Man is to computer as woman is to homemaker? debiasing word embeddings

Aldo Polanco, Aishwarya Varma, Angela Ellis, Darryl York III

Background

The purpose of this research project is to replicate [1].
Word embeddings can be described as objects used to express words for text analysis/generation through **mapping ginormous text corpora into individual word vectors** which can be mathematically manipulated using a tool called **word2vec**. For example:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

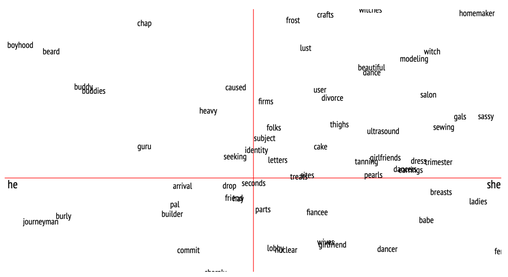

Fig. 1: words in an embedding by gender component.

Research Question

- Can we **quantify the (gender) bias** in a word embedding by generating analogies?
- Are soft and hard debiasing algorithms **effective at removing said bias**?
- Can we debias word embeddings, while **keeping word relationships**?

Methods

There are two methods used by the paper to debias the Google News word2vec:

- **Hard debiasing**
- **Soft debiasing**

We will only use hard debiasing, as soft debiasing was not shown to be effective.

Finding gender

In order to neutralize the gender bias found within the word vectors, we need to find that bias. We do this by conducting **principal component analysis (PCA)**, to find which direction the gender component is in. We establish defininings sets, say {“man”, “woman”} or {“she”, “he”} which capture the difference in gender. We then use PCA to find the direction/vector of most variance in the difference between these defining sets.

Neutralizing gender

Now that we found our **gender direction (B)**, we can reassign every word (that is not definitionally gendered) as follows:

$$\vec{w} := (\vec{w} - \vec{w}_B)/||\vec{w} - \vec{w}_B||$$

Where w is the word vector, and w_B can be defined as the projection of w into the gender direction, B. The projection w_B can then be defined:

$$\vec{w}_B := (\vec{w} \cdot \vec{B})\vec{B}$$

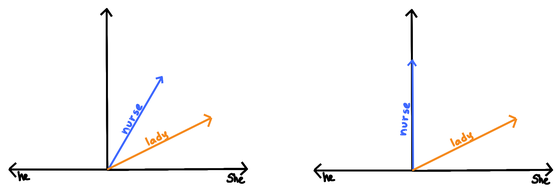

Fig. 2: Visual representation of equalize.

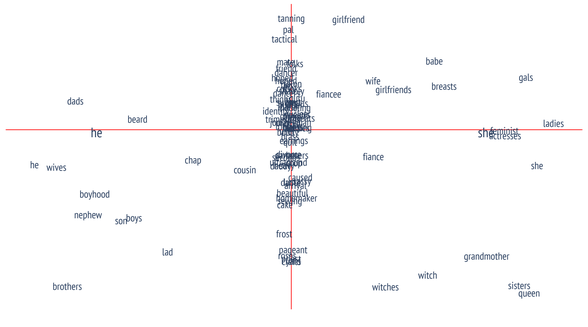

Fig. 3: Fig. 1 after neutralize, note gender-neutral is in the middle to denote equal distance to both genders.

Equalizing gender

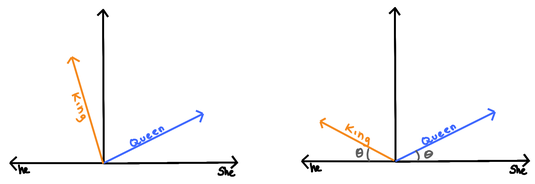

Fig. 4: Visual representation of equalize. Gender-specific words ‘king’, ‘queen’ have equal distance to their gender.

Equalize takes the pairs of words that are gender specific such as **E = {“actor”, “actress”}** and averages their projection onto the gender direction, such that “actor” is just as “male” as “actress” is female. We follow the following formula:

$$\mu := \sum_{w \in E} w/|E|$$

$$v := \mu - \mu_B$$

$$\text{For each } w \in E, \vec{w} := v + \sqrt{1 - ||v||^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||}$$

Results

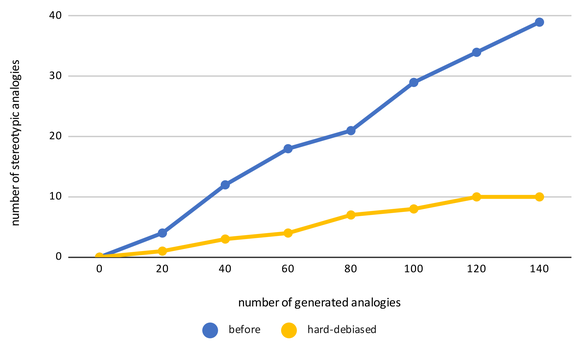

Fig. 5: Analogy analysis results before and after debiasing. Note the results are not free of biased analogies.

Conclusion

Before debiasing:

“she” is to:	as ‘he’ is to:
registered_nurse	physician
feminism	conservatism
queen	king
sewing	carpentry

After debiasing:

“she” is to:	as ‘he’ is to:
registered_nurse	nursing
feminism	fundamentalism
queen	king
sewing	yarn

After debiasing, **we lost a lot of biased analogies**. The biased ones that remained had a worse analogy score than before. It is **possible to debias** using these methods, and the paper’s results are **replicable**.

Limitations

- **Only gender bias** is evaluated, when other kinds are possible.
- We had **limited computing power**, even using Google Collab
- **10 weeks is no where near enough** to fully test the study’s robustness
- Evaluating on **other data sets** would also be helpful, from **other sources**
- Some of the **metrics used were subjective** or context dependent
- We **did not have experience in the field**, allowing us to change certain things in the paper.

References & Acknowledgments

[1] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

Thanks to Dr. Anna Rafferty, the Carleton College Computer Science Department, my teammates, Aishwarya, Darryl and Angela.