Impact of Debiasing Word Embeddings on Information Retrieval

Emma Gerritse
Institute for Computing and Information Sciences
Radboud University
emma.gerritse@ru.nl

Introduction

We present our work on researching bias in word embeddings. Bolukbasi et al. have shown the presence of gender bias in word embeddings. This bias is, however, not as straight-forward as it seems. We highlight the complications and discuss the possible impact of bias and/or debiasing techniques on Information Retrieval tasks.

Bias in Word Embeddings

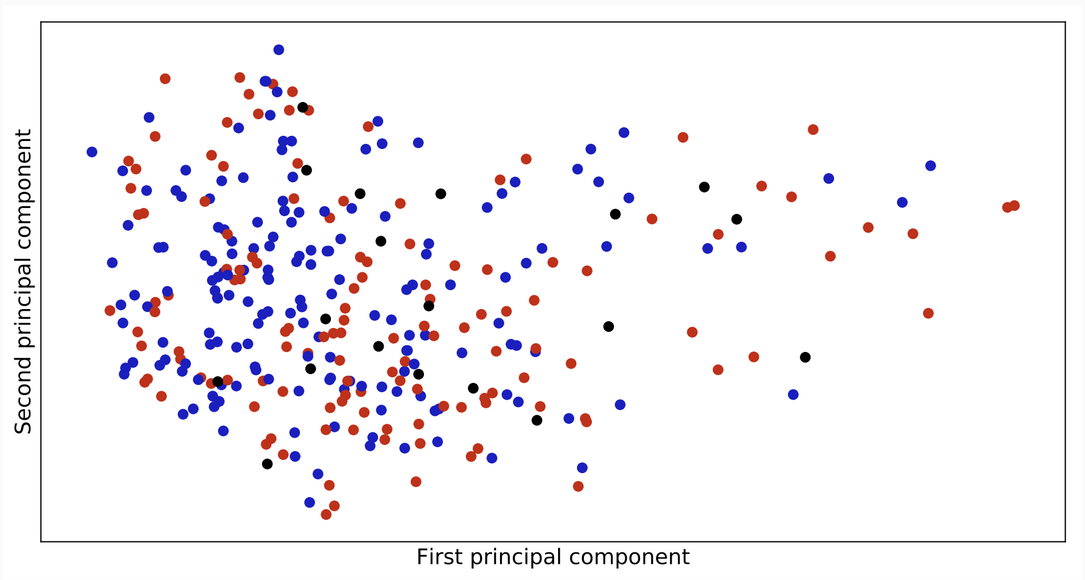

Figure: PCA projection of professions, with blue: $\cos(\vec{w}, \vec{he}) - \cos(\vec{w}, \vec{she}) > 0$, red: $\cos(\vec{w}, \vec{he}) - \cos(\vec{w}, \vec{she}) < 0$, and black the gender neutral words. Here we see that gender is not the first or second principal component.

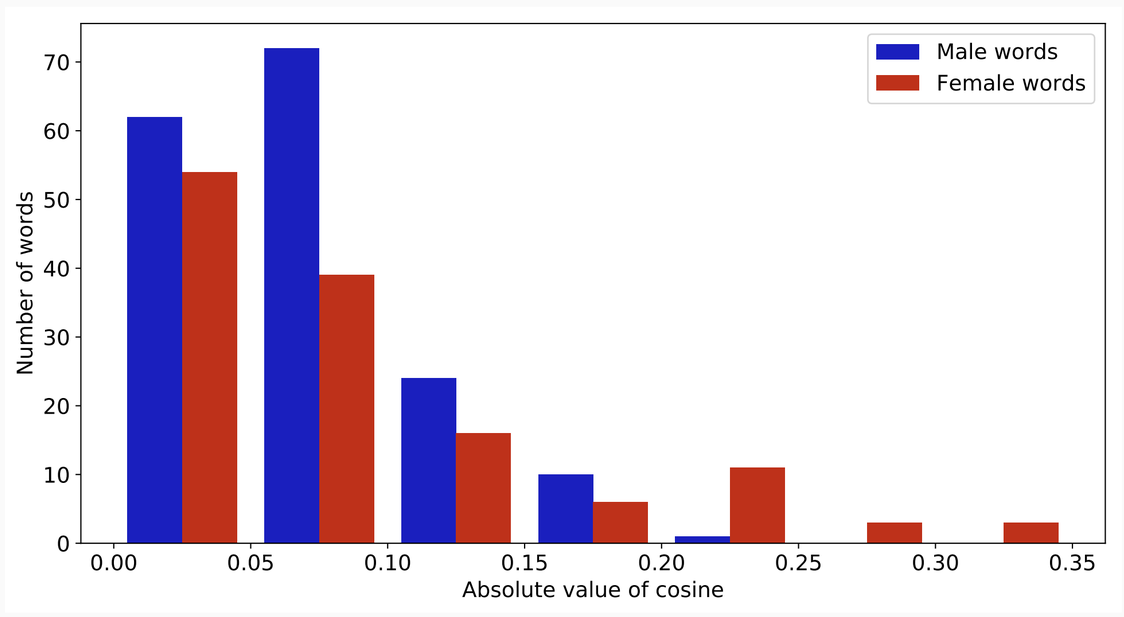

Figure: Frequency of professions, based on $|(\cos(\vec{w}, \vec{he}) - \cos(\vec{w}, \vec{she})|$. Professions with positive value of $\cos(\vec{w}, \vec{he}) - \cos(\vec{w}, \vec{she})$ are considered as male words, and the other way around. Here we see that there are more words with a strong female bias than with a strong male bias.

PCA of Biased Words

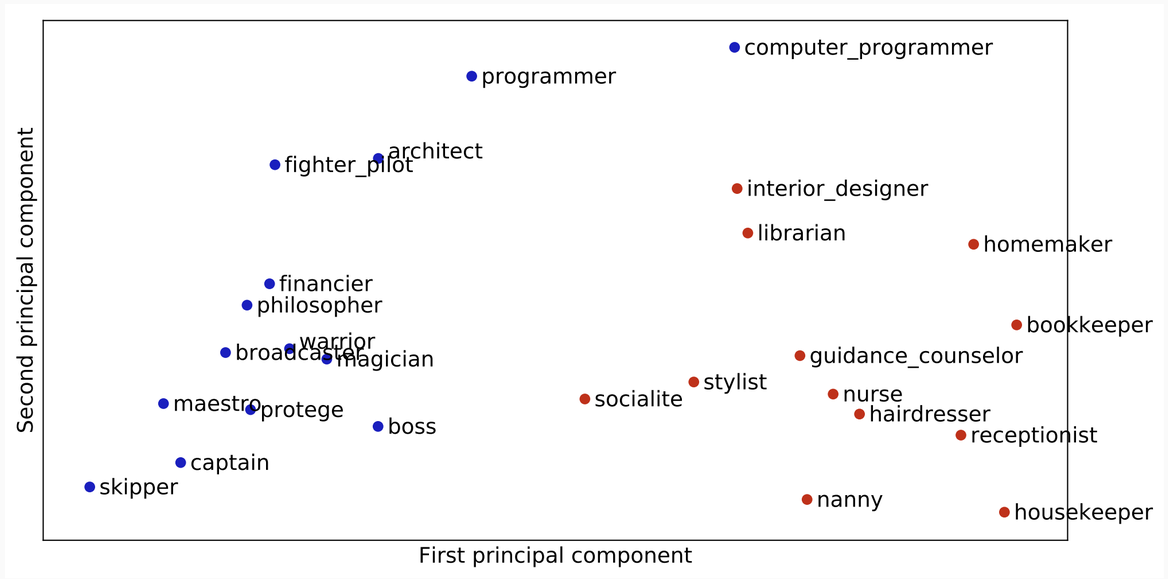

Figure: PCA of words with strong bias, here 'computer_programmer' is seen as a female word.

Bias Convergence

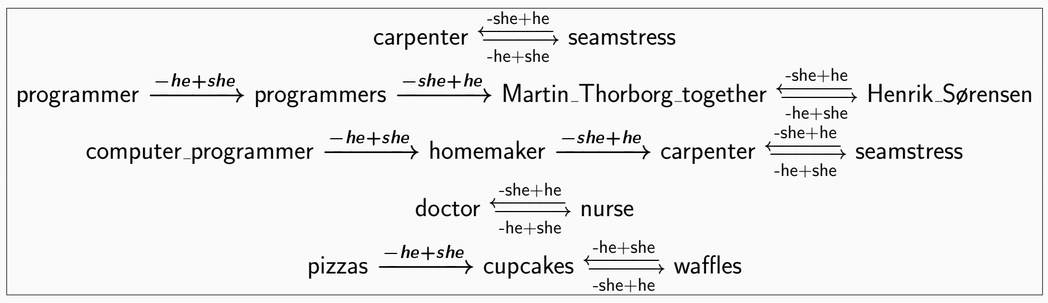

Figure: Analogy convergence over several iterations.

Bias and IR: Research Questions

- Will bias affect search results in e.g. job application search?
- Will debiasing affect results in neural IR systems?
- What if the search intent is biased?
  - For example: Female singers in Eurovision; singer is a 'female word'. Will debiasing give male results as well?

Be careful with embeddings if you do not know whether there is bias!

References

Tolga Bolukbasi et al. (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al., pp. 4349–4357

Paper created as result of following repository:
`https://github.com/informagi/cpo-w2v`

Radboud University