Ethics

CS4248 Natural Language Processing

Week 13

Min-Yen KAN

13

Recap of Week 12

Contextual Word Embeddings

Machine Translation

Question Answering II

Week 13 Agenda

NLP Ethics

Mitigating Word Embedding Bias

Revision (Separate Deck)

NLP Ethics

How I learned to stop worrying and love natural language processing

Why does a discussion about ethics need to be a part of NLP?

The decisions we make about our methods — training data, algorithm, evaluation — are often tied up with its use and impact in the world.

Slide Credits: David Bamman (UC Berkeley)

The common misconception is that language has to do with words and what they mean.

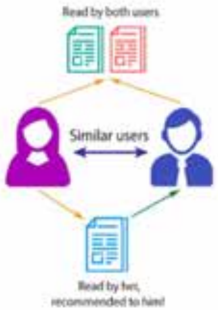
It doesn't.

It has to do with people and what they mean.

Clark & Schober, 1982

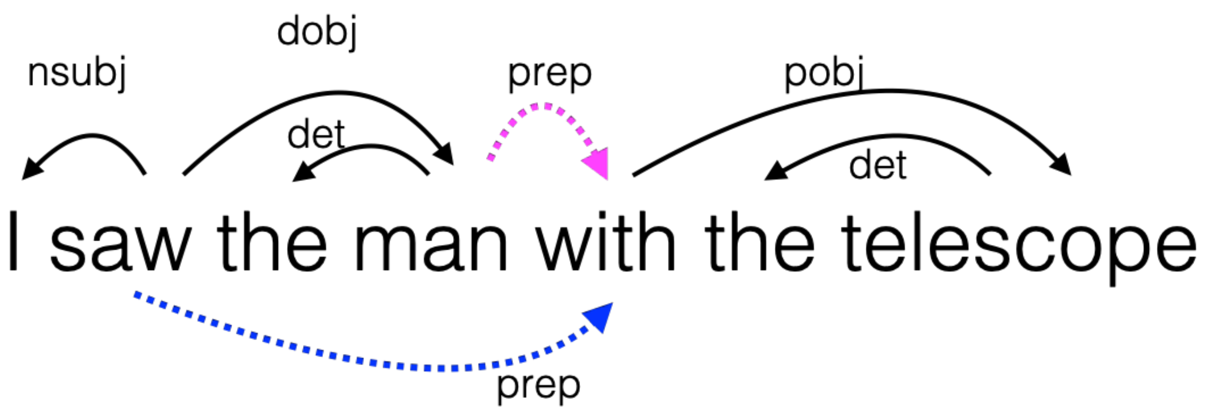
Slide Credits: Diyi Yang (Georgia Tech)

Language, People and the Web



Slide Credits: Diyi Yang (Georgia Tech)

Scope



NLP often operates on text divorced from the context in which it is uttered.

It's now being used more and more to reason about human behavior.

Slide Credit: David Bamman (UC Berkeley)

Learning to Assess Systems Adversarially

- Who could benefit from such a technology?
- Who can be harmed by such a technology?

Representativeness of training data

- Could sharing this data have major effect on people's lives?

- What are confounding variables and corner cases to control for?
- Does the system optimize for the "right" objective?
- Could prediction errors have major effect on people's lives?

Slide Credits: Diyi Yang (Georgia Tech)

Privacy Concerns

- Demographic factors prediction (gender, age, etc)
- Sexual orientation prediction

Dual Use NLP Applications

- E.g., Persuasive language generation
- Socially Beneficial Applications
  - Hate speech detection
  - Monitoring disease outbreaks
  - Psychological monitoring/counseling

  \+ many more

Bias and Fairness Concerns

- Is my NLP model capturing social stereotypes?
- Are my classifiers' predictions fair?

Slide Credits: Diyi Yang (Georgia Tech)

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012

Slide Credit: David Bamman (UC Berkeley)

Facebook fiasco: was Cornell's study of 'emotional contagion' an ethics breach?

A covert experiment to influence the emotions of more than 600,000 people. A major scientific journal behaving like a rabbit in the headlights. A university in a PR tailspin

Slide Credit: David Bamman (UC Berkeley)

Dual Use and Adversarial NLP

Authorship attribution (author of *Federalist Papers* vs. author of ransom note vs. author of political dissent)

Fake review detection vs. fake review generation

Censorship evasion vs. enabling more robust censorship

Slide Credit: David Bamman (UC Berkeley)

Overgeneralization

Managing and communicating the uncertainty of our predictions

Algorithmic Bias: deferring to an automated response.

"The system said so"

Is a false answer worse than no answer?

Slide Credit: David Bamman (UC Berkeley)

Exclusion

Focus on data from one domain/demographic

State-of-the-art models perform worse for young (Hovy and Søgaard, 2015) and minorities (Blodgett *et al.*, 2016)

	AAE	White-Aligned
langid.py	13.2%	7.6%
Twitter-1	8.4%	5.9%
Twitter-2	24.4%	17.6%

Table 3: Proportion of tweets in AA- and white-aligned corpora classified as non-English by different classifiers. (§4.1)

Language identification

Parser	AA	Wh.	Difference
SyntaxNet	64.0 (2.5)	80.4 (2.2)	16.3 (3.4)
CoreNLP	50.0 (2.7)	71.0 (2.5)	21.0 (3.7)

Dependency Parsing

Slide Credit: David Bamman (UC Berkeley)

Biased NLP Technologies

- Bias in Word Embeddings *(Bolukbasi et al. 2017; Caliskan et al. 2017; Garg et al. 2018)*
- Bias in Language ID *(Blodgett & O'Connor. 2017; Jurgens et al. 2017)*
- Bias in Visual Semantic Role Labeling *(Zhao et al. 2017)*
- Bias in Natural Language Inference *(Rudinger et al. 2017)*
- Bias in Coreference Resolution *(Rudinger et al. 2018; Zhao et al. 2018)*
- Bias in Automated Essay Scoring *(Amorim et al. 2018)*

Slide Credits: Diyi Yang (Georgia Tech)

Science AAAS

Home News Journals Topics Careers

Science | Science Advances | Science Immunology | Science Robotics | Science Signaling | Science Translational Medicine

SHARE

f 0

13

REPORT

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

PRE Peer Reviewed ← see details

Slide Credit: David Bamman (UC Berkeley)

Article | Figures & Data | Info & Metrics | eLetters | PDF

Humans are the “Natural” in NLP

Natural language data and annotations will reflect social/cognitive biases

ML algorithms will replicate biases present in their training data

NLP *is* human subject research! (in a way)

Human subject: a living individual about whom a researcher obtains
(1) data through intervention or interaction with the individual or
(2) identifiable private information.

Mitigating Word Embedding Bias

Slide Credits: Diyi Yang (Georgia Tech)

Language Identification: Solved!

"This paper describes … how even the most simple of these methods using data obtained from the World Wide Web achieve accuracy approaching 100% on a test suite comprised of ten European languages"

…or not?

Slide Credits: Diyi Yang (Georgia Tech)

World Englishes

Slide Credits: Diyi Yang (Georgia Tech)

Bias in Word Embeddings

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "*Semantics derived automatically from language corpora contain human-like biases.*" Science 356, no. 6334 (2017): 183-186.

Slide Credits: Diyi Yang (Georgia Tech)

School of Computing

$$\min cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

	Gender stereotype *she-he* analogies	
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant
	Gender appropriate *she-he* analogies	
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 1: **Left** The most extreme occupations as projected on to the *she*−*he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

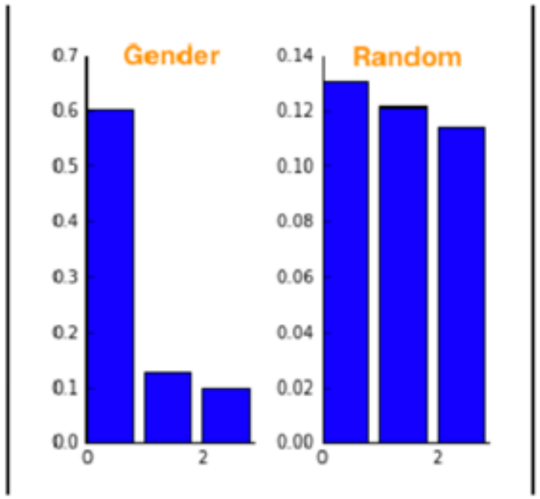
Slide Credits: Diyi Yang (Georgia Tech)

Towards Debiasing

1. Identify gender subspace (direction): B

Bolukbasi et al. (2016) *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*

$\overrightarrow{\text{she}}-\overrightarrow{\text{he}}$
$\overrightarrow{\text{her}}-\overrightarrow{\text{his}}$
$\overrightarrow{\text{woman}}-\overrightarrow{\text{man}}$
$\overrightarrow{\text{Mary}}-\overrightarrow{\text{John}}$
$\overrightarrow{\text{herself}}-\overrightarrow{\text{himself}}$
$\overrightarrow{\text{daughter}}-\overrightarrow{\text{son}}$
$\overrightarrow{\text{mother}}-\overrightarrow{\text{father}}$
$\overrightarrow{\text{gal}}-\overrightarrow{\text{guy}}$
$\overrightarrow{\text{girl}}-\overrightarrow{\text{boy}}$
$\overrightarrow{\text{female}}-\overrightarrow{\text{male}}$

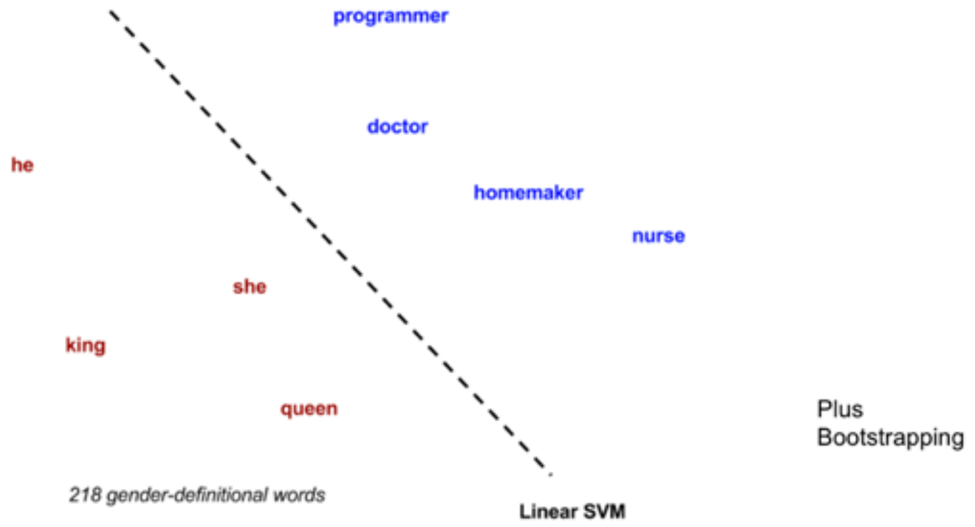


The top PC captures the gender subspace

Slide Credits: Diyi Yang (Georgia Tech)

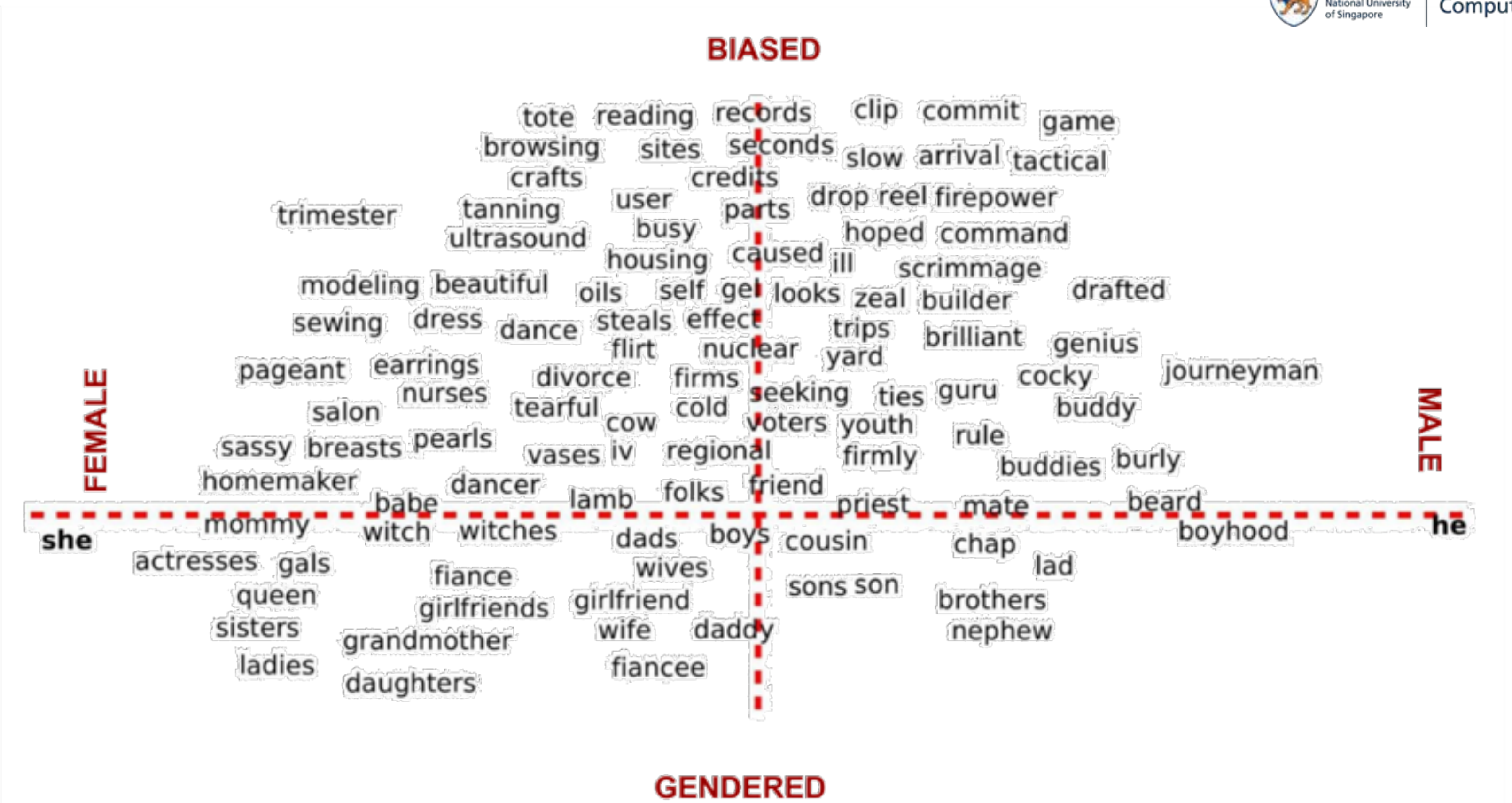
Towards Debiasing

1. Identify gender subspace (direction): B
2. Identify gender-definitional (S) and gender-neutral words (N)



Slide Credits: Diyi Yang (Georgia Tech)

NUS National University of Singapore | School of Computing



Slide Credits: Diyi Yang (Georgia Tech)

Towards Debiasing

1. Identify gender subspace (direction): B
2. Identify gender-definitional (S) and gender-neutral words (N)
3. Apply matrix transformation (T) to the embedding matrix (W) such that:
   - Project away the gender subspace B from the gender-neutral N
   - While not overly changing the embeddings

$$min_T \underbrace{||(TW)^T(TW) - W^TW||_F^2}_{\text{Don't modify embeddings too much}} + \lambda \underbrace{||(TN)^T(TB)||_F^2}_{\text{Minimize gender component}}$$

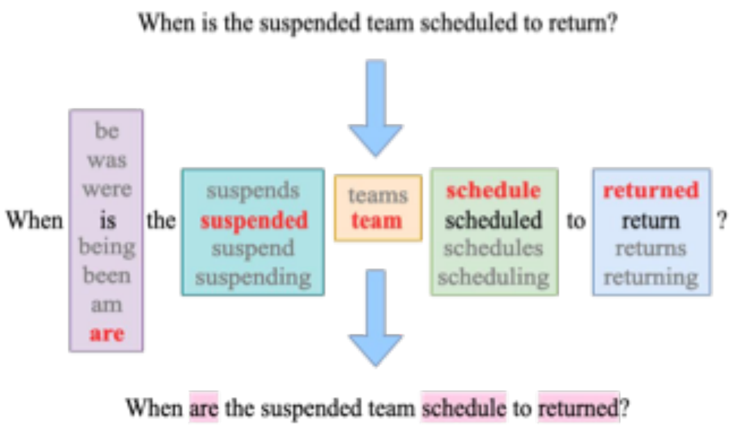
T - the desired debiasing transformation
W - embedding matrix

B - biased space
N - embedding matrix of gender neutral words

Slide Credits: Diyi Yang (Georgia Tech)

Augment the Training Data: Morpheus

Tan et al. (2020) [*It's Morphin' Time! Combating Linguistic Discrimination with Inflectional Perturbations*]



Algorithm 1 Morpheus

Require: Original instance x, Label y, Model f
Ensure: Adversarial example $\hat{x}$

$T \leftarrow \text{TOKENIZE}(x)$
for all $t_i \in T$ **do**
  if $\text{POS}(t_i) \in \{\text{NOUN, VERB, ADJ}\}$ **then**
    $I \leftarrow \text{GETINFLECTIONS}(t_i)$
    $t_i \leftarrow \text{MAXINFLECTED}(I, y, f)$
  end if
end for
$\hat{x} \leftarrow \text{DETOKENIZE}(T)$

Ethics Summary

- Who could benefit from your technology?
- Who can be harmed by your technology?

Representativeness of your data

- Could sharing your data have major effect on people's lives?

- What are confounding variables and corner cases for you to control for?
- Does your system optimize for the "right" objective?
- Could prediction errors of your technology have major effect on people's lives?