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

Anonymous ACL submission

Abstract

Existing studies addressing gender bias of pre-trained language models, usually build a small gender-neutral data set and conduct a second phase pre-training on th model with such data. However, given the limited size and concentrated focus of the gender-neutral data, catastrophic forgetting would occur during second-phase pre-training. Forgetting information in the original training data may damage the model's downstream performance by a large margin. In this work, we empirically show that catastrophic forgetting occurs in such methods by evaluating them with general NLP tasks in GLUE. Then, we propose a new method, GEnder Equality Prompt (GEEP), to improve gender fairness of pre-trained models with less forgetting. GEEP freezes the pre-trained model and learns gender-related prompts with gender-neutral data. Empirical results show that GEEP not only achieves SOTA performances on gender fairness tasks, but also forgets less and performs better on GLUE by a large margin.

1 Introduction

Pre-trained language models, e.g., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have shown competitive performance in a wide variety of NLP downstream applications. However, such models are often prone to exhibit gender bias (de Vassimon Manela et al., 2021; Zhao et al., 2019; Webster et al., 2020), due to their large scale unsupervised training data from the web (Liu et al., 2019; Brown et al., 2020). Gender bias refers to unbalanced model behaviors with respect to a specific gender (Cheng et al., 2020). Among various gender-biased behaviours of pre-trained models, bias on professions is the most prominent and well-studied (de Vassimon Manela et al., 2021; Vig et al., 2020; Qian et al., 2019; Zhao et al., 2019). For example, in coreference resolution tasks, a pre-trained model would predict female pronoun and names for professions like "nurse" and "housekeeper", while predict male pronouns for "computer programmer" or "doctor" (Kurita et al., 2019).

Given the large model size and tremendous time complexity for language model pre-training, training a gender-neutral model from scratch with manually filtered data seems impossible for most organizations. Due to this limitation, existing studies usually build a relatively small gender-neutral data set (for example building a data set that have more balanced gender pronouns for profession names), and conduct second phase pre-training on the pre-trained model with such data (Webster et al., 2020; de Vassimon Manela et al., 2021). However, given the limited size of the gender-neutral data and its potential distributional mismatch with the original pre-training data, *catastrophic forgetting* can occur during second-phase pre-training of such methods. Catastrophic forgetting (Kirkpatrick et al., 2017) is a long-standing problem which illustrates the tendency of a neural network to forget previously learned information upon learning new information. When it comes to further training a pre-trained model, using the small gender-neutral data to update the entire massive model could potentially make the model forget the diverse information from the original pre-training data, which may damage the model's downstream performance by a large margin.

In this paper, we first empirically verify that further updating a pre-trained model (such as RoBERTa (Liu et al., 2019)) with manually-built gender-neutral data can cause catastrophic forgetting. We follow existing work and build our profession-related gender-neutral data set by filtering out Wikipedia sentences mentioning professions and swapping their gender related pronouns. We find that although our gender-neutral data is from Wikipedia which is part of RoBERTa's pre-training data, the model's performance on downstream tasks in GLUE (Wang et al., 2018) still drops with a considerable margin after second-phase pre-training, due to the smaller size and more concentrated focus of the gender-neutral data.

Therefore, we propose a new method, GEnder Equality Prompt (GEEP), to alleviate gender bias of pre-trained models without catastrophic forgetting. Specifically, inspired by recent prompt-tuning methods (Lester et al., 2021) for fine-tuning large pre-trained models, GEEP freezes the entire model, adds and updates new word embeddings of professions as gender equality prompts, instead of updating all model parameters at second-phase pre-training as previous methods. Since all the pre-trained parameters are frozen during further training, diverse information from the original training data preserved in the pre-trained parameters is not erased. Therefore forgetting can be alleviated to large extent. Moreover, since the embeddings of professions are re-initialized when debiasing training starts, gender bias from previous data that is embedded in such representations is already removed before second-phase pre-training. Therefore, GEEP also improves gender fairness of the model more effectively with much fewer iterations. Empirical results show that GEEP not only achieves state-of-the-art performances with fewer iterations on various gender fairness tasks such as pronoun coreference resolution, but also forgets less and achieves better results on general GLUE tasks.

2 Related Work

Compared with the existing work focusing on quantifying and alleviating gender bias (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2018b; Gonen and Goldberg, 2019; Sun et al., 2019; Garg et al., 2018; Zhao et al., 2018a; Bolukbasi et al., 2016; Zhao et al., 2018b) in standard word embedding models, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), gender bias in large pre-trained language models seems less studied. Recent work on gender fairness of pre-trained language models, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), mostly focus on showing and measuring the gender bias embedded in such models (Zhao et al., 2019; Tan and Celis, 2019). These studies propose metrics to quantify gender bias in pre-trained language models (de Vassimon Manela et al., 2021; Tan and Celis, 2019; Webster et al., 2018; Kurita et al., 2019). In our work, we employ such methods to evaluate GEEP and baseline methods on improving gender fairness. Existing works focusing on mitigating gender bias of pre-trained models usually collect and build gender-neutral data on their own and conduct a second phase pre-training on the released pre-trained model (Webster et al., 2020; de Vassimon Manela et al., 2021; Cheng et al., 2020). For example, Cheng et al. (2020) take advantage of such data augmentation methods and train a fair filter (FairFil) network to maximize the mutual information between the representations of the original sentences and their corresponding augmentations. In this work, we demonstrate empirically that even if the gender-neutral data for second-phase pre-training comes from the original training data set, the performance of the debiased model on general downstream tasks such as GLUE, still drops by a considerable margin after the second-phase pre-training. Then, given this phenomenon, we propose GEEP to alleviate gender bias in pre-trained models without forgetting.

3 Improving Gender Fairness without Forgetting

In this section, we first describe the gender-neutral collection method we adopt from existing methods and the forgetting issue in such methods. Then we describe the proposed method GEnder Equality Prompt (GEEP).

3.1 Profession-Related Gender-Neutral Data Collection

We follow existing work to build a profession-related gender neutral data set since profession-related gender bias is a relatively well-studied aspect of gender bias. To construct profession-related data with equal numbers of references to male and female genders, we adopt the data filtering method by (Zhao et al., 2018a) on the English Wikipedia corpus. Specifically, we filter Wikipedia for sentences containing at least one profession that is supposed to be gender-neutral but generally viewed with gender bias, e.g., nurse, defined by (Bolukbasi et al., 2016). For each of these sentences, we swap the gendered terms with their opposite genders (such as "Man" →"Woman", "he"→"she", and vice-versa). Our dataset includes both the original profession-related sentences and their gender-swapped counterparts. After such processing, we get 6.1GB of profession-related gender-neutral text data. Compared with the original pre-training data of RoBERTa (160GB in text size from various

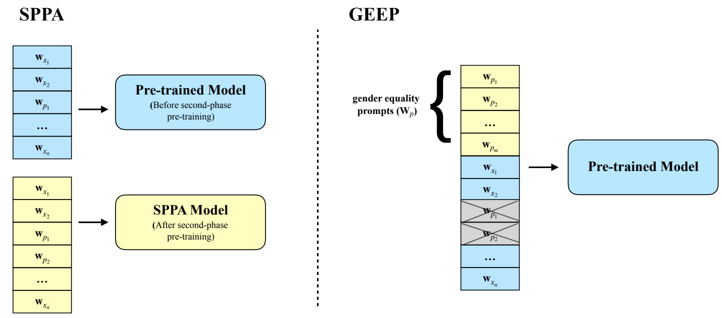


Figure 1: Difference between SPPA and GEEP methods. Blue boxes represent the parameters of the pre-trained model before any further training and yellow boxes show updated parameters during second-phase pre-training (SPPA). SPPA requires updating all the pre-trained model's parameters. In contrast, GEEP only adds and updates new embeddings of biased professions such as $\mathbf{w}_{p_i}$. Gray boxes are the original embeddings of professions which are not updated/used in second phase pre-training or the training/inference after that.

sources), the gender-neutral data we have is smaller and less diverse.

After the gender-neutral data set is built, a common approach to mitigate gender bias in pre-trained language models is to conduct second-phase pre-training to update all model parameters with this data set. We refer to such methods as *SPPA* (Second-Phase Pre-training for All parameters). In Section 4, we empirically show that SPPA methods lead to forgetting and the model's performance on general NLP benchmark GLUE drops by a large margin.

3.2 Gender Equality Prompt Approach

To alleviate forgetting while mitigating gender bias in pre-trained language models, we propose GEnder Equality Prompt (GEEP). In GEEP, instead of updating all model parameters during second-phase pre-training, we freeze all of the pre-trained model parameters and add new trainable embeddings for profession names as gender equality prompts. Since all previous pre-trained parameters are frozen, diverse information from original massive pre-training data that are memorized by the pre-trained parameters wouldn't be erased. Therefore, the forgetting of information from the original training data can be alleviated to the fullest extent.

Let $\mathbf{X} = \{x_1, x_2, ...x_n\}$ denote the original vocabulary of the pre-trained model and $\mathbf{W}_x \in R^{n\times d}$ be the original pre-trained token embedding matrix of the model with dimension of d. Given a set of m profession names, $\{p_1, p_2, ..., p_m\}$, we build an embedding matrix $\mathbf{W}_p \in R^{m\times d}$ where the embedding of each token is initialized randomly. To obtain an integrated word embedding matrix, we concatenate $\mathbf{W}_x$ and $\mathbf{W_p}$ as $\mathbf{W}_{\text{emb}} = \text{Concat}(\mathbf{W}_x, \mathbf{W}_p)$. During both second-phase pre-training and the training/inference after that, once a profession occurs, we only update/use its new embedding in $\mathbf{W}_p$. We show the comparison between GEEP and other second-phase pre-training methods in Figure 1. Given all the pre-trained model's frozen parameters $\mathbf{W}_{\text{whole}}$ that contains $\mathbf{W}_x$, the objective function of second-phase pre-training of GEEP is,

$$\mathcal{L}(\mathbf{x}_{\text{masked}}|\mathbf{x}_{\text{context}}, \mathbf{W}_{\text{whole}}) \tag{1}$$

$$= \frac{1}{N_{\text{mask}}}\left(\sum_{t=1}^{N_{\text{mask}}} -\log p_\theta(x_t|\mathbf{x}_{\text{context}}, \mathbf{W}_{\text{whole}})\right). \tag{2}$$

N_{mask} is the number of masked positions in the input sequence $\mathbf{x}$. With such an objective, $\mathbf{W}_p$ is updated with gender-neutral data. Moreover, since the embeddings of professions are re-initialized when debiasing training starts in GEEP, gender bias from previous data that is embedded in such representations is already erased before second-phase pre-training. Therefore, it is also easier for GEEP to debias the model during further pre-training.

4 Experiments

In this section, we present the results of GEEP and its baselines to show that GEEP achieves state-of-the-art performances on gender fairness tasks while alleviating the forgetting issue of the baselines.

4.1 Experimental Setup

In our experiments, we mainly use the publicly released RoBERTa-base model as the pre-trained model. We have also conducted experiments on publicly released BERT during preliminary explorations. Details on BERT experiments are in Appendix A.5. Given a pre-trained RoBERTa-base model, we compare GEEP with two main baselines.

Table 1: The average accuracy of different models on Coreference Resolution task. The best results are in bold.

Data	RoBERTa	SPPA	GEEP
Winogender	50.9	57.3	**64.5**
WSC	50.1	50.9	**52.7**
DPR/WSCR	50.8	51.1	**53.6**

Table 2: GLUE results. The best results among SPPA and GEEP are in bold.

Task	RoBERTa	SPPA	GEEP
MNLI	87.7	87.2	**87.7**
QNLI	92.4	92.4	**92.4**
QQP	91.8	91.3	**91.7**
SST-2	95.4	94.7	**95.4**
CoLA	64.1	38.9	**50.5**
MRPC	91.4	88.8	**89.8**
RTE	78.4	60.2	**68.7**
STS-B	90.7	88.3	**89.9**
AVG	86.5	80.2	**83.3**

The first baseline is the pre-trained RoBERTa-base model without any further training. The other important type of baselines are SPPA methods. For a fair comparison, our SPPA baseline uses the same gender-neutral data set that we construct for GEEP (details in Section 3.2) to further update all model parameters of the pre-trained RoBERTa-base. The detailed hyper-parameter settings of GEEP and SPPA can be found in Appendix A.1.

4.2 Evaluation Tasks

To assess gender fairness, we conduct pronoun coreference resolution experiments on different data sets, Winogender (Rudinger et al., 2018), Winograd Schema Challenge (WSC) (Levesque et al., 2012), and Definite Pronoun Resolution (DPR) (Rahman and Ng, 2012). Pronoun coreference resolution is the task of linking the pronouns with their references in a text. In order to resolve a pronoun accurately, a model needs to overcome the biased link between gender and profession (e.g. the assumption that nurses are female) and instead make the decision based on available linguistic cues. Therefore, better performances on pronoun coreference resolution usually indicates that less gender bias is preserved by the model (Kurita et al., 2019). Detailed setups of this experiment can be found in Appendix A.2. Additionally, we also qualitatively and quantitatively evaluate our method on direct pronoun prediction. Details of this experiment are in Appendix A.4.

To evaluate how much each debiased model forgets after second-phase pre-training, we report the performances of the debiased models on GLUE benchmark (Wang et al., 2018). Detailed setups of this experiment can be found in Appendix A.3.

4.3 Results

We first show the pronoun coreference resolution results of different models on three datasets in Table 5. Results show that GEEP model obtains the best accuracy compared to other models, especially on the Wingender dataset where the candidate nouns are professions. We also conduct an ablation study to show the effect of total training iterations on the performances of both methods. We find that GEEP can improve the model's performance with significantly fewer number of training iterations. Details are in Appendix A.1.

Then we show in Table 4 the performance of different models on 8 GLUE tasks, to see how severe the forgetting issue is after the second-phase training of SPPA and GEEP. Compared with RoBERTa, SPPA suffers from forgetting issue in 7 out of 8 tasks except QNLI. For tasks like CoLA and RTE, the model's performance drops significantly (more than 10 points) after SPPA. For tasks with larger data set for fine-tuning, such as MNLI, QQP and SST-2, they are less vulnerable to the quality of pre-training (Wu et al., 2020; Joshi et al., 2020). Therefore, SPPA's performance drop on such data sets is less significant. GEEP mitigates the forgetting issue of SPPA in all sub-tasks. Since GEEP ditches the original pre-trained profession embeddings and uses a smaller data set to update new profession embeddings, the forgetting problem cannot be fully avoided. While GEEP still achieves an average GLUE score of 83.3, significantly outperforming SPPA.

5 Closing Remarks

In this paper, we proposed GEEP to improve gender fairness of pre-trained language models with less catastrophic forgetting. For a fair comparison to existing work under the current gender fairness literature, we mainly conduct experiments with profession-related gender neutral data because profession-related gender bias is relatively more well studied so far. The good empirical results indicates it is worth to try applying GEEP to other more challenging and under-explored aspects of gender fairness, which would be our future work.

References

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. *Advances in neural information processing systems*, 29:4349–4357.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2020. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

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. 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 609–614.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

V Kocijan, O-M Camburu, A-M Cretu, Y Yordanov, P Blunsom, and T Lukasiewicz. 2019. Wikicrem: A large unsupervised corpus for coreference resolution. volume D19-1, page 4294–4303. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, New Orleans, Louisiana. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *NeurIPS*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Qiyu Wu, Chen Xing, Yatao Li, Guolin Ke, Di He, and Tie-Yan Liu. 2020. Taking notes on the fly helps language pre-training. In *International Conference on Learning Representations*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

A Appendix

A.1 Hyper-parameters for SPPA and GEEP

For the main results presented in the paper of second-phase pre-training in GEEP and SPPA, we further train RoBERTa-base for $100,000$ steps with our gender-neutral data. We use an AdamW optimizer with a learning rate of $1e-5$, max_seq_length of 128 and batch size 256. In GEEP method, we initialize the embedding of every profession prompt with a normal distribution and standard deviations of 0.2.

Alongside the final results, we also evaluate SPPA and GEEP during the second-phase pre-training. In Table 3 we show SPPA and GEEP's performance on pronoun coreference resolution at the 20k iteration and 50k iteration. From Table 3 we can see that GEEP improves the pre-trained model's gender fairness with much less number of iterations. At 20k iteration, GEEP's performance is already better than SPPA's final performance on all 3 tasks. At 50k iteration, GEEP's performance has almost converged to its final scores on all 3 tasks. While SPPA's performance is still far behind its final performances on Winogender and WSC.

A.2 Pronoun Coreference Resolution Experiment Setup

Pronoun Coreference Resolution is the task of linking the pronouns with their references in a text. Studies show that BERT performance decreases in a text where the gender pronoun is female and the topic is biased towards the male gender (Kurita et al., 2019). To assess the performance of different models in pronoun coreference, we fine-tune our models with GAP data set (Webster et al., 2018) We fine-tune each model for one epoch with a train batch size of 64 and a learning rate of $5.0e-6$. After fine-tuning, we evaluate the performance of different models on three data sets:

- Winogender: This dataset includes $1,584$ sentences with three mentions: a profession, a participant, and a pronoun (where the pronoun is referred to either profession or pronoun)(Rudinger et al., 2018).
- WSC: The Winograd Schema Challenge (WSC) incorporates 273 sentences used for commonsense reasoning for resolution (Levesque et al., 2012).
- DPR: The Definite Pronoun Resolution (DPR) corpus with 131 test sentences contains examples with two noun phrases and a pronoun or possessive adjective referring to one of the noun phrases (Rahman and Ng, 2012).

A.3 GLUE Experiment Setup

To evaluate how much each debiased model forgets after second-phase pre-training, we fine-tune the pre-trained models on GLUE (**G**eneral **L**anguage **U**nderstanding **E**valuation) to evaluate the performance of the pre-trained models. We follow previous work to use eight tasks in GLUE, including CoLA, RTE, MRPC, STS, SST, QNLI, QQP, and MNLI. For evaluation metrics, we report Matthews correlation for CoLA, Pearson correlation for STS-B, and accuracy for other tasks. We use the same optimizer (Adam) with the same hyper-parameters as in pre-training. Following previous work, we search the learning rates during the fine-tuning for each downstream task. For a fair comparison, we do not apply any published tricks for fine-tuning. Each configuration is run five times with different random seeds, and the *average* of these five results on the validation set is calculated as the final performance of one configuration. We report the best number over all configurations for each task.

A.4 Pronoun Prediction Experiment Setup and Results

Different approaches have been proposed to quantify and analyze the gender bias in contextual language models (de Vassimon Manela et al., 2021; Webster et al., 2020; Kurita et al., 2019). For BERT, we choose one approach that can be directly applied to a model pre-trained with Masked Language Modeling (MLM) loss without further fine-tuning. In this approach, we first define a template containing a pronoun and a profession. The profession is supposed to be gender-neutral however it is currently viewed with gender bias to a large extent. By masking the pronoun, the model is queried to predict the pronouns at the masked position given the context, including the profession. Here is an example, "[MASK]" is a registered nurse. The difference between the probabilities of filling the masked position in each sentence with "he" and "she", is used to show gender bias in the model,

$$\text{Pronoun Bias Score} = \tag{3}$$

$$\text{Prob}(\textbf{"he"}) - \text{Prob}(\textbf{"she"}). \tag{4}$$

To assess fairness in BERT model, we consider 303 of professions used by (Bolukbasi et al., 2016). In

Table 3: The average accuracy of different models on Coreference Resolution task. The best results are in bold.

Data	RoBERTa	SPPA-20k	GEEP-20k	SPPA-50k	GEEP-50k	SPPA-100k	GEEP-100k
Winogender	50.9	51.6	64.3	54.6	64.5	57.3	**64.5**
WSC	50.1	50.1	52.1	50.5	52.3	50.9	**52.7**
DPR/WSCR	50.8	50.9	52.1	51.1	53.4	51.1	**53.6**

our study, we analyze a public available pre-trained BERT-Base model [1] that contains 12 layers, 768 hidden nodes, 12 heads, and 110M parameters. Figure 2 shows gender bias of 60 of such professions in BERT-base model. Positive values mean that the professions are biased towards male and vice versa. As the plots show, the contextual representations of professions in BERT-base model exhibits strong gender bias. Professions such as nurse and housekeeper are viewed as jobs for females while surgeon and mathematicians are assumed to be jobs for males.

To find the reference of each pronoun in the template sentences, we follow (Kocijan et al., 2019) approach. Specifically, during the evaluation for every data set, in each sentence there are two candidate nouns (such as "nurse" or "surgeon") and a pronoun. The pronoun is replaced with a [MASK] token, and the model makes a prediction at the masked pronoun position from the two candidate nouns. In order to resolve a pronoun accurately, a model needs to overcome the biased link between gender and profession (e.g. a normative assumption that nurses are female) and instead make the decision based on the available linguistic cues. We report the prediction accuracy of all 3 methods on the aforementioned three data sets.

Figure 3 displays the pronoun prediction bias score (defined in Equation 5) of all methods for 60 biased professions defined in (Bolukbasi et al., 2016). Specifically, in both sub-figures, blue dots show the pronoun prediction bias score from BERT-base model for each profession. In Figure 3 (a), the pink dots are the bias scores from BERT-SPPA model. We can see from this sub-figure that compared with BERT-base, the bias scores from BERT-SPPA model are indeed closer to 0, indicating that BERT-SPPA can mitigate gender bias of such professions to some extent. In Figure 3 (b), the blue dots are the bias scores from GEEP model. Compared with both BERT-SPPA and BERT-base, GEEP's bias scores are significantly closer to 0, indicating that GEEP is more effective at removing gender bias from such biased professions compared with BERT-SPPA. Moreover, we also calculate the average absolute pronoun prediction bias score for all 303 gender-neutral profession words in (Bolukbasi et al., 2016). We obtain 0.44 for BERT-base, 0.16 for BERT-SPPA and 0.13 for GEEP. GEEP model gets the lowest average bias with 70% reduction compared to the BERT-base model.

[1] https://github.com/google-research/bert

Table 4: GLUE results. The best results are in bold.

Task	BERT-base	BERT-SPPA	GEEP
MNLI	84.3	84.0	**84.1**
QNLI	91.4	90.0	**91.3**
QQP	90	90.1	**90.4**
SST-2	93	92.2	**92.4**
CoLA	54.0	52.0	**53.0**
MRPC	85.7	84.1	**84.9**
RTE	69.4	**69.8**	69.1
STS-B	88.0	**88.0**	87.0
AVG	82.0	81.3	**81.6**

A.5 Experiment Results on BERT

During the preliminary exploration on this problem, we have also applied SPPA and GEEP on publicly released BERT and conducted pronoun coreference resolution and GLUE experiments on them. In this experiment, we only further trained the released BERT model for 10k iterations with our gender-neutral data. Moreover, our gender-neutral data set (7.1 GB) is not significantly smaller than the original pre-training data of BERT (16 GB), and the two data sets both come from Wikipedia. Due to these two reasons, the forgetting problem on this BERT experiment is not as obvious for SPPA.

Table 4 shows the performance of different methods on 8 GLUE tasks. Although the forgetting is less server, SPPA still suffers from forgetting issue in the following 6 tasks out of the total 8 tasks, CoLA, MRPC, STS-B, MNLI, QNLI, and SST-2. As for the average GLUE score, SPPA is 0.7 point lower after its second-phase pre-training, which is not a small margin considering it is the average score of 8 tasks. GEEP mitigates the forgetting issue of SPPA in all sub-tasks except in RTE. GEEP also gets the average GLUE score of 82.8, which outperforms SPPA and is similar to the original

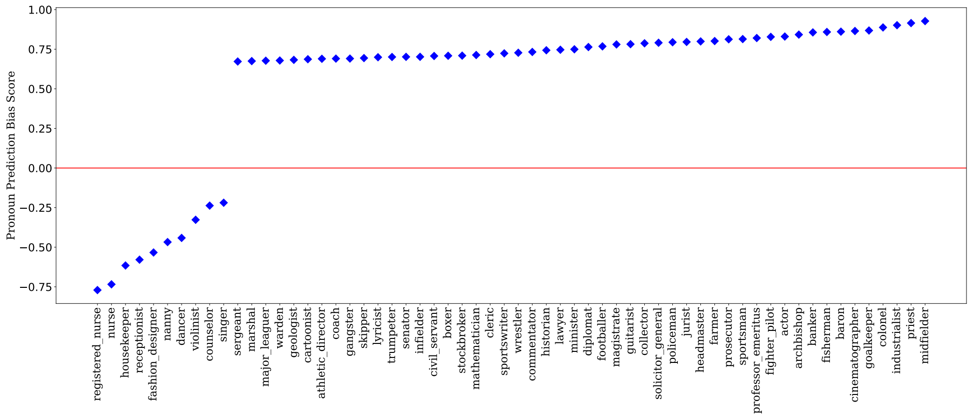
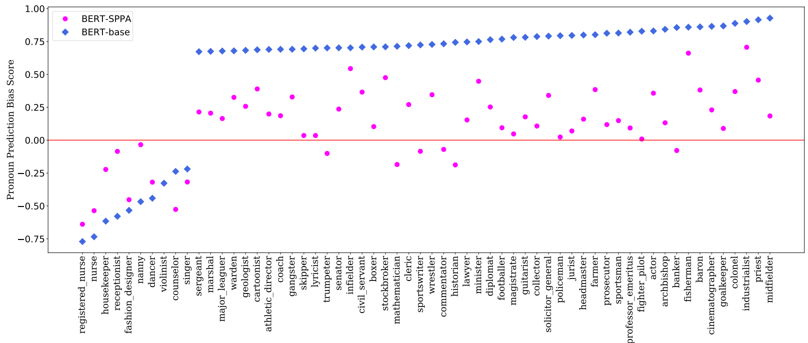
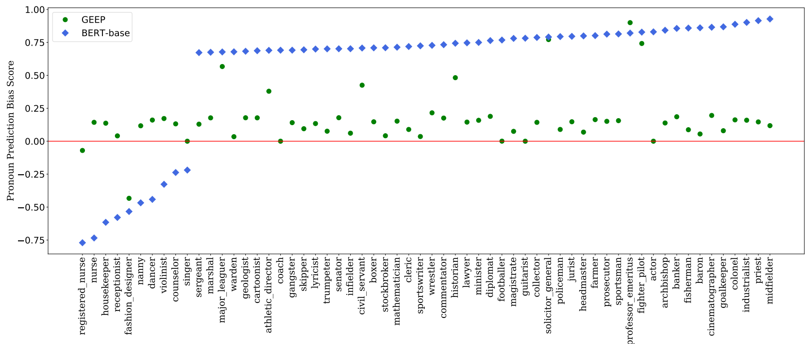


Figure 2: An example of gender bias in 60 most biased profession words in BERT-base model. For each profession, we measure the difference between the probability of filling the masked pronoun in each template sentence with "he" and "she" tokens. Some words such as nurse (-0.73) and receptionist (-0.57) are supposed to be gender neutral by definition but BERT-base model consider them as female professions. On the other hand, lawyer (0.74) and prosecutor (0.81) are considered as jobs for males.



(a) Comparison between pronoun prediction bias in SPPA and BERT-base models



(b) Comparison between pronoun prediction bias in GEEP and BERT-base models

Figure 3: Difference between the probabilities of filling a masked pronoun with "he" and "she" tokens in the template sentences containing 60 most biased professions. GEEP method outperforms the two other methods. For example, the bias score for "nurse" token decreases from -0.7 in BERT-base to -0.5 in BERT-SPPA and 0.1 in GEEP model.

Table 5: The average accuracy of different models on Coreference Resolution task. The best results are in bold.

Data	BERT-base	BERT-SPPA	GEEP
Winogender	50	50.7	**62.9**
WSC	50.1	50.2	**50.5**
DPR/WSCR	50.7	50.9	**52.8**

GLUE score of the pre-trained BERT.

Table 5 shows the coreference resolution results of different models on three data sets. Results show that GEEP model obtains the best accuracy compared to other models, especially in Wingender dataset where the candidate nouns are professions. We observe that the SPPA method also can help improve coreference resolution performance of the pre-trained model, but not as effective as GEEP.