Reducing gender bias in word embeddings

Tuhin Chakraborty (tuhin), Gabrielle Badie (gab47), Brett Rudder (brudder)

Abstract

Word embedding is a popular framework that represents text data as vectors of real numbers. These vectors capture semantics in language, and are used in a variety of natural language processing and machine learning applications. Despite these useful properties, word embeddings derived from ordinary language corpora necessarily exhibit human biases [6]. We measure direct and indirect gender bias for occupation word vectors produced by the GloVe word embedding algorithm [9], then modify this algorithm to produce an embedding with less bias to mitigate amplifying the bias in downstream applications utilizing this embedding.

1 Introduction

Word embeddings represent words as n-dimensional vectors, $\overrightarrow{w} \in \mathbb{R}^n$ as learned from co-occurence data in a large corpus of ordinary language text (news articles, webpages, etc.). A desirable property of these vectors is that they geometrically capture intrinsic relationships between words, making them valuable inputs for applications such as search/result ranking [4] and sentiment analysis [11]. For example, analogies such as *"man* is to *king* as *woman* is to *queen"*, are captured by the equality $\overrightarrow{man} - \overrightarrow{king} \approx \overrightarrow{woman} - \overrightarrow{queen}$ [7].

Recent research[1] into quantifying and mitigating gender stereotypes in word embeddings focuses on reducing bias in pre-trained vectors [2]. We alternatively mitigate bias by updating the GloVe algorithm itself [9], one of the most popular word embedding frameworks. Our modifications focus on gender bias among gender neutral occupation words (doctor, nurse, programmer, etc.) that would otherwise be considered gender-neutral. We tested our changes using the 1 Billion Word Language Model Benchmark [3] dataset as our input data.

2 Identifying and Measuring Gender Bias

Though there is research discussing discrimination bias in various machine learning results in general, there is little literature focused on quantifying and mitigating discrimination bias in word embeddings. To quanitfy bias, we applied the metrics defined by Bolukbasi, et. al.'s foundational work in this area [1], which is based on the broader definition and analysis of direct and indirect discrimination in data-mining defined by Pedreschi et. al. [8]. Before we consider our metrics for direct and indirect bias, we observe the following equality derived from the word vectors trained by GloVe

$$\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computer\_programmer} - \overrightarrow{homemaker}$$

Although neither of the terms *computer_programmer* nor *homemaker* are gendered nouns, the geometry indicates that the word *computer_programmer* is more closely related to the

[1] just presented at NIPS this last month!

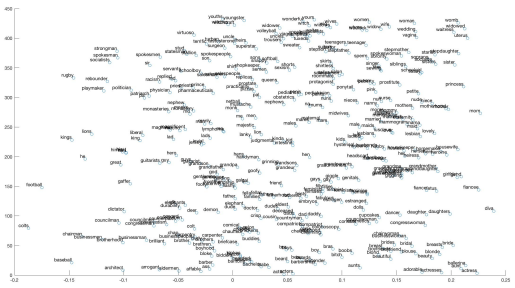

Figure 1: Projection of select words along the *she – he* axis. Words to the left are extreme *he* words. Words on the right are extreme *she* words.

word *he* than it is to the word *she*. To look at potentially stereotyped words we projected all of our words w along the "she-he" axis[2] as $w \cdot \frac{(\vec{he}-\vec{she})}{\|\vec{he}-\vec{she}\|_2}$ and observed this result. Selected words shown in Figure 1.

2.1 Direct Bias

In the specific context of word embeddings, direct bias is measured as the association between a gender neutral word and a gendered word pair such as (she, he), $(sister, brother)$, $(woman, man)$. In order to distinguish between gendered words such as *beard* and *breasts*, and gender neutral words such *mayor*, *seamstress*, and *architect*, Bolukbasi et. al. crowdsource gender analysis of words and word word pairs from random Amazon Mechanical Turkers and have published their results, which we apply directly for our analysis. We define a metric for this in our given word embedding by taking the following 10 'gender pairs', *she:he*, *her:his*, *woman:man*, *Mary:John*, *herself:himself*, *daughter:son*, *mother:father*, *gal:guy*, *girl:boy*, *female:male*, and performing principal component analysis on these different vectors, observing that the top component explains 64.5% of their variance. Given this result, we use this top component (which we label g) as our "gender direction".

Given g, we more quantitatively measure direct bias as

$$DirectBias = \frac{1}{|N|} \sum_{w \in N} |\cos(\vec{w}, g)| \tag{1}$$

which intuitively represents the the average of the cosine similarity between the word vectors in N, a set of gender neutral words (we used gender-neutral occupation words) and the gender direction g.

2.2 Indirect Bias

Targeting words and analogies that express direct bias still excludes indirect biases that may exist in the data, such as the fact that the word *receptionist* is closer to the word $softball$ than it is to the word $football$, which largely derives from the respective relationships between *receptionist* and $softball$ to words such as *she* and *woman*.

To measure this type of indirect bias we first decompose all of our normalized word vectors v into the parts v_g and $v_\perp$, where $v_g = (v \cdot g)g$ represents the gender component and $v_\perp = v - v_g$. We measure indirect bias β between any gender neutral word pair (w, v) as follows

$$\beta(w, v) = \frac{w \cdot v - \frac{w_\perp \cdot v_\perp}{\|w_\perp\|_2 \|v_\perp\|_2}}{w \cdot v}$$

[2]we normalize the word embedding inputs, as is common practice

We looked at β for the words most extremely "softball" and those most extremely "football" when projected on the softball-football direction. This produced the following results for β, which at a glance give us some confirmation of the presence of indirect bias in our word embedding.

top football words	gender portion (%)	top softball words	gender portion (%)
midfielder	1	seamstress	18
captain	4	ballerina	10
footballer	0.01	podiatrist	42
mayor	12	salesperson	35
publisher	24	caregiver	26
architect	18	receptionist	62

3 Methods for reducing bias

First we establish some notation. Let the co-occurrence matrix be denoted by X, whose entries X_{ij} count the number of times word j occurs in the context of word i. Let $X_i = \sum_k X_{ik}$ be the number of times any word appears in the context of word i. Let $P_{ij} = P(j|i) = X_{ij}/X_i$ be the probability that word j appears in the context of word i.

Next, some context for our methodology. The main intuition behind the GloVe algorithm is that co-occurrence probabilities between words encode some kind of meaning in language. In an example given by Pennington et. al. [9], in a 6 billion word corpus we consider $P(k|ice)/P(k|steam)$, with $k \in \{solid, gas, water, fashion\}$. For $k = solid$ we find the ratio to be 8.9, because $solid$ is related to the word ice but not to the word $steam$. For $k = gas$, as we expect, the ratio is low (8.5×10^{-2}), because gas is related to $steam$ and not ice. For words k like $water$ or $fashion$, that are either equally related to both ice and steam, or equally unrelated, we expect the ratio to be closer to one.

This intuition is represented in the weighted least-squares objective of the GloVe algorithm by the ijth entries of the co-occurrence matrix.

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b_j} - log(X_{ij})\right)^2$$

Method 1 - Scaling X_{ij}

Consider, instead of ice and $steam$, examining the gender pair words defined by Bolukbasi. Now we have $P(k|she)/P(k|he)$, where k is a gender neutral occupation word, like $receptionist$. In an unbiased dataset, we expect $receptionist$ to occur as frequently in the context of he as it does in she. The ratio $P(k|she)/P(k|he)$ should then be close to one. When the data is biased this ratio will be greater or less than one. Below we show some examples of biases in co-occurrence probabilities between gender-neutral occupation words and the $he - she$ gender pair.

occupation word	ratio
receptionist	5.301
homemaker	8.628
leader	0.318
boss	0.064

To correct for this, we scale the entries of the co-occurrence matrix in the objective function by some β. Our new objective is as follows.

$$J = \sum_{i,j=1}^{V} f\left(\beta_{ij} X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b_j} - log(\beta_{ij} X_{ij})\right)^2$$

The question, then, is how we compute β? Consider two entries in the co-occurrence matrix, X_{ik} and X_{jk} where i and j are gender pair words like she and he, and k is a gender-

	Direct Bias	MSR-Analogy	RG65	WS353
Baseline	0.113	0.372	0.594	0.430
Method 1: Scaling	0.0551	0.340	0.469	0.418
Method 2: Regularization	**0.0192**	**0.317**	**0.450**	**0.400**

Table 1: Results showing direct bias metric and three word embedding tests

neutral occupation word. We know we want to shift these entries by some s to make them equal. Because *she* and *he* do not necessarily occur with the same frequency in the dataset, this shift needs to be normalized.

$$\frac{X_{ik} + s}{X_i} = \frac{X_{jk} - s}{X_j}$$

.

Notice the sign of s will account for shifts in either direction. Solving for s gives

$$s = \frac{X_i X_{jk} - X_j X_{ik}}{X_i + X_j}$$

Now that we have s, we can compute β as a pair of weights.

$$\beta_{ik} = \frac{X_{ik} + s}{X_{ik}} \qquad \beta_{jk} = \frac{X_{jk} - s}{X_{jk}}$$

Recalling β_{ij} from our new objective function, if i is a female gender pair word like *she*, and j is an occupation word, we have $\beta_{ij} = \beta_{ik}$. If i is a male gender pair word like *he* and j is an occupation word, we have $\beta_{ij} = \beta_{jk}$. Otherwise, $\beta_{ij} = 1$.

Method 2 - Regularization

Recall we measure direct bias by measuring cosine similarity between word vectors and the gender direction, g. We reduce bias in the embedding by adding a regularization term to the objective function that penalizes cosine similarity to the gender direction g. Note that we will only do this for occupation words in the context of our gender pairs.

$$J = \sum_{i,j=1}^{V} f\left(X_{ij}\right)\left(w_i^T \tilde{w}_j + b_i + \tilde{b_j} - log(X_{ij})\right)^2 + \lambda \cos(w_i, g) + \gamma \cos(\tilde{w}_j, g)$$

where λ and γ are weights for the main and context word vectors respectively. We show a sample calculation of one of the derivatives as follows (absorbing the 2 into f):

$$\nabla_{w_i} J = f(X_{ij})\, w_j \cdot \left(w_i^T w_j + b_i + b_j - \log(X_{ij})\right) + \frac{g}{\|w_i\| \cdot \|g\|} - \cos(w_i, g)\frac{w_i}{\|w_i\|^2}$$

4 Results

There are two categories of methods we used to quantify our results. The first is the direct bias measurement described above. The second is a series of three standard analogy tests we used to ensure the word embeddings retained their desirable properties (i.e. that they still do a good job representing the semantics of the English language): contextual correlates of synonymy [10], the WordSimilarity-353 test collection[5], and analogy solving [7].

Overall, the regularization method showed the greatest decrease in direct bias, without a major decrease in performance on the analogy tests. The scaling method showed about a 50% reduction in direct bias without as much loss in performance on the analogy tests.

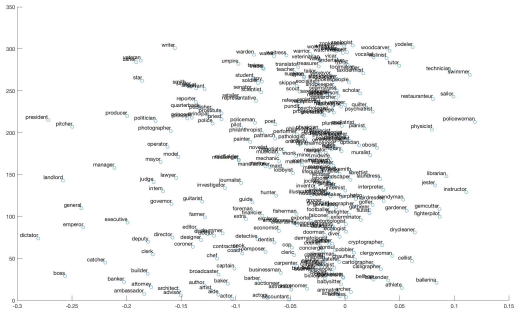

Figure 2: Projection of occupation words along the "she-he" axis after scaling method. Now, occupation words like *homemaker*, *babysitter*, and *footballer* are now gender neutral. Words like *policewoman* retain their inherent gender. Certain words like *president* and *dictator* are still biased, though less so. This "soft" debiasing reduces gender bias but performs better on the analogy tests.

While we do not have a ground truth for indirect bias, we observe some desirable qualitative improvements. Focusing again on occupation words projected on the $softball - football$ direction, the most extremely softball words are now $journeyman$ and *game* which are relevant and do not exhibit gender bias. Stereotypical relationships between softball and words such as $receptionist$, *seamstress*, and *ballerina* have become more neutral on this axis.

In determining the best weights for λ and γ in the regularization method, we found a tradeoff between reduction in direct bias and performance on the analogy tests as the values for λ and γ were increased. We ended up using a value of 10 for each hyperparameter.

In future iterations we would experiment with foregoing g in favor of the $he - she$ direction, or whichever gender direction was being optimized in a particular iteration. This regularization would be more precise, and would likely lead to better results in the analogy tests and lower overall direct bias.

5 Conclusion

We have shown modifications can be made to the GloVe framework to make it more robust to gender bias. We believe this is a promising step to reducing gender discrimination resulting from the use of word embeddings produced today, which exhibit significant, measureable gender bias.

The regularization method was most successful in reducing direct bias, though also had a measurable impact on the analogy scores. If an application of the embeddings requires strict scores on the analogy tests, we would prefer the scaling method, which gives less reduction in direct bias but boasts better performance in the analogy tests.

The efficacy of word embeddings depends on the size of the corpus and the dimension of the vectors. Due to constraints on time and computational power, we worked on a 1 billion word dataset and produced 100-dimensional word embeddings. Given more time and resources, our qualitative and quantitative results could be repeated with a 5 or 6 billion word dataset and 500-dimensional word vectors.

Interesting extensions of this research might focus on other sources of bias in word embeddings, such as racial bias, which have been targeted in other areas of machine learning, or by focusing on other popular word embedding frameworks like word2vec [7].

References

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[3] Mike Schuster Qi Ge Thorsten Brants Ciprian Chelba, Tomas Mikolov. One billion word benchmark for measuring progress in statistical language modeling. *arXiv:1312.3005*, 2013.

[4] Nick Craswell Rich Caruana Eric Nalisnick, Bhaskar Mitra. Improving document ranking with dual word embeddings. In *WWW'16*. WWW World Wide Web Consortium (W3C), April 2016.

[5] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414. ACM, 2001.

[6] Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187, 2016.

[7] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751, 2013.

[8] Turini F Pedreschi D, Ruggieri S. Discrimination-aware data mining. *ACM SIGKDD*, 2008.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[10] Herbert Rubenstein and John B Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[11] Melanie Tosik, Carsten L Hansen, Gerard Goossen, and Mihai Rotaru. Word embeddings vs word types for sequence labeling: the curious case of cv parsing. In *Proceedings of NAACL-HLT*, pages 123–128, 2015.