Natural Language Understanding

Bias in NLP

Adam Lopez

April 3, 2018

School of Informatics
University of Edinburgh
alopez@inf.ed.ac.uk

Reading: Caliskan et al. (2017), Bolukbasi et al. (2016)
Background: Hovy and Spruit (2016).

The social impact of NLP

Technology that impacts lives requires ethical discussion

Modern NLP originated in laboratory experiments with machine learning methods on linguistically annotated public text.

But modern NLP has escaped the lab, and the outcome of an NLP experiment can have a direct effect on people's lives, e.g.

- A sequence-to-sequence RNN implementing an Alexa chatbot responded to "Should I sell my house?" with "Sell sell sell!"
- The same chatbot responded to "Should I kill myself?" with "Yes."
- Facebook's "emotional contagion" experiment.
- NLP used to recommend products, services, jobs...

Also includes wider ethical concerns about ML/ data science, e.g. privacy concerns. We'll focus on NLP here.

Who is affected by an NLP experiment?

If your language data is newspaper articles or novels... perhaps the journalist or author is unaffected by experiments.

What if the language you study is from, e.g. social media?

- Both consciously and unconsciously, people use language to signal group membership.
- Language may convey information about the author and situation.
- Language can predict author demographics, which affect model performance, and can be used to target users.
- Language is political, and an instrument of power.

All of these properties suggest that the authors 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. speech technology works better for white men from California.

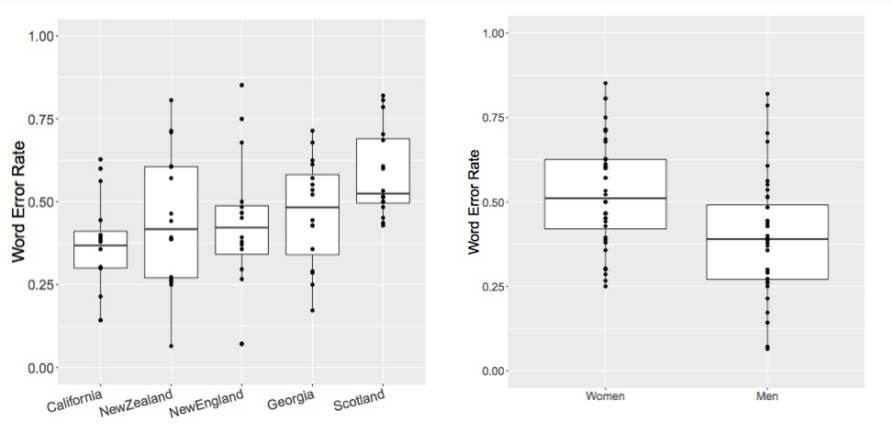
E.g. State-of-the-art NLP models are significantly worse for younger people and ethnic minorities.

Example: The accent challenge

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



Details: Rachael Tatman, Gender and Dialect Bias in YouTube's Automatic Captions (2017)

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.

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

Word embeddings contain human-like biases

Dual-use problems

Even if we intend no harm in experiments, they can still have unintended consequences that negatively affect people.

- 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.
- Text classification and IR can help identify information of interest, but also aid censors.
- NLP can be used to generate fake reviews and news, and also to generate them.

These types of problems are difficult to solve, but important to think about, acknowledge and discuss.

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.*
>
> — Firth (1957)

Example: word2vec learns semantic/ syntactic relationships

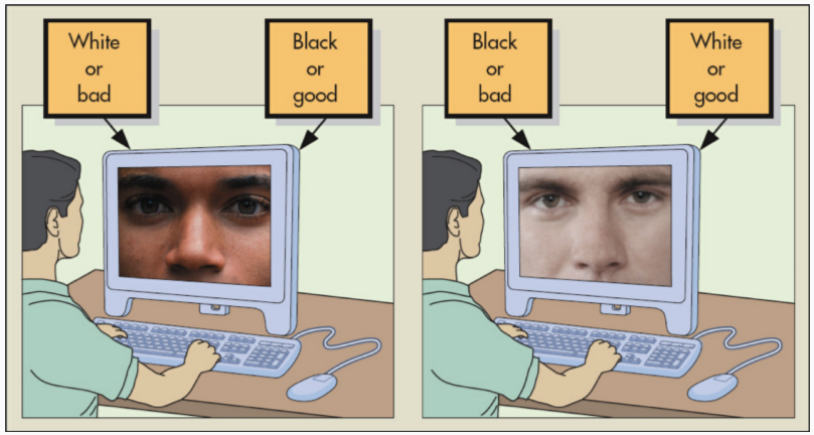
- king - man + woman = queen
- bananas - banana + apple = apples

But what if your words also keep company with unsavoury stereotypes and biases?

- doctor - man + woman = nurse
- computer programmer - man + woman = homemaker

We can measure bias using implicit association tests

Measures 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 use to measure association (in place of reaction time).
2. Compute similarity of group1 and stereotype 2 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.
5. Repeat for group 2.

Experimental details and caveats

- Uses 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.

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 *Similar to previous experiment.*

Unpleasant *Similar to 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.

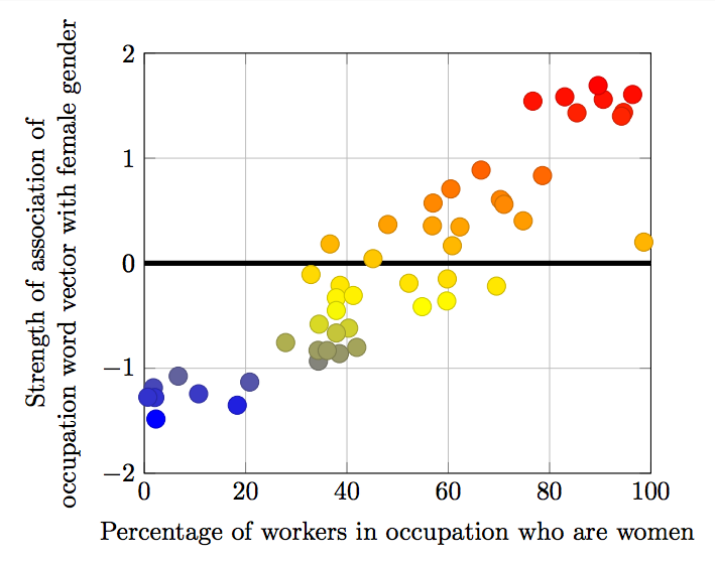
Family home, parents, children, family, cousins, marriage, wedding, relatives.

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 ($p < .10^{-2}$).

Gender biases in data reflect real-world associations



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.

Debiasing word embeddings

Can we remove 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.

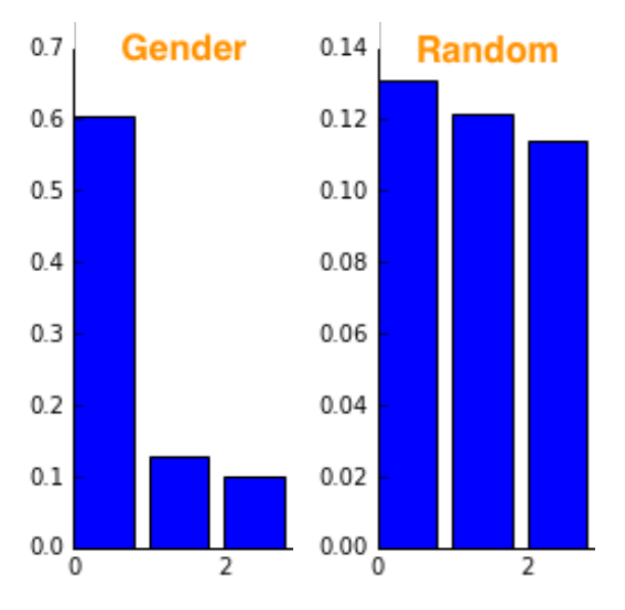
Identifying the "gender subspace"

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

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%

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

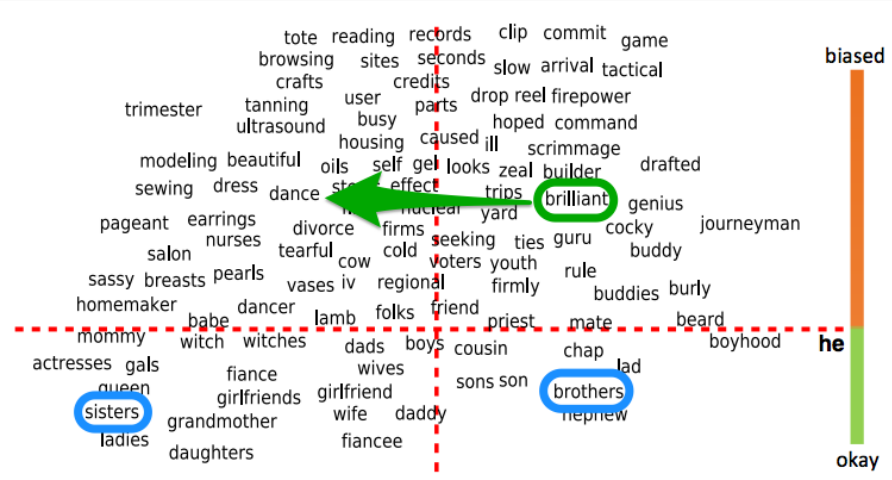
A single direction explains most of the variance of seed pairs



Gender subspace show where words exhibit biases

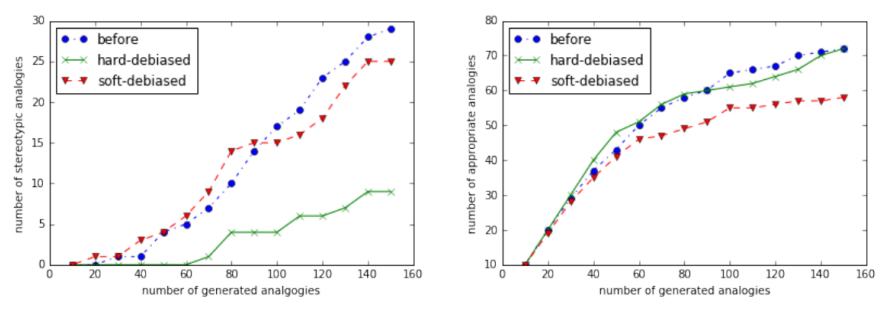
x is projection onto he-she subspace. y captures neutrality.

Neutralize and equalize embeddings



Also possible to trade off between hard neutralization and original embeddings.

Debiasing reduces prevalance 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 affect 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.