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

Hila Gonen & Yoav Goldberg

NAACL19

Gender Bias in Embeddings

- Geometry of word embeddings
  - Closer words have closer meanings
  - Analogy task $\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$
  - Biased analogy $\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$?
  - Cosine similarities $\vec{w} * (\overrightarrow{man} - \overrightarrow{woman})$

- WEAT test
  - IAT test in psychology
  - Criterion $s(w, A, B) = mean_{a \in A} \cos(\vec{w}, \vec{a}) - mean_{b \in B} \cos(\vec{w}, \vec{b})$

Debiasing Methods

- Hard-debiasing
  - Recognizing gender subspace and remove it
- Modifying objective function (GN-Glove)
  - Split dimensions into 2 parts

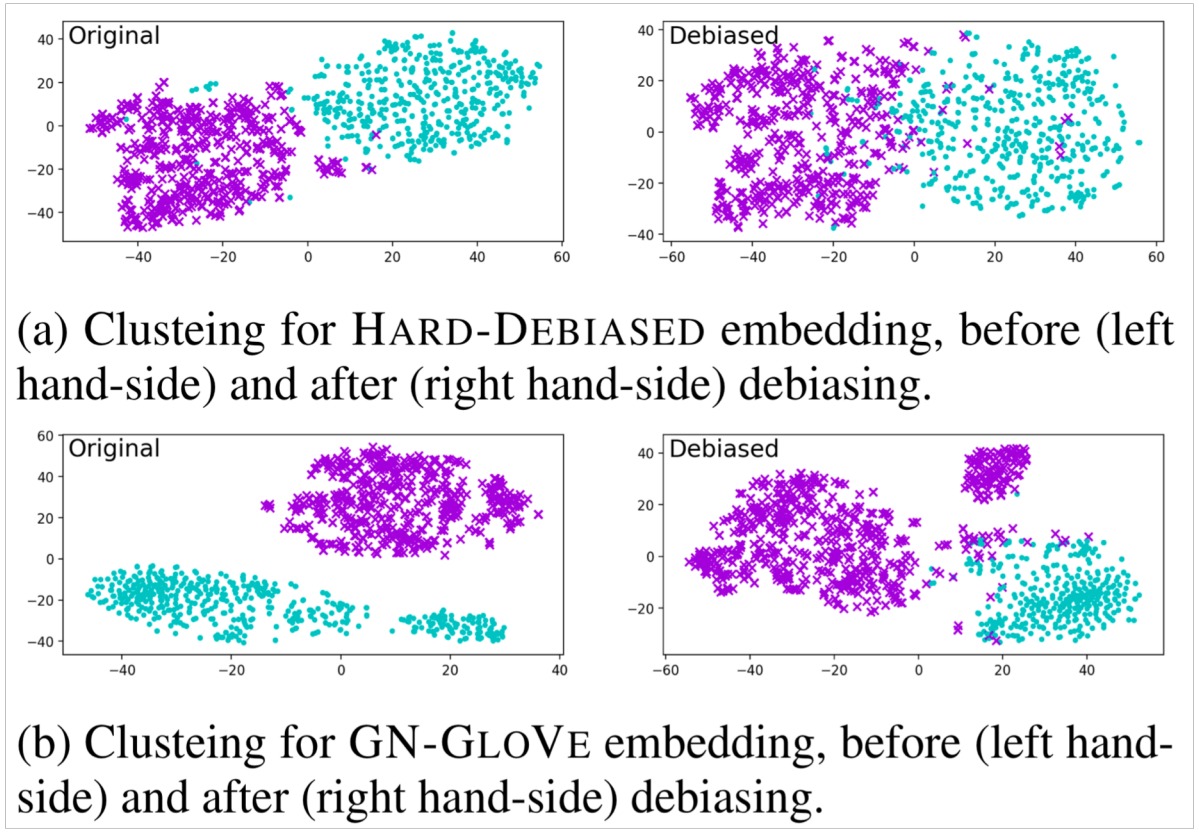
Problems in Definition

- Implicitly define what is good debiasing
- Biases are more profound than previous definitions
- Geometry stays largely the same

Experiments

- Cluster
  - Highest & lowest biased words
  - Cluster into 2 clusters by k-means

	Hard-Debiased	GN-Glove
Original	92.5%	85.6%
Debiased	99.9%	100%



(a) Clusteing for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.

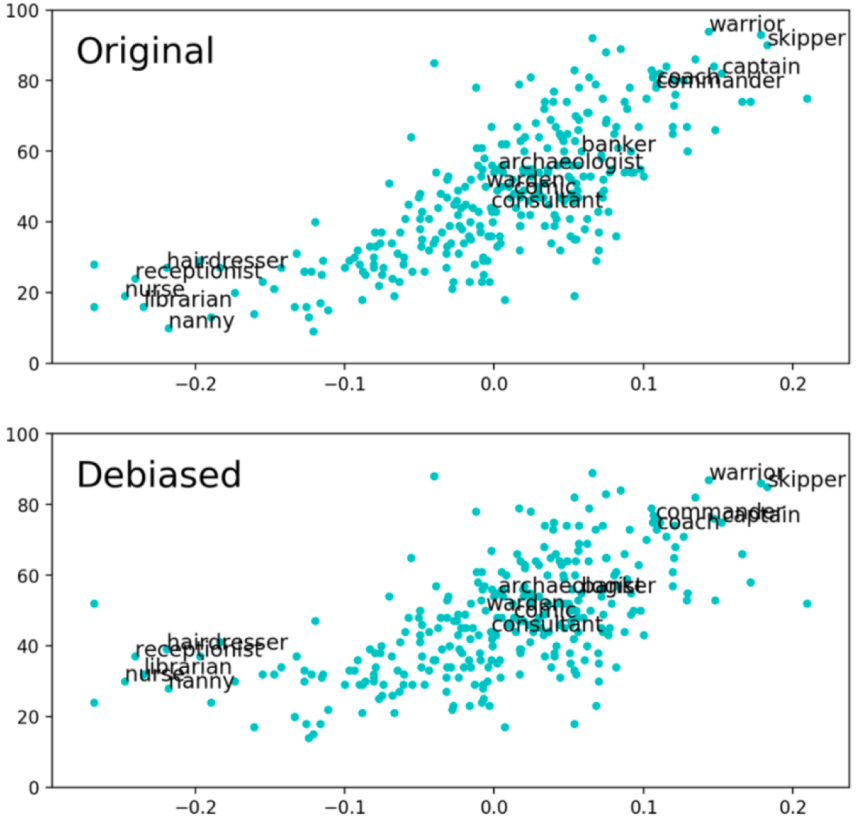
(b) Clusteing for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

Experiments

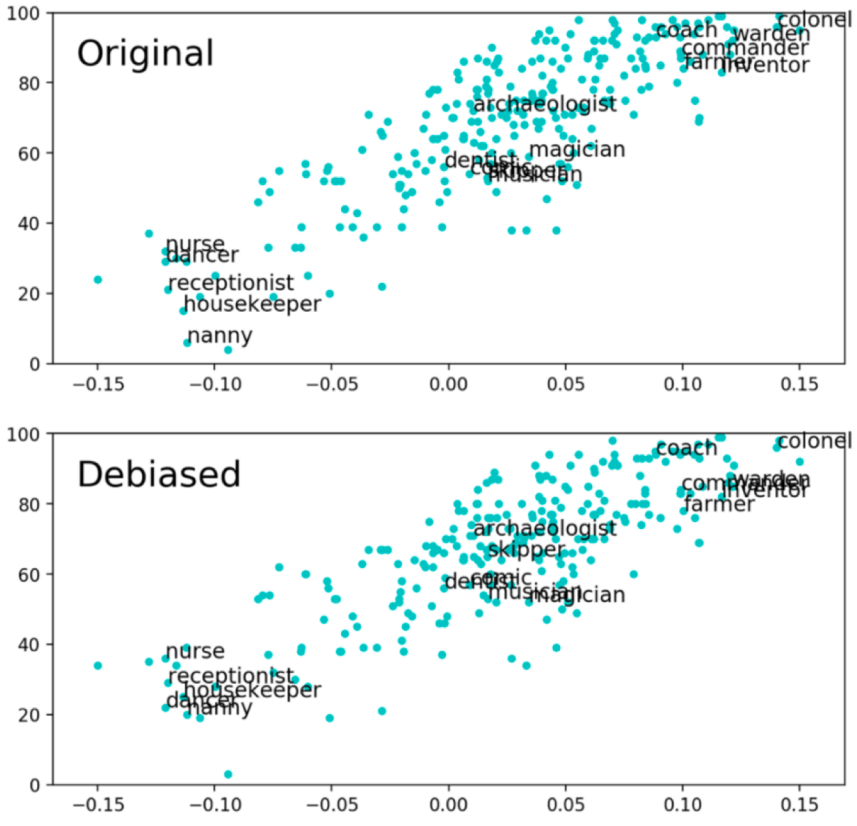
- New measurement on bias
  - Fraction of socially-biased words among neighbors
  - Correlation between different measures

	Hard-Debiased	GN-Glove
Original	0.820	0.747
Debiased	0.792	0.606

Experiments



(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.

Experiments

- Repeat WEAT
  - Female and male names
  - Family words and career words (FC)
  - Arts words and math words (AM)
  - Arts words and science words (AS)

	Hard-Debiased	GN-Glove
FC	0	7.7e-5
AM	0.00016	0.00031
AS	0.0467	0.0064

Experiments

- Classifier on gendered words
  - 5000 most biased words
  - SVM trained on 1000 words and evaluate on the rest

	Hard-Debiased	GN-Glove
Original	98.25%	98.65%
Debiased	88.88%	96.53%