings
  - 3 metrics for quantifying embedding stereotypes [Bolukabasi et. al, 2016]
- Debiasing algorithms [Bolukabasi et. al, 2016]
- Embedding as a lens to study history [Garg et al, 2018]

Previously: Word Embedding is a Dictionary

Training data: corpus of text

Context window

The dog is chasing a cat.

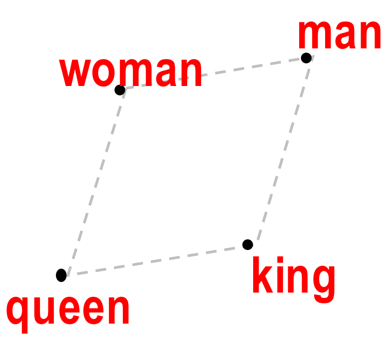
v_{dog} $v_{chasing}$ v_{cat}

find v's to $\max \log P(chasing|dog) + \log P(cat|dog)$

where $P(cat|dog) \propto \exp(v_{cat} \cdot v_{dog})$

Previously: Word Embedding

- Word embedding captures relationships among words
  - Semantic relationship: *woman:man::queen:king*
  - Syntactic relationship: *they: their:: he:his*
  - More complicated knowledge-base like relationship:
    - *Beijing:China :: Paris: France*
  - Standard metric to evaluate a word embedding



Word Embeddings Also Capture Bias [Bolukabasi, 16]

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

- He:Brother :: She:
- He:Blue :: She
- He:Doctor :: She:
- He:Realist :: She:
- She:Pregnancy :: He:
- She:Baking::He:
- She:Blonde::He:
- He:Computer :: She:

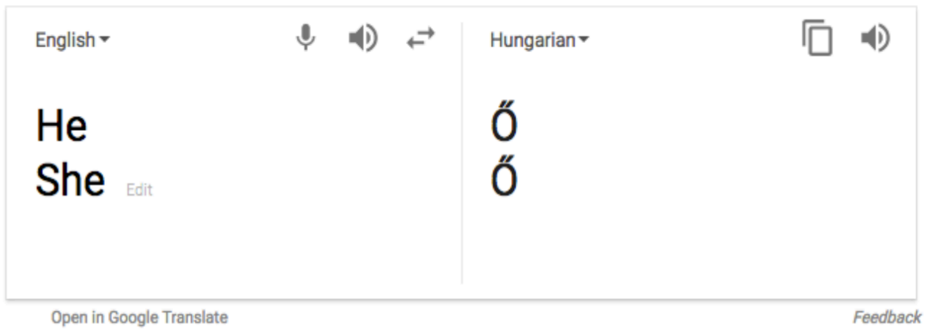
Word Embeddings Also Capture Bias [Bolukabasi, 16]

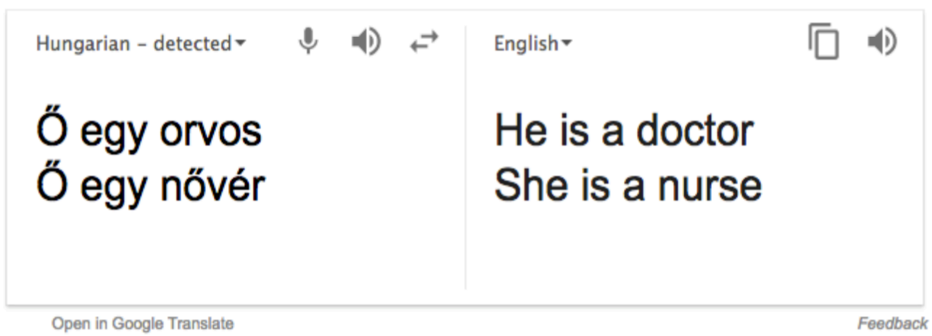
- He: Computer Programmer :: She: Homemaker
  - Equivalent to having a biased dictionary:

nurse ('nərs)
1. A *woman* trained to care for the sick or infirm, especially in a hospital.

computer programmer (kəmˈpjuːtə ˈprəʊgræmə)
1. A *man* who writes programs for the operation of computers, especially as an occupation.

Bias in Downstream Applications: Machine Translation





Metrics to Quantify Gender bias in WE

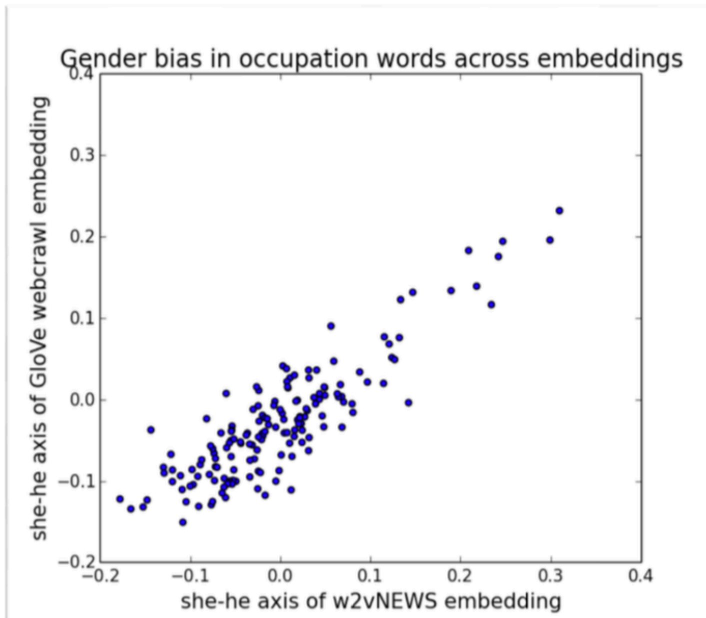
- Metric 1: Occupations
  - 327 gender neutral occupations. Project on to *she—he* direction

Crowdworkers rate each occup. for gender stereotype

$$\mathrm{Corr}(\mathrm{projection}_{she-he}, \mathrm{crowd\ rating}) = 0.51$$

Consistency of embedding stereotype

GloVe trained on web crawl

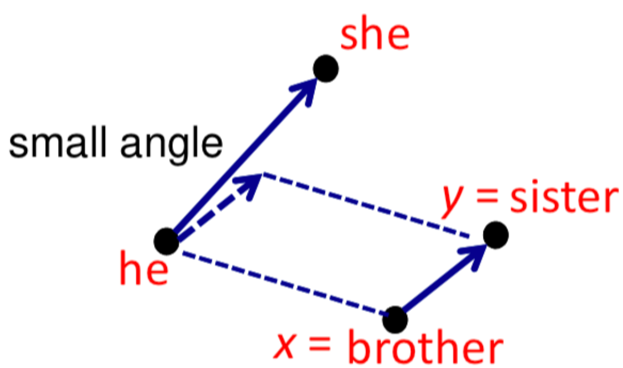


word2vec trained on Google news

Each dot is an occupation; Spearman = 0.8

Metrics to Quantify Gender bias in WE

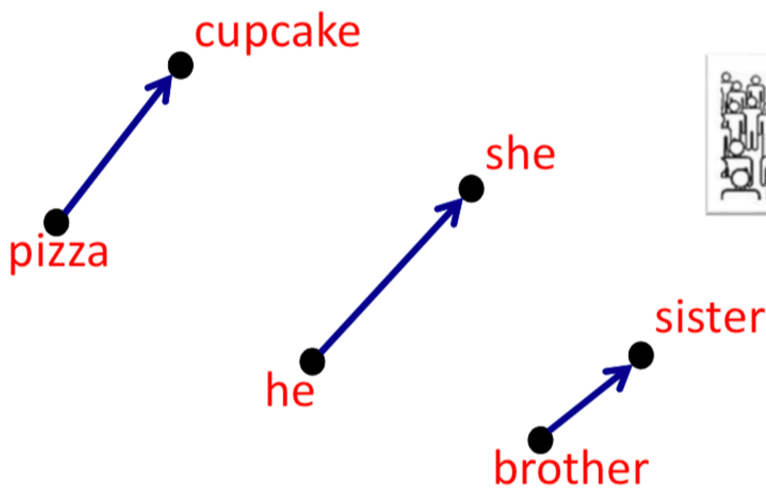
- Metric 2: Analogies
  - Automatically generate he : *x* :: she : *y* analogies.



$$\min \cos(\text{he} - \text{she}, x - y) \text{ such that } ||x - y||_2 < \delta$$

Metrics to Quantify Gender bias in WE

- Metric 2: Analogies
  - Automatically generate he : *x* :: she : *y* analogies.

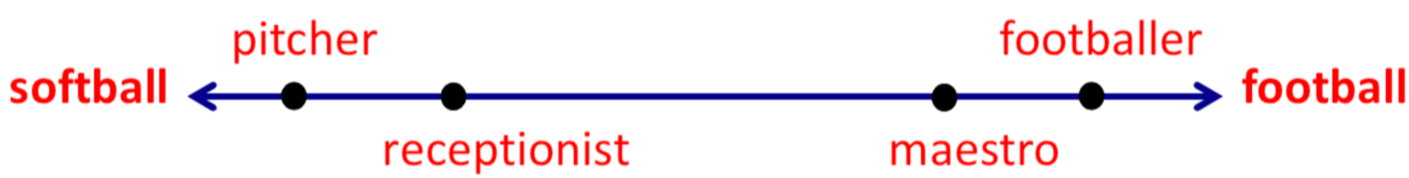


29/150 analogies rated as gender stereotypic by majority of crowdworkers

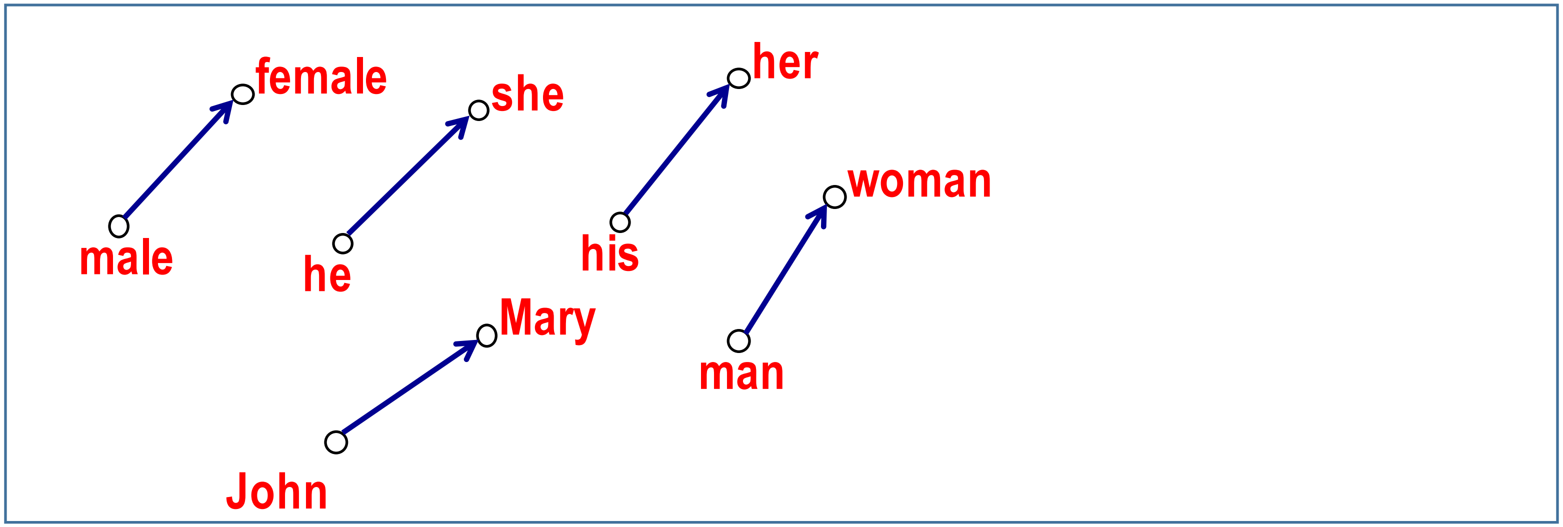
$$\min \cos(\text{he} - \text{she}, x - y) \text{ such that } ||x - y||_2 < \delta$$

Metrics to Quantify Gender bias in WE

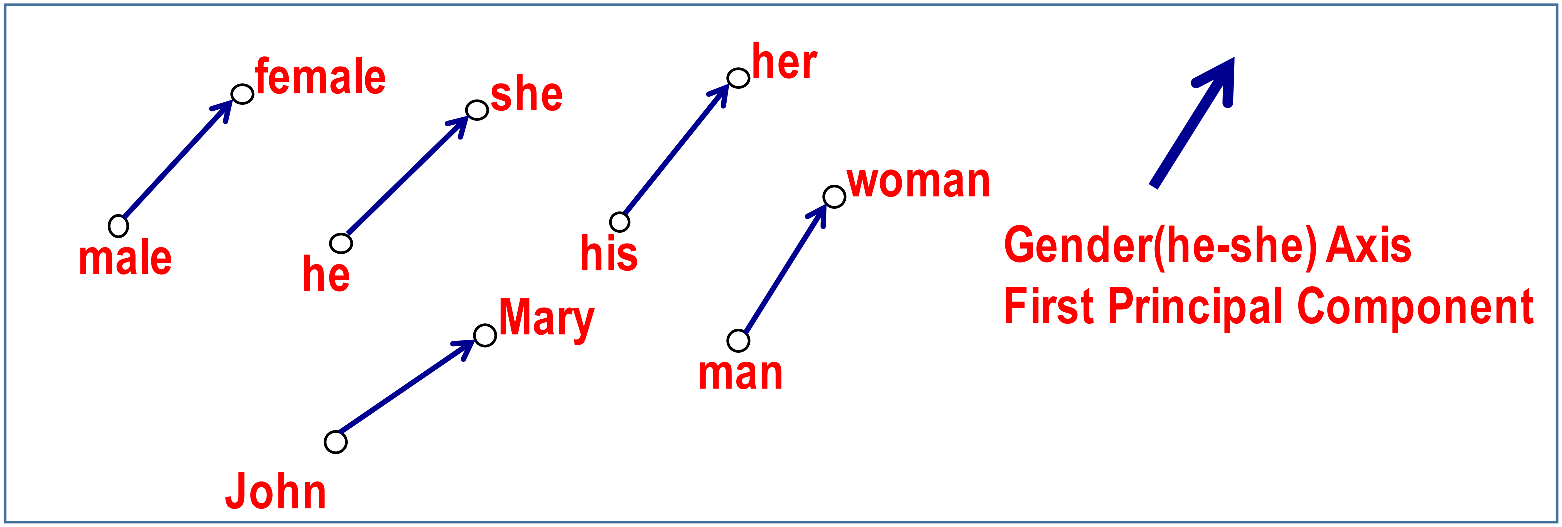
- Metric 3: Indirect Bias
  - Gender stereotype could affect the geometry between words that should be gender-neutral.
  - Project occupations onto softball—football axis.



The Geometry of Gender

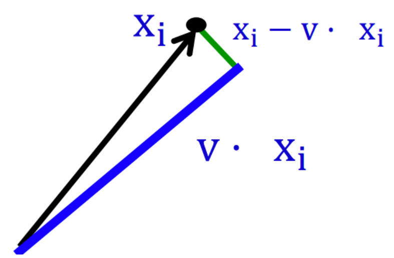


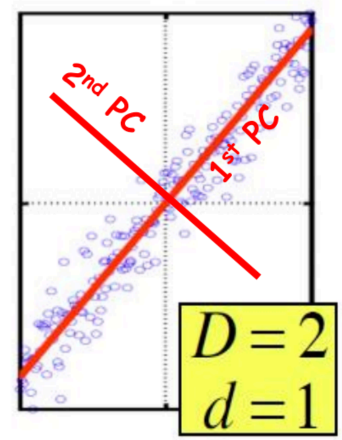
The Geometry of Gender



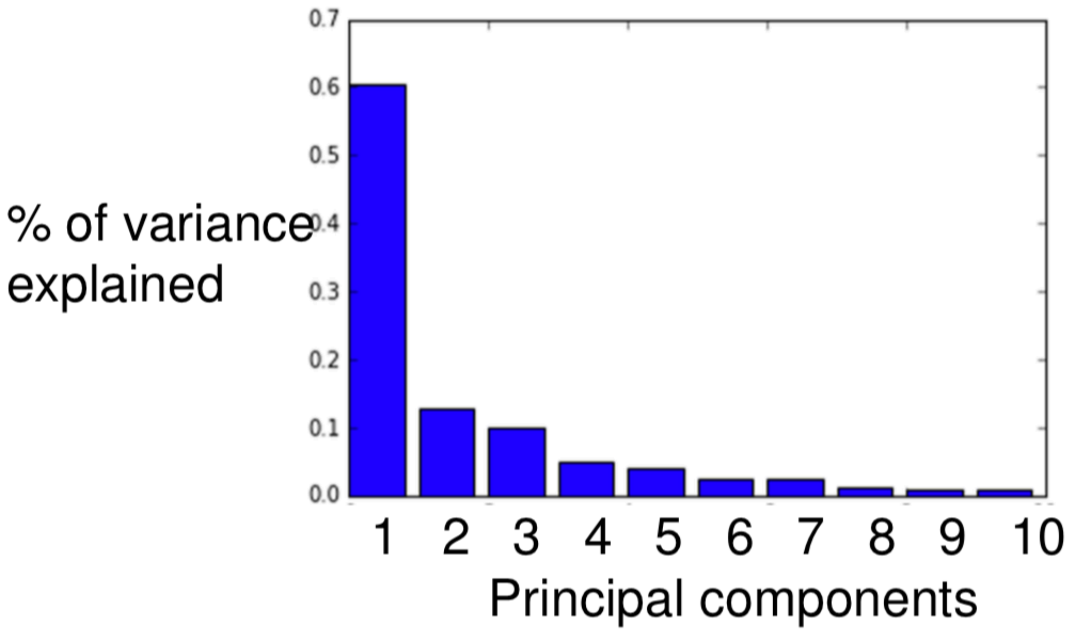
Principal Component Analysis

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





Geometry of Gender

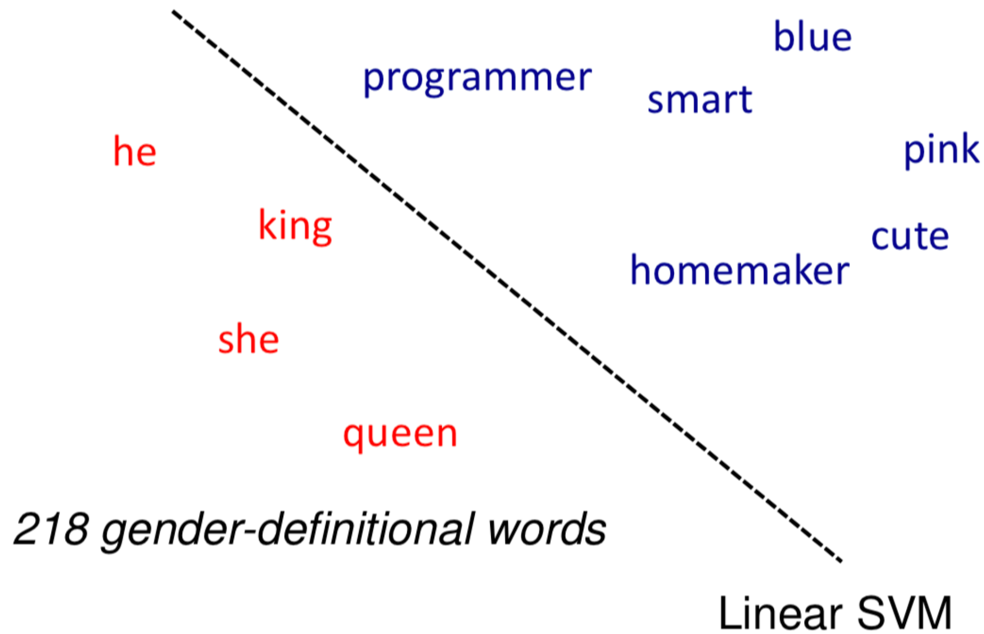


The top PC seems to capture the gender subspace *B*.

Debiasing Algorithm (Hard-debiased)

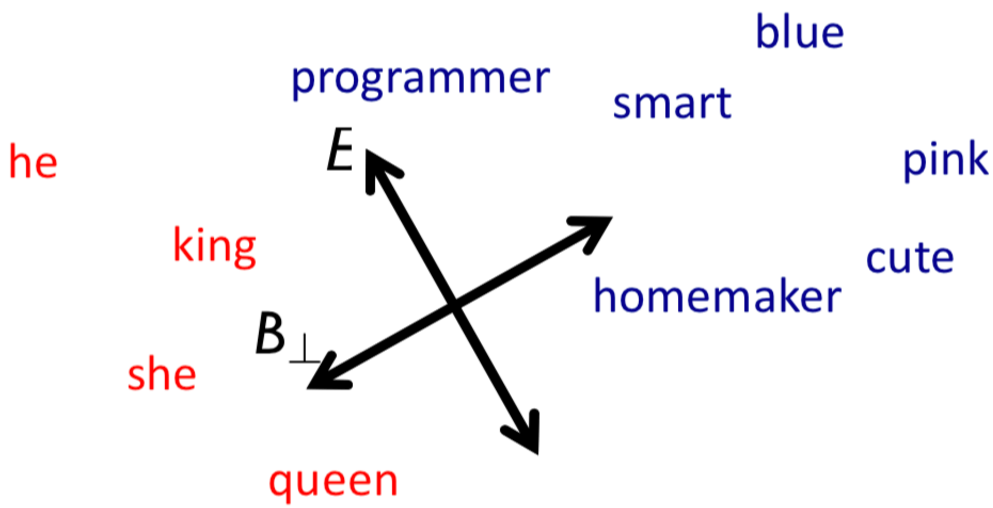
- Identify words that are gender neutral N and gender-definitional S
- Project away the gender subspace from the gender-neutral words
  - w := w- w.B B is the gender subspace
- Normalize vectors

Identify gender-definitional words



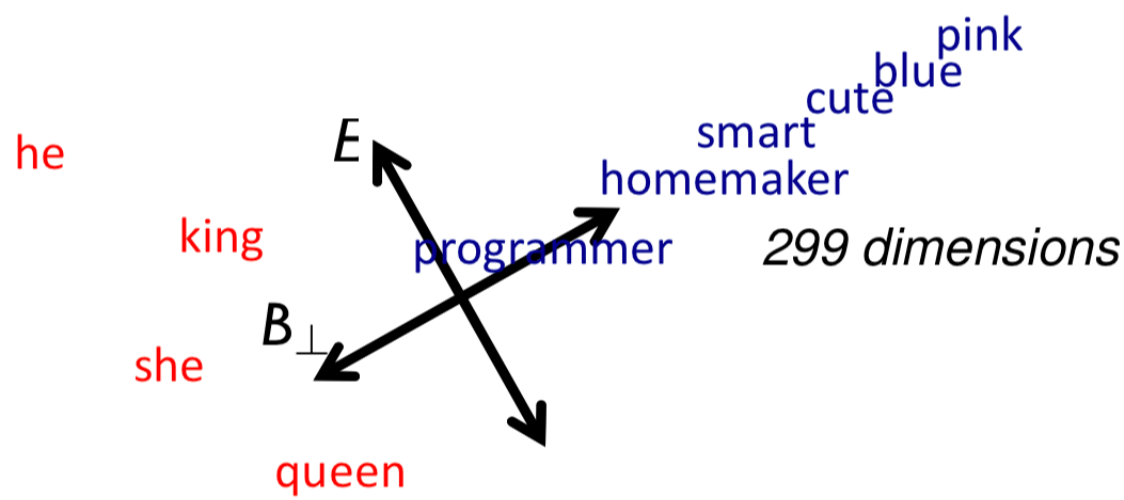
Projecting away gender component

Projecting away gender component



Projecting away gender component

“hard debiasing”



Advanced debiasing (soft debiasing)

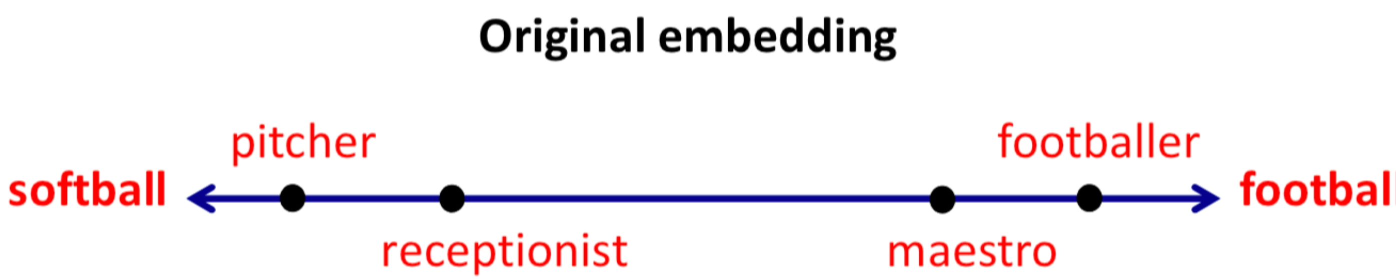
- Find a linear transformation T of the gender- neutral words to reduce the gender component while not moving the words too much.

$W =$ matrix of all word vectors.

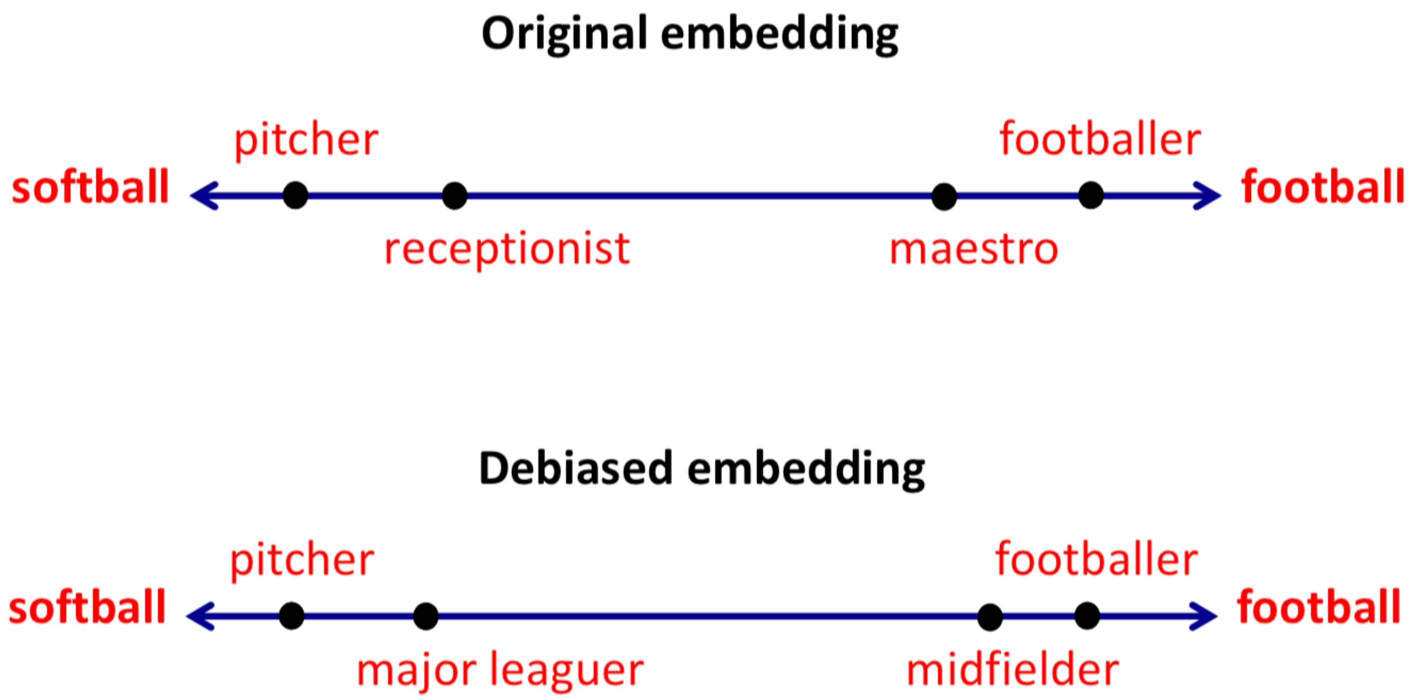
$N =$ matrix of neutral word vectors.

$$\min_{T} \underbrace{||(TW)^T(TW) - W^TW||_F^2}_{\text{don't move too much}} + \lambda \underbrace{||(TN)^T(TB)||_F^2}_{\text{minimize gender component}}$$

Debiasing results: indirect bias

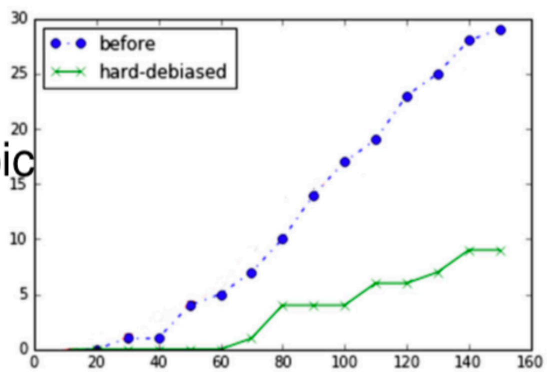


Debiasing results: indirect bias



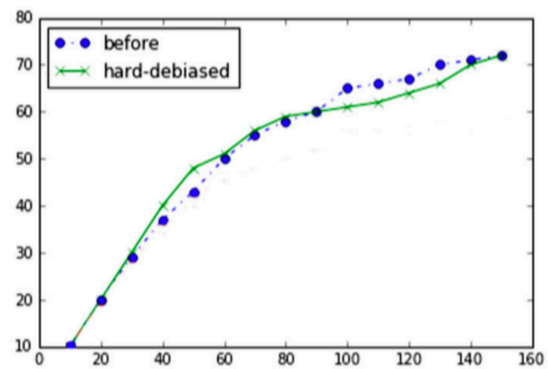
Debiasing result analogies

stereotypic analogies



analogies generated

appropriate analogies



analogies generated

Debiasing result: Appropriate Analogies

	RG	WS	analogy
Before	62.3	54.5	57.0
Hard-debiased	62.4	54.1	57.0

- He:King :: She:Queen
- He:Doctor::She:Doctor

Natural Questions

- 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?**
- To be answered in a later lecture

Summary

- Geometry of word embedding captures bias
  - Who’s responsible: data, algorithm or user?
- Can effective debias algorithms for sensitive applications

Thanks!

- References
  - *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings.* NIPS'16