Summary 2: You Are Grounded!

Sam Showalter
University of California, Irvine (showalte)
showalte@uci.edu

1 Content Summary

This is a paper summary of (Shwartz et al., 2020) that explores the biases of pretrained language models (LMs) on named entities. In particular, it focuses on the representation of given names (generic) that may be strongly associated with sentiment tied to specific entities. The authors demonstrate that this conflation of generic and specific entities in language grounding can be detrimental to model output, performance, and bias.

The central argument of LM bias the authors explore is that while LMs have transformed NLP, they also conflate generic facts with grounded knowledge of specific entities or situations. This behavior is undesired and can lead to biased or stereotypical language. First, the authors explore the prediction of last name for a provided given name, and disover that with high probability the result is a known entity often mentioned on the web or news. Politicians were particularly susceptible and connoted negatively, as found by sentiment analysis.

Moreover, when given question-anwser objectives where two named entities are mentioned, pretrained LMs demonstrated acute sensitivity when the order of these names was swapped. In some cases, performance dropped substantially. Another ablation found that, in the event of a name flip, the prediction of many models changed. The authors state this is of large concern, given how many downstream tasks LMs are used for. They note several additional studies that discover word embeddings from pretrained models encode bias that can then be perpetuated.

In a discussion of these findings of bias and named entity sensitivity, the authors claim that since people ”rarely state the obvious,” the frequency of uncommon events in language is disproportionally present. Similarly, generic terms with the same name as a specific ”newsworthy” entity also tend to be charged with sentiment. In closing, the authors qualify that their findings are restricted to English, their names were somewhat male-skewed, as well as other caveats. Nonetheless, they feel the consequence of their work is that generic entities are not anonymous or without connotation, which has ethical consequences.

2 Analysis

The authors did an excellent job of communicating their ideas clearly and effectively. Their motivation and its relation to existing work was clear, and their experiments were well defined and reported. Sections were organized methodically, moving from motivation to experiments and finally a discussion of their findings and theoretical implications. I particularly enjoyed their section on the limitations of their work, and wish other authors more commonly qualified their findings so thoroughly

One criticism I had of this paper was that its experiments seemed less comprehensive than their claims. However, they address exactly this point in their closing arguments, making their paper very well qualified and their findings interpreted correctly. The exploration of bias in LMs is fundamental to maintaining equality today, with these systems impacting the lives of thousands on a daily basis (CV checkers, spam filters, etc.). It appears there is a strong interest in this field already, and the authors contextualize their research within a larger body of work.

Though well-defined and convincing, I would consider the findings in this paper to be incremental and in-need of further exploration. Based on the closing sections, the authors themselves appear to agree. While the findings were convincing, the experiments were relatively simple and, for skeptics, fairly easy to discredit given their limited scope. However, their message that generic entities are influenced by named counterparts is an incredibly insightful finding. More should be done to tease out the full extent of this relationship.

Some potential future experiments could include probing other entities beyond given name to tease out the interaction of named entities on other part of speech. Moreover, an analysis of how these trends are effected by the corpus used to pre-train LMs would be particularly insightful. One approach could be to adversarially simulate such a dataset, then infer from those findings the extent of the problem in real-world corpi.

References

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. ” you are grounded!”: Latent name artifacts in pre-trained language models. *arXiv preprint arXiv:2004.03012*.