Measuring and Mitigating Gender Bias in BERT

Marion Bartl
University of Malta | University of Groningen
Supervised by Dr. Malvina Nissim and Dr. Albert Gatt

Background

Just like traditional word embeddings, contextualized word embeddings, such as BERT [3], incorporate social biases [6,10]. This work focuses on **gender bias.** Since BERT is used as semantic resource for multiple downstream NLP tasks, its biases trickle into these systems.

Therefore, we need to ask:

1. How can gender bias be **measured** across different languages?
2. Can gender bias be **removed** from contextualized word embeddings?

https://muppet.fandom.com/wiki/Bert_Through_the_Years

Quantifying Bias

One of the most popular methods for quantifying gender bias in standard word embeddings is the Word Embedding Association Test (WEAT) [2]. The test uses cosine similariy to measure the association between two sets of target words that carry an implicit bias (e.g. stereotypical professions), and two sets of attribute words carrying explicit bias (e.g. gender-denoting words). Kurita et al. [6] adapted this method for BERT using sentence templates and the BERT language model.

De-biasing (contextualized) word embeddings

Previous research on de-biasing contextualized embeddings has adapted methods from standard word embeddings, such as projection onto the gender subspace [1,10]. However, Gonen and Goldberg [4] have shown that this debiasing approach is mostly superficial. A more favorable method is to interfere on the training data directly, instead of modifying the embedding space. Counterfactual Data Augmentation (CDA), which was introduced by Lu et al. [7], is a method to augment a corpus with sentences in which gendered words are changed to their opposites. CDA already showed promising results for mitigating gender bias in a coreference system based on ELMo embeddings [6].

Methodology

Measuring Bias

Following Kurita et al. [6], we measure gender bias by calculating the **association** between a gendered word (target) and a profession term (attribute). Probabilities are obtained from a BERT language model.

How to calculate the association

Take a sentence	“My brother (target) is a kindergarten teacher (attribute).”
Mask the target	"My [MASK] is a kindergarten teacher.”
Probability of the target	$p_{target} = P(brother = [MASK] \mid \text{sentence})$
Mask the target and attribute	“My [MASK] is a [MASK] [MASK].”
Prior probability	$p_{prior} = P(brother = [MASK] \mid \text{sentence})$
Association	$\log \frac{p_{target}}{p_{prior}}$

Data

We created a **Bias Evaluation Corpus,** which contains 5,400 sentences for one language and is composed of:

- ✓ **60 professions**, divided into 3 groups: statistically female, male and balanced
- ✓ **18 gendered words**, 9 female, 9 male, taken from the Equity Evaluation Corpus [?]
- ✓ **5 sentence patterns**, of the form "<person> works as <profession>"
- ✓ **2 languages,** English and German

Mitigating Bias

We apply name-based Counterfactual Data Substitution (CDS) [8] on a corpus and fine-tune BERT language model on the gender-swapped data.

Kurt has worked with various charities and organizations to raise awareness and donations for childhood cancer. Since overcoming **his** illness, **Kurt** has teamed with **Paul** Newman's Hole in the Wall Gang Camp and One Lap of America for fundraising events.	**Isabelle** has worked with various charities and organizations to raise awareness and donations for childhood cancer. Since overcoming **her** illness, **Isabelle** has teamed with **Emily** Newman's Hole in the Wall Gang Camp and One Lap of America for fundraising events.

Example extracted from the GAP corpus [9].

Results & Discussion

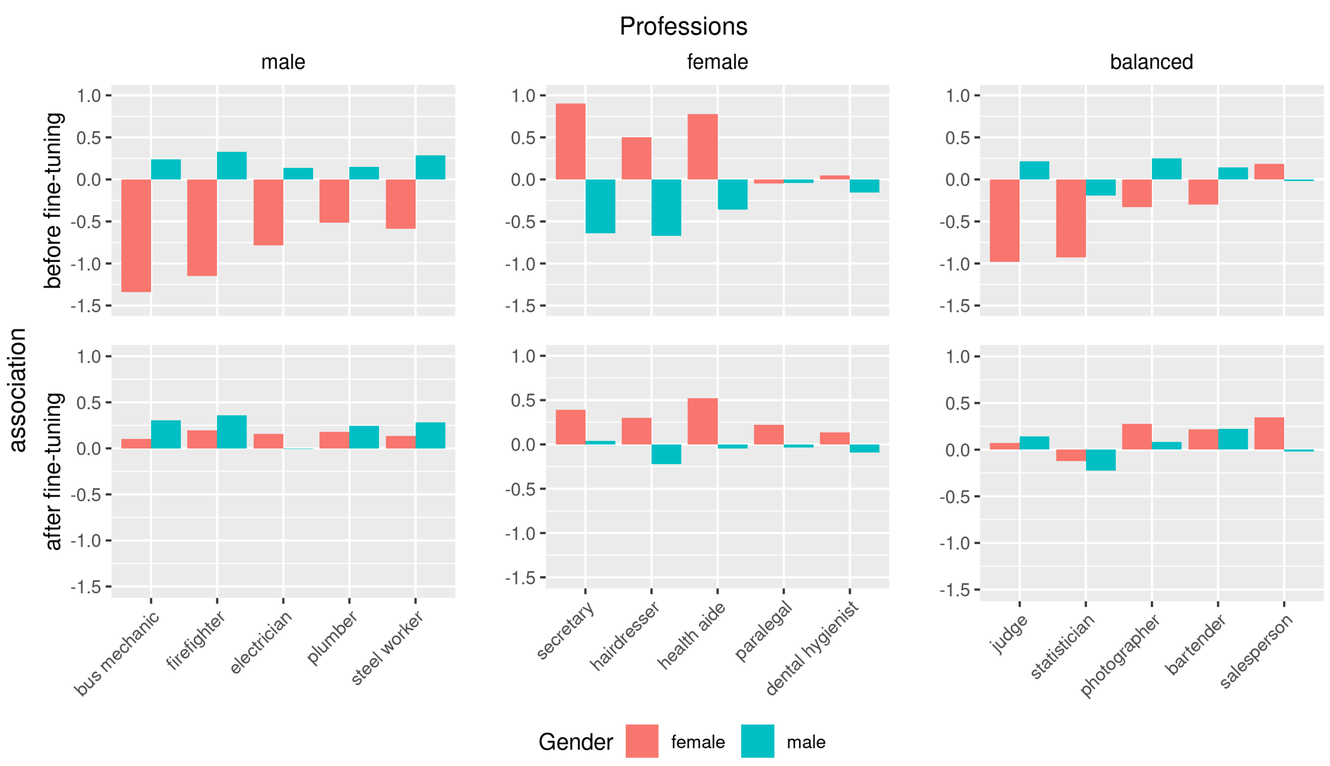

Figure 1: Associations across profession groups before and after fine-tuning

What can we see?

Figure 1 shows the mean association values for male and female target words across the three profession groups for five individual professions each. Before fine-tuning, we observe that the associations mainly follow expected stereotypical patterns for male and female professions: female words have a negative association with male-dominated professions and vice versa. However, in context with statistically balanced professions, female words still show a negative association.

What changes after fine-tuning?

After applying CDS to the GAP corpus [9] and fine-tuning on these gender-swapped data, the absolute association values generally decrease. While female words now show a small positive association with statistically male professions, the association of male words with female-dominated professions admittedly decreases but remains negative. This suggests that fine-tuning caused a larger semantic shift for female than male gendered words. In the balanced group, association values for male and female words are for the most part approaching each other, except in the case of 'salesperson'.

Conclusion

- ✓ **Shortcomings of new NLP technology** – Contextualized word embeddings led to improvements in many NLP tasks and were quickly adapted by the NLP community. Therefore, it is important to assess potential shortcomings and correct them.
- ✓ **The black box effect** – Due to its complexity, BERT is often used as a black-box resource in NLP tasks. Finding a reliable method to measure bias contributes to efforts of opening the black box.
- ✓ **Correspondence to the real-world** – By using employment statistics we establish which biases correspond to real-world findings and which are based on stereotypes.
- ✓ **Work outside of English** – Since most research in NLP focuses on English, there is still considerable need for research on other languages. This especially includes languages whose grammatical structure complicates a simple modification of English models.

References

[1] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in neural information processing systems, pages 4349–4357.

[2] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding.CoRR,abs/1810.04805.

[4] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.arXiv preprintarXiv:1903.03862.

[5] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias intwo hundred sentiment analysis systems .arXiv preprint arXiv:1805.04508.

[6] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. arXiv preprint arXiv:1906.07337.

[7] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018.Gender bias in neural natural language processing. arXiv preprint arXiv:1807.11714.

[8] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution.arXiv preprint arXiv:1909.00871.

[9] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. Transactions of the Association for Computational Linguistics, 6:605–617.

[10] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings.arXiv preprintarXiv:1904.03310.