Attenuating Bias in Word Vectors

Sunipa Dev
University of Utah

Jeff Phillips
University of Utah

Abstract

Word vector representations are well developed tools for various NLP and Machine Learning tasks and are known to retain significant semantic and syntactic structure of languages. But they are prone to carrying and amplifying bias which can perpetrate discrimination in various applications. In this work, we explore new simple ways to detect the most stereotypically gendered words in an embedding and remove the bias from them. We verify how names are masked carriers of gender bias and then use that as a tool to attenuate bias in embeddings. Further, we extend this property of names to show how names can be used to detect other types of bias in the embeddings such as bias based on race, ethnicity, and age.

1 BIAS IN WORD VECTORS

Word embeddings are an increasingly popular application of neural networks wherein enormous text corpora are taken as input and words therein are mapped to a vector in some high dimensional space. Two commonly used approaches to implement this are Word-ToVec [15,16] and GloVe [17]. These word vector representations estimate similarity between words based on the context of their nearby text, or to predict the likelihood of seeing words in the context of another. Richer properties were discovered such as synonym similarity, linear word relationships, and analogies such as `man` : `woman` :: `king` : `queen`. Their use is now standard in training complex language models.

However, it has been observed that word embeddings are prone to express the bias inherent in the data it is extracted from [3,4,7]. Further, Zhao *et al.* (2017) [18] and Hendricks *et al.* (2018) [6] show that machine learning algorithms and their output show more bias than the data they are generated from.

Word vector embeddings as used in machine learning towards applications which significantly affect people's lives, such as to assess credit [11], predict crime [5], and other emerging domains such judging loan applications and resumes for jobs or college applications. So it is paramount that efforts are made to identify and if possible to remove bias inherent in them. Or at least, we should attempt minimize the propagation of bias within them. For instance, in using existing word embeddings, Bolukbasi *et al.* (2016) [3] demonstrated that women and men are associated with different professions, with men associated with leaderships roles and professions like doctor, programmer and women closer to professions like receptionist or nurse. Caliskan *et al.* (2017) [7] similarly noted how word embeddings show that women are more closely associated with arts than math while it is the opposite for men. They also showed how positive and negative connotations are associated with European-American versus African-American names.

Our work simplifies, quantifies, and fine-tunes these approaches: we show that very simple linear projection of all words based on vectors captured by common names is an effective and general way to significantly reduce bias in word embeddings. More specifically:

1a. We demonstrate that simple linear projection of all word vectors along a bias direction is more effective than the Hard Debiasing of Bolukbasi *et al.* (2016) [3] which is more complex and also partially relies on crowd sourcing.

1b. We show that these results can be slightly improved by dampening the projection of words which are far from the projection distance.

2. We examine the bias inherent in the standard word pairs used for debiasing based on gender by randomly flipping or swapping these words in the raw text before creating the embeddings. We show that this alone does not eliminate bias in

Thanks to NSF CCF-1350888, ACI-1443046, CNS-1514520, CNS-1564287, IIS-1816149, and NVidia Corporation.
Part of the work by JP was done while visiting the Simons Institute for Theory of Computing.

word embeddings, corroborating that simple language modification is not as effective as repairing the word embeddings themselves.

3a. We show that common names with gender association (e.g., `john`, `amy`) often provides a more effective gender subspace to debias along than using gendered words (e.g., `he`, `she`).

3b. We demonstrate that names carry other inherent, and sometimes unfavorable, biases associated with race, nationality, and age, which also corresponds with bias subspaces in word embeddings. And that it is effective to use common names to establish these bias directions and remove this bias from word embeddings.

2 DATA AND NOTATIONS

We set as default the text corpus of a Wikipedia dump (`dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2`) with 4.57 billion tokens and we extract a GloVe embedding from it in $D = 300$ dimensions per word. We restrict the word vocabulary to the most frequent 100,000 words. We also modify the text corpus and extract embeddings from it as described later.

So, for each word in the Vocabulary W, we represent the word by the vector $w_i \in \mathbb{R}^D$ in the embedding. The bias (e.g., gender) subspace is denoted by a set of vector B. It is typically considered in this work to be a single unit vector, v_B (explained in detail later). As we will revisit, a single vector is typically sufficient, and will simplify descriptions. However, these approaches can be generalized to a set of vectors defining a multi-dimensional subspace.

3 HOW TO ATTENUATE BIAS

Given a word embedding, debiasing typically takes as input a set $\mathcal{E} = \{E_1, E_2, \ldots, E_m\}$ of equality sets. An equality set E_j for instance can be a single pair (e.g., {`man`, `woman`}), but could be more words (e.g., {`latina`, `latino`, `latinx`}) that if the bias connotation (e.g, gender) is removed, then it would objectively make sense for all of them to be equal. Our data sets will only use word pairs (as a default the ones in Table 1), and we will describe them as such hereafter for simpler descriptions. In particular, we will represent each E_j as a set of two vectors $e_i^+, e_i^- \in \mathbb{R}^D$.

Given such a set $\mathcal{E}$ of equality sets, the bias vector v_B can be formed as follows [3]. For each $E_j = \{e_j^+, e_j^-\}$ create a vector $\vec{e}_i = e_i^+ - e_i^-$ between the pairs. Stack these to form a matrix $Q = [\vec{e}_1\ \vec{e}_2\ \ldots\ \vec{e}_m]$, and let

{`man`,`woman`}, {`son`,`daughter`}, {`he`,`she`}, {`his`,`her`}, {`male`,`female`}, {`boy`,`girl`}, {`himself`,`herself`}, {`guy`,`gal`}, {`father`,`mother`}, {`john`,`mary`}

Table 1: Gendered Word Pairs

v_B be the top singular vector of Q. We revisit how to create such a bias direction in Section 4.

Now given a word vector $w \in W$, we can project it to its component along this bias direction v_B as

$$\pi_B(w) = \langle w, v_B \rangle v_B.$$

3.1 Existing Method : Hard Debiasing

The most notable advance towards debiasing embeddings along the gender direction has been by Bolukbasi *et al.* (2016) [3] in their algorithm called Hard Debiasing (HD). It takes a set of words desired to be neutralized, $\{w_1, w_2, \ldots, w_n\} = W_N \subset W$, a unit bias subspace vector v_B, and a set of equality sets $E_1, E_2, \ldots, E_m$.

First, words $\{w_1, w_2, \ldots, w_n\} \in W_N$ are projected orthogonal to the bias direction and normalized

$$w_i' = \frac{w_i - w_B}{||w_i - w_B||}.$$

Second, it corrects the locations of the vectors in the equality sets. Let $\mu_j = \frac{1}{|E|}\sum_{e \in E_j} e$ be the mean of an equality set, and $\mu = \frac{1}{m}\sum_{j=1}^{m} \mu_j$ be the mean of of equality set means. Let $\nu_j = \mu - \mu_j$ be the offset of a particular equality set from the mean. Now each $e \in E_j$ in each equality set E_j is first centered using their average and then neutralized as

$$e' = \nu_j + \sqrt{1 - \|\nu_j\|^2}\frac{\pi_B(e) - v_B}{\|\pi_B(e) - v_B\|}.$$

Intuitively ν_j quantifies the amount words in each equality set E_j differ from each other in directions apart from the gender direction. This is used to center the words in each of these sets.

This renders word pairs such as `man` and `woman` as equidistant from the neutral words w_i' with each word of the pair being centralized and moved to a position opposite the other in the space. This can filter out properties either word gained by being used in some other context, like `mankind` or `humans` for the word `man`.

The word set $W_N = \{w_1, w_2, \ldots, w_n\} \subset W$ which is debiased is obtained in two steps. First it seeds some words as definitionally gendered via crowd sourcing and using dictionary definitions; the complement – ones not selected in this step – are set as neutral.