Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations

Anonymous COLING 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 (2016; **?**; **?**). Contextual word embedding models, such as BERT (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 (2019; 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 (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 §2.3 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) (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 (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.

2 Methodology

2.1 Debiasing Conceptor

In this paper we will compare our work with the debiasing conceptor (Karve et al., 2019). Given a set of gendered words $V := \{v_1, v_2, \ldots, v_n\}$ and their embeddings $E \in R^{n \times d}$, gender bias can be mitigated by multiplying the debiasing conceptor matrix$\neg C = I - C$, where C is the conceptor matrix that minimizes the objective

$$||E - CE||_F^2 + \alpha^{-2}||E||_F^2 \tag{1}$$

where α^{-2} is a scalar parameter. C has an analytical solution

$$C = \frac{1}{n}EE^T(\frac{1}{n}EE^T + \alpha^{-2}I)^{-1} \tag{2}$$

Intuitively, C is a soft projection matrix on the linear subspace where the word embeddings have the maximum bias.

2.2 Hard Debasing

We will also compare with the hard debiasing method proposed by (Mu and Viswanath, 2018), which is originally a postprocessing technique for improving word representations. (Karve et al., 2019) adopted it as a straightforward debiasing method. Hard debiasing relies upon the assumption that the first principal component of the considered vectors is a meaningful gender direction. The implementation is very simply, it uses PCA to get the first principal component from the embeddings of gendered words set, and then completely project it off.

2.3 DensRay

DensRay is an analytical method for identifying the embedding subspace of certain linguistic features. 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. 3 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 ($L_{\neq}$).

$$\max_q \sum_{(v,w)\in L_{\neq}} \alpha_{\neq}||q^T d_{vw}||_2^2 - \sum_{(v,w)\in L_=} \alpha_=||q^T d_{vw}||_2^2 \tag{3}$$

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. $\alpha_{\neq}, \alpha_= \in [0, 1]$ are hyperparameters. Observing that $||x||_2^2 = x^T x$, objective Eq. 3 can be simplified to:

$$\max_q q^T(\sum_{(v,w)\in L_{\neq}} \alpha_{\neq}||d_{vw}d_{vw}^T||_2^2 - \sum_{(v,w)\in L_=} \alpha_=||d_{vw}d_{vw}^T||_2^2)q =: \max_q q^T Aq \tag{4}$$

This objective 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 (1990). Thus the matrix of k eigenvectors of A ordered by the corresponding eigenvalues yields the matrix Q.

Here we compare the difference among DensRay, debiasing conceptor, and hard debiasing. Figure 1 shows artificially created two dimensional embeddings. The lines show the gender directions identified by hard debiasing, debiasing conceptor and DensRay.

DensRay separates the gendered words into two gender groups to produce a meaningful gender direction by objective Eq. 4, while debiasing conceptor and hard debiasing compute the subspace with the whole gendered words set. Theoretically they may fail to identified the correct gender direction in some cases.

1 <<< hs: since this is a crucial point of the paper, we need a strong summary sentence here >>> 2 <<< sl: I'm not sure about what kind of strong summary sentence we need here >>>

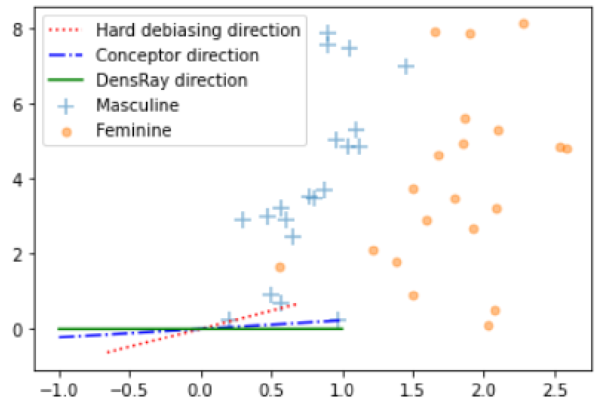

Figure 1: Gender direction on gendered words.

2.4 Removing Gender Information

All prior methods yield a gender dimension $q \in R^d$. In a contextualized language model like BERT each layer yields a contextualized embedding matrix $X \in R^{t \times d}$ where t is the length of the sentence. To debias representations we simply zero out the projected values on q for each position, that is we set $X_i^{\text{debiased}} = X_i - (X_i^\intercal q)q$ for each position i.

2.5 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 $t_1, \ldots, t_j, \ldots, t_n$ (where $t_j \in V$) and compute the contextualized representations $e^m, 1 \leq m \leq M$ of t_j, one for each layer. We compute an orthogonal rotation matrix Q_m for the mth BERT layer using Eq. 4. 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. In this paper, we take the first dimension of the rotated space as the gender subspace.

3 Experiments

3.1 Setup and Data

In the experiments we downcase all text 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 the labels of a gendered word list and a corpus. For the word list, we get 23 masculine words and 23 feminine words from the "family" category,[1] of the Google analogy test set (2013). As the input corpus, we collect text data from Wikipedia that contains 5,000 (resp. 10,000) occurrences of words in the gendered list for the BERT base (resp. 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.

We adapted DensRay to all BERT layers for debiasing. We compare with the hard debiasing method (Mu and Viswanath, 2018) and the debiasing conceptor (Karve et al., 2019) to eliminate gender bias as adapted to contextualized embeddings by (Karve et al., 2019). In the experiments we found that hard debiasing adapted to all BERT layers yields better results than only adapted to the contextual embeddings, while the debiasing conceptor destroyed the language modelling performance. So we adapted debiasing conceptor only to the contextual embeddings.

3.2 Evaluations

3.2.1 OCCTMP

We use two evaluation datasets to measure gender bias: WEAT (Section 3.2.2) and OCCTMP.

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

OCCTMP is a new evaluation dataset based on occupation templates that we created specifically 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[2] 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 probability of "he" is higher than "she", which qualitatively indicates that gender bias can be identified using 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.

3.2.2 Association Tests

Association tests are originated from sociological research. (Greenwald et al., 1998) proposed the Implicit Association Test (IAT) to quantified societal bias. In 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 IAT, (Caliskan et al., 2017) proposed the Word Embedding Association Test (WEAT), which used word similarities between targets and attributes instead of the response times to get rid of the requirement of human subjects. 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. The null hypothesis in the statistical test of WEAT is: there is no difference in the similarity between X_1, X_2 and A_1, A_2. In the prior literature it has been argued that if the null hypothesis can't be rejected, there is no significant gender bias. **3 <<< sl: as far as I saw, all the association tests(IAT,WAET,SEAT,CEAT) use this hypothesis setup, I think it's just a 'tradition'. >>>** The WEAT test statistic is defined as

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

where

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

in which $cos(\vec{x}, \vec{a})$ 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

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

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

$$d = \frac{s(X_1, X_2, A_1, A_2)}{std_{x \in X_1 \cup X_2} s(x, A_1, A_2)}.$$

To extend WEAT to contextual embeddings, (Karve et al., 2019) extracted contextual embeddings from the template '[MASK] is *word*.' (May et al., 2019) proposed Sentence Embedding Association Test (SEAT), which designed more complex templates to extract word embeddings.

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

Dispensed with templates, (Guo and Caliskan, 2020) proposed Contextualized Embedding Association Test (CEAT), which extracted the embeddings of the stimulus' occurrences from the corpus, and computed the weighted mean of effect sizes and statistical significance by a random-effects model. The combination effect size is

$$d_c(X_1, X_2, A_1, A_2) = \frac{\sum_{i=1}^{N} v_i d_i}{\sum_{i=1}^{N} v_i}$$

where v_i is the weights in the random-effects model. Two-tailed p-value is used to measure the statistical significance:

$$p_c = 2[1 - \phi(|\frac{d_c}{std(d_c)}|)]$$

In these association tests we measure the effect size d-value and the one-sided 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 use the three categories C6: career/family, C7: math/arts, C8: science/arts, from (Karve et al., 2019)'s WEAT word list setup.

3.2.3 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 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]

4 Results

4.1 Debiasing Results

Table 1 gives results for OCCTMP. Two OCCTMP examples are given in Table 4. It shows that DensRay can mitigate the gender bias in BERT: the average difference between predicting he/she drops to around two third (e.g., for bert-base from 0.47 to 0.11). It's shown that the debiasing performance of the three methods are comparable. While the probabilities of debiasing conceptor predictions are significantly lower then the other two methods, which indicates that debiasing conceptor will affect the performance of the model.

model	prob(he)	prob(she)	diff	var
bert-base	0.66	0.19	0.47	0.16
bert-base-hard	0.35	0.42	-0.07	0.03
bert-base-conceptor	0.18	0.11	0.08	0.01
bert-base-densray	0.48	0.37	0.11	0.02
bert-large	0.63	0.19	0.44	0.13
bert-large-hard	0.40	0.23	0.17	0.02
bert-large-conceptor	0.05	0.03	0.02	0.00
bert-large-densray	0.47	0.31	0.16	0.02

Table 1: BERT debiasing results on OCCTMP. *bert-base* and *bert-large* are the original models without debiasing. *prob(he)* is the average probability predicted for *he* as the [MASK] in OCCTMP. *var* is the variance of the differences between the probabilities of predicted for *he* and *she*.

Table 2 shows the results on association tests. The debiasing performance of the three methods are comparable.

4.2 Model Performance

Table 3 shows that DensRay debiasing gets comparable results with the original models on Wikitext-2 and GLUE tasks. In most tasks on bert-base and all tasks on bert-large, DensRay performs better than hard debiasing, so DensRay affects model performance less.

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

		WEAT		SEAT		CEAT	
category	model	d	p	d	p	d	p
C6	bert-base	0.66	0.08	1.04	$< 10^{-2*}$	0.94	$< 10^{-2*}$
	bert-base-hard	0.15	0.38	-0.08	0.67	0.88	$< 10^{-2*}$
	bert-base-conceptor	0.07	0.46	0.77	$< 10^{-2*}$	0.94	$< 10^{-2*}$
	bert-base-densray	0.62	0.12	0.36	0.02^*	0.75	$< 10^{-2*}$
C7	bert-base	0.60	0.11	0.17	0.15	0.99	$< 10^{-2*}$
	bert-base-hard	-0.07	0.56	-0.06	0.64	0.99	$< 10^{-2*}$
	bert-base-conceptor	0.54	0.14	-0.25	0.93	0.99	$< 10^{-2*}$
	bert-base-densray	0.09	0.45	-0.47	0.99	0.99	$< 10^{-2*}$
C8	bert-base	0.78	0.08	0.81	$< 10^{-2*}$	0.99	$< 10^{-2*}$
	bert-base-hard	-0.29	0.68	-0.10	0.71	0.99	$< 10^{-2*}$
	bert-base-conceptor	0.62	0.14	0.50	$< 10^{-2*}$	0.99	$< 10^{-2*}$
	bert-base-densray	0.03	0.47	0.41	0.01^*	0.99	$< 10^{-2*}$

Table 2: BERT debiasing results on WEAT. * shows significant gender bias.

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-hard	3.95	45.53	91.74	82.48	82.60	63.54	56.34
bert-base-conceptor	4.46	48.31	91.43	84.08	81.37	59.57	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-hard	3.85	47.45	93.95	85.01	82.33	67.12	63.02
bert-large-conceptor	4.13	49.44	93.87	87.67	83.44	62.45	56.34
bert-large-densray	3.35	48.91	94.02	88.84	85.63	67.78	64.48

Table 3: Language modelling perplexity and GLUE tasks performance.

4.3 Examples

In Table 4 we show two OCCTMP examples. In the first sentence, hard debiasing reverses the male bias and creates female bias. The sum probabilities of debiasing conceptor predicted ”he” and ”she” are lower than 0.6 the other two methods, indicates that the probabilities of words unrelated to these sentences becomes higher and make the model unstable.

sentence	model	prob(he)	prob(she)	diff
[MASK] is a professor.	bert-base	0.84	0.13	0.71
	bert-base-hard	0.37	0.55	-0.18
	bert-base-conceptor	0.28	0.23	0.05
	bert-base-densray	0.53	0.37	0.16
[MASK] is a dancer.	bert-base	0.22	0.72	-0.50
	bert-base-hard	0.27	0.64	-0.37
	bert-base-conceptor	0.20	0.33	-0.13
	bert-base-densray	0.42	0.52	-0.10

Table 4: OCCTMP examples with prediction probabilities.

4.4 Discussions

4.4.1 Debiasing on Attention Heads

We now apply DensRay to 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 in [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 no single attention head which is responsible for processing gender information.

4.4.2 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

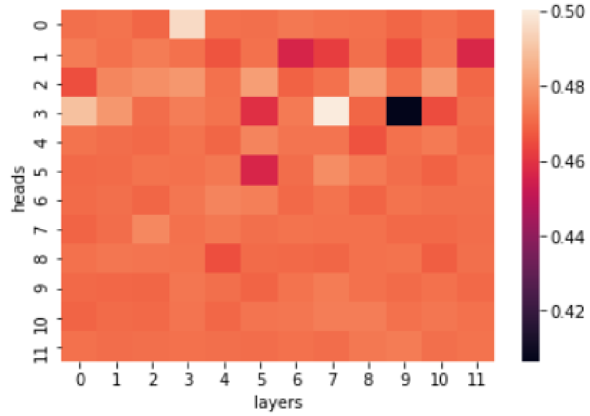

Figure 2: DensRay debiasing on each single attention head in BERT base, measured by diff on OCCTMP.

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.

Similar to other projection-based debiasing methods (Bolukbasi et al., 2016; Zhao et al., 2019; Dev and Phillips, 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 and reversing the gender bias. For example, if there are more masculine words in unbalanced text data, then the embeddings will be biased towards female after debiasing. The figure also shows that a balanced training sample improves the debiasing performance.

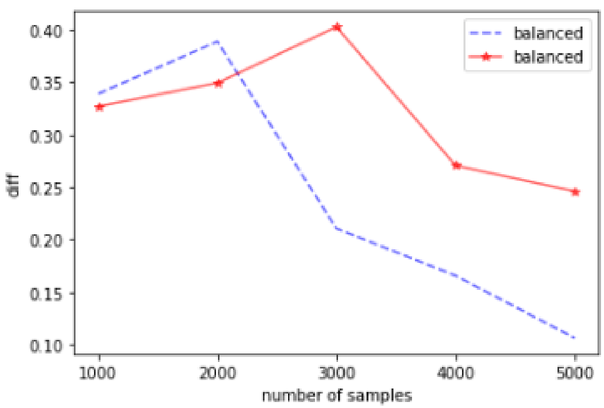

Figure 3: DensRay debiasing results on OCCTMP with different number of samples and unbalanced/balanced data.

4.4.3 Balancing Gender Bias

4 <<< 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.

4.4.4 Debiasing across different layers

So far we have applied DensRay to all BERT layers simultaneously. Figure 4 illustrates the effect of debiasing a single layer on our templates and the three WEAT categories. We see that the debiasing effect is stronger in layers 7–10 than in the other layers in BERT base.

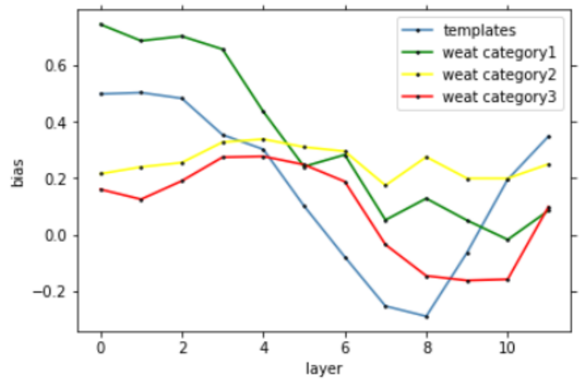

Figure 4: Debiasing on each single layer on BERT base. Bias is measured by diff on the templates and d-value on WEAT categories.

4.4.5 Multilingual Debiasing

We now show that, in a multilingual contextualized language model like mBERT, we can use DensRay for zero-shot debiasing. Specifically, we train a DensRay model on English and use it to debias Chinese. We use bert-multilingual-uncased from (2019). We use the same setup as for bert-base-uncased in our previous experiments. **5 <<< 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. >>>** **6 <<< sl: modify this later >>>**

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

Since Chinese is a language that does not mark gender, we can construct the OCCTMP templates by directly translating from the English templates. We use the following form: "[MASK]是一个*occupation*。" We translate the occupation name based on Tencent Translation[4] and make some manual adjustments to the translation. After removing duplicates, 302 Chinese templates remain.

Table 5 gives results for the Chinese templates. Two examples are given in Table 5. We see that DensRay trained with English can mitigate gender bias in mBERT: the average difference drops from 0.17 to 0.08 on Chinese templates.

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 5: 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.

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

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 6: Sanity check on the Chinese templates, where means *he* and means *she*. The two sentences are translated from Table 4.

5 Related Work

5.1 Quantifying Gender Bias

A typical way to measure gender bias is to evaluate on **downstream tasks**. For coreference resolution, (Zhao et al., 2018) designed Winobias and (Rudinger et al., 2018) designed Winogender schemas. (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 (2019; **?**) making it hard to compare debiasing systems. For sentiment analysis, Equity Evaluation Corpus (EEC) (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.

7 <<< 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 >>> 8 <<< sl: so should we remove this section? just talk about Debiasing Methods in related work >>>

5.2 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. The (Bolukbasi et al., 2016) hard debiasing used 2 groups of gendered words for definition and another 2 groups for alignment to produce a gender direction. It has been found to work better then soft debiasing. (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) proposed the debiasing conceptor, which shrink each principal component of the covariance matrix of the embeddings to achieve a soft debiasing. They also introduced a simple hard debiasing method proposed by (Mu and Viswanath, 2018), which identified the gender subspace by PCA and projected the first principal component off. both the debiasing conceptor and hard debiasing (Mu and Viswanath, 2018) produce gender direction by one word list with male and female, while hard debiasing (Bolukbasi et al., 2016) produces gender direction by male-female pairs. The method we use , DensRay, is similar to hard debiasing (Bolukbasi et al., 2016) in this aspect, but DensRay uses only word lists and can be solved efficiently in closed form, so it's more stable.

6 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.

9 <<< hs: three parts of the paper should be in sync:
(i) the abstract
(ii) the contributions at the end of the introduciton
(iii) the conclusino
make sure to check that during final editing >>>

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. 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, Apr.

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.

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.

Wei Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases.

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, Dec.

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, Vicente Ordonez, and Kai-Wei Chang. 2018. 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, 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.