Paper Title: Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting

Paper Link: https://aclanthology.org/2023.acl-short.108.pdf

1. Summary

1.1 Motivation/Purpose

On a number of natural language processing (NLP) tasks, pretrained language models (PLMs) have attained state-of-the-art performance. However, it has been demonstrated that PLMs are biased, including gender bias. PLMs are trained on enormous datasets of text and code, which frequently reflect the biases of the real world. Several methods for debiasing PLMs have been proposed. Using a gender-neutral dataset to fine-tune the PLM is a prevalent practice. Nonetheless, this strategy may result in catastrophic forgetting, in which the PLM forgets the information it learned from the initial training data. This can degrade the PLM's performance on downstream duties.

GEEP solves the problem of catastrophic forgetting by freezing the PLM's parameters and learning gender-specific prompts using data that is gender-neutral. The prompts are then used to guide the PLM in making predictions that are more gender-neutral. GEEP is influenced by recent prompt-tuning techniques for PLMs. In prompt-tuning, the PLM is provided with a prompt that assists it in comprehending the current mission. GEEP employs a similar strategy, but utilizes gender-specific prompts to assist the PLM in making more gender-neutral predictions. For instance, if the objective is to predict a person's gender based on their profession, the question could be, "Is this profession more commonly associated with men or women?"

1.2 Contribution

1. The authors show that catastrophic forgetting happens when PLMs are fine-tuned on gender-neutral data to make them less biased.
2. They suggest a brand-new method, the GEnder Equality Prompt (GEEP), which will make PLMs more gender-fair without forgetting anything important.
3. They test GEEP on several gender equality tasks, such as predicting pronouns, resolving coreferences, and classifying professions, and discover that it works better than previous methods.
4. They also test GEEP on general NLP tasks and find that it works about the same as past techniques, showing that it doesn't cause catastrophic forgetting.

1.3 Methodology

GEEP involves a two-stage process:

1. Freeze the parameters of the PLM. This prevents the PLM from forgetting the training data-learned information.

2. Learn gender-related prompts with gender-neutral data. This allows the PLM to make more gender-neutral predictions. GEEP utilizes a method known as prompt learning to acquire the gender-related prompts. Prompt learning is a method for training a model to learn prompts that helps it to complete the work.

1.4 Conclusion

Overall, "Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting" is a substantial contribution to the field of NLP. GEEP is a novel and effective technique for enhancing the gender equality of PLMs without compromising performance on general NLP tasks.

2 Limitations

2.1 First Limitation

The effectiveness of GEEP is contingent upon the accessibility of gender-neutral data for the purpose of training gender-related prompts. The availability of this data may vary across different activities or datasets (no available data).

2.2 Second Limitation

The researchers assessed the performance of GEEP on a restricted set of tasks related to gender equity and general natural language processing. There exists a potential for GEEP to exhibit worse performance when applied to alternative tasks or datasets.

3 Synthesis/ Future work

The authors of the paper note that GEEP is still under development. They plan to explore the following areas of future work:

1. Applying GEEP to other facets of gender equality in order to create methods for debiasing PLMs against occupational bias, social role bias, ethnic bias, age bias, and disability bias.
2. Developing more efficient methods for prompt learning.
3. Assessing GEEP on an expanded set of NLP tasks and data sets.
4. Developing question-answering systems that are more gender-sensitive.
5. Creating gender-inclusive educational tools and resources, such as personalized learning experiences and interactive educational activities.