Instructions for EMNLP 2020 Proceedings

Anonymous EMNLP submission

Abstract

As one of the representatives of context word embedding, BERT has achieved the most advanced performance on many NLP tasks. Due to the strong feature extraction ability and the high demand for the amounts of training data, BERT can hardly avoid learning many of the human-generated stereotypes in the text data, including gender bias. In this study, we (1) proposed a template based on language models to quantify gender bias in BERT; (2) proposed a DensRay-based word vector space projection analysis method, which is used in Eliminate gender bias. (3) The method of English training is extended to multilingual BERT, which reduces the gender bias of Chinese on our Chinese template.

1 Introduction

Word embeddings, which represent the semantic meaning of text data as vectors, are used as input in natural language processing tasks (Goldberg, 2017). It has disclosed that word embeddings exhibit unexpected social biases, such as gender bias, present in their training corpora (Bolukbasi et al., 2016; Caliskan et al., 2017; Garg et al., 2018). An example is that man is associated with computer programmer on the embedding space, and woman is associated with homemaker (Bolukbasi et al., 2016). Contextual word embedding models, such as BERT (Devlin et al., 2018), have become increasingly common and achieved new state-of-the-art results in the many NLP tasks. Researches have also found gender bias in contextualized embeddings (Zhao et al., 2019; May et al., 2019).

In this work, we aim to mitigate gender bias on BERT embedding in a straight-forward and interpretable way. We introduce a debiasing method on BERT using the DensRay (Dufter and Schütze, 2019), which is a computational method to get the interpretable dimensions by rotating the word embedding spaces. We show that gender information is captured in every BERT layer. We applied DensRay to every BERT layer and evaluated on a set of simple templates we constructed and the Word Embedding Association Test (WEAT) (Caliskan et al., 2017), our experiments find that the DensRay debiasing method effectively mitigates gender bias, and do little harm to the performance of the BERT model on language modeling and GLUE tasks (Wang et al., 2018). As an extension, we also applied this debiasing method to the multilingual-BERT (mBERT) model: we use English gender label for computing the rotation matrix, and debias on our Chinese templates. Our contributions are summarized as the following:

- We introduce the DensRay debiasing method on BERT, and demonstrates the debiasing effectiveness by our templates and the Word Embedding Association Test.
- We show that the DensRay debiasing method can be applied to mBERT for zero-shot debiasing for other languages.

2 Related Work

2.1 The measurements of 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. Different from the WinoBias, Winogender schemas include gender-neutral pronouns which WinoBias doesn't, and one Winogender schema has one occupational mention and one "other participant" mention while WinoBias has two occupational. Webster et al. (2018) released GAP, a balanced corpus of Gendered Ambiguous Pronouns. Gender bias can be measured as the ratio of F1 score on masculine to F1 score on feminine, however the bias ratios are too close to 1 (Chada, 2019; Attree, 2019), so that the bias can't be presented obviously. For sentiment analysis, Equity Evaluation Corpus (EEC) dataset (Kiritchenko and Mohammad, 2018) was designed to measure gender bias by the difference in emotional intensity predictions between gender-swapped sentences.

Another kind of method to measure gender bias is based on the association test, it's originated from sociological research. Greenwald et al. (1998) proposed the Implicit Association Test (IAT) to quantified 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 used word similarities between targets and attributes instead of the response times to get rid of the reequipment of human subjects. Later, 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.

2.1.1 Word Embedding Association Test

Here we introduce WEAT in detail. Consider two sets of target words X_1 and X_2 where $|X_1| = |X_2|$ (equal size) and two sets of attribute words A_1 and A_2 where $|\mathcal{A}| = |\mathcal{B}|$. As a statistical test, the null hypothesis of WEAT is: There is no difference in the cosine similarity between the 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 can be used as the two attribute sets, such that the null hypothesis means science and art are equally similar to each masculine and feminine names, so there is no gender bias.

The test statistic is defined as,

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

where

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

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 $X \cup Y$. The one-sided p-value of the permutation test is defined as

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

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

$$\frac{s(X_1, X_2, A_1, A_2)}{std_{x \in X_1 \cap X_2} s(x, A_1, A_2)}$$

2.2 Debiasing methods

Researchers proposed various methods to remove gender bias, in which 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 post-processing projecting. Bolukbasi et al. (2016) proposed a hard debiasing method where they used the gendered words to compute the difference embedding vector as the gender direction, and a soft debiasing method which combined the inner-products objective of word embedding and an objective to project the word embedding into a subspace that orthogonal to the gendered words while it performed not so good as the hard debiasing method. 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 hard debiasing method of Bolukbasi et al. (2016) to mitigate gender bias on ELMo (Peters et al., 2018). Karve et al. (2019) introduced the debiasing conceptor, in which they shrined each principal component of the covariance matrix of the word embeddings to achieved a soft debiasing.

2.3 DensRay

DensRay is an analytical method proposed to identify the embedding subspace of linguistic features. Here we introduce DensRay for gender debiasing. Same as the methods mentioned in the previous section, we aim to identify the gender bias subspace using a set of gendered words with vocabulary $V := \{v_1, v_2, \ldots, v_n\}$ and embedding matrix $E \in R^{n \times d}$, thus for word v_i we have the corresponding embedding vector e_{v_i}. We denotes the gendered words into a map $l : V \rightarrow \{-1, 1\}$ (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”.

Consider $L_= := \{(v, w) \in V \times V | l(v) = l(w)\}$ and $L_{\neq} := \{(v, w) \in V \times V | l(v) \neq l(w)\}$, we defined $d_{vw} := e_v - e_w$ as the difference vector of v and w. DensRay solves the following optimization problem,

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

where $q \in R^d$ and $q^T q = 1$ since Q is orthogonal, $\alpha_{\neq}, \alpha_= \in [0, 1]$ are hyperparameters. Intuitively the objective tries to maximize the distance of the word pairs from the same gender group (male words or female words) and minimize the distance of the word pairs from different gender group. Regard that $||x||_2^2 = x^T x$, objective 1 can be simplified to:

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

The objective 2 is maximizing the Rayleigh quotient of A and q. Since A is symmetric, instead of training the model by gradient decent, 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 DensRay debiasing on BERT

Here we introduce our methodology to apply DensRay on BERT. With a given gendered word list $V := \{v_1, v_2, \ldots, v_n\}$, we put the sentences which contains the gendered words into BERT model without masking. Due to the different contexts, for each word we regard every occurrence in a corpus v_{i_j} of the word v_i has an independence embedding e_{i_j}. For each sentence contains word v_{i_j}, the output of each BERT layer yields a contextual embedding $e_{i_j}^m, m \in \{1, 2, \ldots, M\}$, where M is the number of layers in the BERT model. Then an orthogonal rotation matrix Q_m can be computed by formula 2.

We believe that by identifying the interpretable gender subspace and replacing its parameters by 0, the gender information that may cause bias can be mitigated. When a new sentence or corpus is sent into the inference process of BERT, we can rotate its original embedding on the output of the first BERT layer E_1 by the pre-computed Q_m to an interpretable embedding $\hat{E}_1 := E_1 Q_1$. Then we replace the first dimension of $\hat{E}_1$ by 0 and rotate it back to get a new $E_1 := \hat{E}_1 Q_1^T$, since $Q^{-1} = Q^T$ for orthogonal matrix Q. This operation can be applied after every layer in the inference process to yield a debiased BERT output.

3.2 Evaluations

In this paper we will use WEAT to measure gender bias. Besides, to quickly test the results of debiasing, we also constructed a set of templates with 320 occupation names[1] provided by Bolukbasi et al. (2016) for masked language modeling ”[MASK] is a *occupation*.” The gender bias in the templates is measured by the average difference between the probability of BERT predicts [MASK] as ”he” and ”she”,

$$diff = \underset{i\in templates}{mean} (prob(he) - prob(she))$$

Through case study, we find that for most templates the probability of ”he” is higher than ”she”, which qualitatively indicates that gender bias exists in these templates. We also find that for most sentences the sum of the two probability is higher than 0.7, which means that the predictions will be stable. This templates set can be easily extended to other languages that don’t have genus.

4 DensRay Debiasing Experiments on BERT Layers

4.1 Experiments Setup

In the experiments we use the BERT models ”bert-base-uncased” and ”bert-large-uncased” from the huggingface trasnformers library (Wolf et al., 2019).

To compute the rotation matrices by DensRay, there needs a the gendered word list as label, and an input corpus to get the embeddings from BERT. For the word list, we get 23 masculine words and

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

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, for BERT base (large) model we collect text data from Wikipedia that contains 5,000 (10,000) occurrences of words in the gendered list, in which the number of masculine and feminine samples are equal. The hyperparameters are set to $\alpha_{\neq} = \alpha_{=} = 0.5$, since we have balanced the training samples from the corpus.

4.2 Results on Templates

Results about our experiments on the templates are summarized in table 1. Two example templates are given in table 7. The evaluation on our templates shows that DensRay can mitigate the gender bias on BERT.

4.3 Results on WEAT

In WEAT we measure the effect size d-value and the oneside p-value of the permutation test. A higher absolute value of the d-value indicates larger gender bias between the target words with respect to the attribute words. So, for the d-value, the closer to zero, the less gender bias. Refer to the definition of the null hypothesis, if the p-value is less than 0.05 we will reject the null hypothesis so that there will be a significant gender bias. So, we would prefer a high p-value (at least 0.05) to indicate the lack of gender bias. Follow the same WEAT word lists setup as Karve et al. (2019), the results on WEAT is shown on table 3. For all the three categories, DensRay decreased absolute value of d-value and increased the p-value, although on bert-large still showd strong bias in *(Career, Family) vs (Male, Female)*.

4.4 Impact on BERT Model

Here we want to evaluate the performance of BERT model after applied DensRay. We test the perplexity of language modeling on Wikitext-2 dataset (Merity et al., 2016) which is a subset of Wikipedia with 2 million words, the results in table 4 show that DensRay caused a small increase in perplexity on Wikitext-2 for both BERT base and large model.

Follow the same setup as Wolf et al. (2019)[3], we also evaluate on the GLUE tasks (Wang et al., 2018), results are summarized in table 5.

[2] http://download.tensorflow.org/data/questions-words.txt
[3] https://huggingface.co/transformers/

4.5 Discussions

4.5.1 the impact of training samples

Through evaluation and inspection of the impact on the performance of downstream tasks, experiments show that DensRay is an effective debiasing method on BERT. Although DensRay is an analytical solution, the effect still depends on the label data. In the experiments, we regarded the occurrences of the same word in the corpus as independent words with the same gender label, and used balanced samples for masculine and feminine words. Now we analyze the impact of these processes.

Since there are only 46 words in the rendered word list, if we average their embedding under different contexts, there will be only 46 training samples left for DensRay to calculate. 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, the debiasing results improved when we increase the 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 calculated 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.

4.5.2 the ways of debiasing

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.

In figure 2, we tried two other ways to remove bias. The first is to replace the first dimension of the gender interpreteble subspace with the mean value of the first dimension of the training samples. The second way is to standardize the first dimension. The results showed that both of these methods did not perform well. We further checked the mean and found that the mean of the different layers is

model	prob(he)	prob(she)	diff	var
bert-base	0.6594	0.1874	0.4720	0.1600
bert-base-densray	0.5106	0.3447	0.1658	0.0119
bert-large	0.6287	0.1907	0.4380	0.1262
bert-large-densray	0.4751	0.2923	0.1827	0.0150

Table 1: BERT debiasing results on templates. *bert-base* and *bert-large* are the original model without debiasing. *prob(he)* is the mean probability that model predict *he* as the [MASK]in all templates. *var* is the variance of the differences between the probability of BERT predicts [MASK] as *he* and *she*.

sentence	model	prob(he)	prob(she)
[MASK] is a adjunct professor.	bert-base	0.7231	0.1942
	bert-base-densray	0.4423	0.4740
	bert-large	0.7181	0.2212
	bert-large-densray	0.3974	0.5316
[MASK] is a administrator.	bert-base	0.6296	0.2337
	bert-base-densray	0.5045	0.3762
	bert-large	0.6456	0.2269
	bert-large-densray	0.4536	0.3716

Table 2: Sanity check on the templates.

not stable around 0, which is a problem worthy for further exploring.

As shown in figure **??**, we try to delete more dimensions. The results show that removing more dimensions does not improve the debiasing results significantly.

4.5.3 debiasing on each BERT layer

Here we only apply DesnRay on one BERT layer at a time. We constructed a table to illustrate the top three layers with the best performance on our templates and the three WEAT categories.

5 DensRay Debiasing multilingual-BERT

5.1 Setup

As an extension, we apply DensRay to mBERT for zero-shot debiasing on Chinese. Here we use the ”bert-multilingual-uncesed” model from (Wolf et al., 2019), we also use the same setup as the ”bert-base-uncased” model in our previous experiments.

Still, we compute the rotation matrices by the English gendered words from the ”family” category of the Google analogy test set (Mikolov et al., 2013).

Since Chinese is a language that doesn’t contain genus, we can construct the templates by simply translating from the English templates. After removing the duplicates, we get 302 Chinese templates.

5.2 Results on Templates

Results about our experiments on the templates are summarized in table 7. Two example templates are given in table 7. The evaluation on our templates shows that DensRay can mitigate the gender bias on BERT.

We also checked the perplexity for mBERT on Wikitext-2, see table 8. Results show that DensRay can be extended to mBERT as a zero-shot debiasing method for some other languages.

6 Conclusion

We believe that the irregularity of the central point of the gender dimension found in the experiment is worthy of further study. 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.

category	model	d	p
(Career, Family) vs (Male, Female)	bert-base	0.6581	0.08
	bert-base-densray	0.6397	0.11
	bert-large	1.5705	0.00*
	bert-large-densray	0.9980	0.02*
(Math, Arts) vs (Male, Female)	bert-base	0.6017	0.11
	bert-base-densray	0.0739	0.45
	bert-large	0.2239	0.35
	bert-large-densray	-0.0145	0.48
(Science, Arts) vs (Male, Female)	bert-base	0.7762	0.08
	bert-base-densray	0.0167	0.49
	bert-large	0.816	0.04*
	bert-large-densray	0.6743	0.10

Table 3: BERT debiasing results on WEAT. Number with * shows significant gender bias.

Figure 1: Here should be a graph.

model	ppl
bert-base	3.7714
bert-base-densray	3.8051
bert-large	3.2928
bert-large-densray	3.3503

Table 4: Language modeling performance on BERT after applied DensRay.

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

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 Inter-*

model	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI
bert-base	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714
bert-base-densray	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714
bert-large	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714
bert-large-densray	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714	3.7714

Table 5: GLUE tasks performance on BERT after applied DensRay.

Figure 2: Here should be a graph.

national Conference on Learning Representations (ICLR).

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. 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.

Figure 3: Here should be a graph.

Table 6: Here needs a table.

model	prob(he)	prob(she)	diff	var
bert-multi-en	0.5064	0.1427	0.3637	0.0619
bert-multi-densray-en	0.3344	0.1207	0.2137	0.0277
bert-multi-cn	0.2370	0.0718	0.1652	0.0223
bert-multi-densray-cn	0.1247	0.0407	0.0840	0.0055

Table 7: Results of templates on mBERT after applied DensRay. Models with *-en* are tested on our English templates, and those with *-cn* are tested on our Chinese templates.

model	ppl
bert-multi	3.5788
bert-multi-densray	3.7216

Table 8: Language modeling performance on mBERT after applied DensRay.

sentence	model	prob(he)	prob(she)
[MASK] is a adjunct professor.	bert-multi-en	0.6823	0.1611
	bert-multi-densray-en	0.5073	0.1762
	bert-multi-cn	0.5189	0.1108
	bert-multi-densray-cn	0.3046	0.0831
[MASK] is a administrator.	bert-multi-en	0.5344	0.1670
	bert-multi-densray-en	0.3496	0.1318
	bert-multi-cn	0.6823	0.1611
	bert-multi-densray-cn	0.5073	0.1762

Table 9: Sanity check on the Chinese templates. Here we only present their translation for convenience.