Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations

Anonymous EMNLP submission

Abstract

Pretrained language models (PLMs) learn stereotypes held by humans and reflected in text from their training corpora, including gender bias. When PLMs are used for downstream tasks such as picking candidates for a job, people's lives can be negatively affected by these learned stereotypes. Prior work usually identifies a linear gender subspace and removes gender information by removing the subspace. Following this line of work, we use DensRay, an analytical method for obtaining interpretable ultradense subspaces. We show that DensRay performs on-par with prior approaches, but is more stable. In addition, DensRay can be used to obtain interpretable gender scores on token level for all representations. Finally, we demonstrate that we can remove bias multilingually, e.g., from Chinese, using only English training data.

1 Introduction

Word embeddings, which represent the semantic meaning of text data as vectors, are used as input in natural language processing tasks. It has been found that word embeddings exhibit biases such as gender bias, which are present in their training corpora (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018). Contextual word embedding models, such as BERT (Devlin et al., 2018), have become increasingly common and achieved new state-of-the-art results in many NLP tasks. Researchers have also found gender bias in contextualized embeddings (Zhao et al., 2019; May et al., 2019).

A common approach for removing gender information in static embeddings is to identify a linear gender subspace (e.g., a gender direction) and subsequently setting all values on the gender direction to 0. Successful approaches rely on simple principal component analysis (Bolukbasi et al., 2016; Mu and Viswanath, 2018). Bolukbasi et al. (2016) require pairs of gendered words to compute a direction (e.g., "man"-"woman") and Mu and Viswanath (2018) rely on computing a PCA of a set of gender words hoping that the main variation occurs across gender. We propose to use DensRay (Dufter and Schütze, 2019): the main advantage is that DensRay only requires two or multiple groups of gendered words. In contrast to (Bolukbasi et al., 2016), it does not require explicit pairs. Compared to (Mu and Viswanath, 2018), it has explicit supervision with gender labels. We show in §5.5.1 that DensRay is more stable.

In summary our contributions are: i) We adjust DensRay to work on contextualized embeddings. We apply DensRay to every BERT layer and evaluate two tasks: a set of templates we constructed and the Word Embedding Association Test (WEAT) (Caliskan et al., 2017). Our experiments find that debiasing with DensRay effectively mitigates gender bias and performs on par with prior approaches. ii) We show that DensRay is more robust and interpretable than prior approaches. iii) We investigate whether debiased models maintain the performance of BERT on language modeling and GLUE (Wang et al., 2018). iv) We apply our debiasing method to the multilingual-BERT (mBERT) model: we show that English training data can be used to effectively debias Chinese.

1 <<< pd: I would suggest a different structure for the paper: Introduction Methods Hard Debiasing DensRay (potentially add the robustness figure here) Debiasing Contextualized Embeddings Experiments Setup and Data Evaluation Templates WEAT GLUE Results Debiasing Results (OCCTMP, WEAT) Model Performance Examples Analyses (Layer, Attention Heads, Requried samples) Multilingual Debiasing Related Work Conclusion

>>>

2 <<< hs: i like this proposed structure! >>>

2 Background

2.1 Quantifying Gender Bias

A typical way to measure gender bias is to evaluate on **downstream tasks**. For coreference resolution, Zhao et al. (2018a) designed Winobias and Rudinger et al. (2018) designed Winogender schemas. In contrast to WinoBias, Winogender schemas include gender-neutral pronouns. One Winogender schema has one occupational mention and one "other participant" mention while WinoBias has two occupational mentions. **3 <<< pd: is the difference between Winobias and Winogender relevant to this work? >>>** Webster et al. (2018) released GAP, a balanced corpus of Gendered Ambiguous Pronouns, which measures gender bias as the ratio of F1 score on masculine to F1 score on feminine. However the ratio is very close to 1 (Chada, 2019; Attree, 2019) making it hard to compare debiasing systems. For sentiment analysis, Equity Evaluation Corpus (EEC) (Kiritchenko and Mohammad, 2018) was designed to measure gender bias by the difference in emotional intensity predictions between gender-swapped sentences.

An alternative way to measure gender bias is based on **association tests**, which originated from sociological research. Greenwald et al. (1998) proposed the Implicit Association Test (IAT) to quantify societal bias. In the IAT, response times were recorded when subjects were asked to match two concepts. For example, subjects were asked to match black and white names with "pleasant" and "unpleasant" words. Subjects tended to have shorter response times for concepts they thought associated. Based on the IAT, Caliskan et al. (2017) proposed the Word Embedding Association Test (WEAT), which uses word similarities between targets and attributes instead of the response times to get rid of the requirement of human subjects. May et al. (2019) extended WEAT to the Sentence Embedding Association Test (SEAT); Kurita et al. (2019) proposed a template-based log probability bias score to measure the association between targets and attributes in BERT.

4 <<< hs: for many of the papers you discuss above it's not clear what the realtinship to the current work is. this hsould always be clear >>>

2.1.1 Word Embedding Association Test

Here we introduce WEAT in detail. Consider two sets of target words X_1, X_2 with equal size $|X_1| = |X_2|$, and two sets of attribute words A_1, A_2 with $|A_1| = |A_2|$. The null hypothesis in the statistical test of WEAT is: there is no difference in the cosine similarity between X_1, X_2 and A_1, A_2. Taking the measurement of gender bias as an example, word sets about science and art can be used as the two target sets, masculine and feminine names as the two attribute sets. Intuitively, the null hypothesis means science and art are equally similar to each masculine and feminine name. In the prior literature it has been argued that if the null hypothesis cannot be rejected, there is no significant gender bias. **5 <<< pd: we should critisize this reasoning. The null and alternative hypothesis should be swapped. Has other work criticized this setup? Maybe we can do the test in addition in an alternative way? >>>** The WEAT test statistic is defined as

$$\begin{aligned} s(X_1, X_2, A_1, A_2) = & \sum_{x \in X_1} s(x, A_1, A_2) \\ & - \sum_{x \in X_2} s(x, A_1, A_2), \end{aligned}$$

where

$$\begin{aligned} s(x, A_1, A_2) = & \text{mean}_{a \in A_1} cos(\vec{x}, \vec{a}) \\ & - \text{mean}_{a \in A_2} cos(\vec{x}, \vec{a}) \end{aligned}$$

$cos(\vec{x}, \vec{a})$ denotes the cosine similarity between embedding vector $\vec{x}$ and $\vec{a}$. Intuitively, $s(x, A_1, A_2)$ measures the association of a word with the attributes, so the test statistic measures the differential association of the two target sets with the attributes.

Let $\{(X_{1i}, X_{2i})\}_i$ denote all the partitions of $X_1 \cup X_2$. The one-sided p-value of the permutation test is defined as

$$Pr_i[s(X_{1i}, X_{2i}, A_1, A_2)] > s(X_1, X_2, A_1, A_2)$$

6 <<< pd: I do not understand the notation fully. Is the p-value computed with respect to a single partition i? >>> The effect size d-value is a normalized measure of how separated the two distributions of associations between the target and attribute are. It is defined as

$$d = \frac{s(X_1, X_2, A_1, A_2)}{\text{std}_{x \in X_1 \cap X_2} s(x, A_1, A_2)}.$$

7 <<< hs: above: $x \in X_1 \cap X_2$ or $x \in X_1 \cup X_2$ >>>

8 <<< hs: i think there is a summary sentence missing here: how do we use this to evaluate / compare debiasing methods? >>>

2.2 Debiasing Methods

Many methods to remove gender bias have been proposed. The most common way is to define a gender direction (or, more generally, a subspace) by a set of gendered words, and debias the word embeddings in a post-processing projection. Bolukbasi et al. (2016) propose (i) *hard debiasing*: they use the gendered words to compute the difference embedding vector as the gender direction; and (ii) *soft debiasing*, a machine learning based method that combines the inner-products objective of word embedding and an objective to project the word embedding into an orthogonal gender subspace. Hard debiasing has been found to work better. **9 <<< pd: should we mention hard-debiasing by mu et al here and explain the difference to bolukbasi?** >>> Dev and Phillips (2019) explored partial projection and some simple tricks to improve the hard debiasing method. Zhao et al. (2019) applied the data augmentation and debiasing method of Bolukbasi et al. (2016) to mitigate gender bias on ELMo (Peters et al., 2018). Karve et al. (2019) introduce the debiasing conceptor: they shrine each principal component of the covariance matrix of the word embeddings to achieve a soft debiasing. Besides the above post-processing methods, (Zhao et al., 2018b) propose GN-Glove: it debiases during training to learn word embeddings with protected attributes. The method we use here, DensRay, is similar to hard debiasing in that we find and eliminate a gender subspace in post-processing. But DensRay can be solved efficiently in closed form and it is more stable than hard debiasing.

10 <<< hs: above: what does “shrine” mean? >>>

2.3 DensRay

DensRay is an analytical method for identifying the embedding subspace of certain linguistic features. Similar to the methods mentioned in the previous section, we aim to identify the “gender subspace” using a set of gendered words $V := \{v_1, v_2, \ldots, v_n\}$ and their embeddings $E \in R^{n\times d}$, thus for word v_i we have the corresponding embedding vector e_{v_i}. We introduce a function l for the gender attribute: $l : V \to \{-1, 1\}$; e.g. $l(father) = 1, l(sister) = -1$. The objective of DensRay is to find an orthogonal matrix $Q \in R^{d\times d}$ such that EQ is gender-interpretable, specifically, the first k dimensions can be interpreted as the gender subspace.

Let $L_= := \{(v, w) \in V \times V | l(v) = l(w)\}$ and define $L_{\neq}$ analogously. The DensRay objective in Eq. 1 is to maximize the distance of the word pairs from the same gender group ($L_=$) and minimize the distance of the word pairs from the different gender group ($L_{\neq}$).

$$\max_q \sum_{(v,w)\in L_{\neq}} \alpha_{\neq} ||q^T d_{vw}||_2^2 - \sum_{(v,w)\in L_=} \alpha_= ||q^T d_{vw}||_2^2 \quad (1)$$

where we define $d_{vw} := e_v - e_w$. We also have $q \in R^d$ and $q^T q = 1$ since Q is orthogonal, and $\alpha_{\neq}, \alpha_= \in [0, 1]$ are hyperparameters. Regard that $||x||_2^2 = x^T x$, objective **??** can be simplified to:

$$\underset{q}{max}\, q^T (\sum_{(v,w)\in L_{\neq}} \alpha_{\neq} ||d_{vw} d_{vw}^T||_2^2 \sum_{(v,w)\in L_=} \alpha_= ||d_{vw} d_{vw}^T||_2^2) q =: \underset{q}{max}\, q^T A q \quad (2)$$

The objective 2 is maximizing the Rayleigh quotient of A and q. Since A is symmetric, we can get an analytical solution q by the eigenvector with the max eigenvalue of A (Horn et al., 1990). Thus the matrix of k eigenvectors of A ordered by the corresponding eigenvalues yields the matrix Q.

3 Methodology

3.1 Adapting DensRay to Contextualized Language Models

We now describe how we adapt DensRay to contextualized language models. Given a set of gendered words V, we extract sentences containing a word in V from a corpus. We run a contextualized language model with M layers on each sentence $t_1, \ldots, t_j, \ldots t_{n-1}, t_n$ (where $t_j \in V$) and compute the contextualized representations $e^m, 1 \le m \le M$ of t_j, one for each layer. We compute an orthogonal rotation matrix Q_m for the mth BERT layer using Eq. 2. Finally, for debiasing, we set the dimensions of the gender subspace to 0 with the goal of eliminating or at least reducing gender information that may cause bias; for measuring bias, we use the distance to the zero point of the gender subspace as the measurement. In this paper, we take the first dimension of the rotated space as the gender subspace.

3.2 Evaluation

We use two evaluation datasets to measure gender bias: WEAT (Section 2.1.1) and OCCTMP.

OCCTMP is an evaluation dataset based on occupation templates that is tailored for the evaluation of contextualized language models. It has the added advantage that results are easier to interpret than those for WEAT.

To construct OCCTMP, we start with 320 occupation names[1] provided by Bolukbasi et al. (2016). Each occupation name is converted into a template of the form "[MASK] is an *occupation*." We measure gender bias in the templates as the average difference between the probability of BERT predicting [MASK] as "he" vs. "she"

$$\text{diff} = \frac{1}{|\mathcal{T}|}\sum_{T\in\mathcal{T}}(p(\text{he}|T) - p(\text{she}|T))$$

where $\mathcal{T}$ is the set of 320 templates. We find that for most experiments and most templates the probabilities of "he" is higher than "she", which qualitatively indicates that gender bias exists in these templates. We also find that in most cases the sum of the two probability is higher than 0.7; thus, this evaluation task is a good fit for BERT because it has learned that a pronoun is likely to occur in the masked position. Our templates can be easily extended to other languages as we later show for Chinese.

4 Quantifying Gender Bias with DensRay

DensRay can be used to quantify gender bias for any sentence and token, here use the distance to the zero point of the gender subspace as the measurement. Table 1 compared DensRay with the log probability score (Kurita et al., 2019) which can quantify gender bias on specific format '[TARGET] is a [ATTRIBUTE].' We use the bias score of [CLS] as aggregation of the sentence. For example, These examples show that DensRay is more versatile, it can measure the bias on each token.

11 <<< pd: why can't we do the same with hard-debiasing? >>>

DensRay						log score
[CLS]	[MASK]	is	a	professor	.	
-0.79	-0.69	-0.97	-0.9	-0.15	0.45	0.64
[CLS]	[MASK]	is	a	doctor	.	
-0.64	-0.04	-0.11	-0.12	0.31	0.29	-0.26
[CLS]	[MASK]	is	a	nurse	.	
-0.13	2.43	1.34	1.7	1.93	0.5	-5.44
[CLS]	The	professor	asked	.		
-0.79	-2.12	-0.64	0.03	0.52		-

Table 1: Examples for quantifying bias on bert-based. One can see that "doctor" and "nurse" have higher scores and thus potentially a female association. In the bottom sentence no pronoun is existing, but one can see that "professor" has a male association.

5 DensRay Debiasing Experiments on BERT Layers

5.1 Setup

In the experiments we process all the text data into lower case and use the BERT models "bert-base-uncased" and "bert-large-uncased". We implemented all experiments using the transformers library (Wolf et al., 2019).

To compute the rotation matrices by DensRay, we need a gendered word list as label, and some corpus. For the word list, we get 23 masculine words and 23 feminine words from the "family" category[2] of the Google analogy test set (Mikolov et al., 2013), and label them as 1 and -1. As the input corpus, we collect text data from Wikipedia that contains 5,000 (10,000) occurrences of words in the gendered list for BERT base (large) model. We carefully balance the occurrences such that the number of male and female samples are equal. We set $\alpha_{\neq} = \alpha_{=} = 0.5$, as we have balanced the training samples from the corpus.

As to compare with prior works, we compared with the post-processing method proposed by Mu and Viswanath (2018) to eliminate gender bias as Karve et al. (2019) introduced.

5.2 Results on OCCTMP

Results about our experiments on OCCTMP are summarized in Table 2. Two OCCTMP examples are given in Table 3. It shows that DensRay can

[1] https://github.com/tolga-b/debiaswe/blob/master/data/professions.json

[2] http://download.tensorflow.org/data/questions-words.txt

mitigate the gender bias in BERT: the average difference between predicting he/she drops to around two third (e.g., bert-base from 0.47 to 0.11).

model	prob(he)	prob(she)	diff	var
bert-base	0.66	0.19	0.47	0.16
bert-base-Mu	0.35	0.42	-0.07	0.03
bert-base-densray	0.48	0.37	0.11	0.02
bert-large	0.63	0.19	0.44	0.13
bert-large-Mu	0.40	0.23	0.17	0.02
bert-large-densray	0.47	0.31	0.16	0.02

Table 2: BERT debiasing results on OCCTMP. *bert-base* and *bert-large* are the original model without debiasing. *prob(he)* is the average probability that model predict *he* as the [MASK] in OCCTMP. *var* is the variance of the differences between the probability of BERT predicts [MASK] as *he* and *she*. **13**
<<< pd: why don't we compare with conceptor anymore? >>>

sentence	model	prob(he)	prob(she)
[MASK] is a adjunct professor.	bert-base	0.72	0.19
	bert-base-densray	0.44	0.47
	bert-large	0.72	0.22
	bert-large-densray	0.40	0.53
[MASK] is a administrator.	bert-base	0.63	0.23
	bert-base-densray	0.50	0.38
	bert-large	0.65	0.23
	bert-large-densray	0.45	0.37

Table 3: OCCTMP examples with prediction probabilities.

5.3 Results on WEAT

In WEAT we measure the effect size d-value and the oneside p-value of the permutation test. A d-value closer to zero indicates less gender bias. We also prefer a high p-value (at least 0.05) to not reject the null hypothesis, i.e., we do not reject that there is no gender bias. We follow the same WEAT word lists setup as Karve et al. (2019), the results on WEAT is shown in Table 4.

		WEAT		SEAT	
category	model	d	p	d	p
C6	bert-base	0.66	0.08	0.11	0.25
	bert-base-Mu	0.15	0.38	0.48	0.10
	bert-base-densray	0.62	0.12	-0.11	0.75
C7	bert-base	0.60	0.11	0.72	0.01
	bert-base-densray	-0.07	0.56	-0.09	0.72
	bert-base-densray	0.09	0.45	-0.11	0.26
C8	bert-base	0.78	0.08	1.00	0.01
	bert-base-Mu	-0.29	0.68	0.36	0.03
	bert-base-densray	0.03	0.47	0.75	0.01

Table 4: BERT debiasing results on WEAT. * shows significant gender bias.

5.4 Impact on Model Performance

It is crucial that debiasing methods do not harm downstream performance of BERT models.Thus we test the perplexity of language modeling on the Wikitext-2 (Merity et al., 2016), a subset of Wikipedia with 2 million words. We also test on GLUE tasks (Wang et al., 2018). For all the tests we follow the same setup as Wolf et al. (2019)[3]. Table 5 shows that DensRay debiasing gets comparable results with the original models on Wikitext-2 and GLUE tasks.

5.5 Discussions

5.5.1 Compare DensRay and Hard Debiasing

Here we compare the difference between DensRay and hard debiasing by (Mu and Viswanath, 2018). Figure 1 shows artifically created two dimensional embeddings. The black lines show the gender directions identified by hard debiasing and DensRay. Hard debiasing does not use the labels of the male-female word pairs. Instead it relies upon the assumption that the first principal component of the considered vectors is a meaningful gender direction. This can fail in some cases.

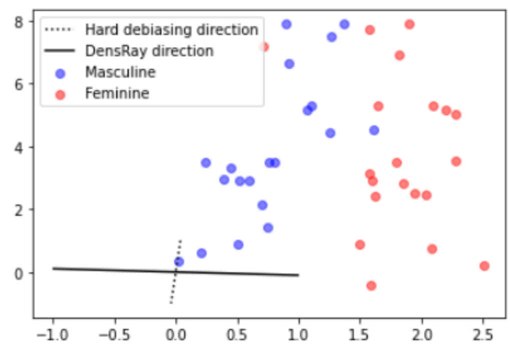

Figure 1: Gender direction on gendered words. **17 <<< pd: We should use the PCA with normalized data to make a fair comparison. Is this the case here? >>>**

5.5.2 Debiasing on Attention Heads

Here we applied DensRay on the attention heads in BERT to debias on OCCTMP, the heatmap Figure 2 shows that the debiasing effect of one single attention head is not obvious, with diff scores all around 0.4 - 0.5. Due to the lack of dimensions and the distribution of gender features in the attention heads, we chose to apply DensRay on layers as debiasing method. We conclude that there is

[3]https://huggingface.co/transformers/

model	Wikitext-2	CoLA	SST-2	MRPC	STS-B	RTE	WNLI
bert-base	3.77	49.15	92.09	85.86	82.66	62.82	52.11
bert-base-mu	3.95	45.53	91.74	82.48	82.60	63.54	56.34
bert-base-densray	3.81	48.04	91.74	84.89	82.43	63.90	53.52
bert-large	3.29	47.93	94.90	89.30	87.60	70.10	65.10
bert-large-Mu	3.85	47.45	93.95	85.01	82.33	67.12	63.02
bert-large-densray	3.35	48.91	94.02	88.84	85.63	67.78	64.48

Table 5: Language modelling perplexity and GLUE tasks performance. **15 <<< pd: it seems on bert-large densray is always better than Mu? Can't we make the argument that DensRay affects performance less? >>>**

no single attention head which is responsible for processing gender information.

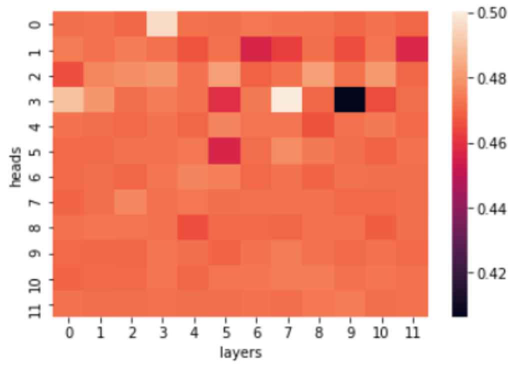

Figure 2: DensRay debiasing on each single attention head in BERT base, measured by diff on OCCTMP.

5.5.3 Number of Training Samples

In the experiments, we collected training samples for DensRay by considering occurrences of the same word in the corpus across different sentences. We collected equally many masculine and feminine words. Now we analyze the impact of these processes. DensRay is essentially a supervised learning method. In the case of insufficient labels, it is difficult for supervised learning to extract useful features. Treating different occurrences as different words greatly enriches training samples. As shown in Figure 3, the debiasing results improve with an increased number of training samples.

The same as other projection-based debiasing methods (Bolukbasi et al., 2016; Zhao et al., 2019; Dev and Phillips, 2019; Karve et al., 2019), the premise of DensRay debiasing is that the bias direction should be correct. If the sample is unbalanced, the bias direction computed by DensRay will be biased towards either the male or the female, resulting in deleting the gender subspace during debiasing will reverse the gender bias (e.g. there are more masculine words in unbalanced text data, thus the embeddings will be biased towards female after biased). The figure also shows that balanced training sample improved the debiasing performers.

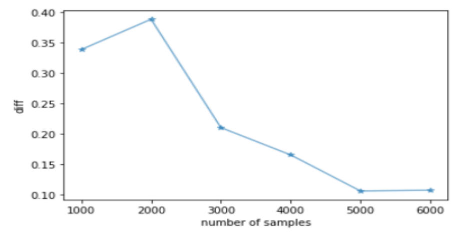

Figure 3: DensRay debiasing results on OCCTMP with different number of samples.

5.5.4 Balancing Gender Bias

18 <<< pd: I think this section can be moved to the supplementary material >>> In this experiment, we used the method of removing the first dimension (replacing its value by 0) of the gender interpreteble subspace to remove gender bias. Here we explore some other ways.

We explored three other ways to remove bias: 1) replace the first dimension of the gender interpreteble subspace with the mean value of the first dimension of the training samples. 2) standardize the first dimension. 3) replace the first dimension with a small random variable sampled from Gaussian distribution. All of them did not perform well. We further checked the mean and found that the mean of the different layers is not stable around 0, which is a problem worthy for further exploring. We also tried to delete more dimensions. However removing more dimensions does not improve the debiasing results significantly, while harming the model performance significantly.

5.5.5 Debiasing across different layers

So far we have applied DensRay on each layer of BERT simultaneously. See Figure 4 to illustrate the results of layers on our templates and the three WEAT categories. It shows that the debiasing effect on the 7-10 layer is more visible than on the other layers in BERT base model.

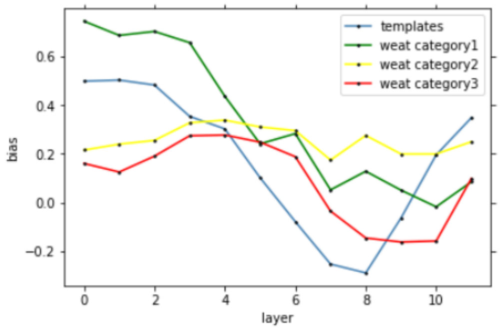

Figure 4: Debiasing on each single layer on BERT base. Bias is measured by diff on the templates and d-value on WEAT categories.

6 DensRay Debiasing multilingual-BERT

6.1 Setup

As an extension, we apply DensRay to mBERT for zero-shot debiasing on Chinese. Here we use the "bert-multilingual-uncased" model from (Wolf et al., 2019), we also use the same setup as the "bert-base-uncased" model in our previous experiments. **19 <<< pd: If I recall correctly the uncased model is not good for Chinese and only the cased model should be used. This is because they used different preprocessing for both models. >>>**

As before, we compute the rotation matrices using the English gendered words from the "family" category of the Google analogy test set (Mikolov et al., 2013).

Since Chinese is a language that does not contain genus, we can construct the OCCTMP templates by directly translating from the English templates. So we got the following form: For the occupation name, we referred to Tencent Translation[4] and made some manual adjustments to the translation. After removing the duplicates, we got 302 Chinese templates.

6.2 Results on OCCTMP

Results about our experiments on the templates are summarized in Table 6. Two examples are given in Table 6. It shows that DensRay trained with English can mitigate gender bias in mBERT: the average difference drops from 0.17 to 0.08 on Chinese templates. Also, the mBERT still gets comparable perplexities on Wikitext-2 after debiasing, see table Table 7.

[4]https://fanyi.qq.com/

model	prob(he)	prob(she)	diff	var
bert-multi-en	0.51	0.14	0.36	0.06
bert-multi-densray-en	0.33	0.12	0.21	0.03
bert-multi-cn	0.24	0.07	0.17	0.02
bert-multi-densray-cn	0.12	0.04	0.08	0.01

Table 6: Results of OCCTMP on mBERT after applied DensRay. Models with *-en* are tested on English templates, and those with *-cn* are tested on Chinese templates.

model	ppl
bert-multi	3.58
bert-multi-densray	3.72

Table 7: Language modeling performance on mBERT after applied DensRay. **21 <<< pd: on which language? Not sure whether this table is necessary in the main paper >>>**

7 Conclusion

We introduced DensRay debiasing on BERT. Our experiments show that this method can effectively mitigate gender bias in BERT on our constructed templates and WEAT. By checking the perplexity on Wikitext-2 and the performers on GLUE tasks, we also found this method causes little loss to the model performance. We also extend this method to mBERT as zero-shot debiasing for Chinese. As to further research, we plan to explore the irregularity of the central point of the gender dimension found in the experiments. In addition, this method can also be extended to other linguistic features, which will also be one of the future works.

References

Sandeep Attree. 2019. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

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.

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

sentence	model		
	bert-multi-en	0.68	0.16
	bert-multi-densray-en	0.51	0.18
	bert-multi-cn	0.52	0.11
	bert-multi-densray-cn	0.30	0.08
	bert-multi-en	0.53	0.17
	bert-multi-densray-en	0.35	0.13
	bert-multi-cn	0.68	0.16
	bert-multi-densray-cn	0.51	0.18

Table 8: Sanity check on the Chinese templates, where means *he* and means *she*. The two sentences are translated from Table 3.

Rakesh Chada. 2019. Gendered pronoun resolution using bert and an extractive question answering formulation. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Philipp Dufter and Hinrich Schütze. 2019. Analytical methods for interpretable ultradense word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1185–1191.

N Garg, L Schiebinger, D Jurafsky, and J Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3635–E3644.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

R.A. Horn, R.A. Horn, and C.R. Johnson. 1990. *Matrix Analysis*, chapter 4.2. Cambridge University Press.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on weat. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*.

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

Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective post-processing for word representations. In *6th International Conference on Learning Representations, ICLR 2018*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

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. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

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.

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