A Causal Inference Method for Reducing Gender Bias in Word Embedding Relation

Zekun Yang, Juan Feng

Department of Information Systems, College of Business
City University of Hong Kong, Hong Kong SAR, China.
`zekunyang3-c@my.cityu.edu.hk, juafeng@cityu.edu.hk`

香港城市大學
City University of Hong Kong

Abstract

Recent research discovers that gender bias is incorporated in neural word embeddings, and downstream tasks that rely on these biased word vectors also produce gender-biased results. While some word-embedding gender-debiasing methods have been developed, these methods mainly focus on reducing gender bias associated with gender direction and fail to reduce the gender bias presented in word embedding relations. In this paper, we design a *causal* and *simple* approach for mitigating gender bias in word vector relation by utilizing the statistical dependency between gender-definition word embeddings and gender-biased word embeddings. Our method attains state-of-the-art results on gender-debiasing tasks, lexical- and sentence-level evaluation tasks, and downstream coreference resolution tasks.

Gender Bias in Word Embedding

Previous research has discovered and defined two types of gender biases in word vectors: **gender bias associated with gender direction** and **gender bias in word vector relation**.

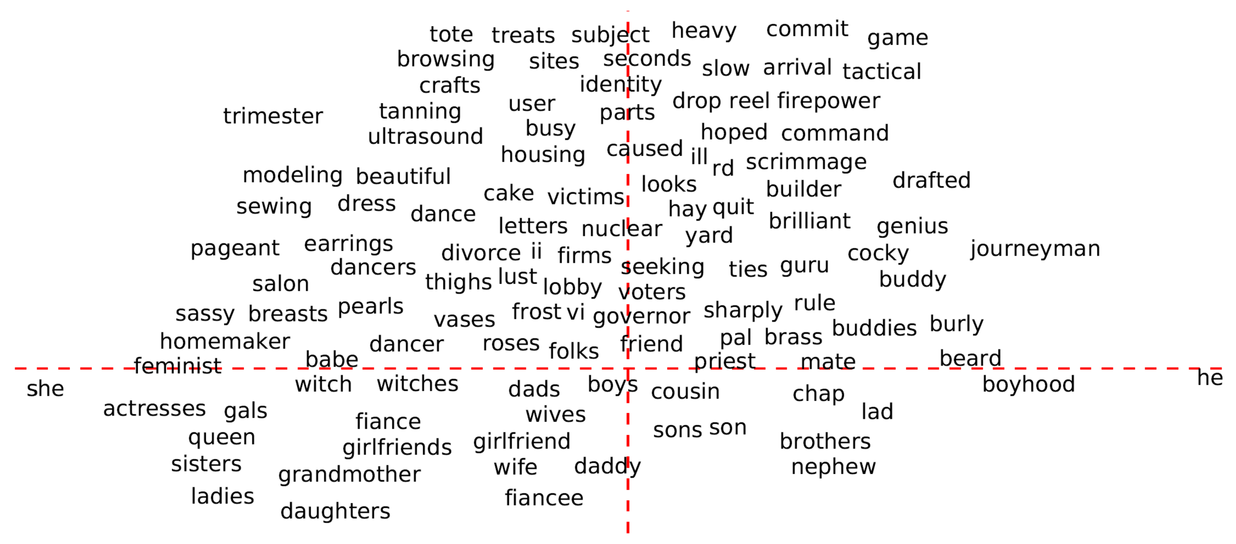


Figure 1: *Gender bias associated with gender direction (reprinted from [1]). This figure shows the projection of word vectors on to the gender direction $\vec{he} - \vec{she}$.*

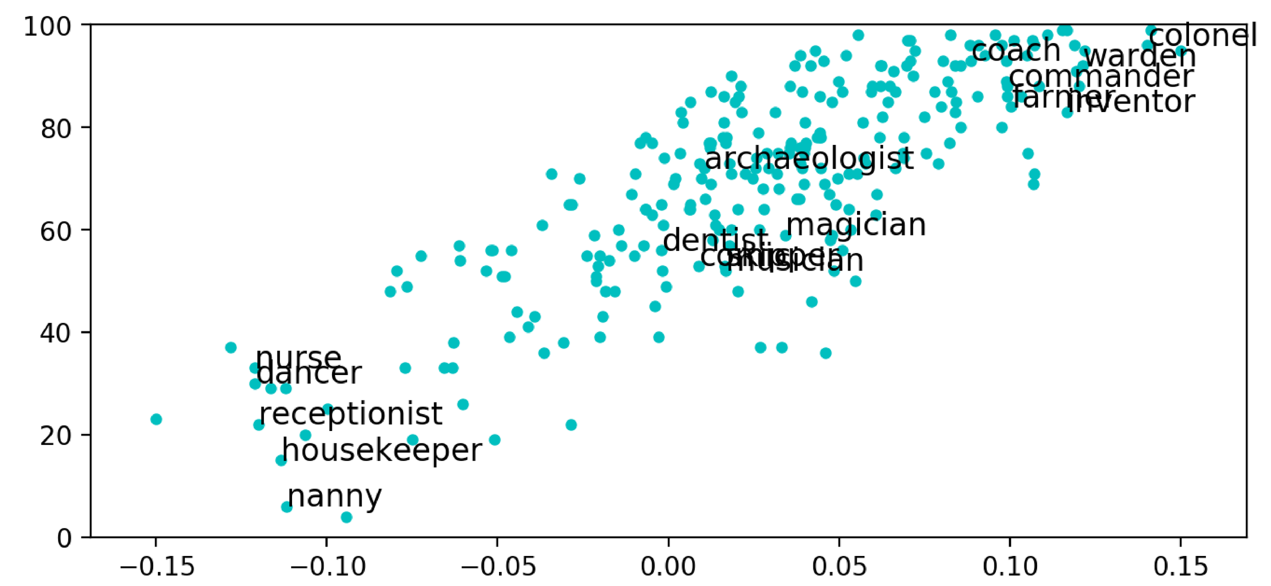


Figure 2: *Gender bias in word vector relation (reprinted from [2]). This figure shows the number of male neighbors for each profession word against its bias-by-projection.*

Causal Gender-Debiasing

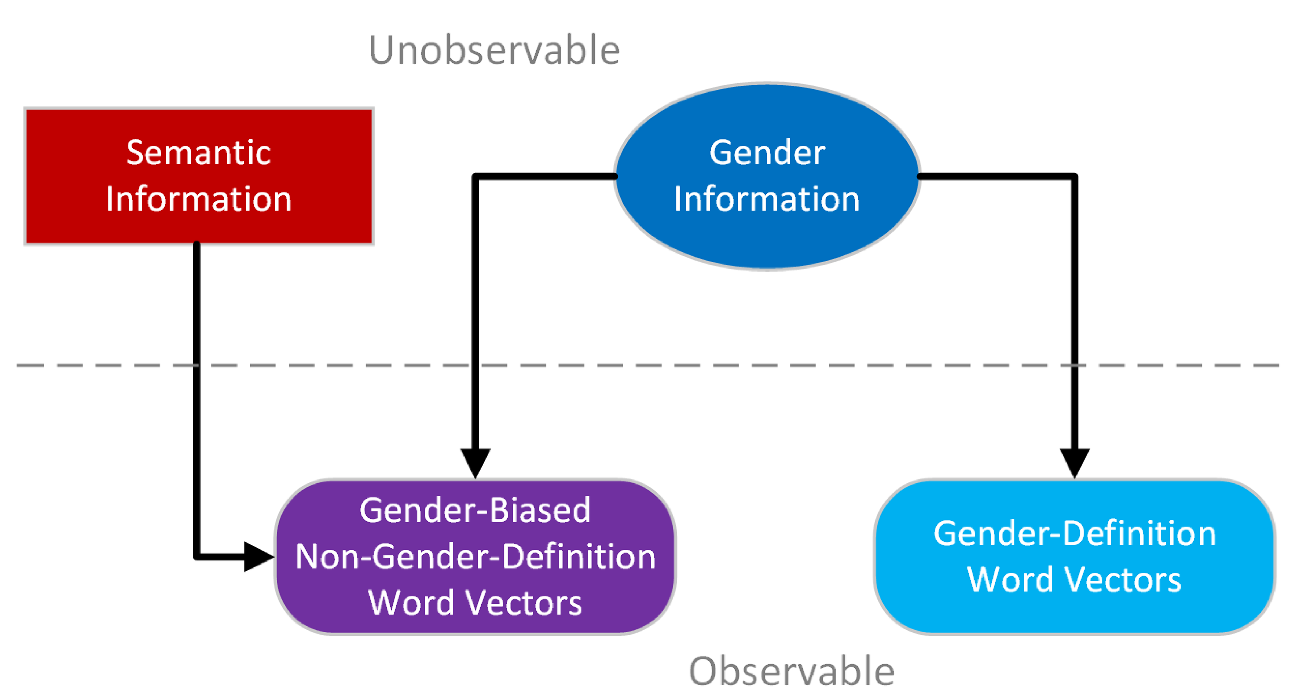


Figure 3: *Relation between gender-definition word vectors and gender-biased non-gender-definition word vectors.*

Based on the half-sibling relationship illustrated in Figure 3, we propose that the debiased non-gender-definition word vectors $\hat{\mathbf{V}}_N$ is learned by subtracting the approximated gender information $\hat{\mathbf{G}}$ from the original non-gender-definition word vectors $\mathbf{V}_N$:

$$\hat{\mathbf{V}}_N \coloneqq \mathbf{V}_N - \hat{\mathbf{G}}, \tag{1}$$

where the approximated gender information $\hat{\mathbf{G}}$ is obtained by predicting $\mathbf{V}_N$ using the gender-definition word vectors $\mathbf{V}_D$:

$$\hat{\mathbf{G}} \coloneqq \mathbb{E}[\mathbf{V}_N | \mathbf{V}_D]. \tag{2}$$

Since $\mathbf{V}_N$ and $\mathbf{V}_D$ embody the same gender information, when predicting $\mathbf{V}_N$ using $\mathbf{V}_D$, the underlying gender information is learned by $\hat{\mathbf{G}}$. Furthermore, as $\mathbf{V}_D$ contains little semantic information apart from the gender information, when approximating $\mathbf{V}_N$ using $\mathbf{V}_D$, the semantic information of $\mathbf{V}_N$ is not learned by $\hat{\mathbf{G}}$. Hence, when we subtract $\hat{\mathbf{G}}$ from $\mathbf{V}_N$, only spurious gender information is eliminated, and the semantic information of $\mathbf{V}_N$ is preserved, which is eventually the gender-debiased word embeddings.

Half-Sibling Regression for Gender-Debiasing

Input: Matrix $\mathbf{V}_D$ of gender-definition word vectors as columns, Matrix $\mathbf{V}_N$ of non-gender-definition word vectors as columns, Ridge Regression constant α.

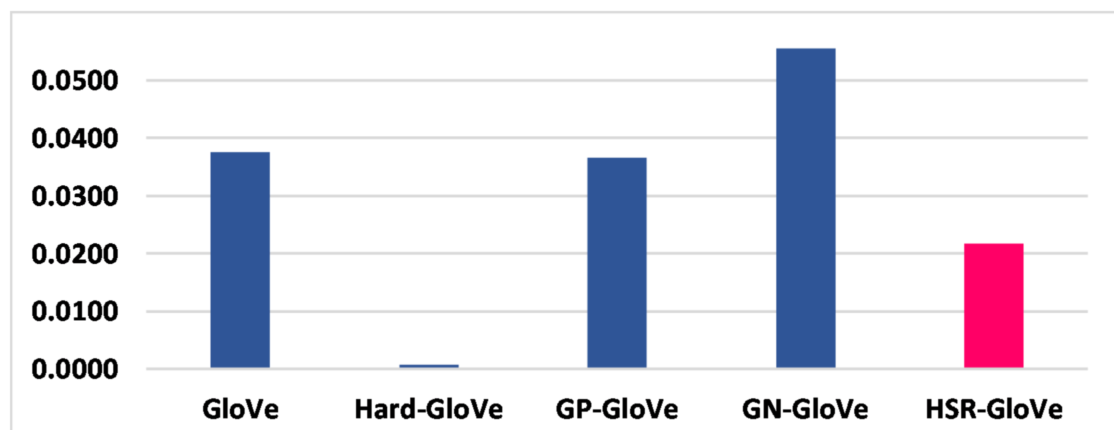
1, Compute the weight matrix of Ridge Regression: $\mathbf{W} \leftarrow \left((\mathbf{V}_D)^\top \mathbf{V}_D + \alpha \mathbf{I}\right)^{-1} (\mathbf{V}_D)^\top \mathbf{V}_N$
2, Compute the approximated gender information: $\hat{\mathbf{G}} \leftarrow \mathbf{V}_D \mathbf{W}$
3, Subtract gender information from the non-gender-definition word vectors: $\hat{\mathbf{V}}_N \leftarrow \mathbf{V}_N - \hat{\mathbf{G}}$

Output: HSR debiased non-gender-definition word vectors $\hat{\mathbf{V}}_N$.

Algorithm 1: *HSR for gender-debiasing*

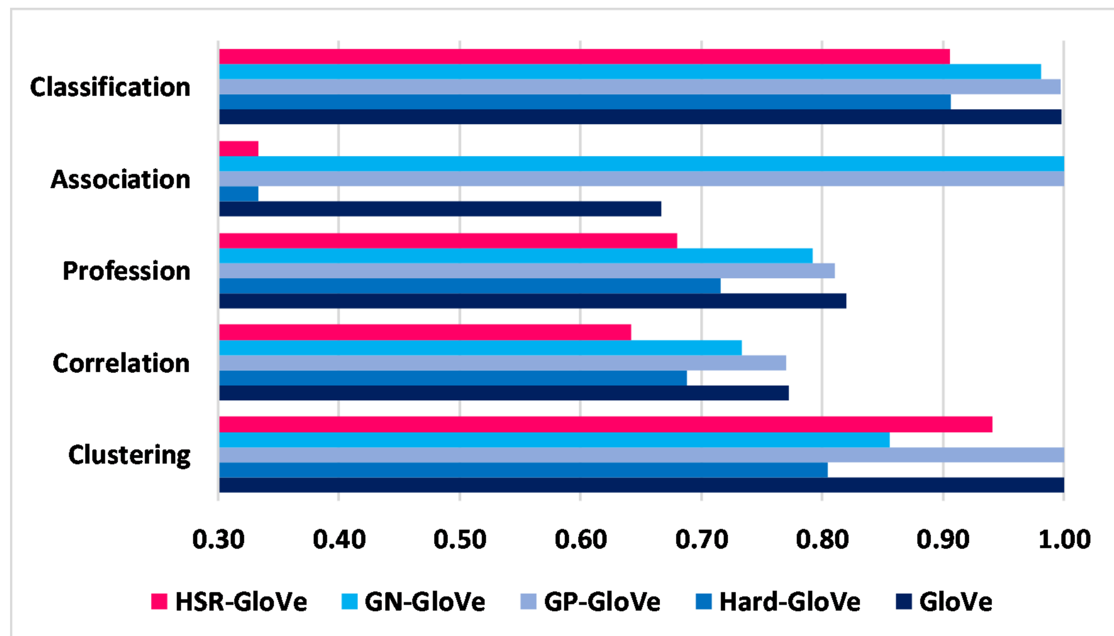
Experiments

Gender Direction Relation Task



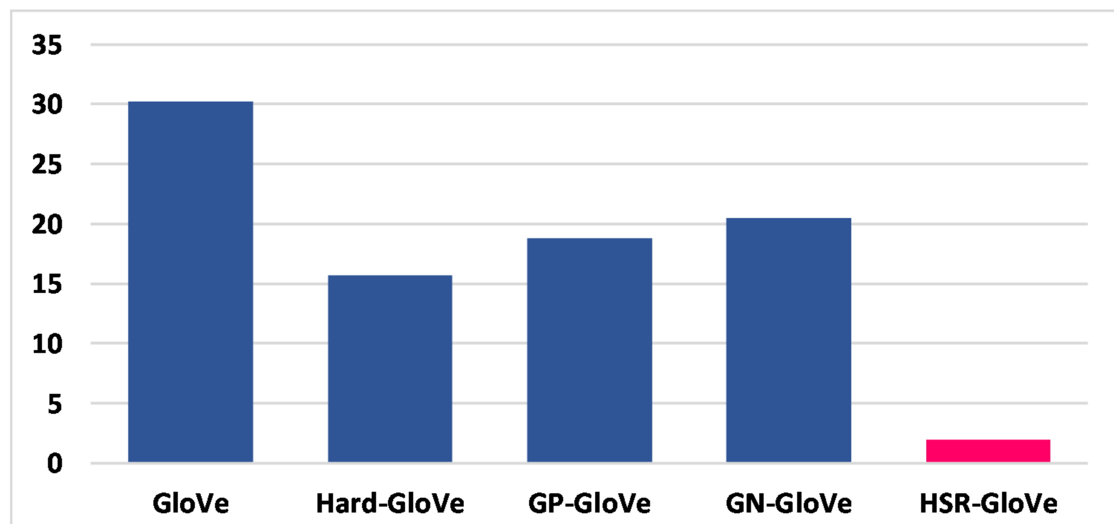
The average absolute bias-by-projection of the embedding of the top 500 male-biased words and the top 500 female-biased words [2]. Bias-by-projection is the dot product between the target word and the gender direction $\vec{he} - \vec{she}$.

Gender-Biased Word Relation Task



The result of five gender-biased word relation tasks proposed by [2]. Smaller results indicate better gender-debiasing performances.

Downstream Task: Gender Coreference Resolution



The difference between the outcomes of WinoBias-PRO and WinoBias-ANTI datasets. WinoBias dataset evaluates the level of gender bias in coreference resolution outcomes [3]. A model passes the WinoBias test when the difference between the outcomes of WinoBias-PRO and WinoBias-ANTI datasets is zero.

References

[1] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 4349–4357, 2016.

[2] H. Gonen and Y. Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the NAACL*.

[3] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2018 Conference of the NAACL*.

Full paper   Code

The 34th AAAI Conference on Artificial Intelligence (AAAI), New York, New York, USA, 2020