CS 6120/CS4120: Natural Language Processing
Bias in NLP

Instructor: Prof. Lu Wang
Guest Lecture by Lisa Fan
College of Computer and Information Science
Northeastern University

[Slides borrowed from Adam Lopez, University of Edinburgh
https://www.inf.ed.ac.uk/teaching/courses/nlu/assets/slides/2018/l12.pdf]

Overview

- The social impact of NLP
- Word embeddings contain human-like biases
- Debiasing word embeddings

WORD	d1	d2	d3	d4	d5	...	d50
summer	0.12	0.21	0.07	0.25	0.33	...	0.51
spring	0.19	0.57	0.99	0.30	0.02	...	0.73
fall	0.53	0.77	0.43	0.20	0.29	...	0.85
light	0.00	0.68	0.84	0.45	0.11	...	0.03
clear	0.27	0.50	0.21	0.56	0.25	...	0.32
blizzard	0.15	0.05	0.64	0.17	0.99	...	0.23
...	...	...	...	...	...	...	...

Readings

- Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. **"Semantics derived automatically from language corpora contain human-like biases."** Science 356.6334 (2017): 183-186.
- Bolukbasi, Tolga, et al. **"Man is to computer programmer as woman is to homemaker? debiasing word embeddings."** Advances in Neural Information Processing Systems. 2016.

The social impact of NLP

Technology that impacts lives requires ethical discussion

Technology that impacts lives requires ethical discussion

- Modern NLP originated in laboratory experiments with machines learning methods on linguistically annotated public text
- But modern NLP has escaped the lab, and outcome of an NLP experiment can have a direct effect on people's lives, e.g.
  - Facebook's "emotional contagion" experiment
  - NLP used to recommend products, services, jobs...

Who is affected by an NLP experiment?

- Both consciously and unconsciously, people use language to signal group membership
- Language may convey information about the author and situation
- Predicting author demographics can affect model performance, and can be used to target users
- All of these properties suggest that the author may be traceable from their data

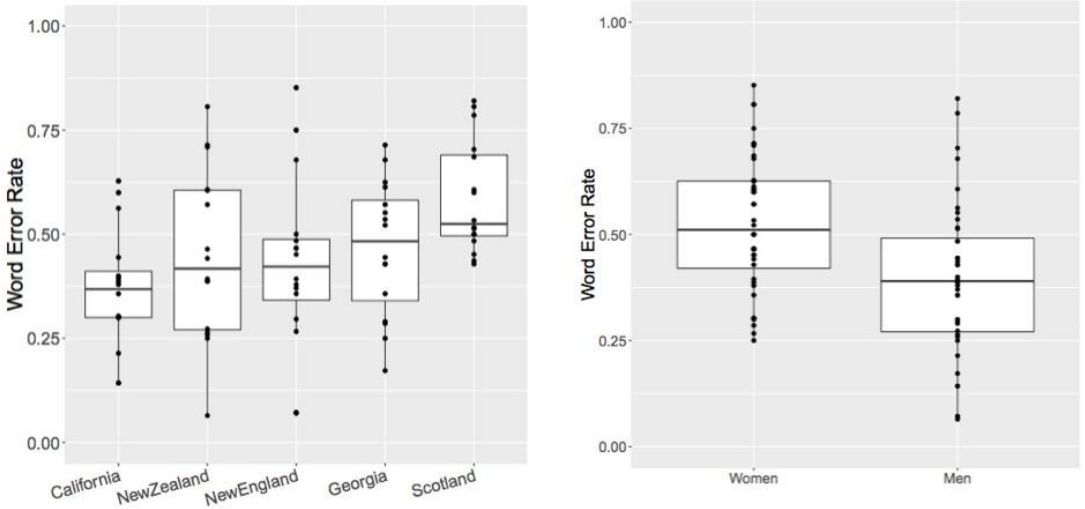
Demographic bias commonly occurs in NLP

- Any dataset carries **demographic bias**: latent information about the demographics of the people that produced it
- Result: **exclusion** of people from other demographics
- E.g. state-of-the-art NLP models are significantly worse for younger people and ethnic minorities
- E.g. speech technology works better for white men from California

Example: The accent challenge

- Details: Rachael Tatman, Gender and Dialect Bias in YouTube's Automatic Captions (2017)
- Youtubers read these words in their native accent: Aunt, Envelope, Route, Theater, Caught, Salmon, Caramel, Fire, Coupon, Tumblr, Pecan, Both, Again, Probably, GPOY, Lawyer, Water, Mayonnaise, Pajamas, Iron, Naturally, Aluminium, GIF, New Orleans, Crackerjack, Doorknob, Alabama
- Compare the read words with youtube's automatic captioning for eight men and eight women across several dialects

The accent challenge reveals differences in access



Topic overexposure creates biases that can lead to discrimination and reinforcement of existing biases. E.g. NLP focused on English may be self-reinforcing

Which is the most populous metropolitan area?

- Lagos
- London
- Paris
- Tianjin

Which is the most populous metropolitan area?

- **Lagos (largest)**
- London
- Paris
- Tianjin

People estimate the sizes of cities they recognize to be larger than the size of cities they don’t know

The **availability heuristic**: the more knowledge people have about a specific topic, the more important they think it must be

Dual-use problems

- Even if we intend no harm in experiments, they can still have unintended consequences that negatively affect people
  - Text classification and IR can help identify information of interest, but also aid censors
  - NLP can be used to detect fake reviews and news, but also to generate them
  - Advanced grammar analysis can improve search and educational NLP, but also reinforce prescriptive linguistic norms
  - Stylometric analysis can help discover provenance of historical documents, but also unmask anonymous political dissenters
- These types of problems are difficult to solve, but important to think about, acknowledge, and discuss

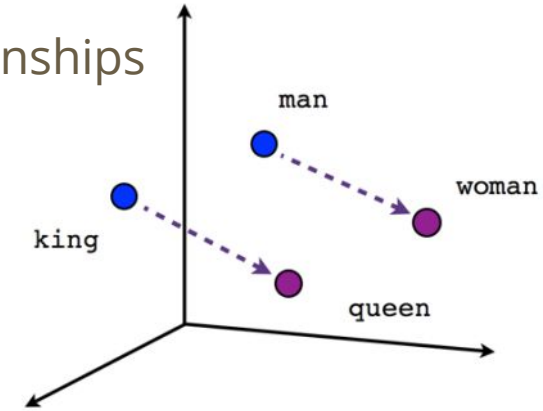
Bad faith actors

- Users may deliberately try to bias your system
    - Trolls took advantage of Microsoft's AI Twitterbot and within 24 hours taught the bot to tweet racist, misogynistic, anti-semitic messages

Word embeddings contain human-like biases

Human language reflects human culture and meaning

- Idea underlying lexical semantics, and word embedding methods like word2vec or neural LMs:
  - *"You shall know a word by the company it keeps."* -- J.R. Firth (1957)
- Example: word2vec learns semantic/syntactic relationships
  - `king - man + woman = queen`
  - `bananas - banana + apple = apples`
  - `walking - walk + swam = swimming`

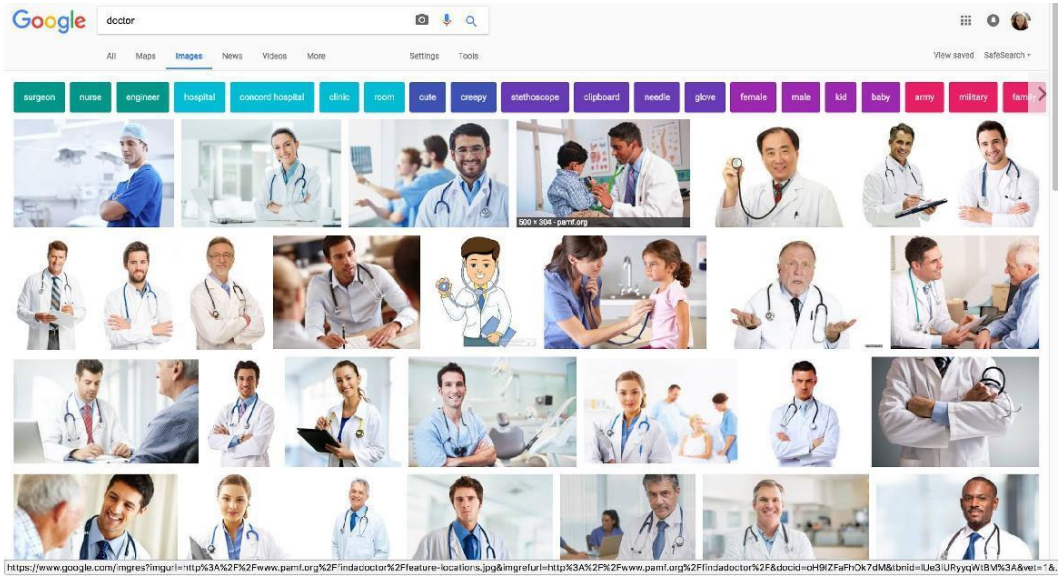


Human language reflects human culture and meaning

- Example: word2vec learns semantic/syntactic relationships
  - `king - man + woman = queen`
  - `bananas - banana + apple = apples`
- But what if your words also keep company with unsavory stereotypes and biases?
  - `doctor - man + woman = nurse`
  - `computer programmer - man + woman = homemaker`

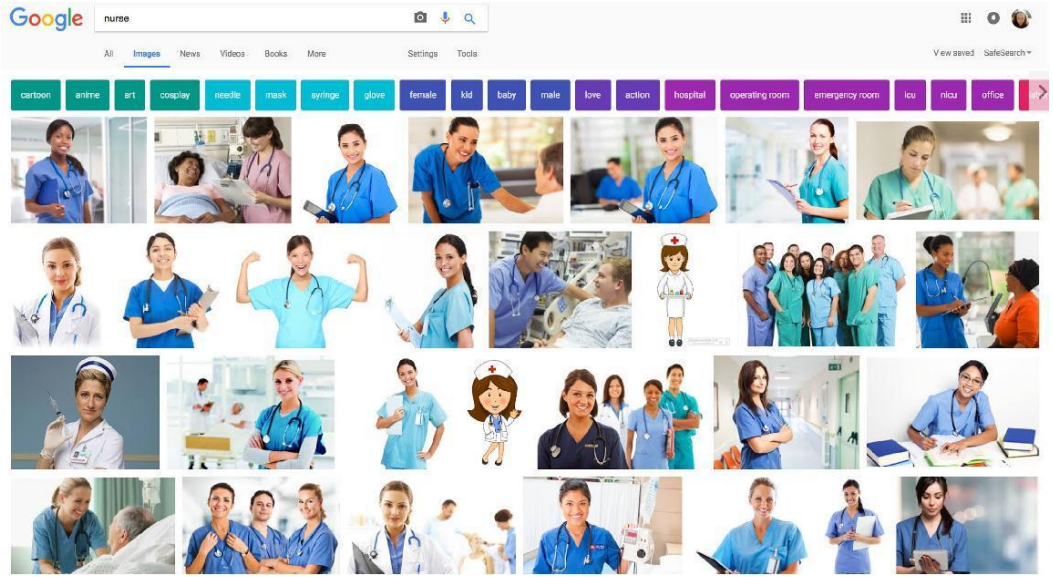
Human language reflects human culture and meaning

- June 2017: image search query for “**Doctor**”



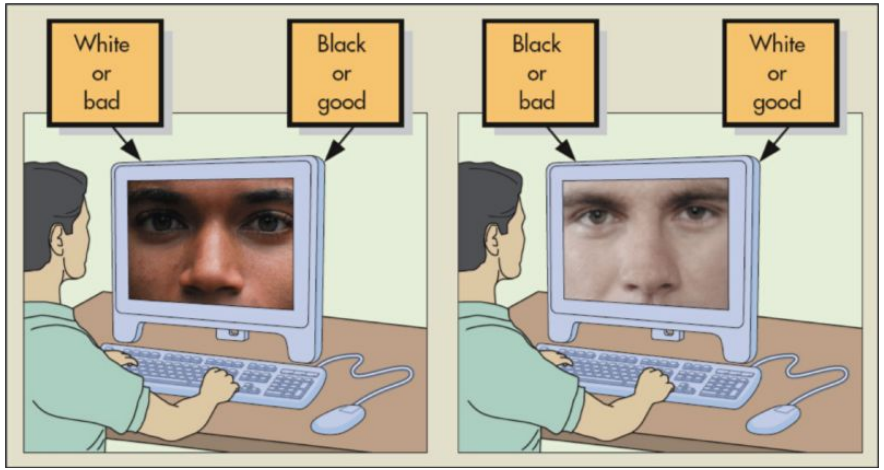
Human language reflects human culture and meaning

- June 2017: image search query for “**Nurse**”



Measuring bias using implicit association tests

- Measure association of groups to stereotype words
- Strong association between a group and a stereotype results in faster reaction times



Implicit Association Test (IAT)

A
GOOD

B
BAD

Love

Implicit Association Test (IAT)

A
GOOD

B
BAD

Hatred

Implicit Association Test (IAT)

A

African American

B

European American

Implicit Association Test (IAT)

A

African American

B

European American

Implicit Association Test (IAT)

A

African American

or

GOOD

Spectacular

B

European American

or

BAD

Implicit Association Test (IAT)

A

African American

or

GOOD

B

European American

or

BAD

Implicit Association Test (IAT)

A

African American

or

BAD

B

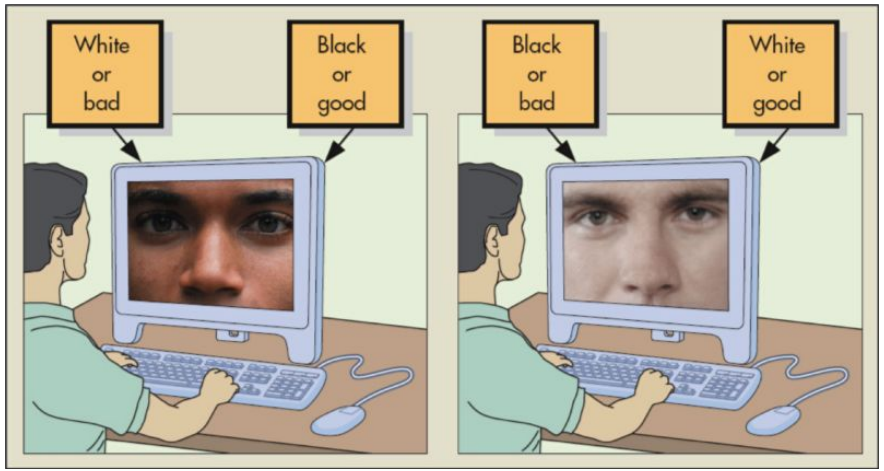
European American

or

GOOD

Measuring bias using implicit association tests

- Measure association of groups to stereotype words
- Strong association between a group and a stereotype results in faster reaction times



- How do we design an IAT for word embeddings?

Designing an IAT for word embeddings

1. Compute similarity of `group1` and `stereotype1` word embeddings. Cosine similarity is used to measure association (in place of reaction time)
2. Compute similarity of `group1` and `stereotype2` word embeddings
3. Null hypothesis: if `group1` is not more strongly associated to one of the stereotypes, there will be no difference in the means
4. Effect size measured using Cohen's d
    - `(M1 - M2) / Pooled SD`
5. Repeat for `group2`

Example:
`group1`: male names
`group2`: female names
`stereotype1`: pleasant
`stereotype2`: unpleasant

Inoffensive associations have strong effects

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

Result: flowers associate with pleasant, insects associate with unpleasant ($p < 10^{-7}$)

Inoffensive associations have strong effects

Instruments	bagpipe, cello, guitar, lute, trombone, banjo, clarinet, harmonica, mandolin, trumpet, bassoon, drum, harp, oboe, tuba, bell, fiddle, harpsichord, piano, viola, bongo, flute, horn, saxophone, violin
Weapons	arrow, club, gun, missile, spear, axe, dagger, harpoon, pistol, sword, blade, dynamite, hatchet, rifle, tank, bomb, firearm, knife, shotgun, teargas, cannon, grenade, mace, slingshot, whip
Pleasant	*As in previous experiment*
Unpleasant	*As in previous experiment*

Result: instruments associate with pleasant, weapons associate with unpleasant ($p < 10^{-7}$)

Names associate with cultural stereotypes

European American names	Adam, Harry, Josh, Roger, Alan, Frank, Justin, Ryan, Andrea, Jack, Matthew, Stephen, Greg, Paul, Jonathan, Peter, Amanda, Courtney, Heather, Melanie, Katie, Betsy, Kristin, Nancy, Stephanie, Ellen, Lauren, Colleen, Emily, Megan, Rachel
African American names	Alonzo, Jamel, Theo, Alphonse, Jerome, Leroy, Torrance, Darnell, Lamar, Lionel, Tyree, Deion, Lamont, Malik, Terrence, Tyrone, Lavon, Marcellus, Wardell, Nichelle, Shereen, Ebony, Latisha, Shaniqua, Jasmine, Tanisha, Tia, Lakisha, Latoya, Yolanda, Malika, Yvette
Pleasant	*As in previous experiment*
Unpleasant	*As in previous experiment*

Result: European American names associate with pleasant, African American names associate with unpleasant ($p < 10^{-8}$)

Names associate with gendered professions

Men's names	John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill
Women's names	Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna
Career	executive, management, professional, corporation, salary, office, business, career
Family	home, parents, children, family, cousins, marriage, wedding, relatives

Result: Men's names associate with career, women's names associate with family ($p < 10^{-3}$)

Other biases appear in the data

- Men's names associate with maths, women's names with arts ($p < .018$)
- Men's names associate with science, women's names with arts ($p < .10^{-2}$)
- Young people's names associate with pleasant, old people's names with unpleasant ($p < .10^{-2}$)

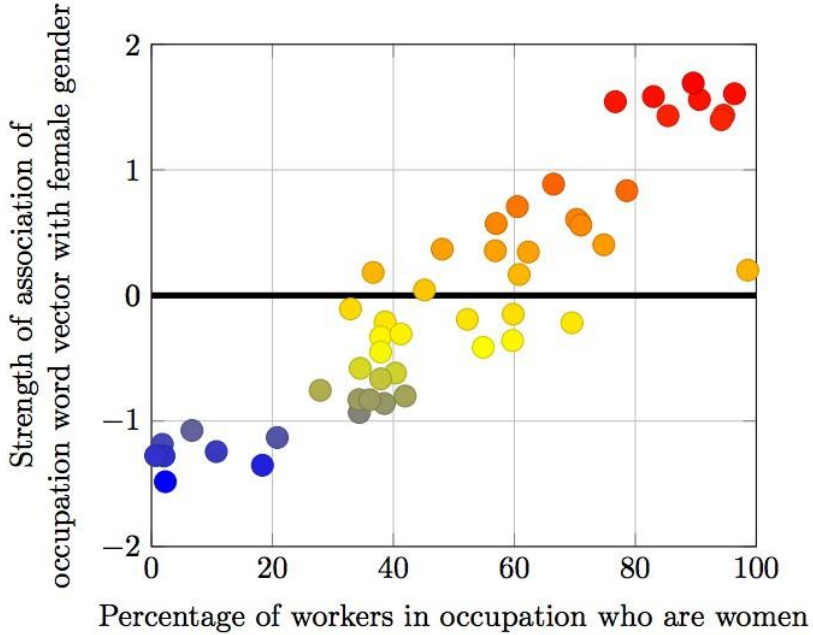
Experimental details and caveats

- Used GloVe (similar to word2vec) trained on Common Crawl (a large-scale crawl of the web)
- Removed names that did not appear with high frequency in data
- Removed names that were least “name-like” (e.g. *Will*) algorithmically
- Each concept is represented using a small set of words, designed for previous experiments in the psychology literature

word2vec analogies from Google exhibit similar biases

Most similar to “he”	maestro, skipper, protege, philosopher, captain, architect, financier, warrior, broadcaster, magician
Most similar to “she”	homemaker, nurse, receptionist, librarian, socialite, hairdresser, nanny, bookkeeper, stylist, housekeeper
Gender “she-he” analogies	*Definitional* queen-king, sister-brother, mother-father, waitress-waiter, convent-monastery
	Stereotypical sewing-carpentry, nurse-surgeon, giggle-chuckle, vocalist-guitarist, diva-superstar, cupcakes-pizzas, housewife-shopkeeper, cosmetics-pharmaceuticals, petite-lanky, charming-affable, lovely-brilliant

Gender biases in data reflect real-world associations



Debiasing word embeddings

Can we remove gender bias from word representations?

- In supervised learning, specific features can be censored from the data by incorporating a term into the learning objective that requires the classifier to be *unable* to discriminate between the censored classes. However, this has many limitations
- In representation-learning systems like word2vec, the classes are not provided *a priori* as features of the data. They are latent in the data

Identifying the "gender subspace"

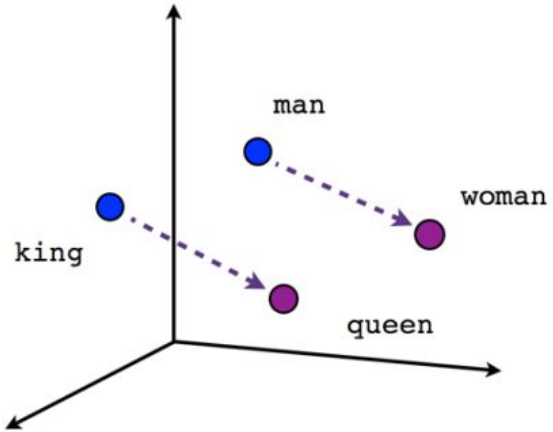
Intuition*:* If analogies reveal a gender dimension, use analogies on specific *seed pairs* to find it.

Classification based on simple test: which element of the pair is the test word closest to in vector space?

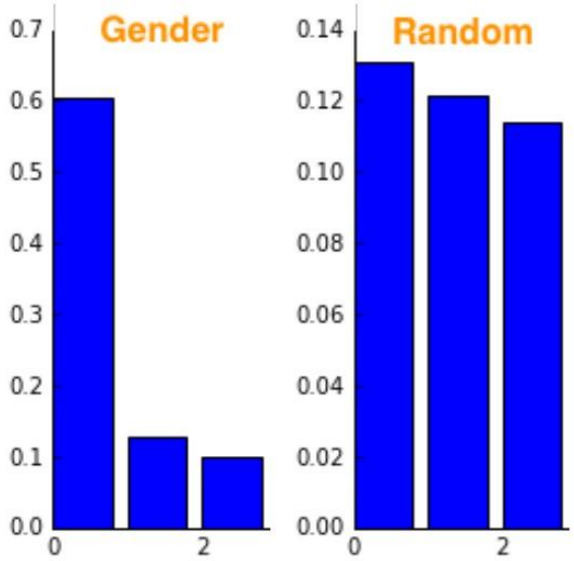
Pair	Classification accuracy on stereotypes*
she-he	89%
her-his	87%
woman-man	83%
Mary-John	87%
herself-himself	89%
daughter-son	91%
mother-father	85%

* based on crowd-sourced words that should be neutral (e.g. "nurse")

A single direction explains most of the variance of seed pairs



Male-Female



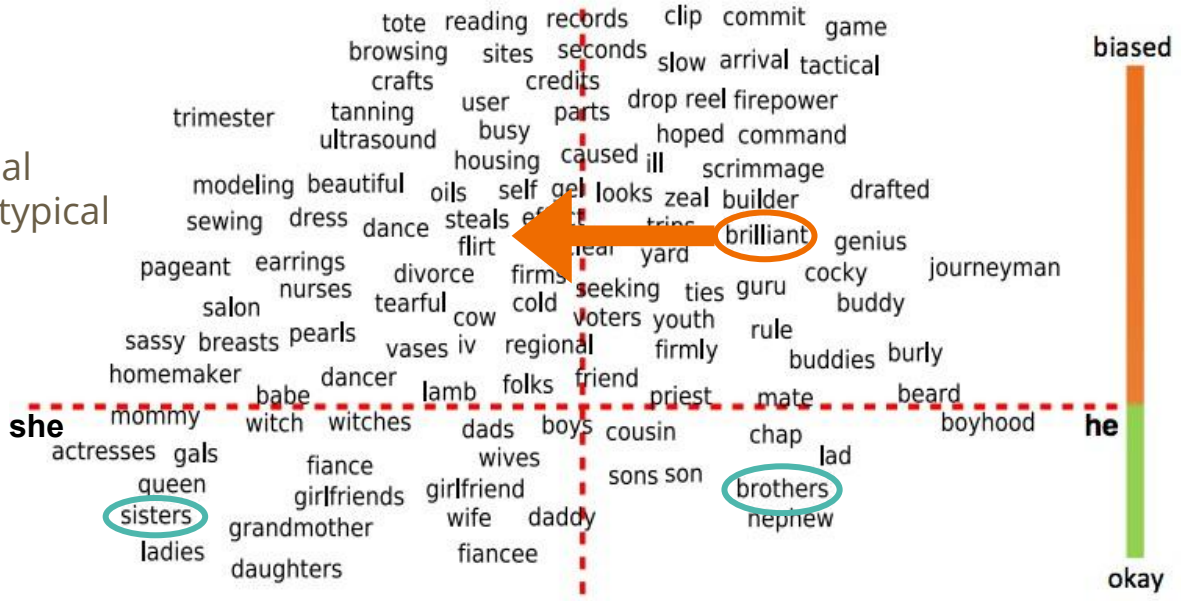
Percentage of variance from PCA components

Gender subspace shows where words exhibit biases

x is projection onto “he-she” subspace. *y* captures neutrality.

Neutralize and equalize embeddings

Blue is definitional
Orange is stereotypical

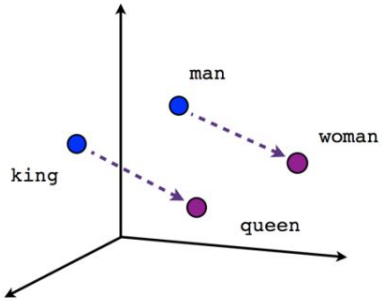


Also possible to trade off between hard neutralization and original embeddings

Debiasing Algorithm

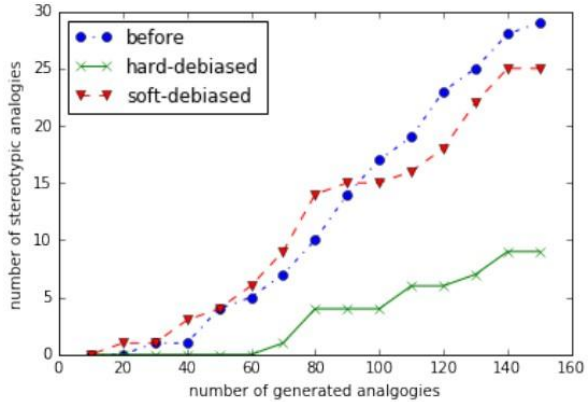
Given a set of gendered pairs (e.g. “mother-father”) and a set of words to neutralize (e.g. “nurse”, “doctor”):

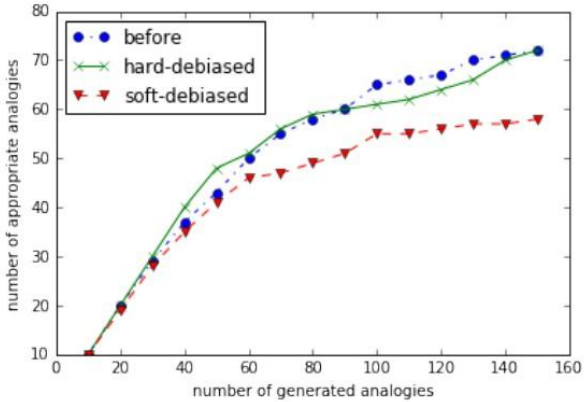
1. Identify the **gender subspace** (“direction” of the embedding that captures the bias)
2. **Neutralize** words to be 0 in the gender subspace
3. **Equalize** words to be equidistant from gendered pairs
4. **Soften** gendered pairs by reducing their difference while maintaining as much of original embedding as possible



Male-Female

Debiasing reduces prevalence of stereotypical analogies





- This is a preliminary result
- How should you choose seed words?
- How should you choose the words to debias?
- Does this actually have the desired effect in downstream applications?

Summary

- NLP is used by millions of people in the real world every day
- NLP developers must be aware of ethical concerns like demographic bias, overgeneralization, topic overexposure, and dual use
- Word embeddings are a basic technology used in many NLP technologies; they are freely available and used by many developers large and small
- Word embeddings empirically exhibit many cultural stereotypes and biases, with strong statistical effects; technology will reflect and *can potentially amplify* these biases
- Substantial ongoing research around the question: how do we design fairer systems?

Closing thought (paraphrasing Herbert Clark)

Language doesn’t have so much to do with words and what they mean.

It has to do with *people* and what *they* mean.