# IN5550: Neural Methods in Natural Language Processing
# Sub-lecture 12.1
# *Bias in static word embeddings*

Andrey Kutuzov

University of Oslo

7 April 2022



# Contents



1. Static word embeddings

# The influence of the training corpus




No vector representation is objectively ‘good’ or ‘bad’. It is all task-dependent.

# Static word embeddings



## Caveats

- Similarity is a complicated notion and can be tricky
  - should '*coffee*' be more similar to '*tea*' than to '*bean*'?

# Static word embeddings

## Caveats

- Similarity is a complicated notion and can be tricky
  - should '*coffee*' be more similar to '*tea*' than to '*bean*'?
- Antonyms are notoriously difficult to capture distributionally
  - '*bad*' and '*good*' generally have the same contexts



# Bias in static word embeddings




*Semantics derived automatically from language corpora contain human-like biases*. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

Slides by R. Rudinger, 2017

# Bias in static word embeddings



- Embeddings reproduce bias and prejudice from the training corpora

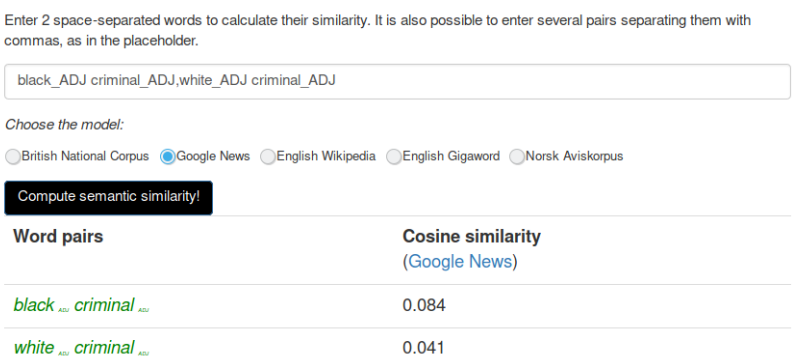
# Bias in static word embeddings



- Embeddings reproduce bias and prejudice from the training corpora
- '*Black*' more similar to '*criminal*' than '*white*', etc.

# Bias in static word embeddings

- Embeddings reproduce bias and prejudice from the training corpora
- '*Black*' more similar to '*criminal*' than '*white*', etc.



# Bias in static word embeddings



- Models learn to pay attention to things we don't want them to pay attention to.

# Bias in static word embeddings



- Models learn to pay attention to things we don't want them to pay attention to.
- 'Protected attributes'

# Bias in static word embeddings



- Models learn to pay attention to things we don't want them to pay attention to.
- 'Protected attributes'
- Is it a problem or a signal that the training algorithm is good?

# Bias in static word embeddings



- Models learn to pay attention to things we don't want them to pay attention to.
- 'Protected attributes'
- Is it a problem or a signal that the training algorithm is good?
- It depends.

# Bias in static word embeddings



- Models learn to pay attention to things we don't want them to pay attention to.
- 'Protected attributes'
- Is it a problem or a signal that the training algorithm is good?
- It depends.
- Imagine an algorithm which helps HR managers to make hiring decisions...
- ...analyzing your CV using biased word embeddings.

# Bias in static word embeddings



- Models learn to pay attention to things we don't want them to pay attention to.
- 'Protected attributes'
- Is it a problem or a signal that the training algorithm is good?
- It depends.
- Imagine an algorithm which helps HR managers to make hiring decisions...
- ...analyzing your CV using biased word embeddings.
- do you want to fall victim to statistical regularities?

# Debiasing

'Debiasing' word embeddings is an important NLP topic now...

**Lipstick on a Pig:**
**Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them**

**Hila Gonen**[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

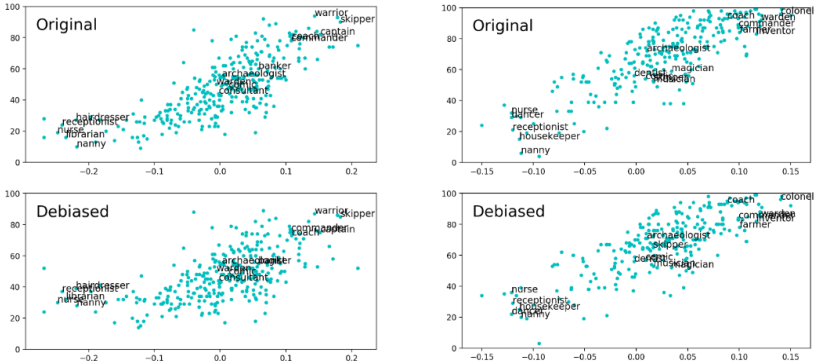
**Abstract**

Word embeddings are widely used in NLP for a vast range of tasks. It was shown that word embeddings derived from text corpora reflect gender biases in society. This phenomenon is pervasive and consistent across different word embedding models, causing serious concern. Several recent works tackle this problem, and propose methods for significantly reducing this gender bias in word embeddings, demonstrating convincing results.

swer the analogy "man is to computer programmer as woman is to x" with "x = homemaker". Caliskan et al. (2017) further demonstrate association between female/male names and groups of words stereotypically assigned to females/males (e.g. arts vs. science). In addition, they demonstrate that word embeddings reflect actual gender gaps in reality by showing the correlation between the gender association of occupation words and labor-force participation data.

# Debiasing

...and this problem is far from being fully solved!




(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.

(b) The plots for GN-GLOVE embedding, before (top) and after (bottom) debiasing.

Figure 2: The number of male neighbors for each profession as a function of its original bias, before and after debiasing. We show only a limited number of professions on the plot to make it readable.

[Gonen and Goldberg, 2019]

# References I


Gonen, H. and Goldberg, Y. (2019).
Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.
In Proceedings of the 2019 Workshop on Widening NLP, pages 60–63, Florence, Italy. Association for Computational Linguistics.