Gender Bias in Machine Learning

Methods for Removing Social Bias in Word Embeddings

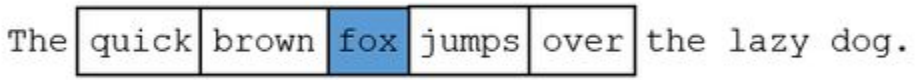
Presented by

Aditya Singh, Nidhi Tiwari, Frank Tranghese

BU Department of Electrical & Computer Engineering

Word Embeddings

- NLP Unsupervised learning that represents words as vectors
- Creation based on dimensionality reduction of co-occurrence matrix
  - N-gram, Skip-gram
- Process can also involve neural nets (e.g. Word2Vec[1]) and probabilistic modeling (GloVe[2]).
- Commonly used in resume screening and web searching/recommendations [3]



(http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/)

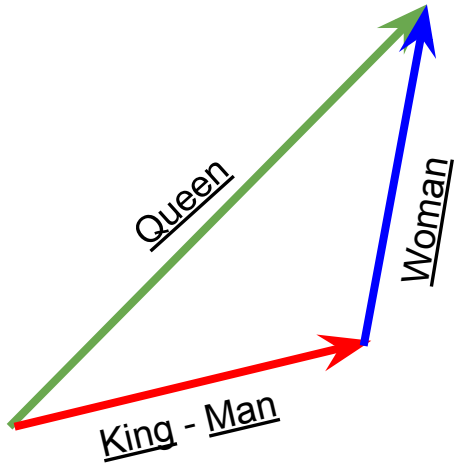
1: Mikolov, Tomas *et al.* (2013). "Distributed Representations of Words and Phrases and their Compositionality"

2: J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation" Computer Science Department, Stanford University

3: C. Hansen *et al.* "How to get the best word vectors for resume parsing," in SNN Adaptive Intelligence / Symposium: Machine Learning 2015.

Representing Words as Vectors

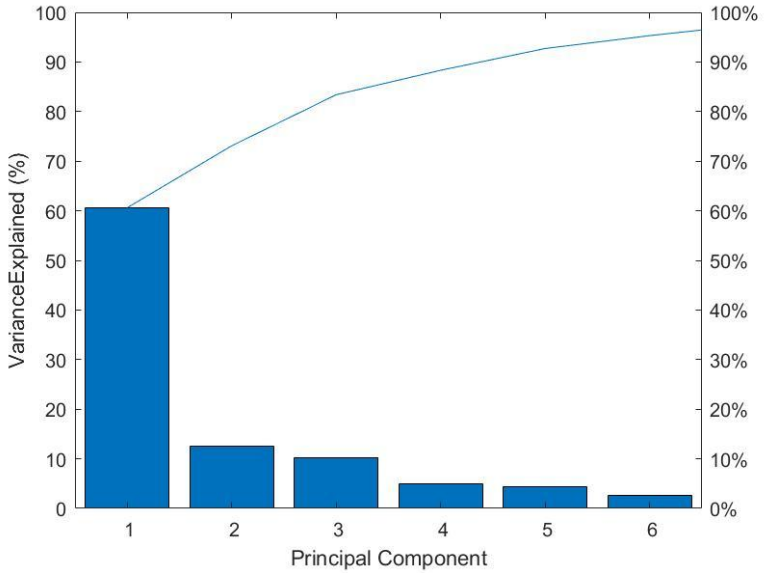
- In word embedding vector space model, direction of a word vector is representative of its semantic meaning
- Relationships and similarity between words can be expressed via vector space math.
- Cosine Similarity now represents how close in meanings two words are



Word Embedding Vector Math

Defining the Gender Subspace

- We create a matrix of gender-defining vector subtractions (he - she, father - mother, etc)
- Performing PCA, the largest component is significantly larger than all others, and accounts for most of the variance
- We define this component as the gender direction g_{dx1}



From: T. Bolukbasi *et al.*. "Man is to computer programmer as women is to homemaker? Debiasing Word Embeddings," *2016*. Advanced in Neural Information Processing Systems

Quantifying Gender Bias

- Direct Bias
  - Definition: The association of a gender-neutral word with a gender-specific word (ex: receptionist to woman, CEO to man).

$$\text{DirectBias}_c = \frac{1}{|N|} \sum_{w \in N} |cos(\overline{w}, g)|^c$$

- Indirect Bias
  - Definition: The association of a gender-neutral word with another gender-neutral word based on their individual association with a gender (ex: football to CEO).

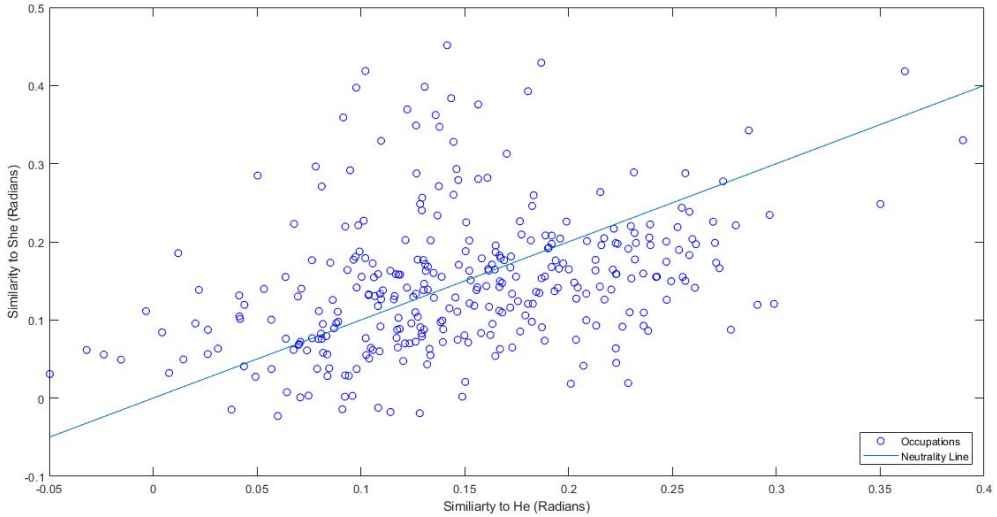
$$\beta(w, v) = (w \bullet v - \frac{w_\perp . v_\perp}{||w_\perp||_2 \, ||v_\perp||_2}) / w \bullet v$$

From T. Bolukbasi *et al.*. "Man is to computer programmer as women is to homemaker? Debiasing Word Embeddings," *2016*. Advanced in Neural Information Processing Systems

Data Preprocessing

- Google News Corpus
  - Pre-trained, 3 million words, 300 feature dimensions
  - Filtered to 26,423 most frequently occuring words (lowercase, less than 20 alphabetical characters)
- 20 NewsGroup Dataset
  - Filtered to the top 6000 frequently occuring words and selecting 30 feature dimensions to train the word2vec model.
  - Obtained the co-occurrence matrix and the trained model from a python script using the ***gensim*** library.
  - Necessary for GloVe method testing due to high dimensionality of Google News dataset. Co-occurrence matrix and retraining computationally intensive.

Existence of Bias in Word Embeddings



Occupations Similarity to He and She

	Occupations	Gender-Specific
Direct Bias	0.0706	0.1552

Average Direct Bias (c=1)

	Football	Softball
Nurse	-0.3976	0.3183
Scientist	0.0764	-0.2306

Indirect Bias Examples

Comparison Term	Avg. Indirect Bias
Football	-0.0210
Softball	0.0624

Average Indirect Bias for all Occupations

Methods We've Explored

- Post-Processing
  - Schmidt Method
  - Soft Debias Method
  - Hard Debias Method
- Pre-Processing
  - GloVe Method

Schmidt Method[1]

- Once the gender subspace, B, is defined, we remove the component in the gender direction from every word in the dataset W_{dxn}
- The aim is to remove the gender subspace entirely, meaning every word in the set will not longer have any association from gender
- This also means gender-specific words (i.e. Mother, Father) will also lose its inherent meaning.

$$\overline{w} = (\overline{w} - \overline{w}_B)^T / \| \overline{w} - \overline{w}_B \|$$

1: B. Schmidt. “Rejecting the gender binary: a vector-space operation,” October 30, 2015. [Online]. Available: http://bookworm.benschmidt.org

The Soft De-bias Method[1]

- Minimize a linear transform T_{dxd} such that the meaning (i.e. dot product) between all words in the set W_{dxn} remain as similar as possible while the the Gender-Neutral words, $N \subseteq W$, as orthogonal as possible to the gender Bias direction, B_{dx1}.

$$\min_T \left|\left|(TW)^T(TW) - W^T W\right|\right|^2_F + \lambda \left|\left|(TN)^T(TB)\right|\right|^2_F$$

- This is a Semidefinite Program, we solved using CVX software for MATLAB (available for free at http://cvxr.com/cvx/)

1 T. Bolukbasi *et al.*. "Man is to computer programmer as women is to homemaker? Debiasing Word Embeddings," *2016*. Advanced in Neural Information Processing Systems

The Hard De-bias Method[1]

- For gender neutral words, we remove the component in the gender direction:

$$\overline{w} = (\overline{w} - \overline{w}_B)^T / \| \overline{w} - \overline{w}_B \|$$

- We then rescale the gender specific words such that words outside of this group have equal distance to the respective genders:

$$\mu := \sum_{w \in E} w/|E| \ , \ v := \mu - \mu_B$$

$$\overline{w} := v + \sqrt{1 - ||v||^2} \ (\overline{w}_B - \mu_B) / \| \overline{w}_B - \mu_B \|$$

1 T. Bolukbasi *et al.*. "Man is to computer programmer as women is to homemaker? Debiasing Word Embeddings," *2016*. Advanced in Neural Information Processing Systems

The GloVe Method[2]

- If a word is gender-neutral, it should have the same probability of co-occurring with words representing both genders
- This method rescaled co-occurrence matrix to reflect this before training

i = “the he word”

j = “the she word”

K = occupation word (to be scaled)

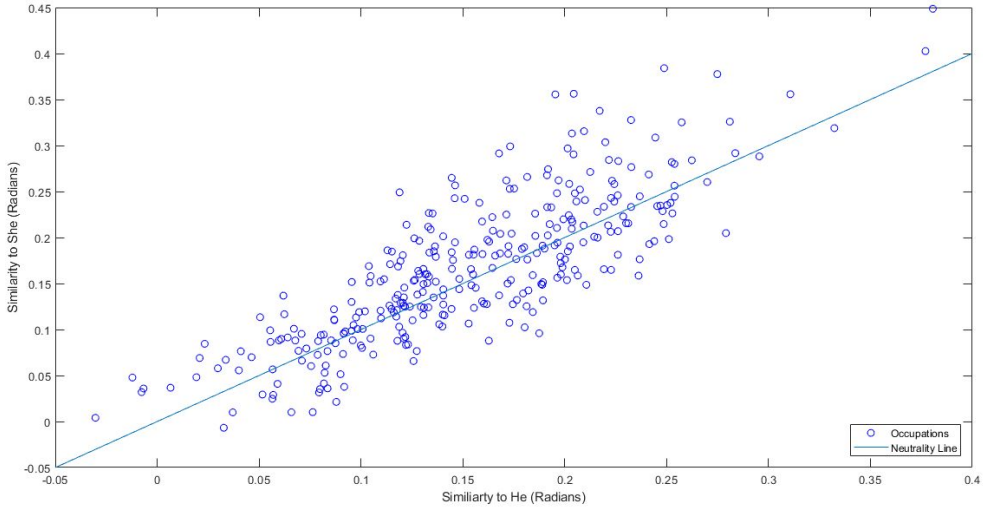
Xi,Xj = Sum of row containing Xij,Xji

Beta = Scaling Factor

$$\beta_{ik} = \frac{X_{ik}+s}{X_{ik}}, \quad \beta_{jk} = \frac{X_{jk}-s}{X_{jk}} \quad \text{here } s = \frac{X_i X_{jk} - X_j X_{ik}}{X_i + X_j}$$

1: J. Pennington, R. Socher, C. D. Manning, “GloVe: Global Vectors for Word Representation” Computer Science Department, Stanford University, Stanford, CA 94305
2: T. Chakraborty *et al*. “Removing Gender Bias in Word Embedding”. Computer Science Department, Stanford University[Online]. Available at: http://cs229.stanford.edu/proj2016/report/

Results: Schmidt Method



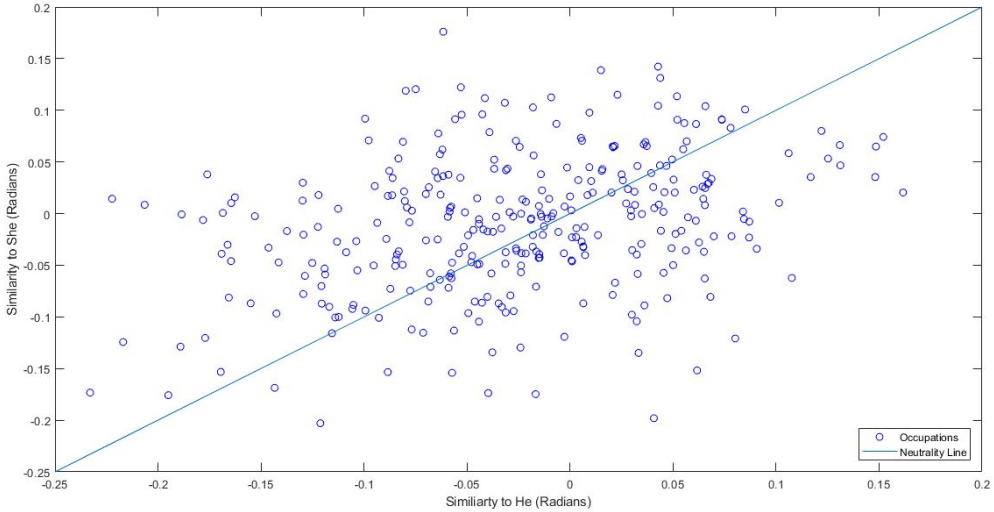
Occupations Similarity to He and She

	Occupations	Gender-Specific
Direct Bias (c=1) (before bias corr.)	0.0706	0.1737
Direct Bias (c=1) (after bias corr.)	0.0040	0.0090

Comparison Term	football		softball	
	before	after	before	after
Nurse	-0.3976	-9.3e-04	0.3183	0.0016
Scientist	0.0764	2.64e-04	-0.2306	-6.10e-04

Comparison Term	Avg. Indirect Bias
Football	-1.2515×10^{-5}
Softball	6.5727×10^{-4}

Results: Soft De-bias



Occupations Similarity to He and She

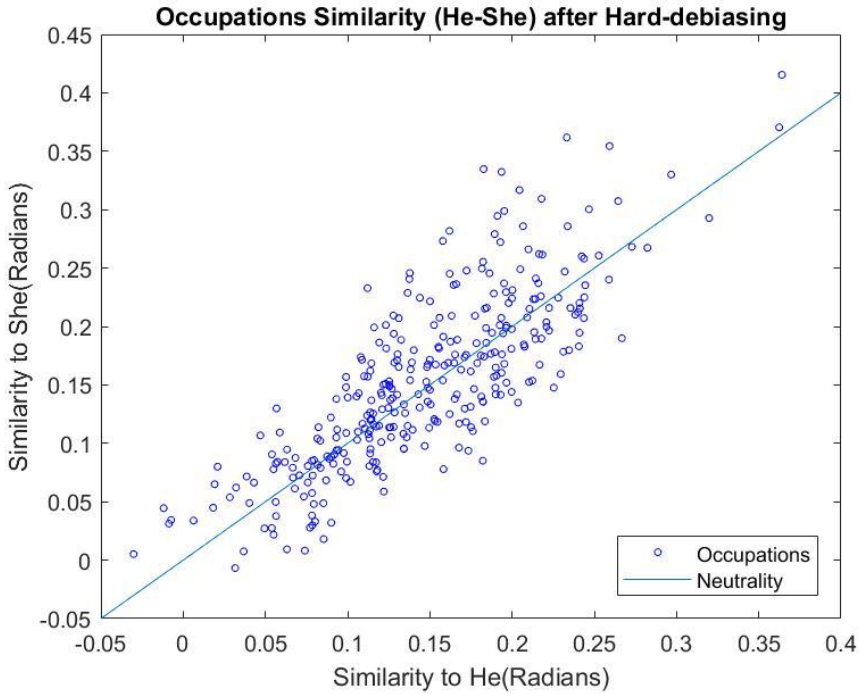
Performed on Largest 50 Components

	Occupations	Gender-Specific
Direct Bias (c=1) (before bias corr.)	0.0706	0.1737
Direct Bias (c=1) (after bias corr.)	0.0396	0.0469

Comparison Term	football		softball	
	before	after	before	after
Nurse	-0.3976	0.0039	0.3183	-0.0090
Scientist	0.0764	0.0092	-0.2306	0.0083

Comparison Term	Avg. Indirect Bias
Football	-0.0022
Softball	-0.0269

Results: Hard De-bias Method



	Occupations	Gender-Specific
Direct Bias (c=1) (before bias corr.)	0.0706	0.1737
Direct Bias (c=1) (after bias corr.)	0.0040	0.0604

Comparison Term	football		softball	
	before	after	before	after
Nurse	-0.3976	-0.0231	0.3183	0.0057
Scientist	0.0764	-0.0029	-0.2306	-0.0344

Results: GloVe Method

- Very recently got new dataset working. Have yet to re-embed the words into vectors for testing in vector space.
- Due to smaller dataset, co-occurrence matrix very sparse.

	Probability Ratio Before	**Probability Ratio After**
Scientist	2.757	1.000

Overview and Conclusions

- Hard debias lowered gender-neutral bias while maintaining some gender-specific meaning with the least computationally intensive method.
- Schmidt's method, while it does get rid of gender meanings completely, may be ideal for applications such as resume parsing, where gender bias should be non-existent.
- Soft-debiasing may have better performance given access to proper resources and deserves further exploration. For similar reasons, GloVe method also requires further exploration.

Questions?