Neutralizing BERT from Gender Bias in Occupation

Junbum Kim
Department of Electrical Engineering
The Cooper Union for the Advancement of Science and Art
New York, NY 10003
kim65@cooper.edu

Abstract

Word embeddings, a crucial model for creating vectorized representation of text, are known to be sexist when it comes to occupation. However, similar analysis has not been pursued on contextual embeddings like BERT, because these models are relatively new and because it operates differently. In this paper, we come up with ways to quantify gender bias in BERT, and propose two algorithms to neutralize it.

1 Introduction

In 2013, the introduction of neural networks into Natural Language Processing has revolutionized text processing. In particular, word embeddings became the new standard to create vectorized representation of words. Models like Word2Vec[1], GloVe[2], ELMo[3], and BERT[4] are all neural network based models that takes texts and transforms it to numerical vectors. Word embedding is an inseparable task for the state-of-the-art natural language processing as it serves three crucial purpose. Firstly, it allows vector operations like cosine-similarity which is at the crux of document querying. Secondly, it reduces the high-dimensional nature of text, which can span to more than 30,000 words, and reduces it into a latent-subspace of less than 1,000 features. Lastly, it embeds semantically related tokens to similar regions. For example, a word emedding through word2vec will produce a similar embedding for the word "happy" and "joyful," although their one-hot-encoded inputs are distant.

[1] Mikolov, et al. (2013)
[2] Pennington, et al. (2014)
[3] Peters, et al. (2018)
[4] Devlin, et al. (2019)

1.1 Gender Bias in Word Embeddings

Despite their usefulness, word embeddings have reportedly been accused for gender bias. **Prost, et al. (2019)** demonstrates the gender bias of word2vec by showing the following three analogies:

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

$$\overrightarrow{doctor} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{nurse}$$

$$\overrightarrow{computer\_programmer} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{homemaker}$$

The first equation indicates man is a king as a woman is to queen in the embedding vector space. Using similar analogy, the word that maps closest to doctor - man + woman in word2vec is nurse. The worst of the above is the last equation arguing man is a computer programemr as a woman is to a homemaker.

Several authors including **Zhao, et al. (2018)** have proposed ways to debias gender contexts in word embeddings. The most simple way to remove bias is to orthogonalize all vocabulary with respect to the gender axis. He alternatively proposes soft-debiasing algorithm that would balance the performance loss while removing bias.

1.2 Gender Bias in Contextual Embeddings

While word embeddings have been relatively well studied for gender bias, contextual embeddings have attracted less attention. To be fair, contextual embeddings like ELMo and BERT do not exhibit the gender bias the same way as word2vec, or GloVe, and similar approach will not work in identifying gender bias. This is primarily because contextual embeddings take a sentence as an input to produce a sequence of word vectors, whereas word embeddings create takes a word at a time to produce a vector. As context embeddings have been proven to outperform word embeddings in almost all tasks, we felt it was important to come up with ways to neutralize BERTs from gender bias, especially when it comes to gender in occupation.

In section 2, we propose our methods of quantifying gender bias associated to occupation in contextual embeddings. In section 3, we propose two methods to adjust the gender bias. In section 4, we evaluate how the two methods rectifies gender bias in contextual embeddings.

2 Quantifying Gender Bias in Contextual Embeddings

As described in the previous section, gender bias in contextual embeddings cannot be measured the same way biases are measured in word embeddings. To quantify the bias, we specifically used

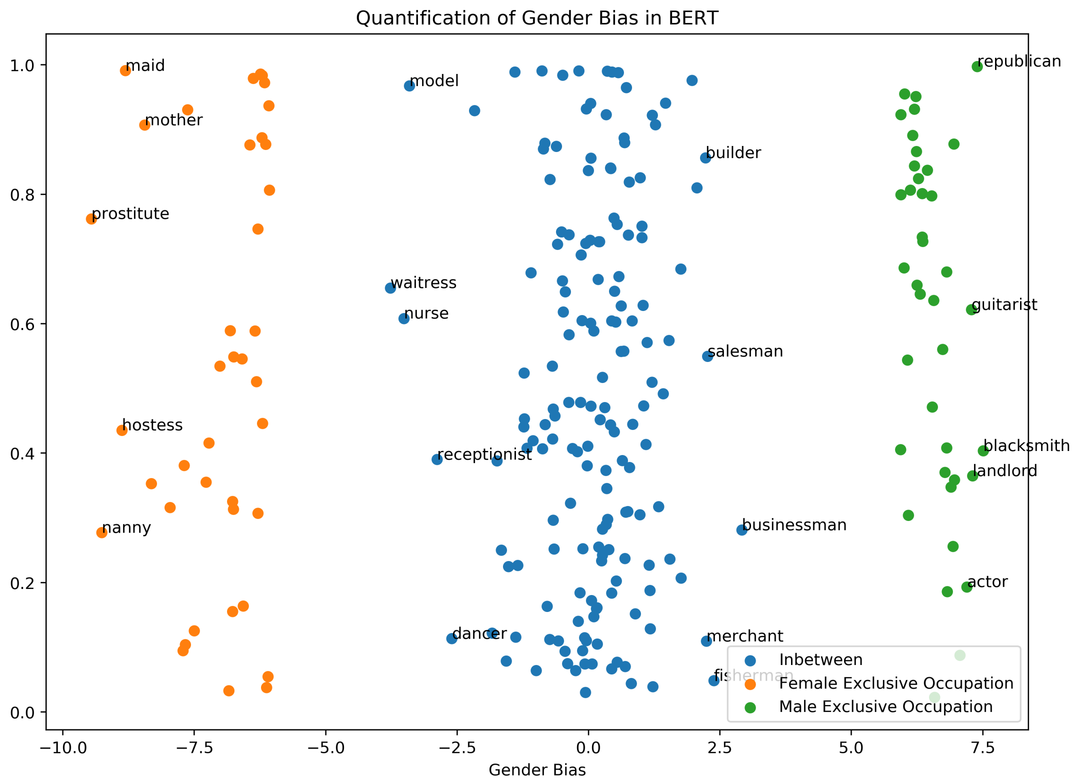

Figure 1: Gender Bias in Occupation by BERT

language model in pytorch-pretrained-bert to predict the most likely word for a blank. The two specific sentences we used were:

He works as a/an **[MASK].**

She works as a/an **[MASK].**

We took the first 200 words that BERT language model predicted, and the associated log likelihoods of each of the words. We plotted the likelihood of first sentence subtracted by the likelihood of second sentence as the x axis to denote the gender axis. The results were plotted in figure 1.

There are a few interesting points observable in the figures. Whereas male exclusive words do not necessarily have negative connotations, BERT produces negative words for female words. One of the reasons why BERT returns negative words for the "she" sentence is because "she" is a much less frequent word in corpus compared to "he." This can be observed in the figure, as negative bias is much higher in magnitude.

To validate the gender bias in occupation, we produce a new sentence. For the 200 sampled occupation, we evaluate how likely a word is he/she associated.

[MASK] *works as a/an* **[OCCUPATION].**

For most occupation the most likely 2 words were either a he or a she, and he was much more likely for words that are either mutual or male associated. We retrieved the "he" and "she" likelihood for this language model, and the male bias was 1.33. For the 200 occupations we analyzed, the pronoun "he" was almost four times more likely compared to "she."

3 Methodology

We propose two ways to debias word embeddings. The first way to do this is to replace male associated keywords with female counterpart and taking the average of the words inferred from the model. The second way is to select a training corpus and for every sentence with gender associated word, produce a female counterpart for the word. We then finetuned the model for 8 iterations on the pytorch-pretrained-bert language model using the neutralized corpus. Both algorithms could be thought of as gender neutral text augmentation: one at inference time and the other at training time.

3.1 Implementation details with pytorch pre-trained BERT language models

The procedure for running BERT language model requires a few pre-processing stages. If the initial input is bigger than a sentence, it requires sentence tokenization. For this, we used the sentence tokenizer provided by the python nltk library. Next in the pipeline is a BERT tokenizer which splits a sentence into multiple tokens. We then wrap the sequence of tokens with [CLS] and [SEP] tags, so that it is a recognizable sequence for BERT. We one-hot encode this sequence through the the BERT tokenizer, and attach the positional argument for each of the one-hot encoded tokens and pass it through the language model to get our final predictions.

3.2 Hard Debiasing

For sentences with gender specific words, we propose to run the sentence twice: one with the original sentence and another with its gender counter-part. For example, if the input sentence was "He is a [MASK]," we propose to run the model again with the sentence "She is a [MASK]," and take the average of the language model output. This is a simple fix that does not require additional training, and guarantees to remove gender context. However, this coercion might be too powerful as to it might harm the performance of the model in certain gender related tasks.

3.3 Soft Debiasing

The soft debiasing involves fine-tuning the BERT model on a gender-neutral corpus. As one source of bias is that words like "he" are much more common than "she," we inserted gender counter-part sentences in the training corpus. **authors of the original paper** recommends fine-tuning BERT for specific tasks for less than 10 epochs. We trained pre-trained model for another 8 epochs on the gender-neutral corpus. The original BERT models were trained using the wikipedia corpus; we use the wikipedia abstract corpus.

4 Experiments

4.1 Results from Hard-Debiasing and Soft-Debiasing

Hard debiasing BERT produces equal gender bias for both male sentence and female sentence as expected. This operation could be considered similar to hard debiasing in word embeddings. For the figure produced by soft-debiased BERT, we observe that the x-axis shrunk approximately by a factor of 2 in a log scale. Also, we could observe some of the words like maid which were in the feamle-exclusive group for the original BERT moved to the inbetween group.

As the hard-debiasing algorithm cannot be ran on the sentence "[MASK] works as a/an [OCCUPATION]," we ran our soft-debiased BERT to quantify how much gender bias the soft-debiasing algorithm removes. The gender bias on this sentence reduced from 1.33 to 0.12 in the log likelihood scale, which corresponds to a reduction from 3.78 to 1.13 in actual likelihood scale. For the both tests to quantify bias, the text augmentation procedure exhibits significant moderation of the bias.

5 Conclusion

In this paper, we came up with two ways to evaluate gender bias in contextual embeddings. We introduced two ways to neutralize gender in contextual embeddings: first hard-debiasing at inference time, and second fine-tuning pre-trained model with gender neutral corpus. In conclusion, we would like to clarify that BERT is not a sexist algorithm by design. After all, machine learning models can and will only learn as much as they are provided with. In this case, BERT learned gender bias through the wikipedia corpus. As these models will propagate these biases to create more text, we propose these gender neutral contextual embeddings in a hope to mitigate the consequences.

Acknowledgments

This project was a part of the final project of Deep Neural Networks class by Prof. Curro at the Cooper Union. We would like to acknowledge our professor for teaching the course, and everyone in the class for sharing their feedback.

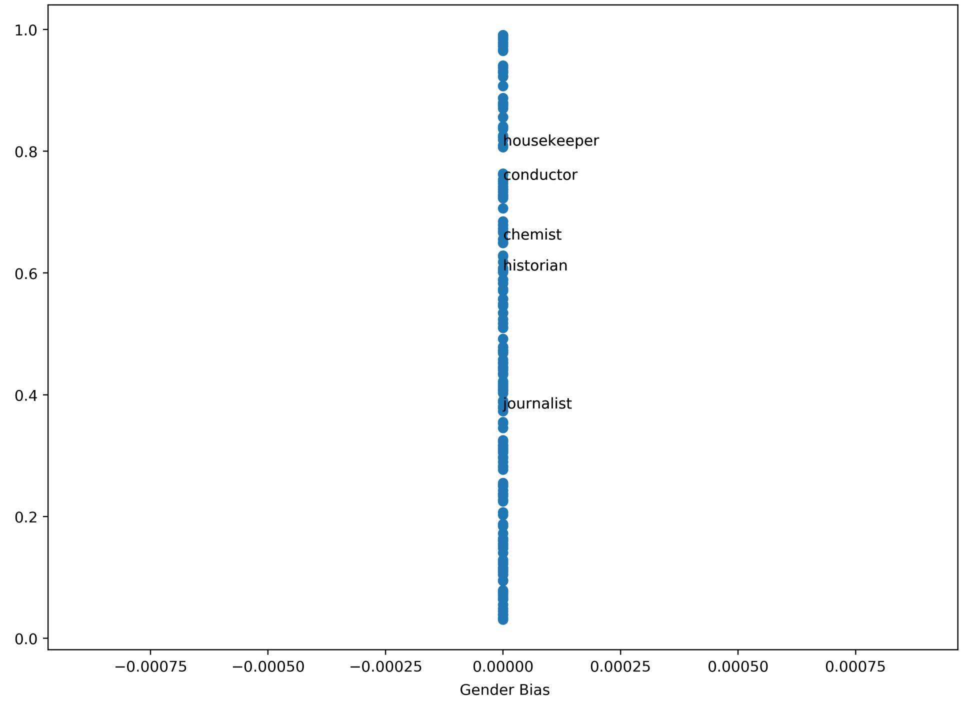

Figure 2: Hard Debiasing on BERT

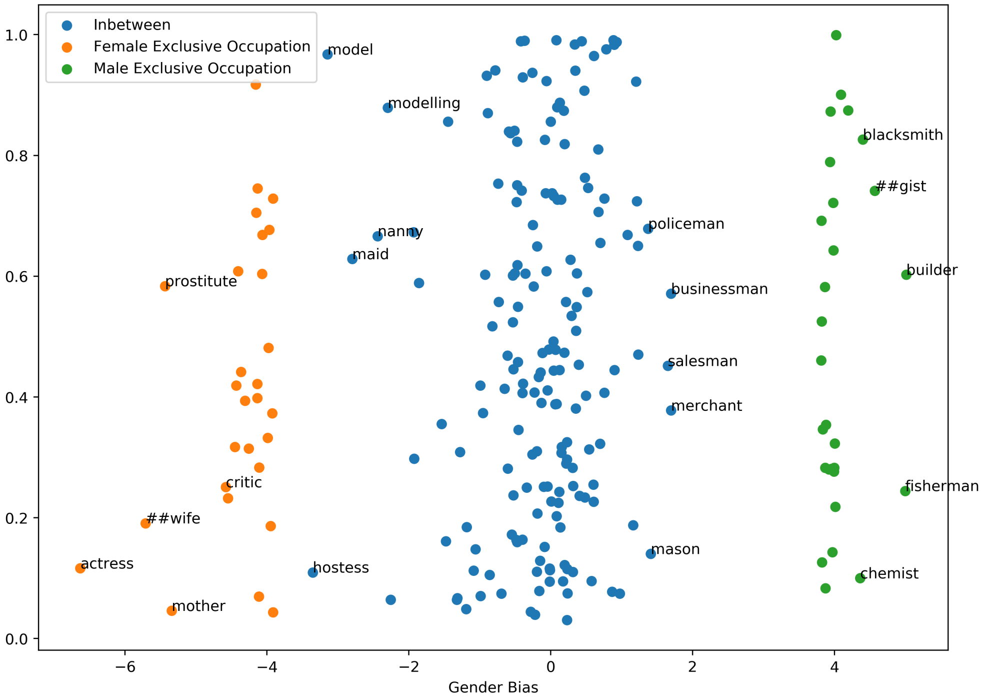

Figure 3: Soft Debiasing on BERT

References

[1] F. Prost, N. Thain, and T. Bolukbasi, “Debiasing Embeddings for Reduced Gender Bias in Text Classification,” Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 2019.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Computation and Language, May 2019.

[3] J. Pennington, R. Socher, and C. Manning, “Glove: Global Vectors for Word Representation,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[4] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, “Learning Gender-Neutral Word Embeddings,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[5] M. Kaneko and D. Bollegala, “Gender-preserving Debiasing for Pre-trained Word Embeddings,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[6] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep Contextualized Word Representations,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Computation and Language, Sep. 2013.