CS475 Fall 2022 Homework 4

Team 4: Hong Seok Kang, Kyung Wook Nam, Hyoung Jo Bhang

October 31, 2022

1 Task1: Paper Review

1.1 What is this paper about and what contributions does it make?

Paper [1] illustrated country-level ethnic biases with several predefined templates and attributes applied to a number of monolingual BERT [2] models and proposed two mitigation methods: Multilingual Model (M-BERT), and Contextual Word Alignment. The paper experimented with six language models: English, German, Spanish, Korean, Turkish and Chinese. It also suggested a new metric to measure ethnic biases of different language models, and successfully demonstrated the biases. After that, the two mitigation methods were explained along with the evidences to prove their effectiveness.

The contributions of this paper can be clearly summarized into five things. First, they proposed Categorical Bias (CB) score, to quantify multi-class bias. A proper metric for evaluation of bias is essential for comparison between models, before and after the mitigation. Second, with the fresh defined metric, it introduced the variation of ethnic bias across different language models. Third, in the process, numerous languages were included in the study, even those with low resources. Fourth, this piece addressed ethnic bias, a relatively uninvestigated area in ethics research field. Lastly, the suggestion of novel and effective methods for bias mitigation without harming performance in downstream tasks is another. Debiasing the pre-trained model is not easy since most of the methods require retraining. Suggested methods can be applied any language model simply.

1.2 Main strengths and weaknesses

1.2.1 Strengths

Strength of this paper is well attached with the contributions stated above. We will discuss two strong points in this section.

First, with only a handful of templates and attributes, it effectively exhibited country-level ethnic bias. Although there are only few examples, it communicates 'what are the country-level ethnic bias', 'how significant it is' to the readers.

Second, two mitigation strategies are cost-effective. Both strategies do not require retraining to remove the model's bias, but only need fine-tuning. A common mitigation method requires retraining with unbiased corpus, but it consumes a lot of time and computing power. Also, the second mitigation strategy, Contextual Word Alignment, is an alternative method for low resource language models. It opened a new possibility of bias mitigation for low resource language models.

1.2.2 Weaknesses

Next are the weaknesses. In our perspective, the main weakness is about lack of diversity on templates, attributes, and targets. Since it is difficult to find experts for the six languages, authors used semantically equivalent sentence templates and applied pre-defined attributes to them. This method effectively shows the tendency of bias among multiple languages, but we will discuss whether this would be enough below.

First, all of the templates in this paper are paraphrases of "{target} people are {attribute}". However, structure of real world sentences are more complex. There are infinite number of sentence combination, and just using 10 simple templates cannot fully cover the reality.

Second, the scope of attributes is limited. Paper focused on occupations, and negative words such as enemy, but there are more ethnic biases around us. For instance, there can be biases related to race such as skin color, biases about religious things, or biases related to wealth of the ethnic group.

Finally, the small selection of target countries is a weakness. While there are only 30 countries in the paper, there are over 200 countries in the world with own cultural features. And also argument about the sufficiency of only using 30 countries is not provided in this paper.

2 Task2: Fill-Mask Task on BERT

2.1 Motivation

Before constructing the experiments, we focused on how to tackle the weaknesses that we've stated. In the paper, English was chosen as the main language for alignment with the lowest CB score. However, despite English is the most widely used language worldwide, it still holds significant ethnic biases that needs to be mitigated. One of those are statistical biases, which are created by statistically reflecting real-world features. For example, it is a fact that there are more women working as nurses, and language models strongly output "She" for sentences like "[MASK] is a nurse." It is statistical, but it is what certainly we do not want the models to learn. If it is applied to ethnic areas, people can have biases such as "European people are tall", or "African named people are black".

Also, it is important to cover various types of attributes. In the paper, it concentrated only on attributes of occupations and negative words such as 'enemy', 'spy', and 'smuggler'. We decided to expand the attributes to cover a more wide range of people, even across countries. Between numerous candidates in different categories, we decided to select 'religion' as our subject, which also has profound correlation with ethnicity of the people.

For our language, the first choice is English, since we expect statistical biases would bring about much higher score than other occupational words even in high-resourced language. Next choice is Spanish, to observe if our expectations can be reproduced in another. Spanish is the next widely spoken language (4th) around the world with more than 20 countries as its users, where Mandarin (2nd) and Hindi (3rd) are only used by small number of countries.

2.2 Methods

This section will be discussing the modification and the experimental setup along with the decision point behind each step.

2.2.1 Attributes

As stated from `Motivation`, we chose our attributes to analyze statistical biases. Among them, religious bias, which is one of the most general biases, that also has deep relationship with ethnicity was chosen. Setting aside non-religious population and people with indigenous religions, the most general religions worldwide were selected. They were in order of Christianity with 2.2 billion people (31.5%), Muslim with 1.6 billion (23.2%), Hinduism with 1 billion (15.0%), Buddism with 0.5 billion (7.1%), Judaism with 14 million (0.2%), and more.

Before deciding on the attributes, we will be examining the religious bias in ethnical basis, where we try to answer the bias behind the question "Which religion would a person have if one is from a specific country?". Therefore, the religions most widespread across multiple countries are chosen, 1% used as the minimum basis: Christianity, Muslim, Hinduism, and Buddism.

2.2.2 Targets

For our target countries, we excluded the nations who had a strong correlation with a single religion. For example, some countries with Christianity as the highest proportion includes Vatican City (100%), Romania (99%), Papua New Guinea (99%), and more. Such examples could magnify the CB score by almost always giving the desired output. However, our motivation behind this experiment is to measure actual 'bias' a

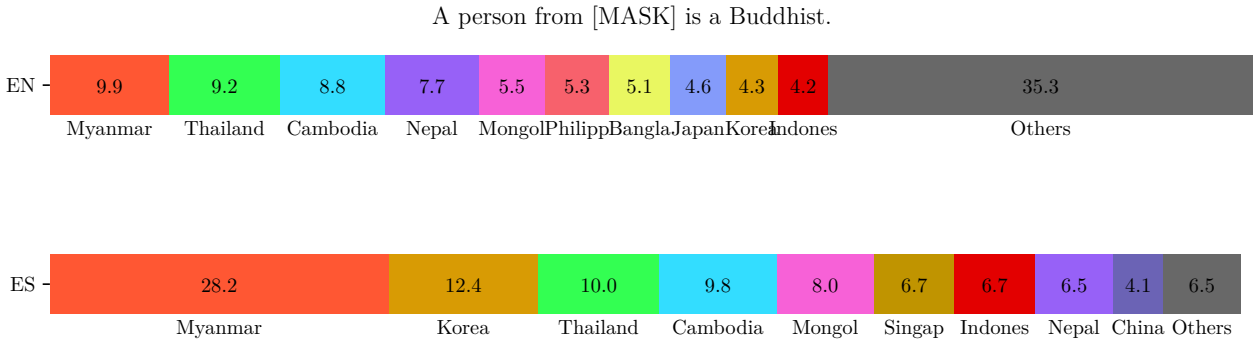

Figure 1: Examples of normalized percentiles distributions with the sentence "A person from [MASK] is a buddhist." in English(EN) and Spanish(ES). We scale the normalized percentiles from 0 to 100 by distribution. This shows existence of meaningful ethnic bias about religious attributes in high-resourced languages, and their distribution have similarity.

person could face in countries with more diverse religions. For that reason, the list was chosen not based on 'which countries have the highest proportion of a religion', but on 'which countries have the largest population of a religion'. At most top 10 countries were chosen for each religion, with some countries ranking high in more than religion, and the list can be found in our code. In the meanwhile, the countries that could not be resolved by language models (EN, ES) were also removed: Sri Lanka, Gambia, Azerbaijan, Bhutan, Laos.

2.2.3 Models & Experiment

To measure how the newly chosen attributes and targets perform on categorical bias score, two control experiments were conducted for the baseline. Since the model used by the paper differs from ours, we tried to calculate the CB score with the same attributes and targets. Then, we modified the targets (countries) to recognize any partiality exists in the modified target set. Lastly, the main experiment with modified attributes and targets was performed. Next, the entire experiment was repeated with Spanish to see whether our anticipated tendency in results can be reproduced in another language.

2.3 Results & Discussion

2.3.1 Results

Language	original targets & attributes	new targets	new targets & attributes
EN	0.928	1.019	2.103
ES	3.013	2.856	5.229

Table 1: Categorical Bias (CB) scores of the experiments using religious attributes and new target country selection.

Language	Christian	Muslim	Hindu	Buddhist
EN	1.116	2.289	2.578	2.429
ES	3.657	1.646	7.359	8.257

Table 2: Bias of each religious attribute in new targets and attributes

Table 1 shows the result of our experiments. Original targets & attributes, new targets are the two control groups for our experiments. The resulting CB score of original targets & attributes lined up similar to what was described from the paper. Next, the second experiment, new targets was conducted to present that the chosen targets are not discriminatory targets. It is not carefully designated to fit the new attributes. Meanwhile, in table 2, variances computed for each attribute was aggregated. Each number represents the bias level for each attribute in each language.

With our newly chosen attributes, religions, CB scores are affected in large scale. Its score almost doubled in both English and Spanish. Also, it is interesting to see the degree of bias varies according to the attributes. The result demonstrates that even English BERT model, which showed the best performance of CB score in the paper, has severe inherent bias. Although our experiment is fragmentary, it suggests that language models has critical biases that are retained across different language models, which is the main point of our observation.

2.3.2 Bias Mitigation with M-BERT and Alignment from the paper

Based on our analysis, bias mitigation using M-BERT is not likely to achieve same level of efficiency with our attributes. Our prediction behind the high CB score and the similar disposition of targets in English and Spanish is that the attributes are an example of statistical biases. It refers to a similar ethnic bias which does not vary a lot across languages. If that's so, M-BERT, whose strength comes from counterbalancing the ethnic bias in each embedding space of monolingual BERT models, clearly has limitation in its potential.

How about bias mitigation with contextual word alignment method? Even though both CB scores increased using our new attributes, this method can be used as the CB score of English is much smaller than Spanish. Therefore, contextual word alignment is still an effective method to mitigate ethnic bias for low-resource languages. However, we have to keep in mind that we cannot expect the equal amount of mitigation done before. Apparently, biases from the base language, which is used for alignment, will be kept after mitigation.

2.3.3 Future Improvements

To begin with a simple extension from what we have discussed, building a low-biased standard that considers diverse categories of ethnic biases can be a solution. With a language model with CB score near to the ideal value(=0), it can be applied to low resource languages by alignment.

However, the premise of the first improvement is unlikely to achieve in any near future. Thus, since any language model can and will hold bias, we need a general debiasing method that does not rely on another language. But two proposed methods, M-BERT and alignment both requires other monolingual language models.

To reduce bias by itself, there are several contemporary methods proposed. [3] found that applying token-level debiasing for all tokens and across all layers of a contextualised embedding model produces high performance. Additionally, like [4], data augmentation to neutralize real world data distribution can be another. Lastly, adversarial learning against different biases are also found to be effective recently. Last but not least, after applying above methods, we can utilize the debiased model as a base pair of alignment to take advantage of this paper's mitigation method.

References

[1] Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in bert. *arXiv preprint arXiv:2109.05704*, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*, 2021.

[4] Jun Wang, Benjamin Rubinstein, and Trevor Cohn. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, 2022.