# Gender-preserving Debiasing for Pre-trained Word Embeddings

Author: Masahiro Kaneko, Danushka Bollegala
Presenter: Haonan Duan

**CS 886: Deep Learning and Natural Language Processing**
**Winter 2020**

# Outline

- Gender bias in word embeddings
- Introduction to the proposed method
- Evaluation

Man is to Computer Programmer as Woman is to Homemaker (Bolukbasi, 2016)

# How do we measure Occupational Stereotypes in a word embedding?

The cosine similarity between the word vector

and the gender directional vector ($\overrightarrow{he} - \overrightarrow{she}$)

# Occupational Stereotypes in word2vec embeddings

| **Extreme *she*** | **Extreme *he*** |
|---|---|
| 1. homemaker | 1. maestro |
| 2. nurse | 2. skipper |
| 3. receptionist | 3. protege |
| 4. librarian | 4. philosopher |
| 5. socialite | 5. captain |
| 6. hairdresser | 6. architect |
| 7. nanny | 7. financier |
| 8. bookkeeper | 8. warrior |
| 9. stylist | 9. broadcaster |
| 10. housekeeper | 10. magician |

We believe that no human should be discriminated on the basis of demographic attributes by an NLP system.

There exist clear legal (European Union, 1997), business and ethical obligations to make NLP systems unbiased (Holstein et al., 2018).

Given the broad applications of pre-trained word embeddings in various down-stream NLP tasks, it is **extremely important** to debias word embeddings before they are applied in NLP systems that interact with and/or make decisions that affect humans.

Ideally, we want a debiasing method for word embeddings that can

- Remove all information related to discriminative bias
- Retain all non-discriminative information for the down streaming NLP tasks
- Do not assume any knowledge about the specific pre-trained word embedding algorithms
- Do not assume the availability or access to the or access to the language resources such as corpora or lexicons that might have been used by the word embedding learning algorithm.

# Main idea

Given a pre-trained set of $d$-dimensional word embeddings $\{w_i\}_{i=1}^{|V|}$, over a vocabulary $V$, we can learn a mapping $E : \mathscr{R}^d \rightarrow \mathscr{R}^l$ that projects the original pre-trained word embeddings to a debiased l-dimensional space.

$E(w)$ is the new vector representation we will use for the later task.

# Partition the vocabulary

Also, it should be noted that removing all gender bias is not necessarily ideal.

For example, one would expect the word ***beard*** to be associated with man and ***skirt*** to be associated with woman.

Gender bias like this is not discriminative and actually would be useful, for example, for a recommendation system

# Partition the vocabulary

Therefore, the proposed method should be able to differentiate between undesirable (stereotypical) biases from the desirable (expected) gender information in words.

# Partition the vocabulary

To achieve this goal, we will first partition the vocabulary $V$ into $4$ sets: $V_m$ (masculine), $V_f$ (feminine), $V_n$ (gender neutral), $V_s$ (stereotypical).

For example, *beard* $\in V_m$, *skirt* $\in V_f$, *desk* $\in V_n$, *nurse* $\in V_s$

Note that $V_m, V_f, V_n, V_s$ are mutually exclusive and $V = V_m \cup V_f \cup V_s \cup V_s$

# Partition the vocabulary

The proposed debiased mapping should satisfy the following four criteria:

- For $w_f \in V_f$, we preserve its feminie properties
- For $w_m \in V_m$, we preserve its masculine properties
- For $w_n \in V_n$, we preserve its gender neutrality.
- For $w_s \in V_s$, we remove its stereotypical bias

# System overview

- Encoder (debiasing function): $E : \mathscr{R}^d \rightarrow \mathscr{R}^l$
- Decoder: $D : \mathscr{R}^l \rightarrow \mathscr{R}^d$
- Feminine regressor: $C_f : \mathscr{R}^l \rightarrow [0,1]$
- Masculine regressor: $C_m : \mathscr{R}^l \rightarrow [0,1]$

In the proposed method, $E, D, C_f, C_m$ are all neural networks. Their parameters will be trained by minimizing the loss function introduced soon.

# Loss function I - $V_f, V_m$

For feminine and masculine words, we require the encoded space to retain the gender-related information.

$$L_f = \sum_{w \in V_f} ||\, C_f(E(w)) - 1 \,||_2^2 + \sum_{w \in V \setminus V_f} ||\, C_f(E(w)) \,||_2^2$$

$$L_m = \sum_{w \in V_m} ||\, C_m(E(w)) - 1 \,||_2^2 + \sum_{w \in V \setminus V_m} ||\, C_m(E(w)) \,||_2^2$$

# Loss function II - $V_n$, $V_s$

For the stereotypical and gender-neutral words, we require that they are embedded into a subspace that is orthogonal to a gender directional vector $V_g$.

How do we define and compute $V_g$ ?

# Loss function II - $V_n, V_s$

We collect a set $\Omega$ of feminine and masculine word-pairs $(w_f, w_m)$, such as, $(he, she), (man, woman)$.
Then $V_g$ is computed as:

$$v_g = \frac{1}{|\Omega|} \sum_{(w_f, w_m) \in \Omega} (E(w_m) - E(w_f))$$

# Loss function II - $V_n, V_s$

We consider the squared inner-product between $v_g$ and the debiased stereotypical or gender-neutral words as the loss $L_g$:

$$L_g = \sum_{w \in V_n \cup V_s} (v_g^T E(w))^2$$

# Loss function III

It is important that we preserve the semantic information encoded in the word embeddings as much as possible when we perform debiasing.

For this purpose, we minimize the reconstruction loss:

$$L_r = \sum_{w \in v} ||D(E(w)) - w||_2^2$$

# Loss function

$$L = \lambda_f L_f + \lambda_m L_m + \lambda_g L_g + \lambda_r L_r$$

Here, $\lambda_f, \lambda_m, \lambda_g, \lambda_r$ are all hyper-parameters

# Implementation Details

- The masculine regressor and the feminine regressor are both implemented as feed forward neural networks with one hidden layer. The sigmoid function is used as the nonlinear activation function.

- Both the encoder E and the decoder D of the autoencoder are implemented as feed forward neural networks with two hidden layers. Hyperbolic tangent is used as the activation function throughout the autoencoder.

# Evaluating debiasing performance

| Embeddings | SemBias | | | SemBias-subset | | |
|---|---|---|---|---|---|---|
| | Definition ↑ | Stereotype ↓ | None ↓ | Definition ↑ | Stereotype ↓ | None ↓ |
| GloVe | 80.2 | 10.9 | 8.9 | 57.5 | 20 | 22.5 |
| Hard-Glove | 84.1 | 9.5 | 6.4 | 25 | 47.5 | 27.5 |
| GN-GloVe | 97.7 | 1.4 | 0.9 | 75 | 15 | 10 |
| AE (GloVe) | 82.7 | 8.2 | 9.1 | $62.5^{\dagger}$ | $17.5^{\dagger}$ | 20 |
| AE (GN-GloVe) | $98.0^{\dagger *}$ | $1.6^{\dagger *}$ | $\mathbf{0.5}^{\dagger *}$ | 77.5 | $17.5^{\dagger}$ | $\mathbf{5}^{\dagger *}$ |
| GP (GloVe) | $84.3^{*}$ | 8.0 | $7.7^{*}$ | $65^{\dagger}$ | $15^{\dagger}$ | 20 |
| GP (GN-GloVe) | $\mathbf{98.4}^{\dagger *}$ | $\mathbf{1.1}^{\dagger *}$ | $\mathbf{0.5}^{\dagger *}$ | $\mathbf{82.5}^{\dagger *}$ | $\mathbf{12.5}^{\dagger *}$ | $\mathbf{5}^{\dagger *}$ |

Table 1: Prediction accuracies for gender relational analogies. $*$ and † indicate statistically significant differences against respectively **GloVe** and **Hard-GloVe**.

**Definition: (man, woman), (waiter, waitress)**
**Stereotype: (doctor, nurse)**
**None: (dog, cat)**

# Preservation of Word Semantics - Analogy Detection

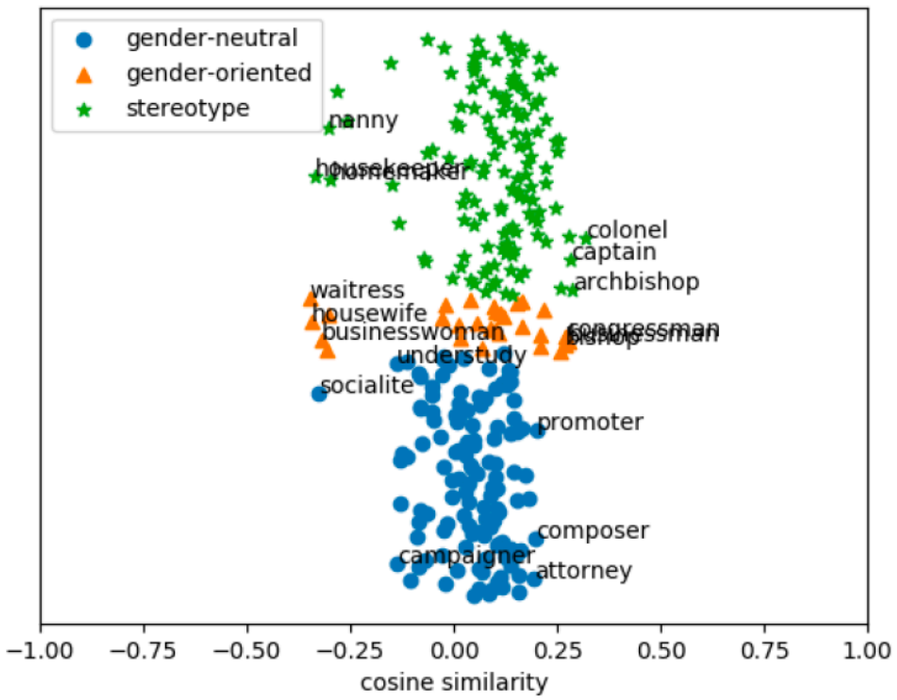
| Embeddings | sem | syn | total | MSR | SE |
|---|---|---|---|---|---|
| GloVe | 80.1 | 62.1 | 70.3 | 53.8 | 38.8 |
| Hard-GloVe | 80.3 | **62.7** | **70.7** | **54.4** | 39.1 |
| GN-GloVe | 77.8 | 60.9 | 68.6 | 51.5 | 39.1 |
| AE (GloVe) | **81.0** | 61.9 | 70.5 | 52.6 | 38.9 |
| AE (GN-GloVe) | 78.6 | 61.3 | 69.2 | 51.2 | 39.1 |
| GP (GloVe) | 80.5 | 61.0 | 69.9 | 51.3 | 38.5 |
| GP (GN-GloVe) | 78.3 | 61.3 | 69.0 | 51.0 | **39.6** |

Table 2: Accuracy for solving word analogies.

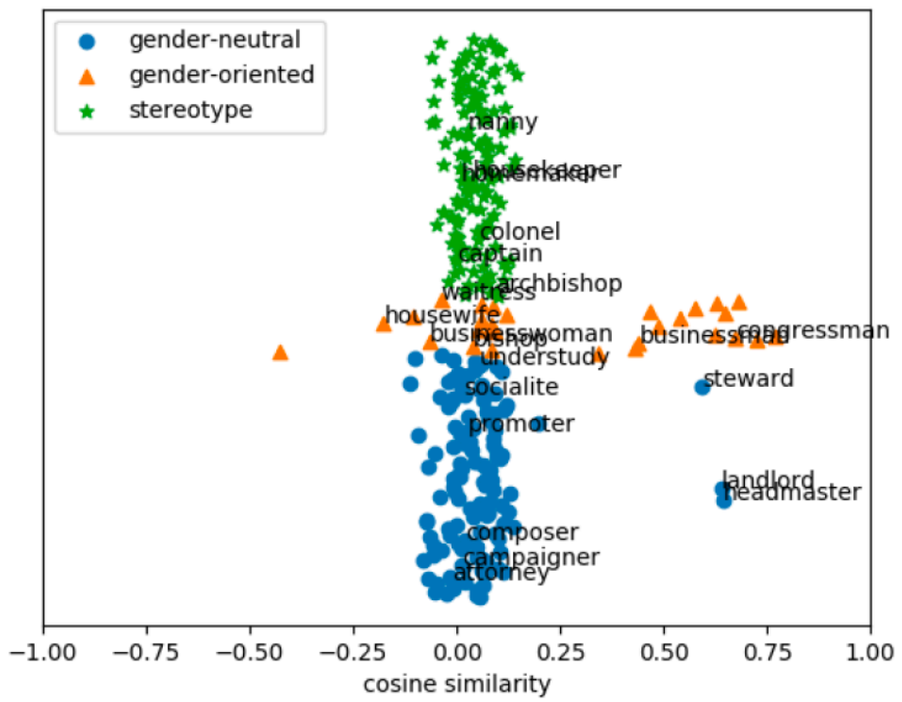
# Preservation of Word Semantics - Semantic Similarity Measurement

| Embeddings | WS | | RG | | MTurk | | RW | | MEN | | SimLex | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Orig | Bal | Orig | Bal | Orig | Bal | Orig | Bal | Orig | Bal | Orig | Bal |
| GloVe | 61.6 | 62.9 | 75.3 | 75.5 | 64.9 | 63.9 | 37.3 | 37.5 | 73.0 | 72.6 | 34.7 | 35.9 |
| Hard-GloVe | 61.7 | 63.1 | 76.4 | 76.7 | 65.1 | 64.1 | 37.4 | 37.4 | 72.8 | 72.5 | 35.0 | 36.1 |
| GN-GloVe | 62.5 | 63.7 | 74.1 | 73.7 | 66.2 | 65.5 | 40.0 | 40.1 | 74.9 | 74.5 | 37.0 | 38.1 |
| AE (GloVe) | 61.3 | 62.6 | **77.1** | **76.8** | 64.9 | 64.1 | 35.7 | 35.8 | 71.9 | 71.5 | 34.7 | 35.9 |
| AE (GN-GloVe) | 61.3 | 62.6 | 73.0 | 74.0 | 66.3 | 65.5 | 38.7 | 38.9 | 73.8 | 73.4 | 36.7 | 37.7 |
| GP (GloVe) | 59.7 | 61.0 | 75.4 | 75.5 | 63.9 | 63.1 | 34.7 | 34.8 | 70.8 | 70.4 | 33.9 | 35.0 |
| GP (GN-GloVe) | **63.2** | **64.3** | 72.2 | 72.2 | **67.9** | **67.4** | **43.2** | **43.3** | **75.9** | **75.5** | **38.4** | **39.5** |

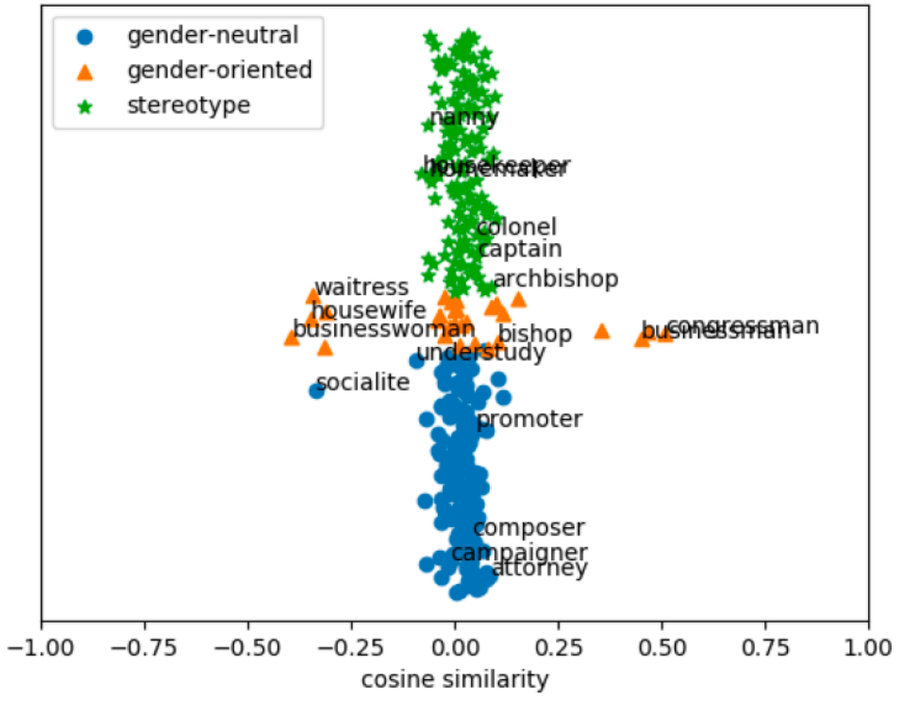
Table 4: Spearman correlation between human ratings and cosine similarity scores computed using word embeddings for the word-pairs in the original and balanced versions of the benchmark datasets.



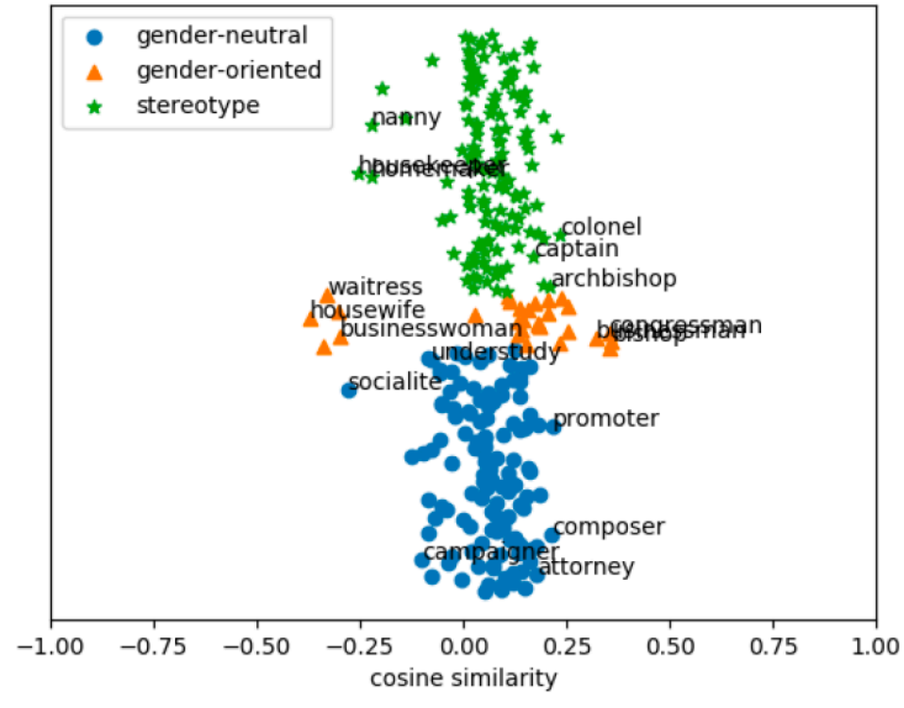

(a) GloVe




(b) GN (GloVe)




(c) Hard-Glove




(d) GP (GloVe)

Figure 1: Cosine similarity between gender, gender-neutral, stereotypical words and the gender direction.

# Reference


- Masahiro, Kaneko, and D. Bollegala. "Gender-preserving Debiasing for Pre-trained Word Embeddings." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

- Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou,Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker?debiasing word embeddings. In NIPS.

- European Union. 1997. Treaty of amsterdam (article 13).

- Kenneth Holstein, Jennifer Wortman Vaughan,Hal Daum´e III, Miro Dud´ık, and Hanna Wallach.2018. Improving fairness in machine learningsystems: What do industry practitioners need?