He is to shorts as she is to skirts

Mateusz Basiak and Dawid Barzyk on word embeddings debiasing.

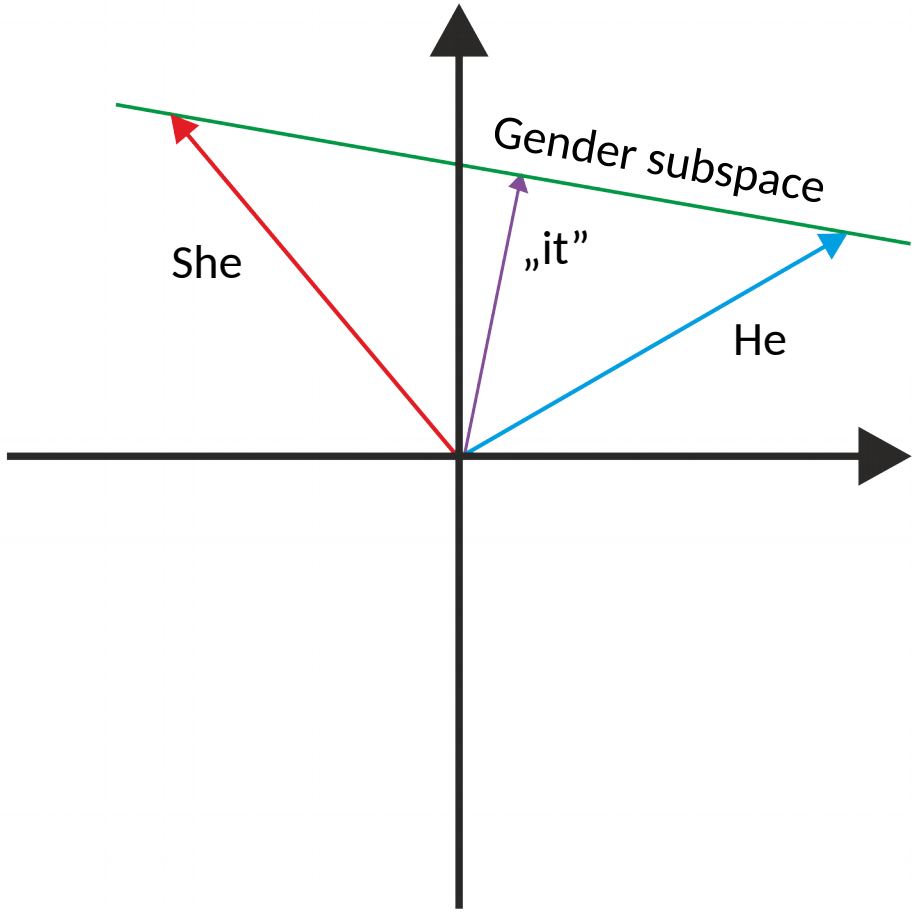
Bias in word embeddings

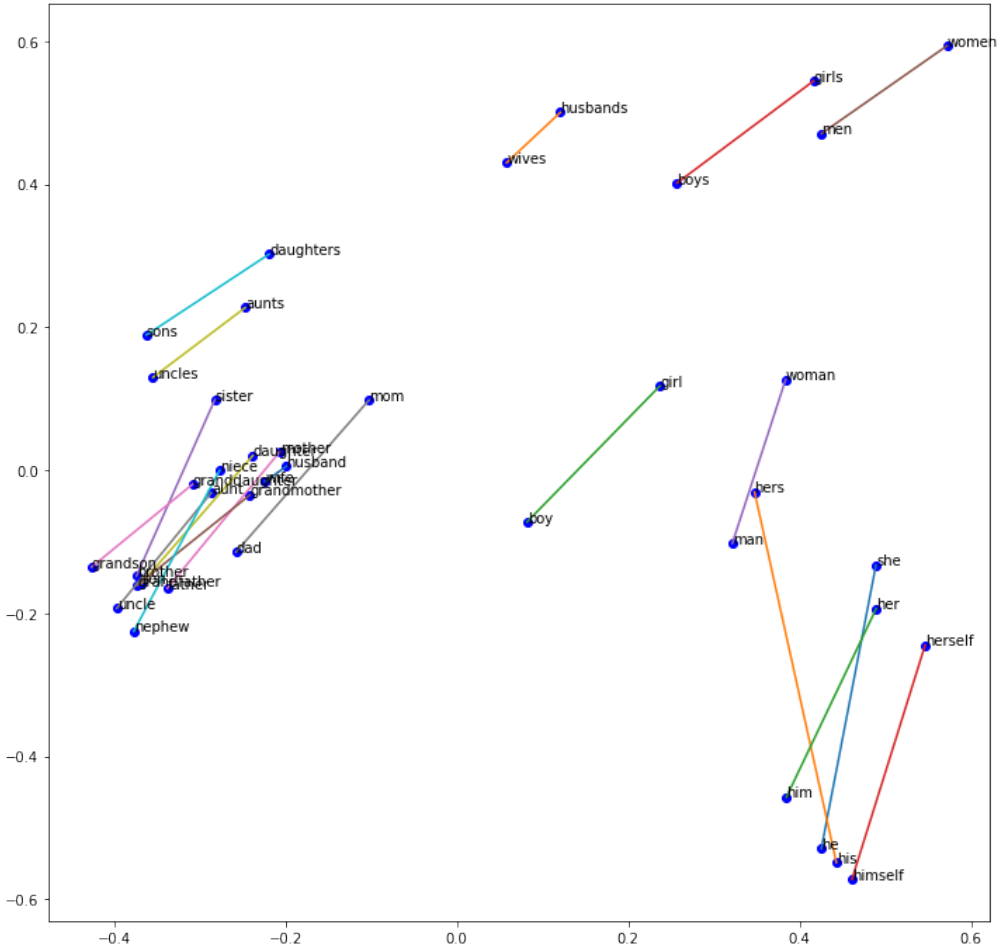
- Depending on what we are doing, some of the biases may have bad impact on the results.
- Sometimes we consider problems where some of the words mean same thing to us.

General idea

- Below we consider gender bias.
- Considering bias geometrically we can say that:
  - $\overrightarrow{man} - \overrightarrow{women} = \overrightarrow{king} - \overrightarrow{queen}$
- To use this observation we can create a set of pairs that define difference between males and females e.g. man & woman, he & she, mom & dad…
- We want to use debiasing only on words that are gender specific, hence we need to define set of gender specific words e.g. dude, bloke, nun, widow…
- All words that are not gender specific we note as gender neutral and we do not debias them

Gender subspace





Gender biased words

Gender subspace

- First using defining sets we will calculate gender subspace that shows us general direction in which the gender-component is spread in our embedding.
- We calculate SVD(C) where:

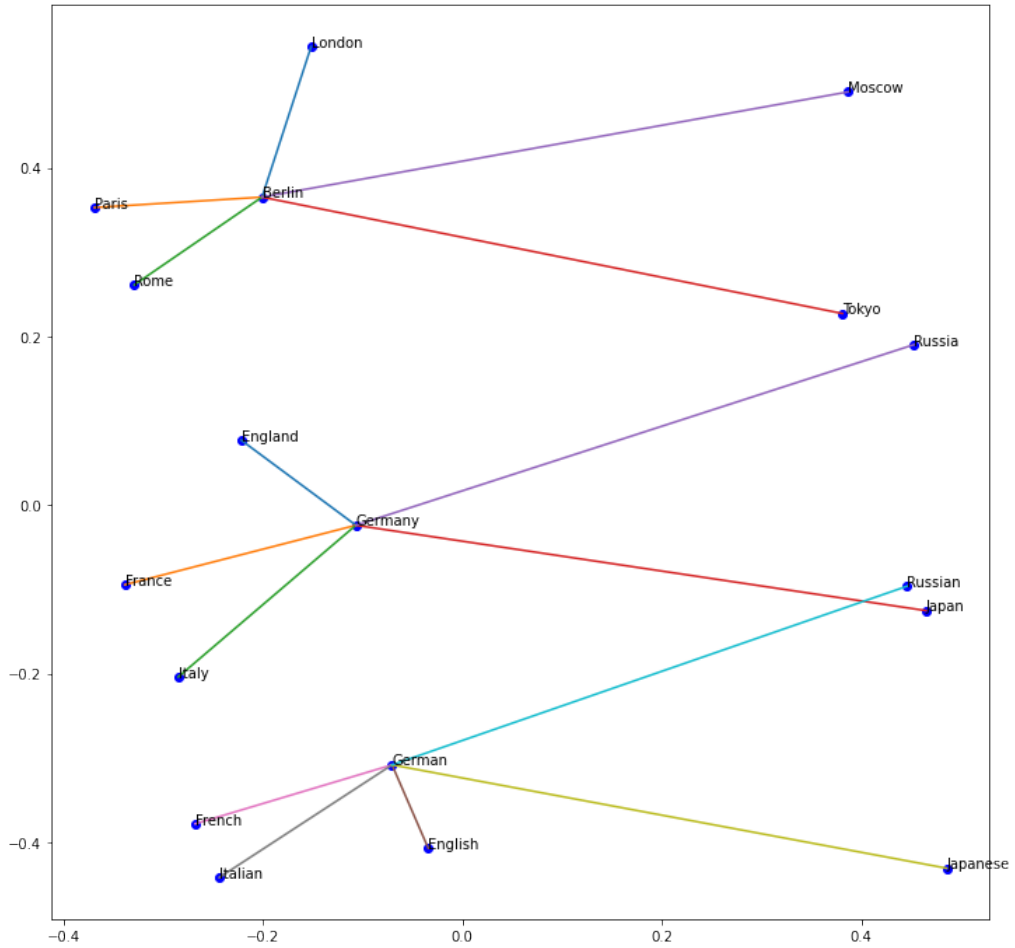
$$\mathbf{C} := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \mu_i)^T (\vec{w} - \mu_i) / |D_i|$$

  - With as ith defining set and as arithmetic mean of ith defining set.
- Our gender subspace will be defined as first k vectors of matrix V where SVD(C) = (In our tests we use k = 1)

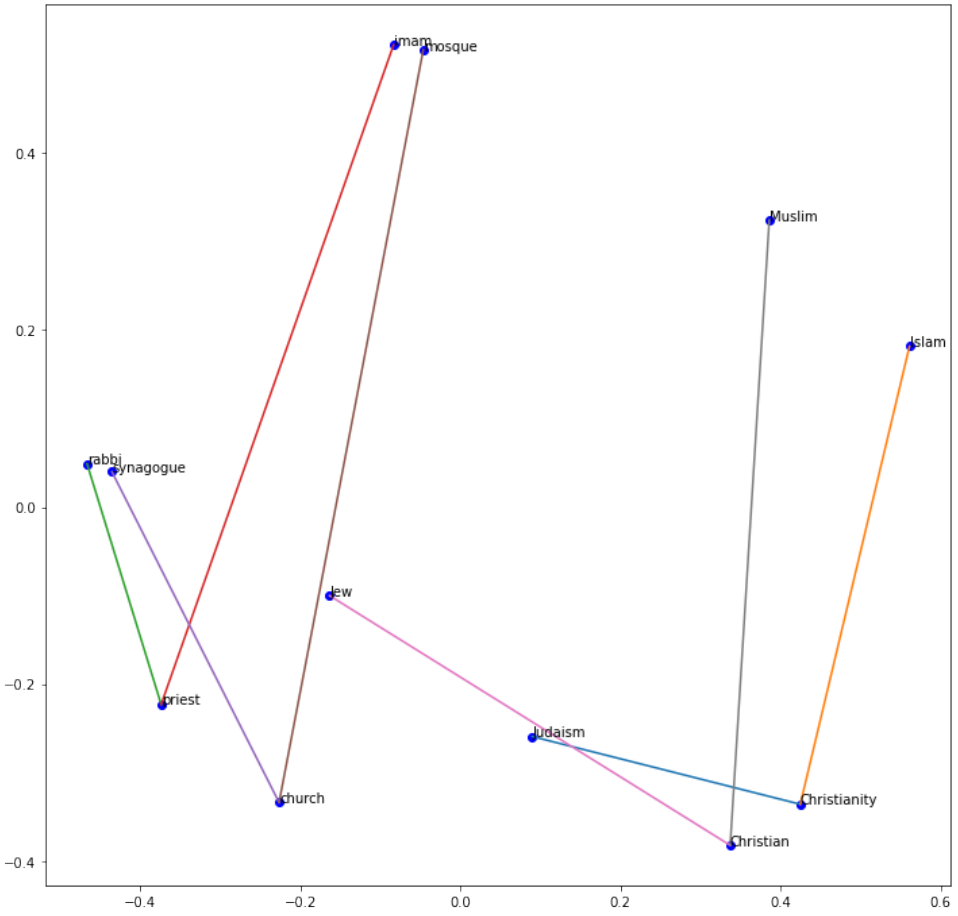
Hard debiasing

- Neutralize:
  - First we neutralize the words that we need to neutralize (gender specific ones)
  - We do that by subtracking the orthogonal projection of each vector on gender subspace
- Equalize:
  - 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”.

Bias directions examples



Nationality bias



Religion biased words

Generating stereotypes for genders

- Here using bias subspaces we will try finding some stereotypes (as in example of gender subspace)

```
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
```

```
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
```

Generalizations

- Testing different biases, often with multiple possibilities: gender, nationality, race, religion.
- Testing embeddings with single debiassing and with all possible debiassings applied.

Testing

- Generating analogies – keeping the right ones.

```
Muslim is to pray as Christian is to praying

Russia is to vodka as Germany is to lager
```

- Testing distance beteween biased points.

```
Old sum of all distances between religions: 18.3754
New sum of all distances between religions: 14.1862
Old embedding sum of distances to christian words: 345818.3718
Old embedding sum of distances to jewish words: 346537.6705
Old embedding sum of distances to muslim words: 347328.4601
Diff: 3020.1766
New embedding sum of distances to christian words: 346628.7609
New embedding sum of distances to jewish words: 346628.7603
New embedding sum of distances to muslim words: 346628.7613
Diff: 0.0020
```

Possible future work

- Trying to use our embeddings inside some neural networks.
- Different debiasing schemas – trying to prevail as many correct connections in embedding and get rid of as many stereotypes as possible.