Attenuating Bias in Word Vectors Literature Survey
Paymon Haddad, Raja Sriramoju, Yijing Zhou

The paper "Attenuating Bias in Word Vectors" serves as an extension to the foundational paper "Man is to Computer Programmer as Woman is to Homemaker" (presented in week 3). This new paper provides a brief review of existing methods presented by the foundational paper used to debias word embeddings and then seeks to improve upon those methods. Specifically, the authors present their own debiasing method built upon existing methods and they additionally present a gender subspace defined over pairs of gendered *names* as opposed to more general gender pairs.

DEBIASING TECHNIQUES

The paper quickly reviews the foundational debiasing method known as "hard debiasing". To review, this method defines a gender subspace over crowd sourced gendered words and then ensures all *non*-gendered words are 0 (i.e. orthogonal) w.r.t. that space:

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

Where w_B is the projection of w onto the gender subspace.

It further centers all gender equality pairs along around the gender subspace axis to ensure neutral words are equidistant to each entry in a gender equality set:

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

The paper then presents 3 alternatives to hard debiasing, all of which rely on the definition of some gender/bias subspace. The first is to simply subtract the gender direction from every vector:

$$w' = w - v_B.$$

This method is convenient, but is flawed because it fails to preserve the semantic meaning of words that are definitionally gendered, such as "he" or "grandma".

The authors also highlight another option is to project every word in the corpus orthogonal to the gender subspace:

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

This is a simplified version of hard debiasing that essentially just removes the gender direction from the embedding space. That is to say, if the embedding space is D=300 originally, with a 1D gender subspace the embedding space will span D=299 after the projection. The authors show that this universal projection actually outperforms hard debiasing.

The authors then present "partial projection", their own debiasing method, as a third option. This option softens the orthogonal projection as the norm of the orthogonal projection of the embedding vector increases. That is to say, larger normed embeddings are allowed to keep more of their gendered component w.r.t. the gender subspace.

They start by defining two values for each embedding. The first value intuitively gives some notion of the bias present in the word:

$$\beta(w) = \langle w, v_B \rangle - \langle \mu, v_B \rangle.$$

Where v_B is the gender direction and μ is the average across all *definitionally* gendered words defined by equality sets the same way as done in the hard debiasing paper.

The second value is just the orthogonal component w.r.t. the gender component:

$$r(w) = w - \langle w, v_B \rangle v_B.$$

From here, they propose debiasing according to the rule:

$$w' = \mu + r(w) + \beta \cdot f_i(\eta(w)) \cdot v_B$$

Where

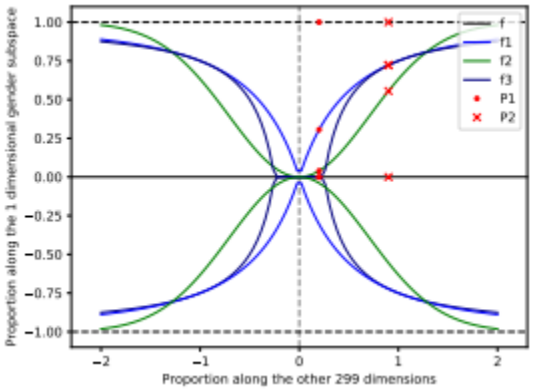
$$\eta(w) = \|r(w)\|$$

And f_i takes one of

$$f_1(\eta) = \sigma^2/(\eta + 1)^2$$
$$f_2(\eta) = \exp(-\eta^2/\sigma^2)$$
$$f_3(\eta) = \max(0, \sigma/2\eta)$$

Empirically, the authors found σ = 1 to be the best choice across the three choices of f.

Intuitively, what this debiasing method does is take the orthogonal component with some constant bias added to it, which is represented by μ + r(w), and then add some additional bias component proportional to the norm of the orthogonal projection of the embedding, which is represented by $\beta f_i(\eta(w))v_B$. This makes sense because all 3 choices of f increases with η(w), which is the norm of r(w), and βv_B represents the bias component of w. Therefore, the greater the norm of the orthogonal component of the bedding embedding, the more bias it is allowed to keep. This is visualized as follows for the different choices of f, which determine how aggressively to tune this bias curve:



The intuition is that words that have large norms are likely to also have large norms in their gender component, which does not necessarily mean that the word itself is biased, but rather that its embedding just has a large norm. We can determine if this is the case using the norm of the

embedding's orthogonal projection. This is because if a word has a small orthogonal projection norm but has a large norm in the gender direction, it should still be heavily neutralized in the gender direction, since this likely means the word is biased.

DEFINING THE GENDER SUBSPACE

The authors observed that when they define a gender subspace according to the method described in the hard debiasing paper, names tended to exert large dot products w.r.t. to the gender direction. Using this observation, they developed a new gender direction using name words as the foundation.

The authors take the top 10 most common male $\{m_1, \dots, m_{10}\}$ and top 10 most common female $\{s_1, \dots, s_{10}\}$ names from a list of 100K words, (excluding names that could potentially have alternate, non-name definitions e.g. "Hope") create random male-female name pairs without replacement, and compute the stop singular vector the same way as done in the hard debiasing paper. However, this was not effective because randomly pairing words does not create a notion of the words being gendered opposites in the same way that using core gender pair words does. For example, the pair "he" "she" can form a pair, because semantically "he" and "she" are opposites. However, "Alison" and "Allen" cannot functionally work as a pair because "Alison" and "Allen" are not strictly speaking opposites.

To combat this, the authors use a simple averaging approach as follows:

$$v_{B,\text{names}} = \frac{s - m}{\|s - m\|},$$

$$\text{where } s = \frac{1}{10}\sum_i s_i \text{ and } m = \frac{1}{10}\sum_i m_i.$$

Intuitively this makes sense because taking the average "female" vector using common names and the average "male" vector using males names and taking the difference between them should provide some notion of a gender direction. The authors state that this gender direction provides similar results to the method described in the hard debiasing paper without the need for any crowdsourcing to define gendered words and equality sets.

In summary, this paper's chief contribution is the introduction of debiasing methods that perform comparably to the original hard debiasing technique, in terms of their ability to maintain the semantic meaning of a word while removing bias, while entirely removing crowdsourcing from the pipeline.