Gendered Language in Resumes and its Implications for Algorithmic Bias in Hiring

Oskar van der Waal, Jaap Jumelet, Katrin Schulz, Willem Zuidema
Institute for Logic, Language and Computation, University of Amsterdam
{o.d.vanderwal,j.w.d.jumelet,k.schulz,w.h.zuidema}@uva.nl

Abstract

Detecting and mitigating harmful biases in modern language models are widely recognized as crucial, open problems. In this paper, we take a step back and investigate how language models come to be biased in the first place. We use a relatively small language model, using the LSTM architecture trained on an English Wikipedia corpus. With full access to the data and to the model parameters as they change during every step while training, we can map in detail how the representation of gender develops, what patterns in the dataset drive this, and how the model's internal state relates to the bias in a downstream task (semantic textual similarity). We find that the representation of gender is dynamic and identify different phases during training. Furthermore, we show that gender information is represented increasingly locally in the input embeddings of the model and that, as a consequence, debiasing these can be effective in reducing the downstream bias. Monitoring the training dynamics, allows us to detect an asymmetry in how the female and male gender are represented in the input embeddings. This is important, as it may cause naive mitigation strategies to introduce new undesirable biases. We discuss the relevance of the findings for mitigation strategies more generally and the prospects of generalizing our methods to larger language models, the Transformer architecture, other languages and other undesirable biases.

This paper has been accepted as a non-archival publication.

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 75 - 75