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

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

Abstract

Word embeddings are widely used in NLP for a vast range of tasks. It was shown that word embeddings derived from text corpora reflect gender biases in society. This phenomenon is pervasive and consistent across different word embedding models, causing serious concern. Several recent works tackle this problem, and propose methods for significantly reducing this gender bias in word embeddings, demonstrating convincing results. However, we argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between "gender-neutralized" words in the debiased embeddings, and can be recovered from them. We present a series of experiments to support this claim, for two debiasing methods. We conclude that existing bias removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling.

1 Introduction

Word embeddings have become an important component in many NLP models and are widely used for a vast range of downstream tasks. However, these word representations have been proven to reflect social biases (e.g. race and gender) that naturally occur in the data used to train them (Caliskan et al., 2017).

In this paper we focus on gender bias. Gender bias was demonstrated to be consistent and pervasive across different word embeddings. Bolukbasi et al. (2016b) show that using word embeddings for simple analogies surfaces many gender stereotypes. For example, the word embedding they use (word2vec embedding trained on the Google News dataset[1] (Mikolov et al., 2013)) answer the analogy "man is to computer programmer as woman is to x" with "x = homemaker". Caliskan et al. (2017) further demonstrate association between female/male names and groups of words stereotypically assigned to females/males (e.g. arts vs. science). In addition, they demonstrate that word embeddings reflect actual gender gaps in reality by showing the correlation between the gender association of occupation words and labor-force participation data.

Recently, some work has been done to reduce the gender bias in word embeddings, both as a post-processing step (Bolukbasi et al., 2016b) and as part of the training procedure (Zhao et al., 2018). Both works substantially reduce the bias with respect to the same definition: the projection on the gender direction (i.e. $\overrightarrow{he} - \overrightarrow{she}$), introduced in the former. They also show that performance on word similarity tasks is not hurt.

We argue that current debiasing methods, which lean on the above definition for gender bias and directly target it, are mostly hiding the bias rather than removing it. We show that even when drastically reducing the gender bias according to this definition, it is still reflected in the geometry of the representation of "gender-neutral" words, and a lot of the bias information can be recovered.

2 Gender Bias in Word Embeddings

In what follows we refer to words and their vectors interchangeably.

Definition and Existing Debiasing Methods Bolukbasi et al. (2016b) define the gender bias of a word w by its projection on the "gender direction": $\overrightarrow{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$, assuming all vectors are normalized. The larger a word's projection is on $\overrightarrow{he} - \overrightarrow{she}$, the more biased it is. They also quantify the bias in word embeddings using this definition and show it aligns well with social stereotypes.

[1] https://code.google.com/archive/p/word2vec/