Gender bias and Natural Language Processing

Marta R. Costa-jussà and Christian Hardmeier
Universitat Politecnica de Catalunya, Barcelona
University of Edinburgh/Uppsala University

work in cooperation with Christine Raouf Basta, Joel Escudé Font,
Eva Vanmassenhove and Andy Way

Outline

- General concepts: NLP and Gender Bias
- (Contextual) Words Embeddings:
  - Concept
  - Debiasing techniques
  - Gender Bias Evaluation

- Machine Translation
  - Dealing with Gender Bias in Machine Translation

Natural Language Processing

- NLP focuses on how to program computers to process and analyze natural language

- NLP is trained on large amounts of data

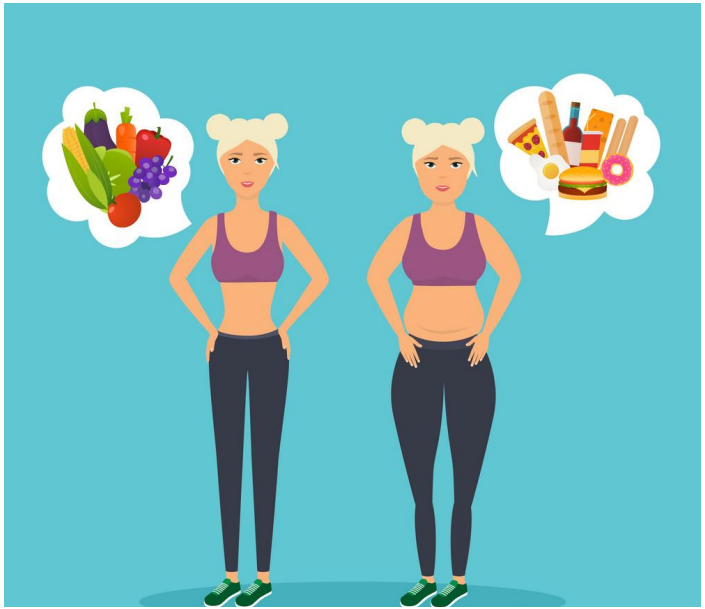
- NLP applications are used in multiple well-known applications: Translation systems, Personal Assistants...

Professions Bias

Programmer vs. Teacher

Stereotypical words

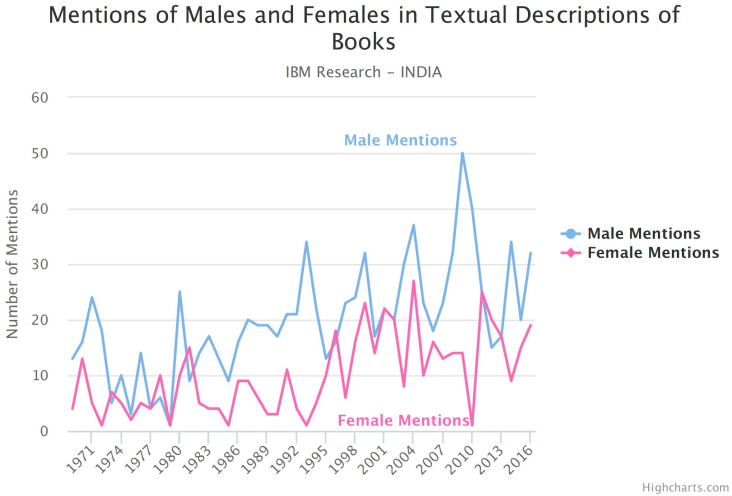
Diet (Female stereotyped) vs Hero (Male stereotyped)



Gender Bias Examples

- Under-representation of females in text books

[Maadan et al., 2018]



- Sportsmen and sportswomen are asked different questions [Fu et al., 2016]

"Hi Rafa, How do you feel playing here?"

"Hi Serena, Who designed your clothes today?"

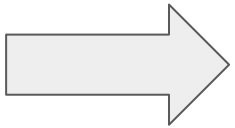
More Gender Bias Examples

- Women's evaluations contain nearly twice as much language about their communal or nurturing style e.g. "helpful" or "dedicated"
- Managers are nearly 7 times more likely to tell their male employees that their communication style is too soft, they tell their female employees 2.5 times as much feedback related to their aggressive communication style
- Men hear nearly twice as many references to their technical expertise and their vision

[Clayman Institute for Gender Research, Stanford, 2015]

Link between examples and NLP

- Women's evaluations contain nearly twice as much language about their communal or nurturing style e.g. "helpful" or "dedicated"



As a consequence, sentiment analysis system trained on such data will link those qualifications to women.

NLP infers gender biases through these data correlations

Since NLP is trained on data, these gender biases are perpetuated and amplified

NLP tasks proven to have some kind of gender bias

- Word Embeddings
- Coreference
- Machine Translation
- Sentiment analysis
- Image captioning
- ...

Gender Bias in (Contextual) Word Embeddings

Words Embeddings

- Learned from raw data based on the Distributional Hypothesis:
  - ***“You shall know a word by the company it keeps” (Firth, 1957)***

- Each word in the vocabulary is represented by a low dimensional vector

Motivation for Contextual Word Embeddings

- Same word can have different meaning depending on the context. Example:
  - *Mary and Joanna **play** basketball in a wonderful way*
  - *John is the protagonist in this year's school **play***

- Classic word embeddings offer the same vector representation regardless of the context.

- Contextual Word Embeddings create **word representations** that **depend on the context.**

Approaches for Contextual Word Embeddings

Model Alias	Org.	Article Reference
ULMfit	fast.ai	*Universal Language Model Fine-tuning for Text Classification* Howard and Ruder
ELMo	AllenNLP	*Deep contextualized word representations* Peters et al.
OpenAI GPT	OpenAI	*Improving Language Understanding by Generative Pre-Training* Radford et al.
BERT	Google	*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* Devlin et al.
xling BERT	Facebook	*Cross-lingual Language Model Pretraining* Lample and Conneau

[credits to Noe Casas]

Word Embeddings encode bias I

- Caliskan et al. 2017 replicate a spectrum of biases from using word embeddings, showing text corpora contain several types of biases:
  - morally neutral as toward insects or flowers
  - problematic as toward race or gender ,
  - reflecting the distribution of gender with respect to careers or first names

Concepts 1	Concepts 2	Attributes 1	Attributes 2
Flowers: buttercup, daisy, lily	**Insects**: ant, caterpillar, flea	**Pleasant**: freedom, health, love	**Unpleasant**: abuse, crash, filth
European American names: Brad, Brendan	**African American names**: Darnell, Lakisha	**Pleasant**: joy, love, peace	**Unpleasant**: agony, terrible
Male attributes: male, man, boy	**Female attributes**: female, woman, girl	**Math words**: math, algebra, geometry	**Arts Words**: poetry, art, dance

[credits to Hila Gonen]

Word Embeddings encode bias II

- Bolukbasi et al. 2016 show analogies like: *Man* is a *Computer Programmer* as *Woman* is to a *Homemaker*
- They **define a gender direction**: check how similar a word is to “he” and “she”(cosine similarity) and care about the difference between the two
- This gender direction can be computed using several pairs together (e.g. man-woman, brother-sister)

Techniques to Debias Word Embeddings

(1) Debias **After** Training [Bolukbasi et al. 2016] ---> Debias WE

Define a gender direction
Define inherently neutral words (nurse as opposed to mother)
Zero the projection of all neutral words on the gender direction
Remove that direction from words

(2) Debias **During** Training [Zhao et al. 2018] ---> GN-Glove

Train word embeddings using GloVe (Pennington et al., 2014)
Alter the loss to encourage the gender information to concentrate in the last coordinate
To ignore gender information –simply remove the last coordinate

All are not enough to remove bias [Gonen et al. 2019]

Evaluation of (Contextual) Word Embeddings

Contextual embeddings get a vector representation for the word according to its context, so we expect a different attitude towards the gender bias. [Zhao et al. 2019] show that contextualized word embeddings may inherit implicit gender bias. This motivates us to study **two main questions:**

- Do contextualized word embeddings **exhibit gender bias** and how does this bias **compare** to standard and debiased word embeddings**?**

- Do different evaluation techniques identify similar biases and what would be the **best measure** to use **for gender bias detection** in contextualized embeddings**?**

Why ELMO?

- Elmo was used for our experiments, as it provides word-level representations, as opposed to BERT's subwords.

- This makes it possible to study the word-level semantic traits directly.

Experiments For Evaluation Bias

Five experiments were carried out in our evaluation:

1. Detecting gender space
2. Direct Bias
3. Male and female biased words clustering
4. K-NN of professions
5. Classification approach of biased words

Results are compared with results taken from (Bolukbasi et al., 2016) for the **first two experiments** and (Gonen and Goldberg, 2019) for the **next three experiments**.

Note for the experiments

- While standard and debiased Word Embeddings are trained on the same sets for all experiments reported, contextual word embeddings are not, so we are not offering a strictly fair comparison.

- Our comparison is based on pre-trained sets of all these options. For experiments, we use the English-German news corpus from WMT18

- We are taking the average of 10 experiments for the clustering, KNN and classification experiments.

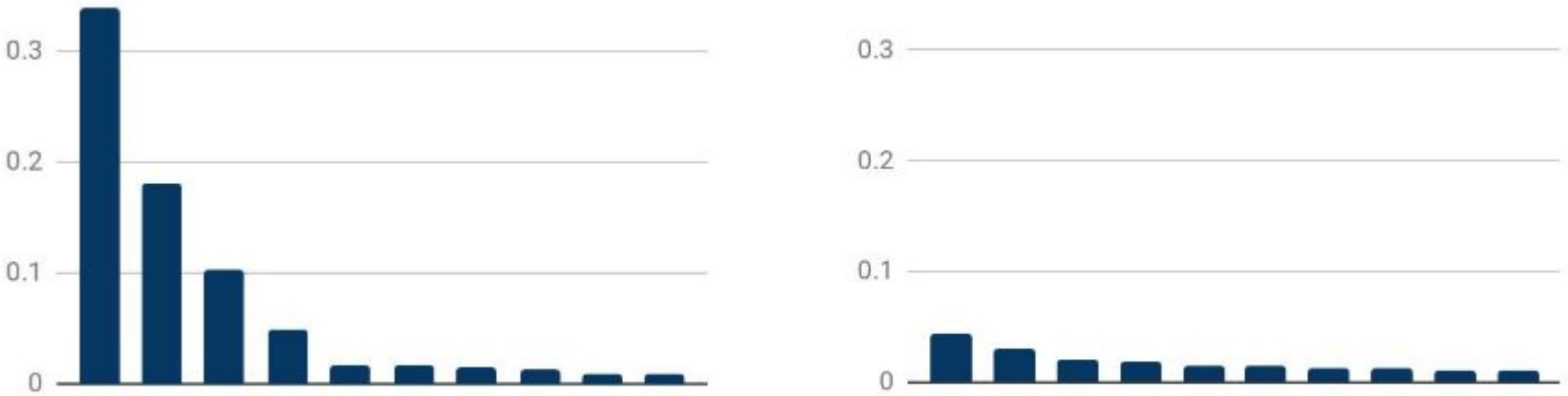
Lists for Definitional, Biased and Professions Terms

- **Definitional List** 10 pairs (e.g. he-she, man-woman, boy-girl)
- **Biased List**, which contains of 1000 words, 500 female biased and 500 male biased. (e.g. breastfeeding, bridal and diet for female and hero, cigar and teammates for male)
- **Extended Biased List**, extended version of Biased List. (5000 words, 2500 female biased and 2500 male biased)
- **Professional List** 319 tokens (e.g. accountant, surgeon)

1. Detecting the Gender Space

1. Randomly sampling sentences that contain words from the Definitional List, swap the definitional word with its pair-wise equivalent from the opposite gender.
2. Get Elmo embeddings for the word and its swapped equivalence, compute their difference.
3. On the set of difference vectors, we compute their principal components to verify the presence of bias.

Percentage of variance in PCA: definitional vs random



(Left) the percentage of variance explained in the PC of definitional vector differences.
(Right) The corresponding percentages for random vectors

2. Direct Bias

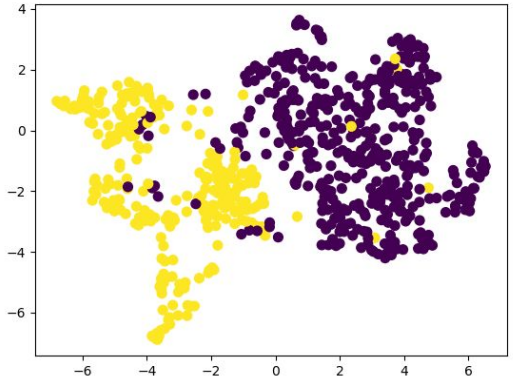
- **Direct Bias** is a measure of how close a certain set of words are to the gender vector.
- Computed on list of professions.

$$\frac{1}{|N|}\sum_{w \epsilon N} |cos(\vec{w}, g)|$$

	Direct Bias
Biased WE	0.08
ELMO	0.03

3. Male and female-biased words clustering

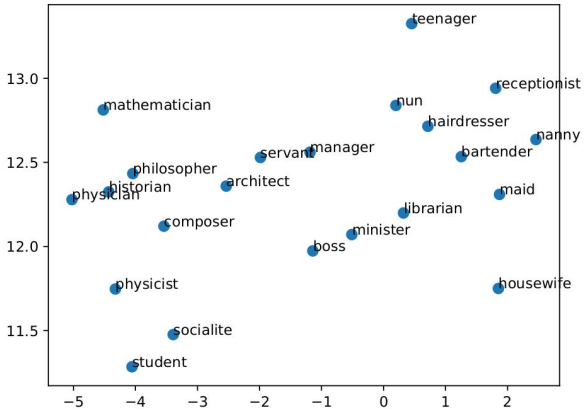
- **k-means** is applied
- Generate 2 clusters of the embeddings of tokens from the **Biased list**



	Accuracy
Biased WE	99,9%
Debiased WE	92,5%
ELMO	70,1%

4. K-Nearest Neighbor

- kNN on the Professional List
- Compute the percentage of female and male stereotyped professions among the kNN of each profession token
- Compute the Pearson correlation of this percentage with the original bias of each profession.



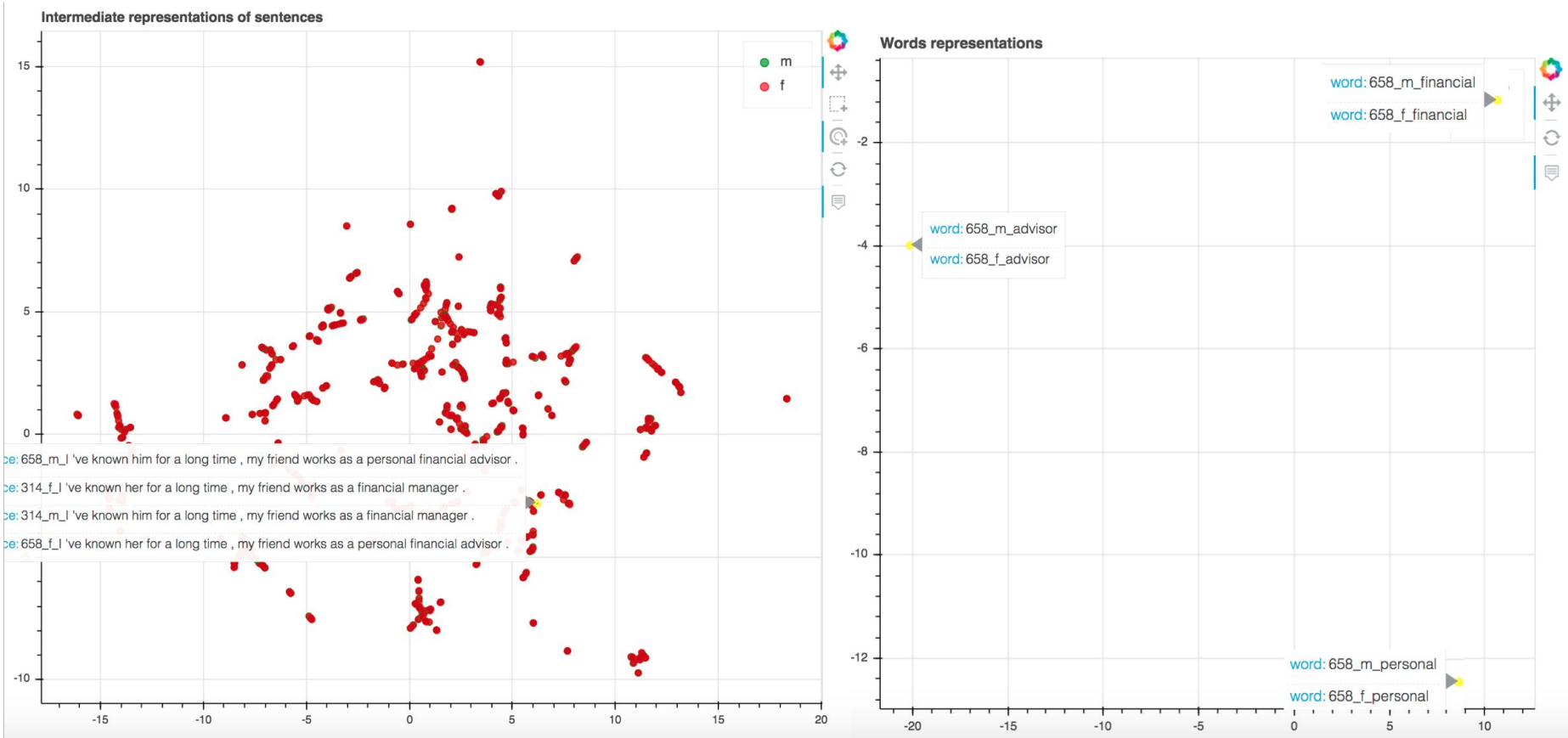
	Pearson Correlation
Biased WE	0.77
Debiased WE	0.60
ELMO	0.89

5. Classification Approach

- Classify Extended Biased List into words associated between male and female
- SVM
- 1000 for training, 4000 for testing

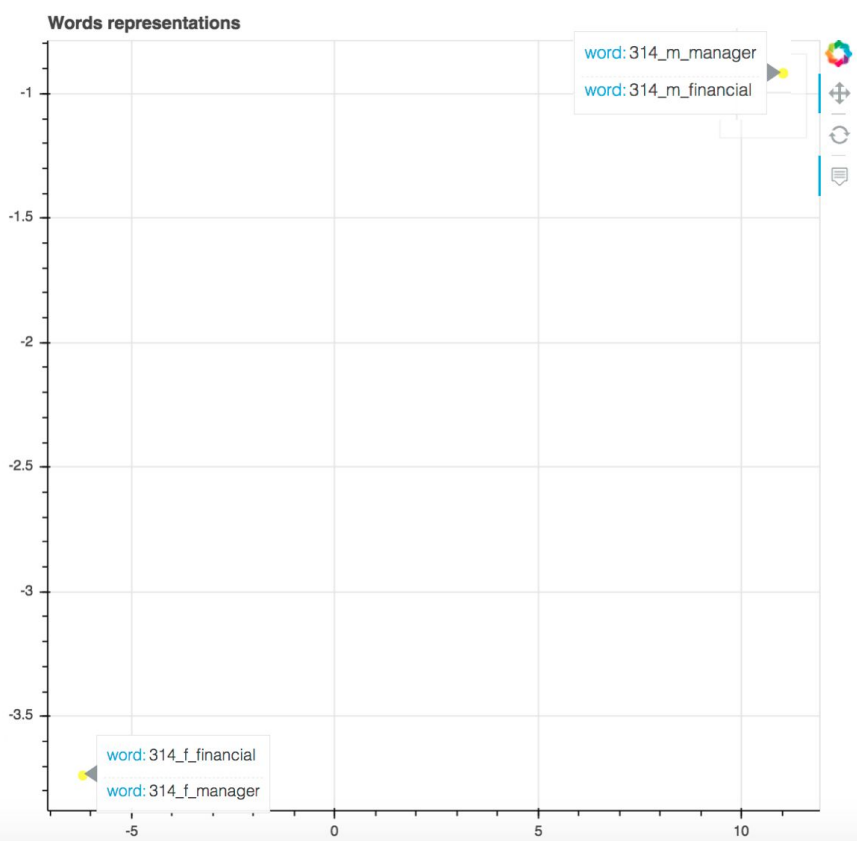
	Accuracy
Biased WE	98.25%
Debiased WE	88.88%
ELMO	85.56%

Visualization tool



Visualization Tool available at:

https://github.com/elorala/interlingua-visualization



Conclusions on evaluating gender bias in contextual word embeddings

😀 Contextualize Embeddings **mitigate bias** in when measuring in the following aspects:

- ⬇ gender **space**
- ⬇ **direct bias**
- ⬇ male/female **clustering**,

😞 Contextualized word embeddings **preserve and even amplify** gender bias:

- ~ **classification** experiment
- ⬆ **KNN** professional experiment

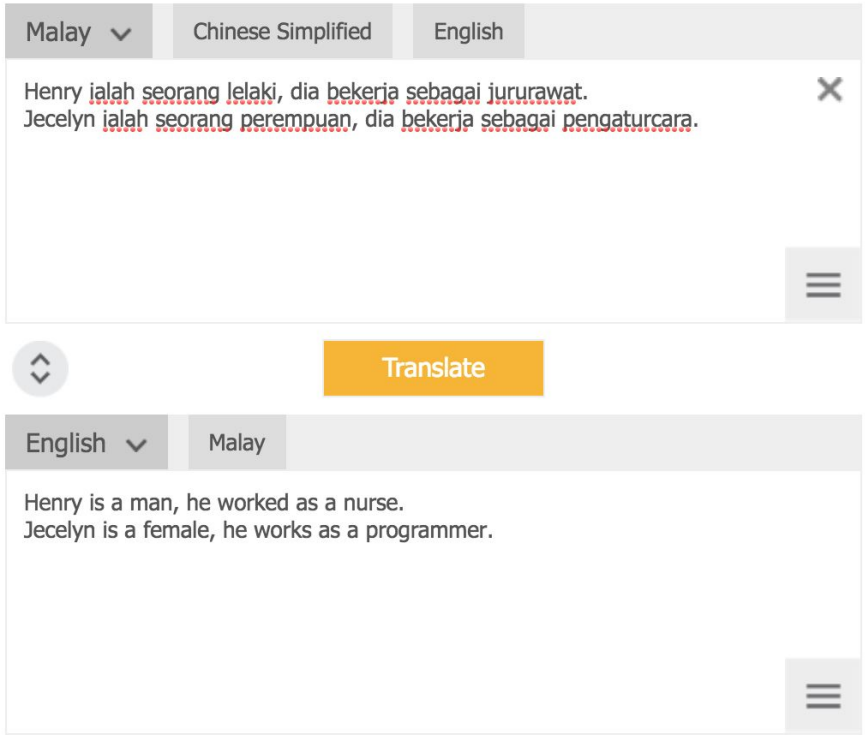
Open questions

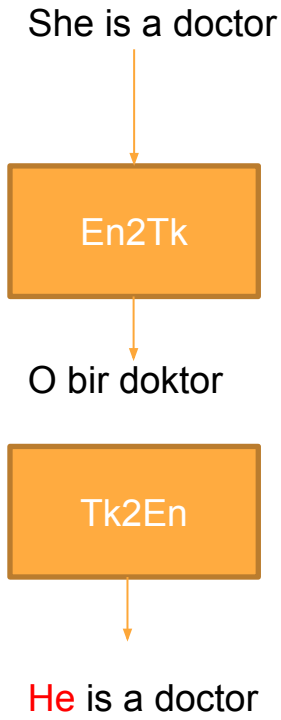
- What is the right measure to evaluate the bias in such embeddings??

- How to evaluate in other languages?

Gender Bias in Machine Translation

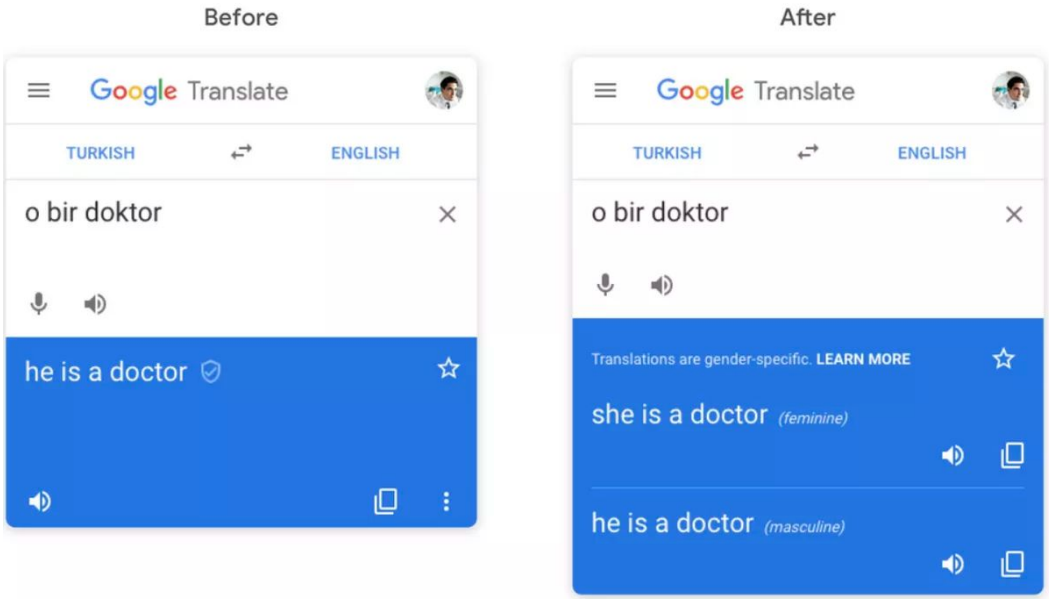
Gender Bias in MT: Example





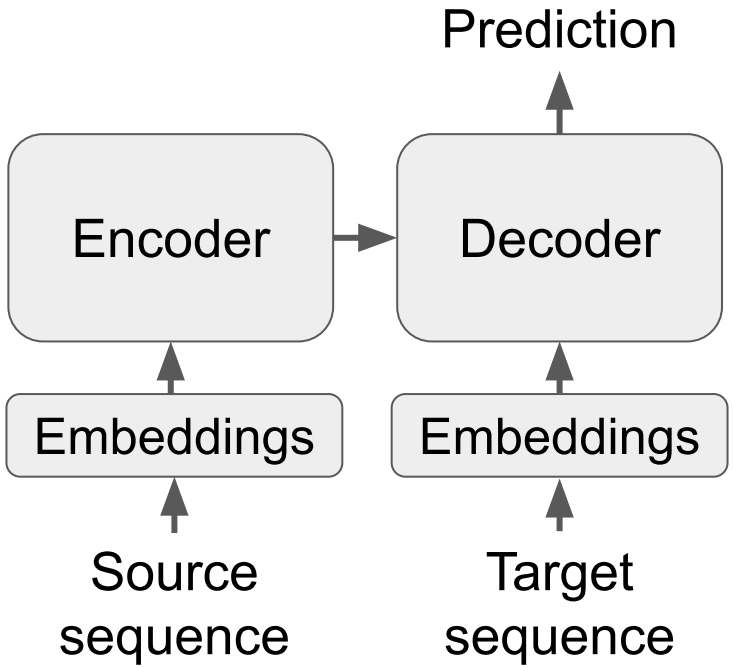
Related work: Providing Gender-Specific Translations

[Johnson et al., 2018]



How to reduce gender bias in a neural MT?

- Neural MT system
  - Transformer
- Word embeddings
  - GloVe
  - GloVe hard-debiased
  - GN-GloVe
- Data
  - EN->ES WMT



Impact on Translation Quality

Pre-trained emb.	BLEU
Baseline	29.78
GloVe	30.62
GloVe Hard-debiased	29.95
GN-GloVe	30.74

DataSet for Testing Improvements on Gender Bias

Test of 1000 sentences, on the patterns

(En) I've known her/him for a long time, my friend works as a/an .... [OCCUPATION]
(Es) La/Lo conozco desde hace mucho tiempo, mi amiga/amigo trabaja como .... [OCCUPATION]
(En) I've known Mary/John for a long time, my friend works as a/an .... [OCCUPATION]
(Es) Conozco a Maria/John desde hace mucho tiempo, mi amiga/amigo trabaja como .... [OCCUPATION]

List of 1000 occupations [U.S. Bureau of Labor Statistics].

(En) accounting clerk : (Es) contable

Impact on Equalizing Gender Bias: Accuracy

Pre-trained emb.	her : amiga	him : amigo	her:Mary	him:John
Baseline	99.8	99.9	69.5	99.9
GloVe	**100.0**	100.0	90.0	100.0
GloVe Hard-Debiased	**99.9**	100.0	**100.0**	100.0
GN-GloVe	99.6	100.0	56.4	100.0

Impact on Equalizing Gender Bias: Examples

Pre-trained word emb.	Prediction
Source	my *friend* works as a *refrigeration mechanic*.
GN-GloVe enc+dec	mi **amiga** trabaja como **mecánica** de refrigeración.
Reference	mi *amiga* trabaja como *mecánica de refrigeración*.

Pre-trained word emb.	Prediction
Source	my *friend* works as a *mine shuttle car operator*.
GN-GloVe enc+dec	mi **amiga** trabaja como **operadora** de transporte de minas.
Reference	mi *amiga* trabaja como *operadora de vagones de minas*.

Conclusions on Equalizing Gender Bias in MT

Using equalized word embeddings on a MT system show:

- Similar translation quality
- Less biased gender predictions

Limitations

- Based on “debiased” word embeddings (Gonen and Goldberg 2019)
- Re-learning biases during MT training

Getting Gender Right in NMT

Eva Vanmassenhove, Christian Hardmeier and Andy Way, EMNLP 2018.

- Creation of a multilingual dataset with utterances labelled for speaker gender and other demographic information.

- Experiments with NMT systems tagged for speaker gender.

The following slides are based in part on a presentation by Eva Vanmassenhove.

Compilation of Datasets

- **Large dataset with speaker information**
  - Europarl source files:
    - Speaker tags
    - Date of session
    - MEPs meta-information available (Rabinovich et al. 2017)

      => Retrieve gender, age …

“Madam President , as a former health care professional…”

20 Languages Paired with English

Languages	# sents	Languages	# sents
EN–BG	306,380	**EN–IT**	1,297,635
EN–CS	491,848	**EN–LT**	481,570
EN–DA	1,421,197	**EN–LV**	487,287
EN–DE	1,296,843	**EN–NL**	1,419,359
EN–EL	921,540	**EN–PL**	478,008
EN–ES	1,419,507	**EN–PT**	1,426,043
EN–ET	494,645	**EN–RO**	303,396
EN–FI	1,393,572	**EN–SK**	488,351
EN–FR	1,440,620	**EN–SL**	479,313
EN–HU	251,833	**EN–SV**	1,349,472

Table 1: Overview of annotated parallel sentences per language pair

Gender Tagging for NMT

- Gender information added as a tag at the beginning of the sentence:
  “**FEMALE** Madam President, as a...”
  “**MALE** ….”

- Similar to earlier experiments on politeness (Sennrich et al., 2016)

- Sequence-to-sequence encoder-decoder with LSTMs
  OpenNMT-py
  BPE to reduce OOV

Hypotheses

- We expect to see stronger improvements for translation from English into languages with **morphologically marked gender agreement**.

  \+ English into French, Italian, Portuguese, Spanish, Greek
  \- English into Danish, German, Dutch, Finnish, Swedish

- We expect to see stronger improvements for **female speakers** due to the overrepresentation of men in the training set.

Results

General Test Sets:

Systems	EN	EN-TAG
FR	37.82	**39.26***
ES	42.47	42.28
EL	31.38	**31.54**
IT	31.46	**31.75***
PT	36.11	**36.33**
DA	36.69	**37.00***
DE	28.28	28.05
FI	21.82	21.35*
SV	35.42	35.19
NL	28.35	28.22

Results mostly consistent with hypothesis, except EN-ES (no improvement) and EN-DA (unexpected improvement).

Results

Male vs Female Test Sets:

Test Sets	EN	EN-TAG
FR (M)	37.58	**38.71***
FR (F)	37.75	**38.97***
FR (M1)	39.00	**39.66***
FR (F1)	37.32	**38.57***

M: Male data
F: Female data
M1: Male 1st person utterances
F1: Female 1st person utterances

Successes and Failures

(Ref)	En tant que **vice-président**...	(M)
(BASE)	En tant que **vice-présidente**...	(F)
(TAG)	En tant que **vice-président**…	(M)
(Ref)	... je suis **heureuse que**...	(F)
(BASE)	... je suis **heureux que**...	(M)
(TAG)	... je suis **heureuse que**…	(F)
(Ref)	je suis **gênée** que...	(F)
(BASE)	je suis **embarassé** que...	(M)
(TAG)	je suis **embarassé** que...	(F)

Side Effects

(Ref)	Je **pense** que …	(Ref)	J’ ai plusieurs **remarques**...
(BASE)	Je **crois** que…	(BASE)	J’ ai un nombre de **commentaires**...
(TAG)	Je **pense** que…	(TAG)	J’ ai plusieurs **remarques**..

- Both correct translations
- Enriched system picks ‘preferred’ variant

  ~ Different preferences: constructions, word choices etc.

  ~ Frequency list:

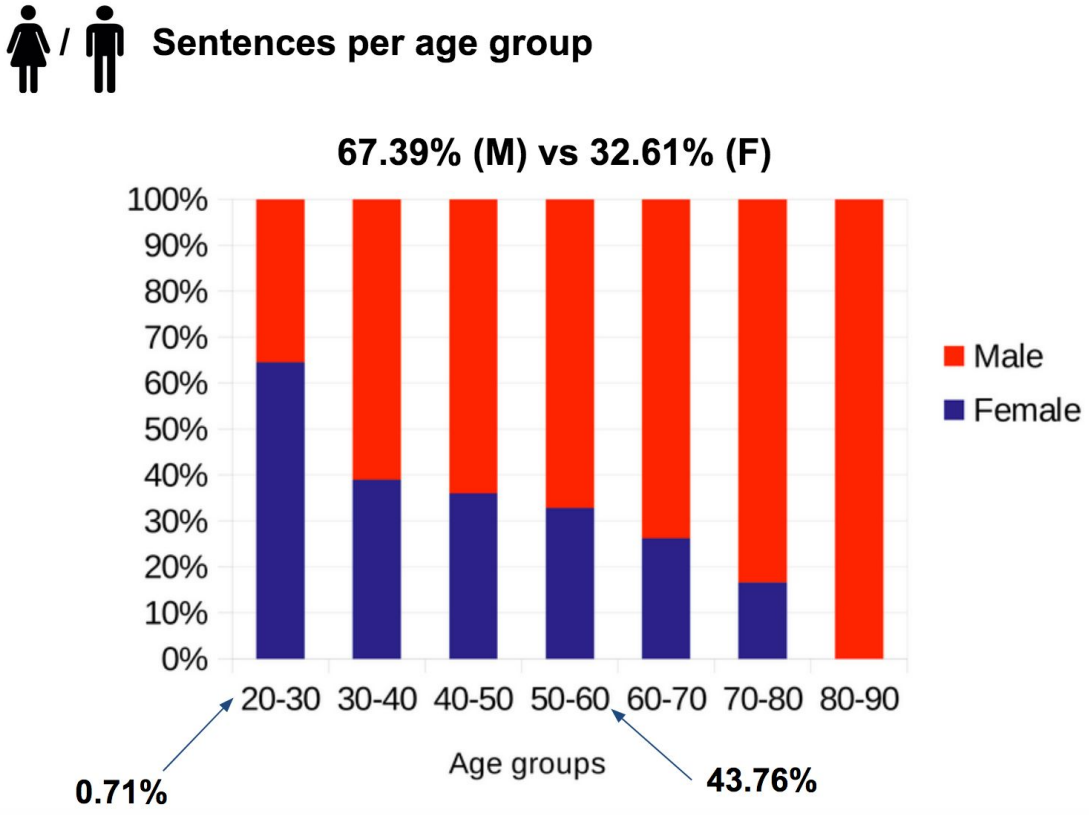
  => “penser” (more neutral) vs “croire” (~male)

  IS THIS SOMETHING WE WANT?
- Gender of the translator?

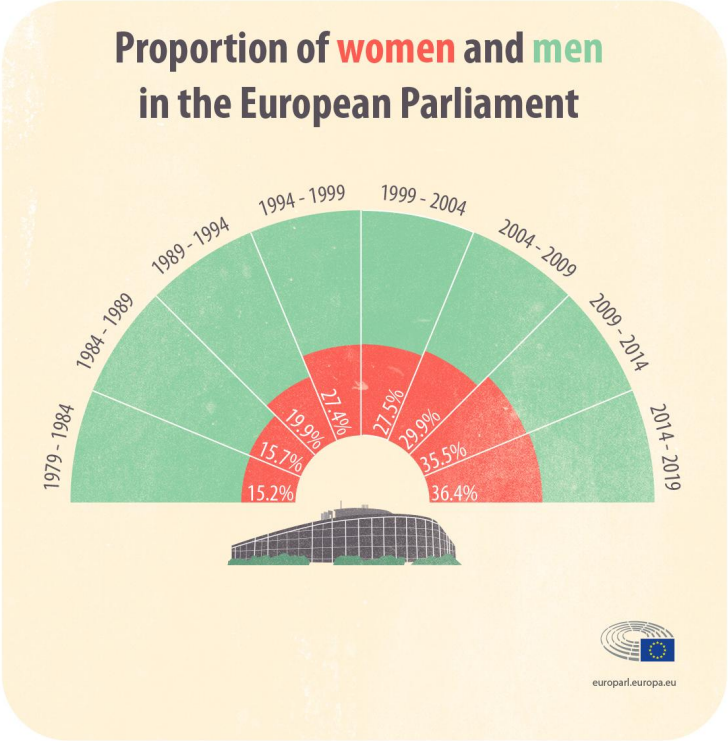
Variable Correlations

- In any specific corpus, gender is correlated with other variables.
- We need to be careful what we measure!
- More analysis is required to understand how “debiasing” and gender-aware methods can have a positive impact on society.

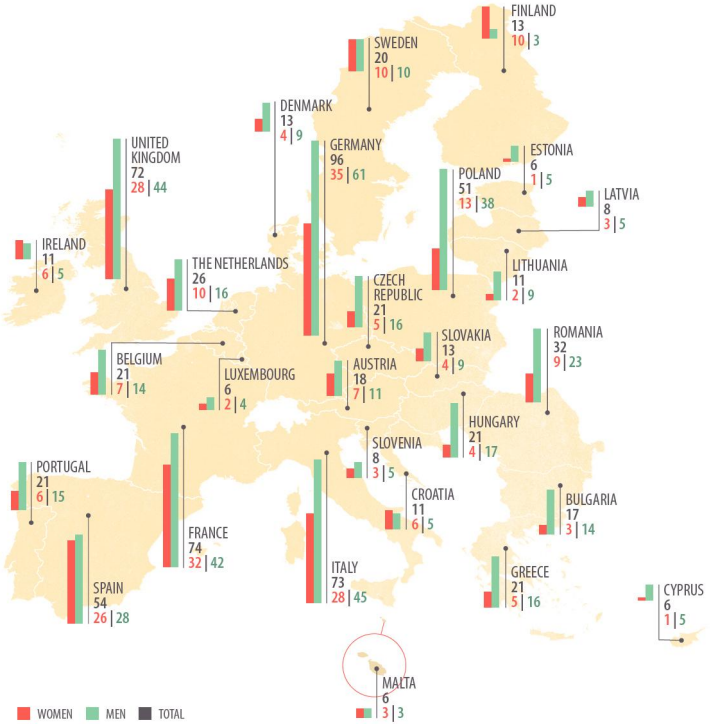
Gender is Correlated with Age



Gender is Correlated with Session Period



Gender is Correlated with Nationality



Arguments for Debiasing

Why detect bias? Unconscious bias can be harmful

Why debias NLP? NLP reflects society since it is trained on OUR data

Debiasing computer systems may help in debiasing society

Biased Embeddings and the Real World

- Study on gender bias in word embeddings and its real-world correlates by Friedman et al. (GeBNLP 2019).
- Gender bias of different word lists computed with “standard” methods.
- Correlated with various statistical indicators of gender equality.
- Data sets covering English-language Twitter data from 99 countries and from 50 US states.

Biased Embeddings and the Real World

	govt	intellect	workplace	excellent	childcare	illness	communal	victim	*"pretty"*	r-1	r-2	r-3	r-4
Index: Overall Gender Gap	**.30**	.11	.17	.12	-.01	-.07	-.20	-.06	-.19	-.01	.02	.03	.02
Sex ratio at birth	.00	.01	.03	.00	-.02	-.01	.00	-.02	.00	.00	.03	-.04	.00
Index: Educational Attainment	.03	.05	.10	.03	-.18	-.12	-.19	**-.23**	-.07	-.04	.02	.00	.00
Literacy rate	.07	.08	.05	.07	-.18	-.13	-.21	**-.23**	-.08	-.03	.02	-.05	-.07
Enrollment tertiary education	.06	.10	.07	.02	**-.24**	-.20	-.12	-.11	-.06	-.01	-.01	-.08	-.03
Enrollment secondary education	.02	.01	.03	.00	-.08	-.01	**-.21**	-.13	-.02	.00	.04	-.02	-.01
Enrollment primary education	.01	.01	.05	.01	-.05	-.06	**-.15**	-.12	-.06	-.01	.04	.01	.03
Index: Political Empowerment	**.28**	.02	.10	.01	.04	.01	-.03	.04	-.14	.07	.01	-.04	.00
Women in ministerial positions	**.25**	.03	.17	.09	.00	-.02	-.02	.00	-.08	.04	.03	-.01	-.07
Women in parliament	**.16**	.04	.07	.02	.01	.03	-.02	.05	-.06	.02	.06	-.09	.00
% 50 years female head of state	.05	-.04	.01	-.07	.02	.01	.00	.00	-.06	.01	-.02	-.01	.01
Index: Economic Participation	.10	.04	.13	.10	-.02	-.12	**-.33**	-.09	-.10	-.07	.04	-.05	.04
Professional and technical workers	.20	.23	**.27**	.23	-.15	-.21	-.19	-.12	-.14	-.05	.08	-.08	.06
Legislators, officials, managers	.08	.15	.10	**.18**	-.05	-.14	**-.18**	-.10	-.01	-.05	.04	.00	.03
Labour force participation	.03	.04	.08	.02	-.03	-.09	**-.21**	-.04	-.14	-.09	.02	-.03	.01
Wage equality (survey)	-.02	-.04	-.02	-.02	.03	.04	.00	.00	-.02	.00	-.02	-.03	-.05
Index: Health and Survival	.03	.09	.08	.06	-.12	**-.13**	-.02	-.06	.00	-.02	.09	.01	.00
Healthy life expectancy	.06	.09	.12	.14	-.16	**-.31**	-.07	-.11	-.01	-.02	.07	.01	-.01

Figure 1: Correlation of themed neutral word sets' gender bias (columns) against categories of gender gaps from worldbank.org (rows). Values are R^2 coefficient of determination, where negation is added to indicate inverse correlation. The rightmost four word sets (*r-1 to r-4*) were randomly sampled from the vocabulary for comparison.

Is Debiasing What We Want?

- Strong focus on geometrical debiasing methods in NLP community
  - Identify directions corresponding to gender in the embedding space
  - Project embeddings to collapse those dimensions
- “Lipstick on a pig”? (Gonen and Goldberg, NAACL 2019)
- Difficult to scale to different forms of bias (but see Karve et al., GeBNLP 2019)
- Is debiasing even (always) desirable?
  - ML is about learning biases. Removing attributes removes information.
  - Often this is what we want.
  - But a healthcare app would be well advised to give consideration to the differences between men and women (without favouring one gender over the other)!
- Strongly entrenched notions of binary gender

Threat Model for Biased NLP

- Gender information in NLP systems becomes harmful when the use of the system has a negative impact on people's lives.
- To counteract harmful bias, we need to **understand the negative impact** that the use of our models has.
- Gender bias is a social phenomenon that can't be solved with mathematical methods alone. Collaborate with social sciences/sociolinguistics.

Gender Bias is about Accuracy

Lest we believe we can ignore this if we don't have strong feelings about gender equality...

- **Gender bias causes NLP systems to make errors.**
- You should care about this even if accuracy is all you care about.
- Moreover: Errors caused by bias are likely to offend people.
- Ultimately, it's not so much about "removing bias" than about **making the right choice in each individual case.**