ACL Paper Summary

fcd180001

In their paper, 'French CrowS-Pairs: Extending a Challenge Dataset for Measuring Social Bias in Masked Language Models to a Language Other than English,' Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort investigate cultural differences between the United States and France in relation to measuring bias in language. The authors argue that what may be considered biased in one country, such as the United States, may not be seen as biased in another country, such as France, and vice versa. They identify that the issue stems from the datasets used and propose creating a new dataset that accounts for the biases specific to France. By doing so, they can accurately measure social bias within the social context of France.

The primary author of this paper, Aurélie Neveol, is a PhD based in France who specializes in biomedical and clinical natural language processing. Her work in this field has earned her 4474 citations on Google Scholar. While her general interests include information extraction and knowledge representation, much of her previous research has focused on the same issue addressed in this paper: dealing with bias in a multilingual context [1]. Although her previous work delves deeper into this issue, this paper presents a relatively simple solution to the problem by expanding upon an existing dataset called CrowS-Pairs.

To provide context, the Crowdsourced Stereotype Pairs (CrowS-Pairs) dataset is designed to train masked language models to recognize various types of bias. The dataset includes nine different categories of bias: Race/Color, Gender, Sexual Orientation, Age, Nationality, Disability, Physical Appearance, and Socioeconomic status. It comprises 1508 examples of bias, which were selected from participants and labeled with the type of bias and the level of confidence (if any) associated with it [2]. The dataset has shown promise as a tool for debiasing models, as it helps to identify and mitigate various forms of bias.

The authors of this paper share the goal of limiting bias in models, which is also the aim of the original Crowdsourced Stereotype Pairs (CrowS-Pairs) dataset. However, not all biases presented in the original CrowS-Pairs dataset are applicable in the same social context in France as they were in the United States. Therefore, the authors saw an opportunity to expand upon the dataset and adapt it for a new social context. Their work aimed to address this gap by developing a new dataset called expanded CrowS-Pairs, which accounts for cultural differences in bias between the languages.

The process the authors developed to expand the dataset is relatively straightforward. They began by identifying the applicability of each bias type to the French social context. After analyzing the original Crowdsourced Stereotype Pairs (CrowS-Pairs) dataset, they concluded that all nine bias types were relevant and could be easily applied in the French context. Therefore, no changes were necessary regarding the types of biases identified.

The second step involved translating a random set of sentence pairs. A French author who is fluent in both French and English performed the translations, which were then validated by another author. During this step, sentences that were deemed untranslatable because they did not apply to the French cultural context were marked as such. For example, the sentence "Deshawns horse reared as he clutched the reins after looking at the KKK members" was deemed

untranslatable because the KKK, although prevalent in the social context of the United States, is not prevalent in the social context of France. Throughout the translation process, notes were taken to document instances where words had to be changed to convey the same bias as the original sentence, or if the translation identified a new social bias. Given that the translation between the two languages is not always straightforward, careful attention was paid to ensure that the translated sentences accurately reflected the original biases, otherwise they were eliminated from the set.

Finally, after the translation of the randomly selected sentences, the authors crowd-sourced new sentences in a similar manner to the original dataset. All identifications made by participants were validated by the authors to ensure that the correct social bias was being identified. These sentences were translated into English to maintain a record of translations throughout the dataset. As a result, the dataset now includes both English and French sentence pairs, each of which is associated with an identified bias.

The authors evaluated their work by testing the expanded CrowS-Pairs dataset on several popular French masked language models including: CamBERT, and FlauBERT. The results showed that bias was consistently higher in English language models than in French ones, although the French models still exhibited a certain level of bias. The revised English sentences were also tested on various English models, and they produced similar results to the original CrowS-Pairs dataset, as expected.

Though the proof of the concept of this working was validating for the research, the most exciting and unique aspect of it was actually is the general strategy it offers for creating a new dataset to measure social bias in any country's language. The process can be simplified into three steps: first, verifying the application of bias types and potentially creating new ones if necessary; second, translating the original dataset into the language of the model being used, which may require tedious revisions of sentences to ensure their meaning is preserved and applicability to the local social context; and third, crowdsourcing new sentences to expand the dataset. By following these steps, a powerful tool for debiasing models can be created that can be applied to any language and social context.

It is important to note that the creation of this dataset will not completely eliminate bias, but rather only serve to identify it. However, the authors hope that by identifying these biases, researchers can take the necessary steps to address the biases present in their models and work towards reducing them.

I initially selected this work because as a native French speaker, I was excited to learn about the progress of natural language processing in languages other than English. In our classes, we primarily focus on English language applications, and I was curious to see how this technology is being applied to other languages. The significance of this work lies in the fact that social biases are often viewed through an "American" lens, and this process will allow language models to identify biases in different social contexts and languages. The creation of a process like this is crucial, as it will allow other countries to expand their datasets to capture bias in their own machine learning models. The authors' work has demonstrated the importance of creating

diverse datasets and expanding natural language processing beyond a single language, and I hope other countries follow suit in expanding their datasets to identify biases in their own languages and social contexts.

Sources

[1] Névéol, Aurélie. "neveol." LIMSI-CNRS, https://perso.limsi.fr/neveol/.

[2] Nangia, Nikita, et al. "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models." arXiv preprint arXiv:2008.13590 (2020).