# He is to shorts as she is to skirts

## On word embedding debiasing

Mateusz Basiak, Dawid Barzyk

July 2, 2020

## 1 Introduction

Very often usage of machine learning comes along with the risk of amplifying biases present in data. There are many reasons for us to want to get rid of these biases, e.g. turning off unnecessary biases for more results or to customize our model in order to represent it in most suitable way. We examine this problem on the example of word embeddings trained on Google News articles.

## 2 General idea

In word embedding one of desired characteristics is that words represented as vectors keep their features what means that vector differences between words in embeddings have been shown to represent relationships between words. For example we can say that $\overrightarrow{he} - \overrightarrow{she} = \overrightarrow{man} - \overrightarrow{woman}$, similarly we can say that $\overrightarrow{Christian} - \overrightarrow{Muslim} = \overrightarrow{Bible} - \overrightarrow{Quran}$.

Main idea of using this feature is to generate the interesting us subspace. We can do it by considering a set of pairs (or bigger sets) that will define our subspace.

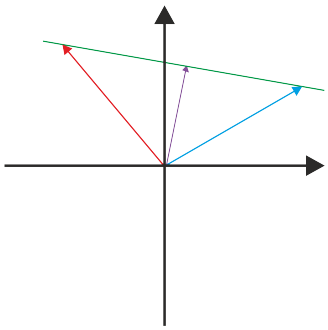


Figure 1: The diagram shows how we would like to interpret each defining set. Blue and red arrows are words that define stereotype, green line is our gender subspace, and purple line is a word (before normalization) we want to transform both blue and red words into

After creating our subspace we will use debiasing only on words that are gender specific, for example: word nurse is mostly connected with women but we can't be sure what is a male stereotype corresponding to being nurse but we would like to debias that word becuase it is gender specific, meanwhile any name of the country is not gender specific so we don't necessarily want to debias it. To do that we need to define set of gender specific words

## 3 Subspace generation

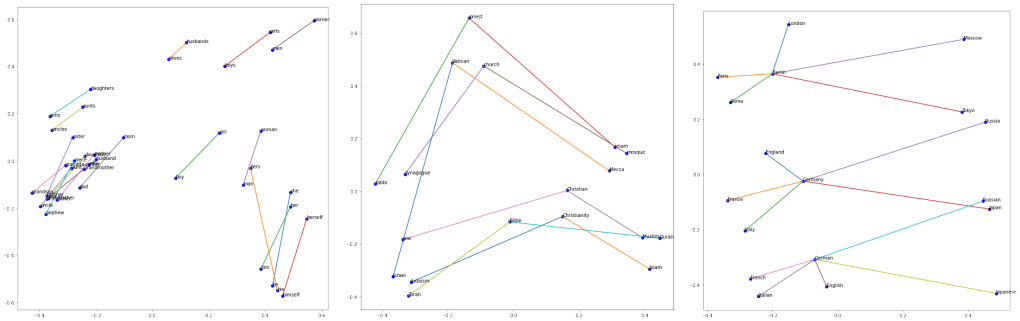
We calculate bias subspace in following way.

1. We create defining sets, so that each set has one representative for each of biases we want to unify.
2. We calculate mean of each defining set ($\mu_i$ is mean of set $D_i$).
3. We calculate the value of matrix $C = \sum_{i=1}^{n} \sum_{w \in D_i} (\overrightarrow{w} - \mu_i)^T (\overrightarrow{w} - \mu_i) / \mid D_i \mid$.
4. Using first k vectors of matrix V of SVD(C) we define our subspace.

## 4 Hard debiasing

This part of algorithm consists of 2 steps. First we neutralize the words that we need to neutralize (gender specific ones) by subtracking from them their orthogonal projection on bias subspace. Second for each defining set we calculate base which is mean of values with subtracked value of its projection on gender subspace. To this base we add individual value of each vector (calculated as in step Neutralize) multiplied by a normalizer.

## 5 Bias directions examples



(a) Gender bias (b) Religion bias (c) Nationality bias

Above are pictures representing how the gender subspace is represented for some of the stereotypes we considered. Each of the drawings represents some of stereotype on our bias subspace. As said before we can notice a pattern in each of the pictures above, e.g. each word for Christianity is more to up of chart than words for Islam and Judaism.

## 6 Generating stereotypes

Here using bias subspaces we will try finding some stereotypes by writing sentences like "He is to x as she is to y" where we will set x as conservatism, shorts, doctor, coward. We get following results:

```
Analogies generated by original embedding:
he is to conservatism as she is to feminism
he is to shorts as she is to skirts
he is to doctor as she is to nurse
he is to coward as she is to bitch
```

(a) Before debiasing

```
Analogies generated by hard debiased embedding:
he is to conservatism as she is to feminism
he is to shorts as she is to pants
he is to doctor as she is to physician
he is to coward as she is to gal
```

(b) After debiasing

As expected words that are gender specific by stereotypes (e.g. skirts and shorts) are debiased so their meaning is now very similar and words that are not included in gender specified set are not unified, e.g. conservatism and feminism. We also got interesting results for nationalities and for religion such as "Russians are to vodka as Germans are to lager" or "Muslim is to pain as Christian is to ache".

## 7 Distance between words

Below are results of calculating sum of distances of each gender biased word from all the words in defining sets both before and after debiasing. Also there is calculated distance between male and female biased words (from defining sets).

```
Old sum of all distances between pairs: 21.4633
New sum of all distances between pairs: 21.9811
Old embedding 'he'  distance: 1666284.2197
Old embedding 'she' distance: 1665323.2772
Diff: 960.9426
New embedding 'he'  distance: 1665931.1125
New embedding 'she' distance: 1665931.1091
Diff: 0.0033
```

Figure 4: Before debiasing

As desired difference between distances between "he-words" and "she-words" is equal to 0 in new embedding while sum of distances between all words is unchanged.

## 8 What now

Further developing of this work could contain testing new debiased embeddings on their solving capabilities for several tasks. Also testing debiasing multiple stereotypes while keeping as much correct connections in embedding as possible.