**Summary**

Word embeddings from human-created corpora inherit significant gender bias that can be further amplified in downstream tasks. Corpus regularities such as word frequency adversely affect the performance of current post-hoc debiasing algorithms.

**Proposition**: Purify the word embeddings against such corpus regularities prior to inferring and removing the gender subspace.

**The Problem: Gender Bias in Embeddings**

Word embeddings: learned vectors representing words; capture semantic + syntactic meanings of words and relations with other words.

**Problem**: Inherit unintended gender bias from training corpus.

A word exposes gender bias when it is gender-neutral by definition but its learned embedding is closer to a certain gender.

**Solution**: debias word embeddings!

**Old approach: Hard-Debias (Bolukbasi et al.)**

**Key idea**: Subtract component associated with gender from word embeddings through post-processing.

**How?** Compute the first principal component of difference vectors of each word pair in a set of gender specific word pairs to get the gender direction in the embedding space.

$$d_{u,v} = \sum_{n=1}^{N} (u_n - v_n) \cdot \phi_i w_i$$

Next, project biased word embeddings into a subspace orthogonal to the inferred gender direction to get rid of gender bias.

$$\psi = \psi + \mu_i$$

The bias subspace $B$ is the first $k$ rows of SVD($C$), where

$$C = \sum_{u,v} (\psi - \mu_i \phi_i)^T (\psi - \mu_i \phi_i)$$

From Bolukbasi et al., $k = 1$, so $B$ is reduced to the gender direction.

Next, transform each word embedding $w$ such that it has zero projection in this gender subspace. Then, re-embed each word as follows.

$$\bar{w} = \bar{w} - \bar{v}_p$$

Bolukbasi et al. demonstrate this method improves gender bias in word analogy tasks.

**Limitations**: Effectiveness limited, as gender bias can still be recovered from the geometry of the embeddings post-debiasing (Gonen et al.).

**Contributions**

Reproduction of Baselines: GloVe, GP-GloVe, GN-GloVe, GP-GN-GloVe, and Hard-GloVe for both English and Spanish corpora.

New Datasets: Investigation of the debiasing effect on Spanish GloVe vectors

New Evaluation Metric: RIPA (Relational inner product association) + generation of Spanish analogy datasets.

**Double-Hard Debias: Eliminate Influence from Frequency**

Mu and Vishwanath and Gong et al. show word frequency significantly impacts geometry of word embeddings, negatively affecting the procedure of identifying the gender direction and degrading ability of Hard Debias for debiasing gender.

**Key idea**: project word embeddings into an intermediate subspace before applying Hard Debias, i.e. all word embeddings are transformed into a frequency-free subspace where a more accurate gender direction can be computed.

**How?** Find the dimension which encodes frequency information that distacts the gender direction computation using clustering of top biased words as a proxy and iteratively test the principal components of the word embeddings.

1. Compute principal components of all embeddings as frequency dimension candidates: $w_i = \langle u_i, v_i \rangle - \langle e_i, w_i \rangle$.
2. Select a set of top biased male and female words.
3. Repeat step 4-6 for each candidate dimension $\phi_i$ independently. ($1 \leq i \leq d$).
4. Project embeddings into an intermediate space which is orthogonal to $\phi_i$ and thus get revised embeddings: $w_i^\prime = w_i - \langle w_i, \phi_i \rangle \phi_i$.
5. Apply Hard Debias on the revised embeddings: $w_i^\prime = HardDebias(w_i^\prime)$.
6. Cluster debiased embeddings from step 5 of the selected top biased words and compute the clustering accuracy: $\phi_i = \text{KMeans}(\{w_i^\prime\}; k)$.
7. Finally, apply Hard Debias on revised embeddings: $\psi = HardDebias(w_i^\prime)$.

If the clustering algorithm in step 6 still clusters biased words into two groups aligned with gender, the $\phi_i$ failed to improve debiasing.

Hence the $u_i$ that leads to the most significant drop in biased word clustering accuracy is removed.

**Baselines**

**GloVe**: pre-trained GloVe embeddings on Wikipedia $i$, non-debiased baseline for comparison.

**GP-GloVe**: debiased GloVe embeddings, attempt to preserve non-discriminative gender information and remove stereotypical gender bias.

**GP-GN-GloVe**: apply Gender-Preserving debiasing to debiased GN-GloVe embeddings.

**Hard-GloVe**: embeddings debiased by the Hard debias method.

**Double-Hard GloVe**: Debia pre-trained embeddings using proposed Double-Hard Debias method.

**Visualisations**

tSNE visualization of top 500 most male and female embeddings

**Evaluation**

Semantic and syntactic accuracies of 80.94% and 61.64% through MSR and Google Word Analogy datasets. Double-hard debias reached similar semantic and syntactic accuracies with its debiased embeddings to GloVe proving that it preserves proximity between among the words. Similarly, concept categorization showed DH-Debias performed similarly to the GloVe embeddings with a score of .795 compared to a .81. There is an insignificant difference in the semantic integrity of the word embeddings.

While we see some of the same semantic and syntactic accuracies, we see a lower Neighborhood Metric score. DH-Debias reaches a score of .665, .742, and .704 compared to the Hard debias scores of .76, .805, and .8025.

An accuracy value closer to 0.5 indicates less biased word embeddings as this metric has k-Means algorithm cluster selected words into a male group and a female group, suggesting the presence of a strong bias.

DH-debias provides the lowest scores from the WEAT test, which is a permutation test used to measure bias in embeddings for different target word sets. It receives scores of 1.531 for Career and Family, -0.094 for Math and arts and -0.149 for Science and arts.

Overall, double-hard debias provides a fair debiasing method that incorporates removing frequency of words to mitigate its effect and reduces the gender bias seen in word embeddings while maintaining the semantic and syntactic integrity of the embeddings.

**Conclusion and future work**

Simple changes in word frequency statistics can have an undesirable impact on the debiasing methods.

Double-Hard Debias demonstrates the negative effects that word frequency features can have on debiasing algorithms.

Although Double-Hard Debias has worse performance on the evaluation metrics, less gender information is encoded in the word embeddings debiased by this method as shown by the proximity of male and female embeddings in the tSNE visualization.

**References**

