Social bias and fairness in NLP

RISE Learning Machines Seminars, 2020-02-20

Olof Mogren, RISE

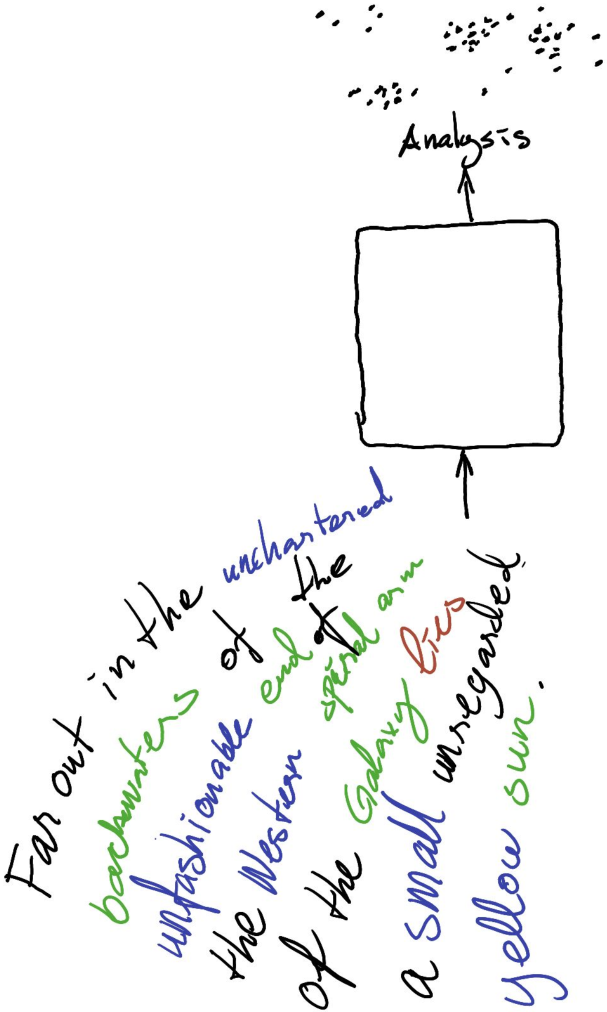
Natural language processing (NLP)

A field of research.

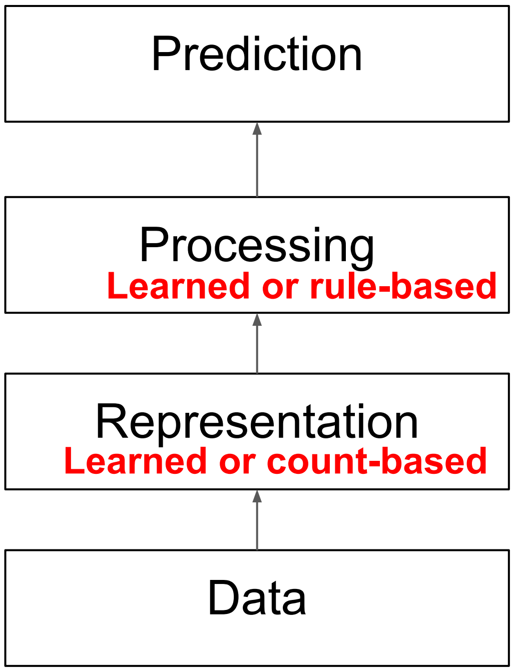
Tasks: classification, translation, summarization, generation, understanding, dialog modelling, etc. (many; diverse)

Data: language: a kind of protocol for inter-human communication; **discrete**

Solutions: many; diverse.

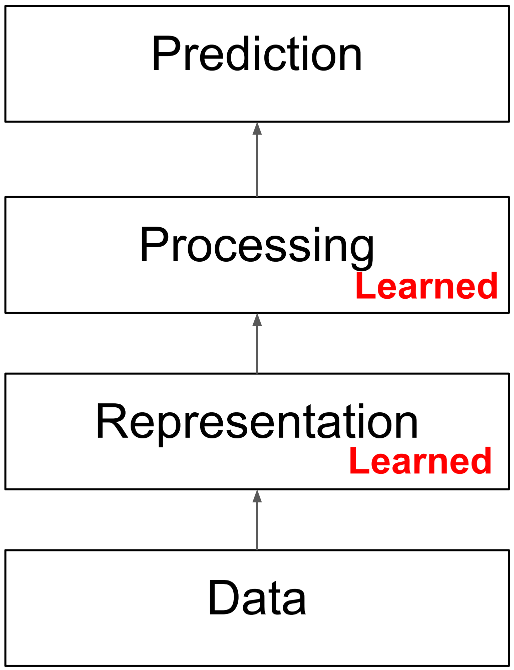


Traditional NLP pipeline



End-to-end NLP pipeline

(deep learning)

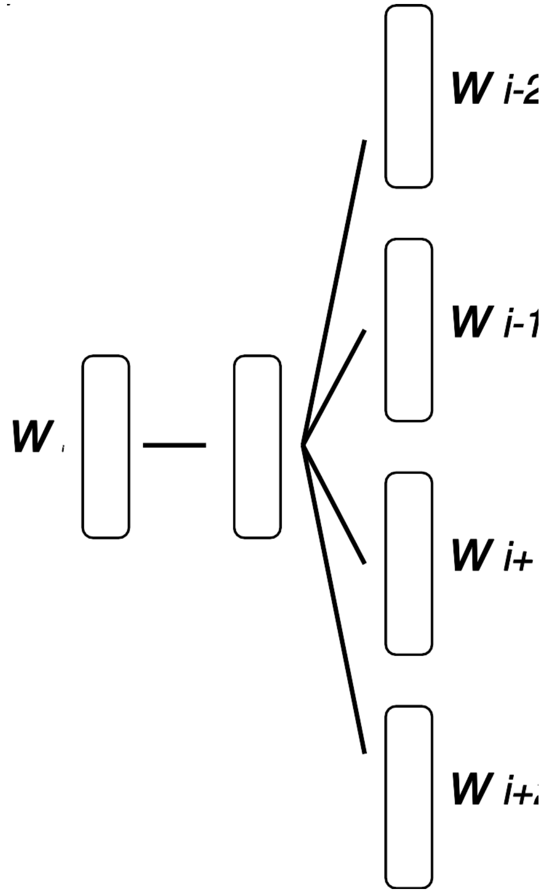


Deep learning

- Sequence of transformations
- Each transformation produce a vector (representation) of increasing abstraction
- Associate an embedding to each datapoint
- Language data:
    - Documents
    - Sentences
    - Words
    - Subword units
    - Characters

Word embeddings

- Word2vec, Glove, etc
- Trained using co-occurrences



Mikolov, et.al., 2013, Pennington et.al., 2014

king

- ('kings', 0.71)
- ('queen', 0.65)
- ('monarch', 0.64)
- ('crown_prince', 0.62)

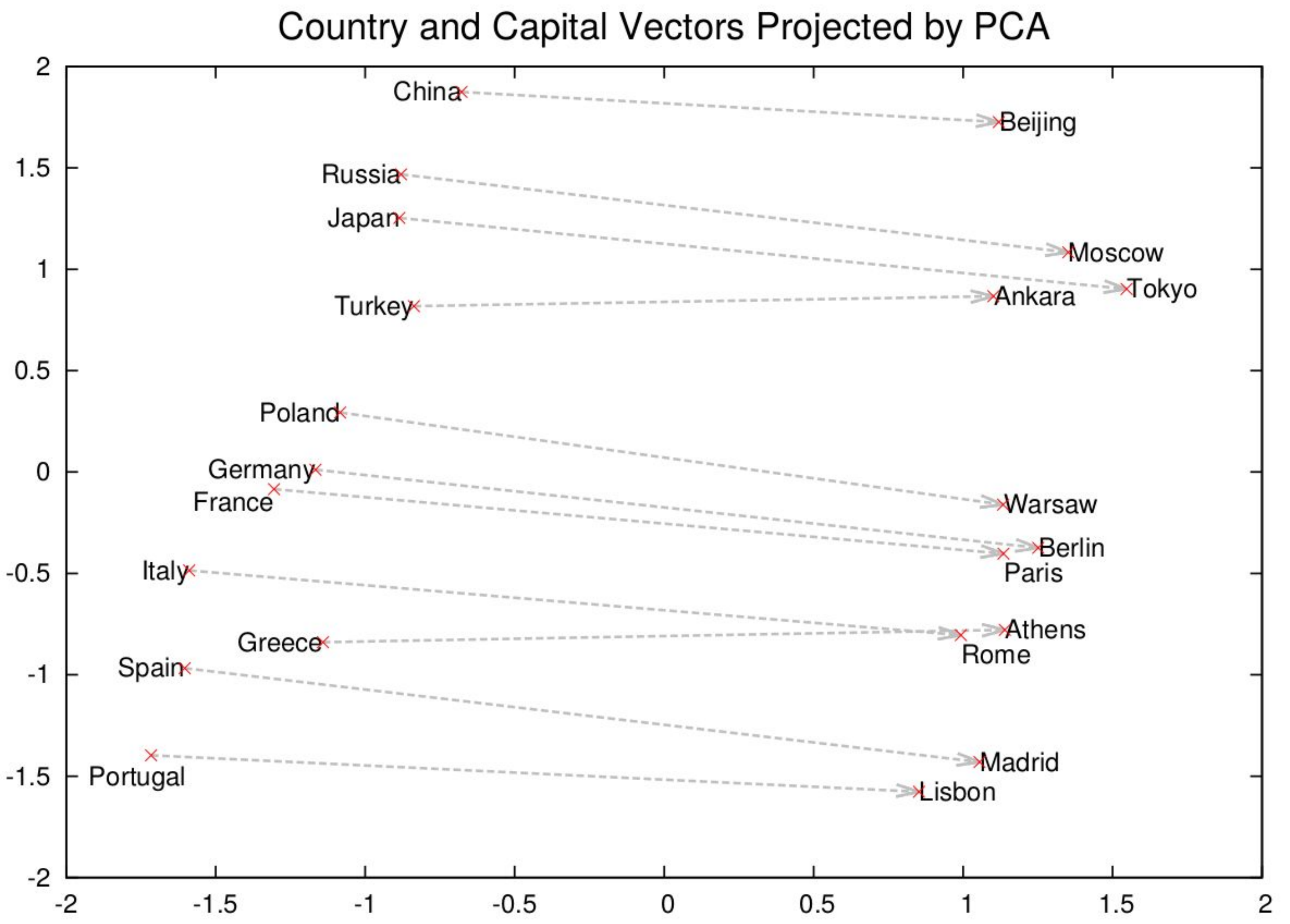
queen

- ('queens', 0.74)
- ('princess', 0.71)
- ('king', 0.65)
- ('monarch', 0.64)

Stockholm

- ('Stockholm_Sweden', 0.78)
- ('Helsinki', 0.75)
- ('Oslo', 0.72)
- ('Oslo_Norway', 0.68)

Distributional hypothesis: words with similar meaning occur in similar contexts.
(Harris, 1954)


Country and Capital Vectors Projected by PCA
China
Beijing
Russia
Japan
Moscow
Tokyo
Turkey
Ankara
Poland
Germany
France
Warsaw
Berlin
Paris
Italy
Greece
Spain
Athens
Rome
Madrid
Lisbon
Portugal
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Word2vec Skipgram analogies

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Word embeddings *was* transfer learning for language

Prediction

Processing
Learned or rule-based

Representation
Learned

Auxilliary data

Data

E.g.

- Multi-document summarization (1)
- Translation
- Text classification

*1: Kågebäck, **Mogren,** Tahmasebi, Dubhashi (2014)*

Deep transfer learning for language

- BERT (Transformer), ELMO (RNN), etc
- Trained using language modelling (word co-occurrences)
- Can compute word embedding that changes according to context

- “NLP’s Imagenet moment”: deep transfer learning for NLP, pretrain deep models.
- E.g. QA, Reading comprehension, Natural language inference, translation, constituency parsing, etc.

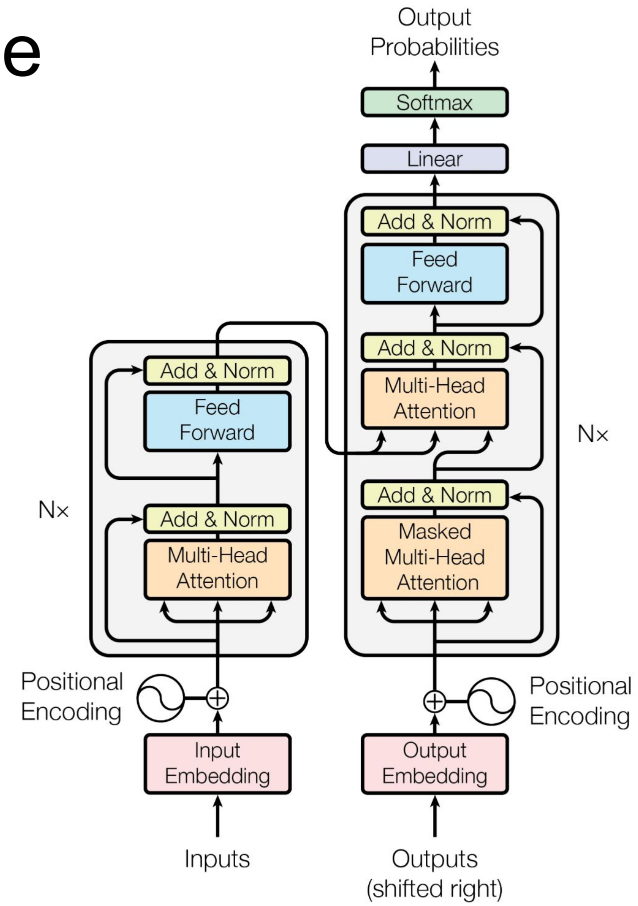


Figure 1: The Transformer - model architecture.

Vaswani, et.al. (2017), Devlin, et.al. (2018), Peters, et.al. (2018)

NLP Models are Brittle

Generating Natural Language Adversarial Examples [ASEHSC(EMNLP 18)]

Original Text Prediction: **Entailment** (Confidence = 86%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A runner wants to head for the finish line.*
Adversarial Text Prediction: **Contradiction** (Confidence = 43%)
Premise: *A runner wearing purple strives for the finish line.* **Hypothesis:** *A **racer** wants to head for the finish line.*

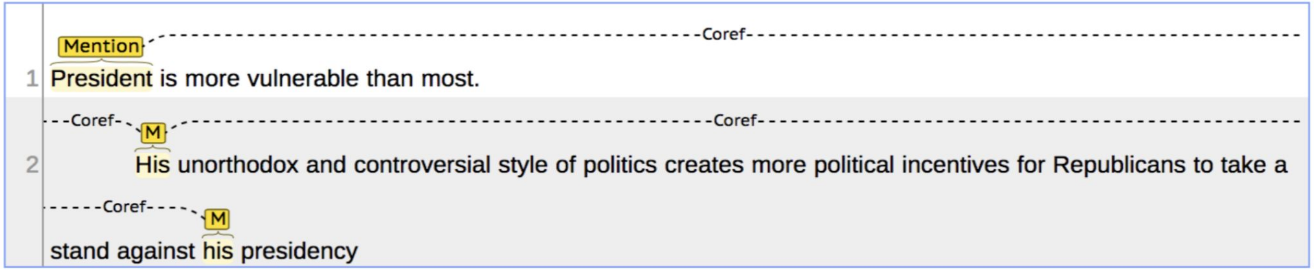
Retrofitting Contextualized Word Embeddings with Paraphrases [SCZC (EMNLP 19)]

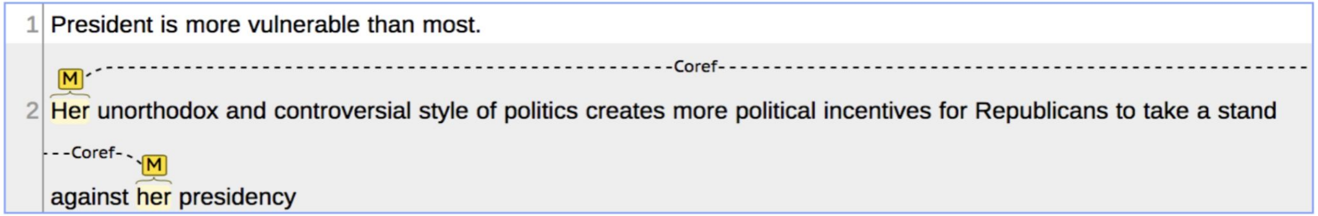
Paraphrased contexts	**L2**	**Cosine**
How can I make **bigger** my arms? How do I make my arms **bigger**?	6.42	0.27
Some people believe earth is **flat**. Why? Why do people still believe in **flat** earth?	7.59	0.46
It is a very **small** window. I have a **large** suitcase.	5.44	0.26

NLP Model is biased

[1,] Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods [ZWYOC NAACL 2018]

- Coreference resolution is biased[1,2]
  - Model fails for female when given same context





[2]Rudinger et al. Gender Bias in Coreference Resolution. NAACL 2018

NLP Model is biased

The Woman Worked as a Babysitter: On Biases in Language Generation [SCNP EMNLP 2019]

- Language generation is biased

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Human-like bias in Glove and Word2vec

- Insects and flowers (pleasantness)
- Musical instruments vs weapons (pleasantness)
- Racial bias: European-American names vs African-American names
- Gender and occupations
- Gender and arts vs sciences/mathematics

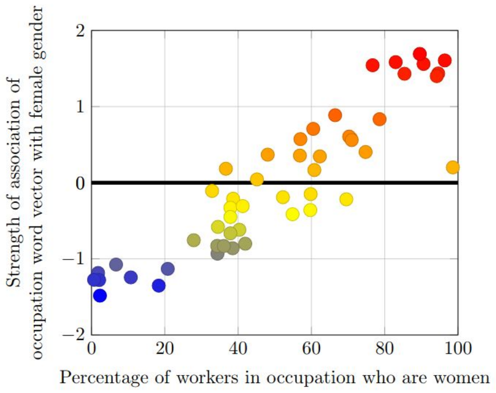


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

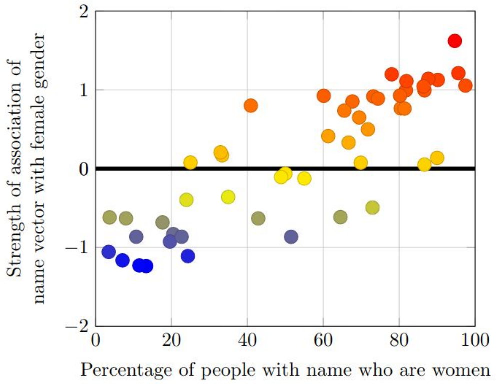


Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

Caliskan, et.al. (2017)

Man is to computer programmer as woman is to homemaker

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

gender bias in Word2vec

Bolukbasi, et.al., (NeurIPS 2016)

Also in Swedish! Also in contextualized embeddings!

- Gender-bias in Swedish pretrained embeddings
- Gender vs occupation
- Word2vec, FastText, ELMO, BERT

Sahlgren & Ohlsson (2019)

Don’t we want to model the data?

All dimensions in an embedding may be desired

But social bias may be problematic for downstream applications eg:

- Resume filtering
- Insurange, lending, hiring
- Next word prediction on your phone
- Some systems may actually perform worse, cf. coreference resolution

We need to know what we are modelling, and how data can be used for this.

Social bias

- E.g. Gender bias, racial bias, etc.
- On what attributes can we base a decision?
- How can we isolate them?

Fairness

- Is an individual treated fair in a decision? (Demographics, etc)

Privacy

- What attributes about myself do I share?

Disentanglement

- Attributes are often correlated
- Underlying factors

How do we make models react to certain information but not to all of it?

Approaches

Data augmentation

- Train models using augmented data.
- he/she
- Anonymization of names

Calibration

- Identify sensitive dimensions
- Modify

Adversarial representation learning

- Train to make it difficult for adversary

What is it that we want to model, and how do we go about it?

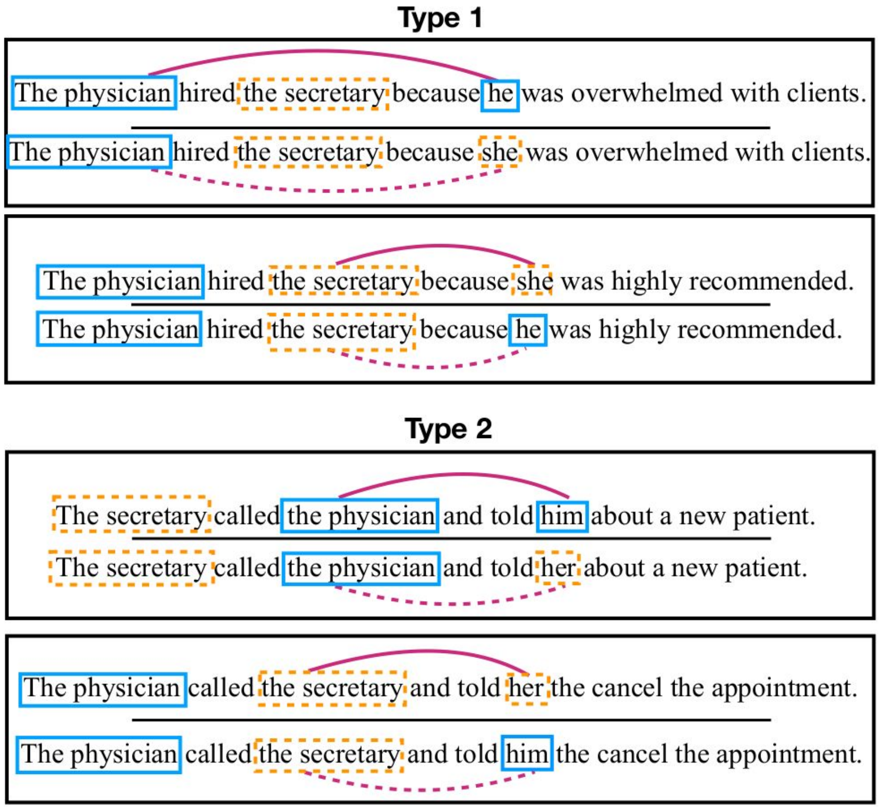
Data augmentation

"Anti-stereotypical" dataset.

Swap biased words, e.g.:

- he/she
- Anonymization of names

- Wino-bias dataset



Zhao, et.al., Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, NAACL 2018

Counterfactual Fairness

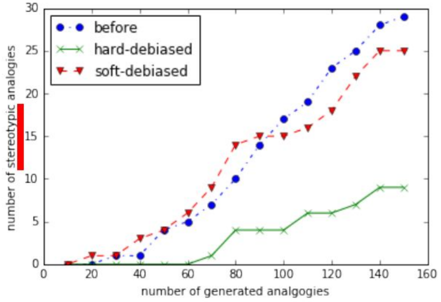
A decision is the same to an individual in

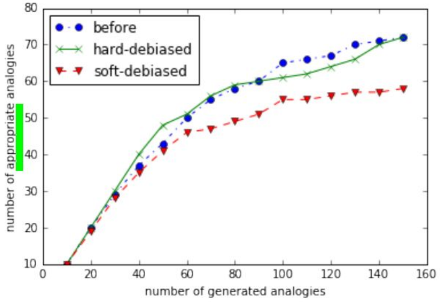
- the actual world and
- in a counterfactual world, belonging to a different demographic group

Kusner, et.al., Counterfactual Fairness, NeurIPS 2017

Calibration

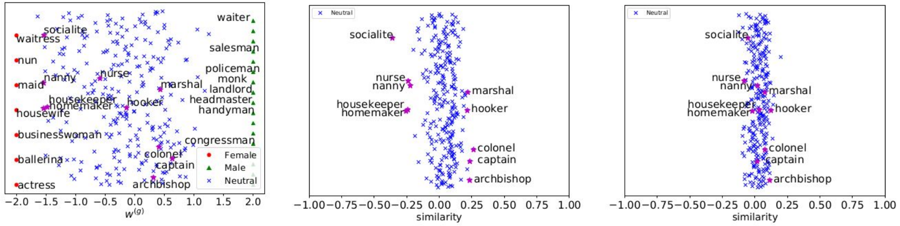
1. Identify “appropriate” gendered words (e.g. *grandfather-grandmother, guy-gal)*
2. Train model to identify these words
3. Identify gender direction
4. Modify vectors
    a. Neutral words: zero gender direction(s)
    b. Acceptable gender words: equidistant to neutral words in gender direction(s)





Bolukbasi, et.al. (NeurIPS 2016)

- Restrict sensitive attributes to specific dimensions of embedding
- Minimize distance between words in the two groups in other dimensions



Zhao, et.al. (EMNLP 2018)

Adversarially learned de-biasing calibration of word-embeddings

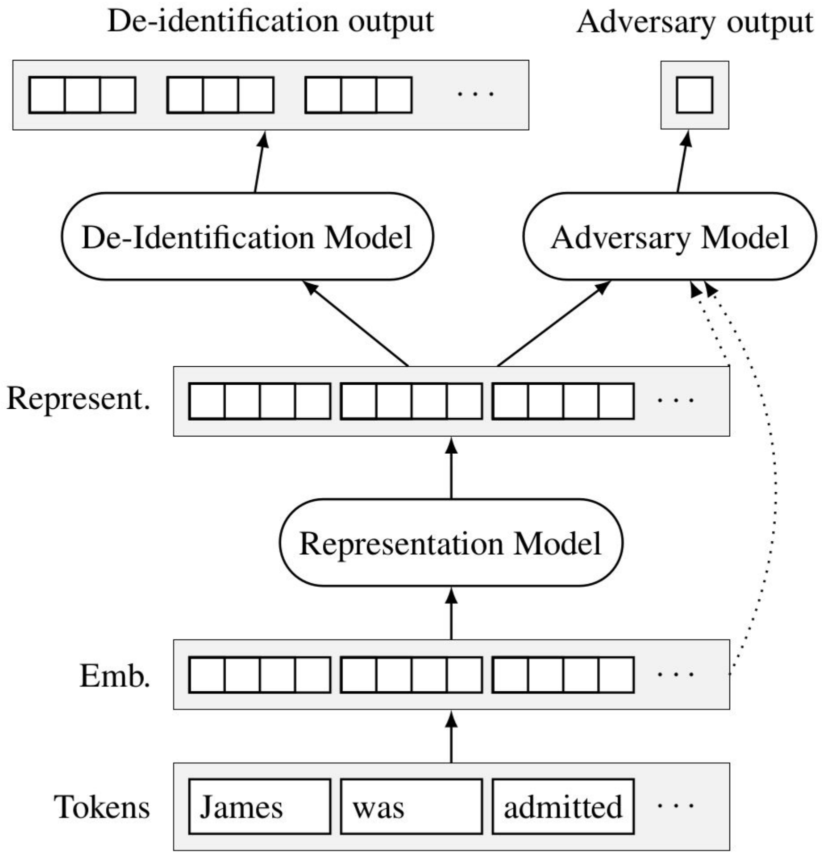
Similar to Bolukbasi, et.al., but:

- Adversary: predicts the gender.
- Transformation network: transforms embeddings to de-biased embeddings

Zhang, et.al. (AIES 2018)

Adversarial representation learning for language

- Adversary: detect privacy leakage in embeddings
- Embeddings: fool adversary
- Privacy preserving embeddings
- (Requires data augmentation)



Friedrich, et.al. (ACL 2019)

Discussion

- When should we trust data?
- Shouldn’t we model how people use language?
- Can we enumerate (think of) all possible sensitive attributes?
- Can we enumerate all correlated attributes?
- What is “appropriate” gender association? What is stereotypical?

References

Bolukbasi, et.al., NeurIPS 2016, Man is to Computer Programmer as Woman is toHomemaker? Debiasing Word Embeddings

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186

Zhao, et.al, EMNLP 2018, Learning Gender-Neutral Word Embeddings

Sahlgren & Ohlsson, 2018, Gender Bias in Pretrained Swedish Embeddings

Kiela & Bottou, EMNLP 2014, Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics

Zhao, et.al., NAACL 2018, Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods

Sato, et.al., ACL 2019, Effective Adversarial Regularization for Neural Machine Translation

Wang, et.al., ICML 2019, Improving Neural Language Modeling via Adversarial Training