Lipstick on a Pig:

Debiasing methods cover up systematic gender biases in word embeddings but do not remove them

Hila Gonen, Yoav Goldberg

Presented by:
Ankit Pant 2018201035
Tarun Mohandas 2018201008
Team: *The Lost Linguists*

Outline

Introduction

Introduction

- Word Embeddings (WE) - crucial component in for NLP
- WEs have been proven to reflect social biases
- The paper focuses on Gender Bias
- Some work done to reduce gender bias in WE
    - Debiasing in post processing step (Bolukbasi et al.)
    - Debiasing during training (Zhao et al.)
- Paper argues that
    - Current debiasing methods mostly hide bias
    - Lot of bias information can be recovered even after debiasing

Gender Bias in Word Embeddings

Word Embeddings and Gender Bias

- Word Embeddings
  - Most popular representation of document vocabulary
  - Captures context of words
  - Various models include *Word2Vec* and *GLoVe*
  - Both use neural network to form word representations
- Gender Bias in Word Embeddings:
  - Gender bias of a word *w* is its projection on the “gender direction”
  - $\vec{w} \cdot (\vec{he} - \vec{she})$ (normalised)
  - Larger projection of $\vec{w}$ on $(\vec{he} - \vec{she}) \implies$ larger bias

Existing Debiasing Methods

- Post-processing debiasing (Bolukbasi et al.):
  - Make change to word vector to reduce encoded gender bias
  - Done by zeroing the gender projection of each word on a predefined gender direction
  - $\vec{W} = \frac{(\vec{w} - \vec{w_b})}{||(\vec{w} - \vec{w_b})||}$
  - Ensure all neutral words are equally close to the two words
- Train word embeddings from scratch (Zhao et al.)
  - Alter the loss of GloVe model
  - Concentrate most gender information to last coordinate of each vector
  - Use word representation excluding the gender coordinate
  - Representation of gender neutral words orthogonal to gender direction

Remaining Bias after using Debiasing methods

- Both methods – good evidence of removing gender bias
- However, they rely on similar specific bias definition:
    - “no gender bias if each non-explicitly gendered word in the vocabulary is in equal distance to both elements of all explicitly gendered pairs” (Bolukbasi et al.)
- Paper argues that bias is still reflected in similarities between “gender-neutral” words
- **Observation:** “most word pairs maintain their previous similarity, despite their change in relation to the gender direction.”
- **Implication:** Words having specific bias are grouped together

Experimental Setup

Experimental Setup

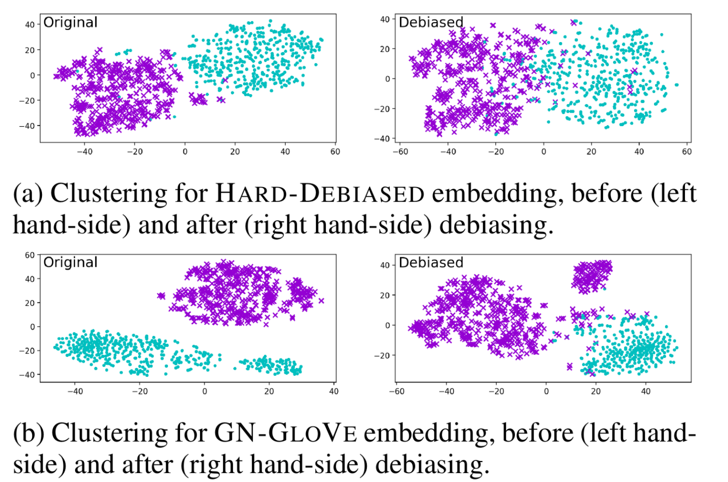
- Approach consists of two steps:
  - For hard-debiasing (Bolukbasi et al.): compare to embeddings before applying debiasing
  - For GN-GloVe (Zhao et al.): compare to embeddings trained with standard GloVe
- Vocabulary:
  - Hard-debiased: 26,189 words
  - GN-GloVe: 47,698 words
- Bias is computed for a word by taking its projection on gender direction $\vec{he} - \vec{she}$,

Experiments and Results

Clustering of Male and Female Biased words

- Most biased words in the vocabulary taken
- Total 100 words taken (500 male-biased, 500 female-biased)
- Clustered into two clusters using *k-means*
- For hard-debiasing, alignment accuracy:
    - 99.9% in original biased dataset
    - 92.5% in debiased dataset
- For GN-GloVe, alignment accuracy:
    - 100% in original biased dataset
    - 85.6% in debiased dataset

Clustering of Male and Female Biased words



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

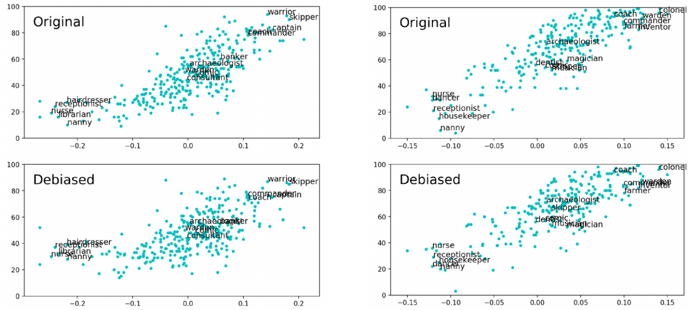
(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

Figure 1: Result of clustering on male and female biased words

Bias-by-neighbours

- Clustering indicates that bias cannot be observed directly
- Social bias associated with a word cannot be observed directly in the new embeddings
- Can be approximated using the gender-direction in non-debiased embeddings
- **New mechanism**: percentage of male/female socially-biased words among the k-nearest neighbours of the target word

- Considered list of professions used by (Bolukbasi et al.)



(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: Result of clustering on male and female professions

Classifying biased words

- Attempt to find if a classifier can be trained to generalise from some gendered words to others
- Vocabulary: 5000 words (2500 per gender)
- Trained SVM classifier on 1000 random words of vocabulary (500 per gender)
- Predict gender for remaining 4000 words
- Prediction Accuracy for heard-debiasing:
  - 98.25% for non-debiased data
  - 88.88% for debiased data
- Prediction Accuracy for GN-GloVe:
  - 98.65% for non-debiased data
  - 96.53% for debiased data

Conclusion

Conclusion i

- A systematic bias is found in word embeddings even after traditional debiasing
- Words with strong bias cluster together
- Words having implicit gender will tend to group with other gender-implicit words
- Implicit gender of words with prevalent previous gender bias can be predicted from vectors alone
- Debiasing methods removes gender direction but it is superficial

Conclusion ii

- Algorithmic discrimination is more likely to happen by associating one implicitly gendered term with other implicitly gendered terms
- Gender-direction measures gender-association of a word but does not determine it
- Popular definitions for quantifying and removing bias are not sufficient

References

References

- Hila Gonen, Yoav Goldberg, *Lipstick on a Pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them*, https://arxiv.org/pdf/1903.03862.pdf
- Tolga Bolukbasi, et al., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, https://arxiv.org/pdf/1607.06520.pdf
- Jieyu Zhao, et al., *Learning Gender-Neutral Word Embeddings*, https://arxiv.org/pdf/1809.01496.pdf