Challenges in Measuring Bias via Open-Ended Language Generation

Afra Feyza Akyürek **Muhammed Yusuf Kocyigit** **Sejin Paik** **Derry Wijaya**
Boston University
{akyurek,koyigit,sejin,wijaya}@bu.edu

Abstract

Researchers have devised numerous ways to quantify social biases vested in pretrained language models. As some language models are capable of generating coherent completions given a set of textual prompts, several prompting datasets have been proposed to measure biases between social groups—posing language generation as a way of identifying biases. In this opinion paper, we analyze how specific choices of prompt sets, metrics, automatic tools and sampling strategies affect bias results. We find out that the practice of measuring biases through text completion is prone to yielding contradicting results under different experiment settings. We additionally provide recommendations for reporting biases in open-ended language generation for a more complete outlook of biases exhibited by a given language model. Code to reproduce the results is released under https://github.com/feyzaakyurek/bias-textgen.

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 76 - 76
July 15, 2022 ©2022 Association for Computational Linguistics