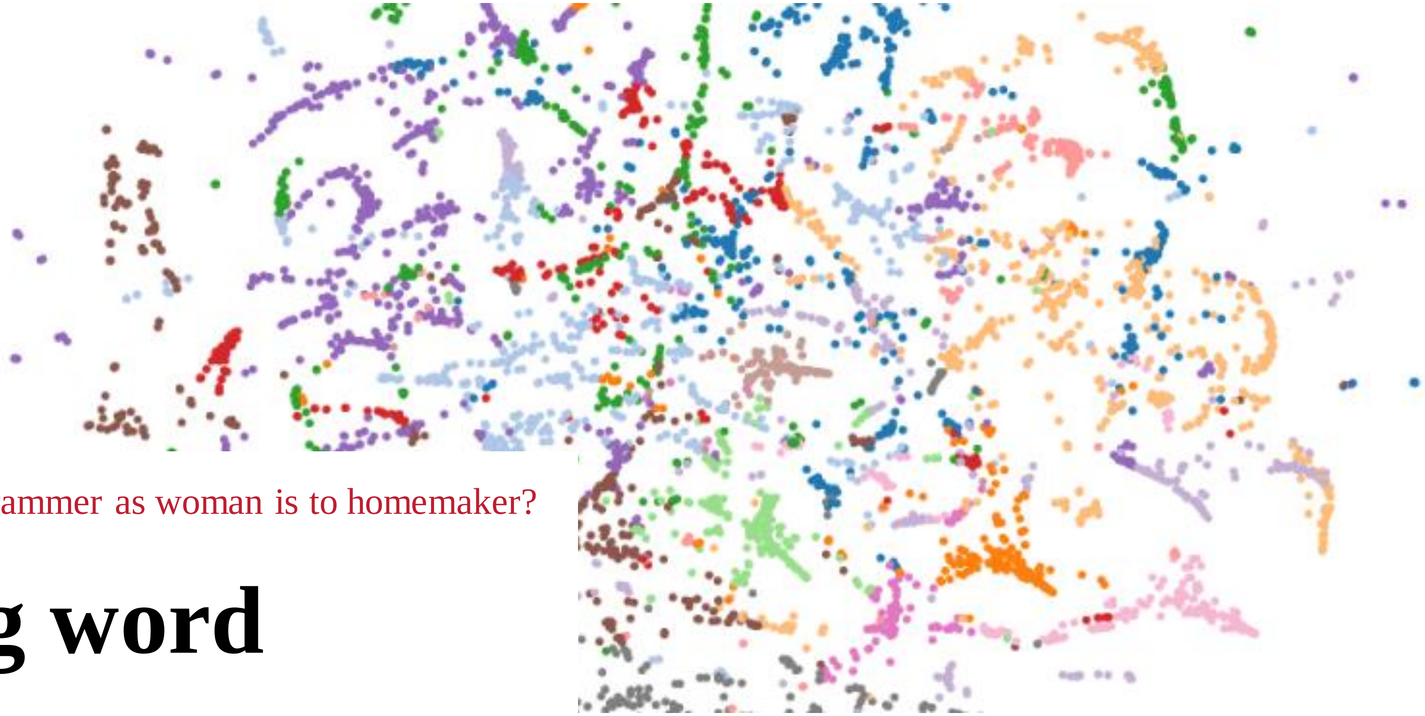
University of Amsterdam



Man is to computer programmer as woman is to homemaker?

Debiasing word embeddings

Ivan Yovchev – Davide Barbieri – Andras Csirik – Balint Hompot 01.31.2019.

Outline

1. **Bias in word embeddings**
2. Removing gender bias
3. Results after debiasing
4. Searching for remaining bias
5. Conclusions

The problem with word embeddings

Increaing popularity of task automation

- Hiring employees
- Evaluating applications
- Granting loans
- Etc.

Increasing application of NLP methods

- Many rely on word embeddings
- Embeddings encode word knowledge, including bias

Scalability of automated processes

- One unfair application can affect many lives

We need to address bias in word embeddings to ensure fair applications

World knowledge is also world bias

Original text

Engineering Phd. student

Wanted association

Engineering **postgrad** student

Unwanted association

Male Phd. student

Project outline

Replicate:

- Removing gender bias in word embeddings
  - **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings**
    - Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai, 2016

Extension:

- Running debiasing on both word2vec and Glove
- Performance assesment on word embeddings benchmark, following:
  - **How to evaluate word embeddings? On importance of data efficiency and simple supervised tasks**
    - Stanisław Jastrzebski, Damian Leśniak, Wojciech Marian Czarnecki
- Quantitative assesment of debiasing following:
  - **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them**
    - Hila Gonen, Yoav Goldberg, 2019
- Searching for the remaining bias

Outline

1. Bias in word embeddings
2. **Removing gender bias**
3. Results after debiasing
4. Searching for remaining bias
5. Conclusions

Intuition

Identify gender subspace

- Select gender defining word pairs (“he”-”she”)
- Calculate gender defining subspace

Neutralize words

- Find words that should be gender neutral
- Ensure that their projection on the gender space is 0

Equalize with respect to word pairs

- Define pairs of equality words (e.g. “grandmother” – “grandfather”)
- Ensure that neutral words in equal distance from each word (e.g. “babysit” – “grandma” and “babysit” – “grandpa”)

Identify gender subspace

Inputs:

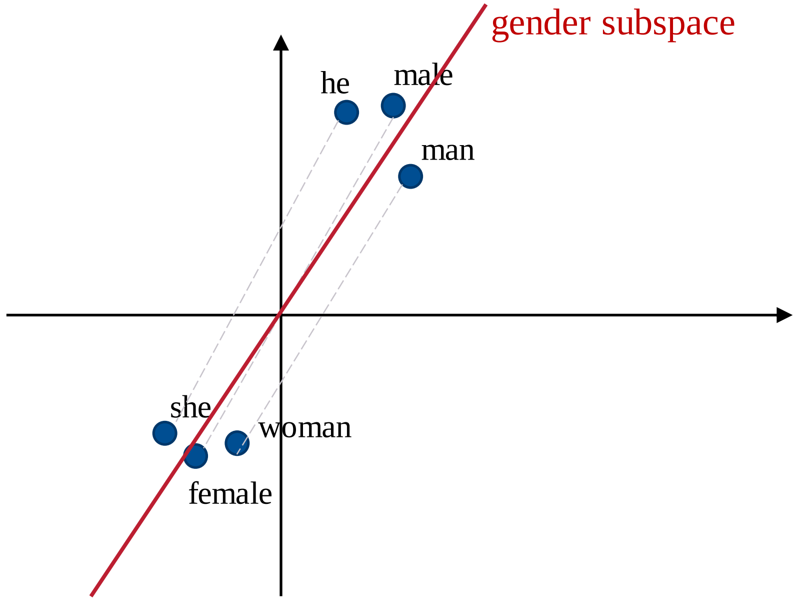
- Word sets W
- Defining sets $D_{1 \ldots}$
- Embeddings $\vec{w}$
- Projection on space B is noted in subscript $(\vec{w} \blacktriangleright \vec{w}_B)$

Output:

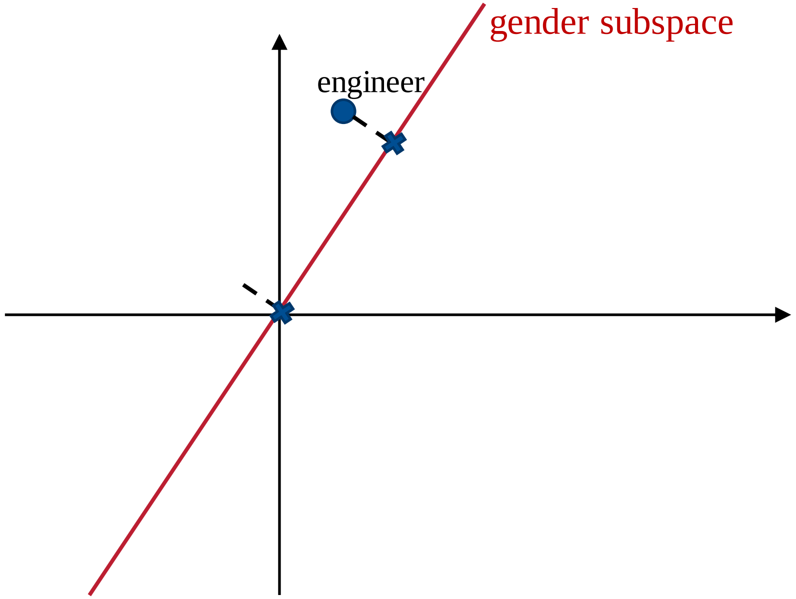
- Gender subspace B as the first k components of SVD(**C**)

$$\mu_i := \sum_{w \in D_i} \vec{w}/|D_i|$$

$$\mathbf{C} := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \mu_i)^T (\vec{w} - \mu_i)/|D_i|.$$



$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|.$$



Neutralize

Additional input:

- Words to neutralize N
- Words embeddings $\vec{w}$ in N

Output:

- Neutralized word embeddings in N

Equalize

Additional inputs:

- Equlization reference pairs $E_{1\ldots}$

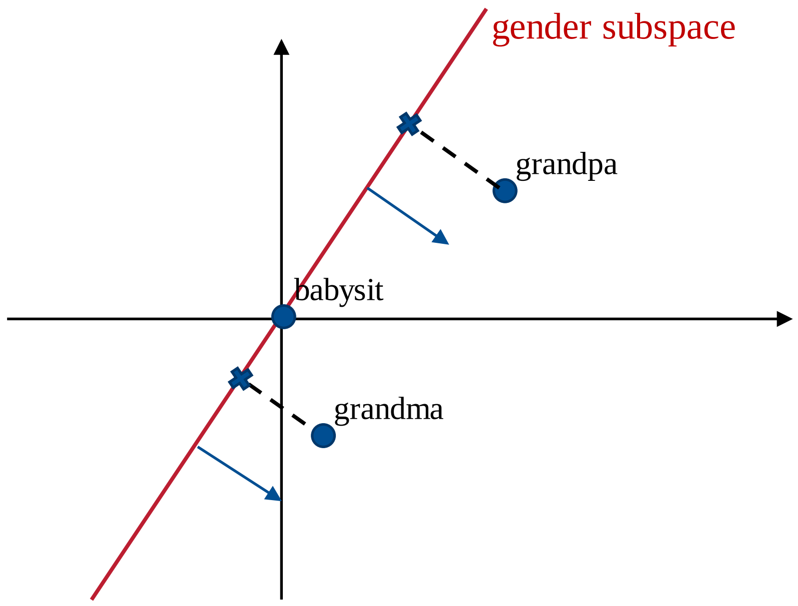
Output:

- Equalized word embeddings in N

$$\mu := \sum_{w \in E} w/|E|$$

$$\nu := \mu - \mu_B$$

$$\text{For each } w \in E, \quad \vec{w} := \nu + \sqrt{1 - \|\nu\|^2}\,\frac{\vec{w}_B - \mu_B}{\|\vec{w}_B - \mu_B\|}$$



Outline

1. Bias in word embeddings
2. Removing gender bias
3. **Results after debiasing**
4. Searching for remaining bias
5. Conclusions

Double goal

Maintaining performance

Removing bias

Maintaining performance – word embeddings benchmark

Measuring performance on 4 sets of tasks:

- Similarity (e.g. identifying close words)
- Analogy (generating analogies as defined by humans)
- Sentence analysis (e.g. sentiment)
- Retaining word properties (e.g. POS-tagging)

Debiasing did not affect the performance

	w2v	dw2v	glove	dglove
AP	0.56	0.56	0.53	0.54
BLESS	0.67	0.68	0.76	0.76
Batting	0.24	0.23	0.27	0.27
ESSLI1a	0.73	0.73	0.77	0.72
ESSLI2b	0.8	0.8	0.75	0.75
ESSLI2c	0.64	0.64	0.62	0.62
MEN	0.7	0.7	0.76	0.76
MTurk	0.51	0.52	0.64	0.63
RG65	0.69	0.69	0.75	0.75
RW	0.28	0.28	0.18	0.18
SimLex999	0.44	0.44	0.4	0.4
WS353	0.65	0.65	0.7	0.7
WS353R	0.58	0.58	0.66	0.66
Google	0.33	0.33	0.39	0.38
MSR	0.57	0.57	0.55	0.55
SemEval2012	0.2	0.2	0.18	0.18

Double goal

Maintaining performance

Removing bias

Qualitative debiasing assessment

Direct

	word2vec	she	he
Before	1.	homemaker	maestro
	2.	registered nurse	skipper
	3.	nurse	protage
	4.	receptionist	philosopher
	5.	librarian	captain
After	1.	socialite	planner
	2.	nurse	mechanic
	3.	homemaker	gangster
	4.	hairdresser	fighter pilot
	5.	registered nurse	pollster

Indirect

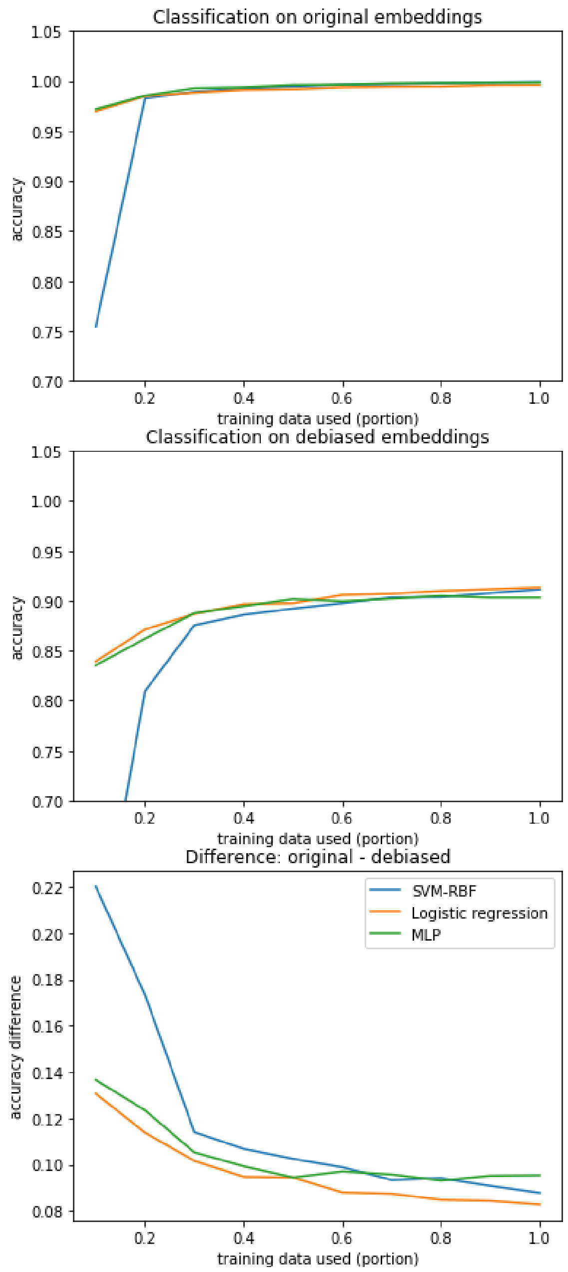
	word2vec	softball	football
Before	1.	bookkeper	footballer
	2.	receptionist	businessman
	3.	registered nurse	pundit
	4.	waitress	maestro
	5.	homemaker	cleric
After	1.	infielder	footballer
	2.	major leaguer	cleric
	3.	bookkeeper	vice chancellor
	4.	clerk	lecturer
	5.	investigator	fashion designer

Quantitative debiasing assessment

- Select the 2500 most biased female and male words (largest projection)
- See if we can classify them as male and female biased words after debiasing
- Iteratively increase training size to see if the classification is harder

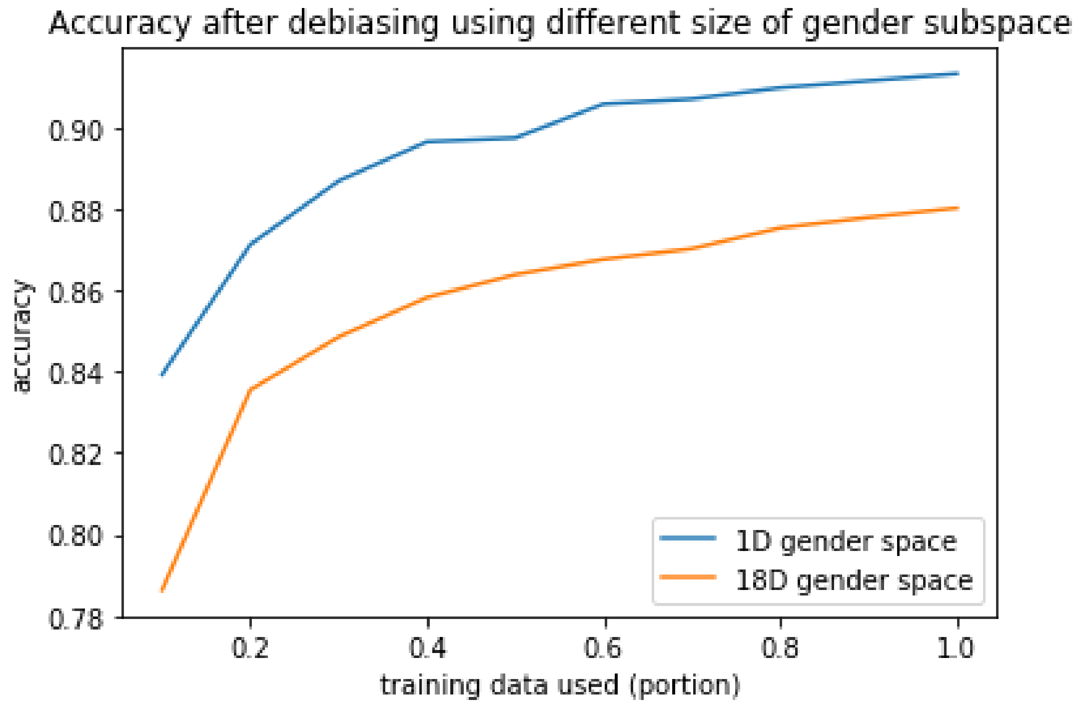
Classifiers need somewhat more data, but the accuracy is almost as good as without debiasing

→ **Bias is hidden, but not removed**



Outline

1. Bias in word embeddings
2. Removing gender bias
3. Results after debiasing
4. **Searching for remaining bias**
5. Conclusions



Using larger gender subspace

- Original paper uses 1 principal component as most variance is explained by it
- Remaining component may contain the remaining bias
- We use maximum size (18 components) gender space for projection in neutralization and equalization

With larger gender subspace, classification becomes harder, but the accuracy is still high

The performance on the benchmark is very close, but slightly smaller

Outline

1. Bias in word embeddings
2. Removing gender bias
3. Results after debiasing
4. Searching for remaining bias
5. **Conclusions**

Conclusions

- The original results are replicated
- Qualitative analysis shows, that the bias still remains
- Extending the gender subspace helps, but the bias is still present
- To improve, we can further extend the subspace, or search for a non-linear one, but it may hurt the performance
- We cannot remove the bias completely in post-processing
- Other biases are even harder to remove: race, country of origin etc.
- Fair word embeddings should come from fair data

Thank you for your attention

Questions?