Multilingual Gender Bias

Keywords: Gender Bias, Multilingual, Clustering, Bias Measures
Skills: Python, PyTorch

It is a well-known phenomenon that language models learn and replicate social biases that are present in the training data. This has sparked a new wave of research in NLP, focusing on measuring and debiasing pretrained language models. However, the vast majority of these proposed methods were developed specifically for English, thus much work still remains to be done in creating techniques that can be applied to other languages.

In this project we will investigate the effectiveness of linear cross-lingual debiasing techniques at removing structural distributional gender biases in word embeddings. Linear debiasing techniques have been shown to be effective at reducing bias scores for many popular bias measures, however, linearly debiased models have been shown to still encapsulate substantial structural distributional biases in the word embedding space after debiasing[1]. In other words, the linearly debiased word embeddings may be superficially debiased along a given dimension, but word embeddings relating to a given gender may still cluster together, indicating that the bias still exists but may became "hidden" to the bias measure.

The first goal of this project is to verify if structural biases exist in Chinese after applying the methodology of DensRay[2], a linear cross-lingual debiasing technique, to English embeddings in mBERT.
The second goal would be to investigate novel measures of structural bias in the word embeddings, taking inspiration from a range fields, including information theory.

[1] Hila Gonen and Yoav Goldberg, 'Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them', *ArXiv:1903.03862 [Cs]*, 24 September 2019, http://arxiv.org/abs/1903.03862.

[2] Sheng Liang, Philipp Dufter, and Hinrich Schütze, 'Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations', in *Proceedings of the 28th International Conference on Computational Linguistics* (Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020), 5082–93, https://doi.org/10.18653/v1/2020.coling-main.446.