CSC2552: Review 7, Paper 1

Due on March 13

Ashton Anderson

498 words

Thomas Hollis

Paper 1

This 2016 paper by Bolukbasi, Chang et. al. is a *crowdsourced computer-assisted observational study* proposing a new word embedding debiasing algorithm based on responses from Amazon MTurk. The main results suggest that word embeddings, like word2vec, reflect existing bias in their input data and the debiasing algorithm proposed does not significantly decrease performance while helping avoid some gender bias in word embeddings.

A major limitation of this methodology is that HITs on online marketplaces are vulnerable to the issue of *weak motivators* [1]. Indeed, in this study the main motivator of turkers is money, so there is no incentive for users to truly give their utmost attention to the task at hand. The turker's goal is rather directed toward completing the task as fast as possible while ensuring that they get paid. Fortunately, this limitation does not seem to harm the results thanks to the *aggregation strategy* employed by the paper. An alternative approach could consist in debiasing the original data from Google News but this would require some sort of selective removal, which has biases of its own, or it could consist in a global cultural shift toward a linguistic overhaul which seems unrealistic. Another compromise of this paper is the use of opinionated and emotive language. While the *research question* is very clear, there is definitely a trade-off between using emotive language to emphasise importance of the problem domain as opposed to using neutral language to strengthen argument objectivity. Indeed, the use of language such as "disturbing" and "none of these papers have recognised how blatantly sexist the embeddings are" seems somewhat out of place for such a high-quality journal like NeurIPS.

In contrast, a major strength of this paper is the methodology's supplementary use of *crowdsourced opinion data* to help ascertain perceived word embedding bias. This *armada strategy* approach, as documented in BitByBit [1], allows the authors of the paper to use opinions other than their own for the purpose of debiasing the word embeddings. Another significant strength of this paper is the identification and frequent referencing of a *mechanism* by which word embeddings become biased in the first place [2].

Implications of these results are promising as they offer a possible debiased alternative for sensitive machine learning applications. While this new algorithm does not present a notable performance decrease in the particular task used for evaluation, the paper's authors do indeed acknowledge that it may nonetheless fail to capture useful statistics. As such, I disagree with the main conclusion of the paper which recommends to "err on the side of neutrality and use the debiased embeddings provided here as much as possible". While I appreciate the debiasing advantage for particular use-cases, I do not believe enough evidence has been provided to justify such a large scale overhaul of tried-and-tested approaches like word2vec that are the bedrock of many existing algorithms. This dependency between statistical performance and fairness is gaining traction in NeurIPS and is further discussed in detail in [3].

[1] Salganik, M. J. (2017). Bit by bit: social research in the digital age. Princeton University Press.

[2] Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In Ninth international AAAI conference on web and social media.

[3] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems (pp. 5680-5689).