Debiasing doesn't come without a bargain
Master Thesis Proposal

Ángela Duarte Pardo
a.duarte-pardo@students.hertie-school.org

1. Motivation

NLP models use word embeddings for various tasks, such as machine translation, word analogies, co-reference resolution, etc. Word embeddings are vector representations of words that encode their relations to other words.

Because similar words occur in similar contexts, these representations can also represent semantic relations.[1] However, these word embeddings may inherit social bias from the corpora they are trained in[2].[1]

To debiase word embeddings, post-processing debiasing algorithms (Hard[1], Double-Hard[5] algorithms) typically aim to remove the bias projection of words on the bias space. This removal has proven to reduce bias without affecting performance on word similarity tasks.[4]

As an example, the gender space is defined on the direction of $\overrightarrow{she} - \overrightarrow{he}$. To remove the "gender component", the algorithm removes the projection of every word vector into this space.

However, these debiasing algorithms are limited because of three reasons:

- They do not necessarily guarantee that individual biases are consistently removed along all the embedded words: while some words can have significantly less bias after debiasing, some others can end up with a stronger bias than before.[2]

- Word embeddings can present multiple biases that are non-trivially correlated.[3] Thus, debiasing for one type of bias –say, gender– can influence other biases –say race. Thus, debiasing methods focusing on individual biases –without an intersectionality approach– can be counterproductive.

- Debiasing algorithms may only mask bias but cannot completely remove it. Some scholars argue that, even after debiasing, language models can pick up biases from other cues, including the proximity to words that carry implicit biases.[4]

The above reasons suggest that bias might not be encoded on a "bias space" to be easily and completely removed from word embeddings. Bias might also appear in other relations, including closeness to words carrying implicit cues to other social biases. Thus, removing the "bias space" from embeddings might not only be a limited approach, but it could potentially imply a change in the mathematical encoding that could affect other relations embedded in the vectors.

2. Research Project

Hard and Double-Hard Debiasing algorithms involve some trade-offs that should be better understood so that researchers and users can make more informed decisions regarding their limitations and benefits. This research project aims to characterize these trade-offs, explain their origin and, if possible, offer some insights into what it would mean for embeddings to be (correctly) "debiased".

The main objective of this research project is to answer the following question:

How can the limitations and benefits of geometric approaches to debiasing word embeddings be weighted to effectively mitigate biases without generating unforeseen and problematic relations between words?

3. Preliminary research outline

To answer the research question, this project will be conducted in three steps:

- Analysis of the mathematical theory underpinning post-processing approaches to debiasing word embeddings. With this I aim to understand the limitations of the geometric approach from the onset.

- Design and implementation of a coding experiment involving two debiasing approaches to identify in practice the what the trade-offs of the geometric debiasing algorithms.

- Drafting of possible ways of weighting these trade-offs to de-bias word embeddings more effectively.

[1] Gender and racial bias are the most studied types of biases encoded in embeddings[4].

[2] This was one interesting result of the tutorial that my group did for the Deep Learning course last semester. The tutorial is available here

References

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

[2] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[3] L. Cheng, N. Kim, and H. Liu. Debiasing word embeddings with nonlinear geometry. *arXiv preprint arXiv:2208.13899*, 2022.

[4] H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

[5] T. Wang, X. V. Lin, N. F. Rajani, B. McCann, V. Ordonez, and C. Xiong. Double-hard debias: tailoring word embeddings for gender bias mitigation. *arXiv preprint arXiv:2005.00965*, 2020.