Unsupervised Mitigation of Gender Bias by Character Components: A Case Study of Chinese Word Embedding

Xiuying Chen[1,2,*], **Mingzhe Li**[3,*], **Rui Yan**[4] , **Xin Gao**[1,2], **Xiangliang Zhang**[5,2,†]
[1]Computational Bioscience Reseach Center, KAUST
[2]Computer, Electrical and Mathematical Sciences and Engineering, KAUST
[3] Ant Group
[4] Gaoling School of Artificial Intelligence, Renmin University of China
[5] University of Notre Dame
xiuying.chen@kaust.edu.sa, li_mingzhe@pku.edu.cn

Abstract

Word embeddings learned from massive text collections have demonstrated significant levels of discriminative biases. However, debiasing on the Chinese language, one of the most spoken languages, has less explored. Meanwhile, existing literature relies on manually created supplementary data, which is time- and energy-consuming. In this work, we propose the first Chinese Gender-neutral word Embedding model (CGE) based on Word2vec, which learns gender-neutral word embeddings without any labeled data. Concretely, CGE utilizes and emphasizes the rich feminine and masculine information contained in radicals, *i.e.,* a kind of component in Chinese characters, during the training procedure. This consequently alleviates discriminative gender biases. Experimental results show that our unsupervised method outperforms the state-of-the-art supervised debiased word embedding models without sacrificing the functionality of the embedding model.

1 Introduction

Investigations into the representation learning revealed that word embeddings are often prone to exhibit discriminative gender stereotype biases (Caliskan et al., 2017). Consequently, these biased word embeddings have effects on downstream applications (Dinan et al., 2020; Blodgett et al., 2020). Mitigating gender stereotypes in word embedding are becoming a research hotspot due to its penitential application, and a number of the existing debias works are dedicated to the English language (Zhao et al., 2018a; Kaneko and Bollegala, 2019). However, debiasing on Chinese, one of the most spoken languages, has drawn less attention these days.

In the Chinese language, "radical" is a graphical component of Chinese characters, which serves as an indexing component in the Chinese dictionary. Radical can suggest part of the meaning of the character due to the phono-semantic attribute of the Chinese language. For example, "氵(water)" is the radical of "河 (river), 湖 (lake)". Consequently, a series of works have shown that radicals can enhance the word embedding quality (Chen et al., 2015; Yin et al., 2016; Chen and Hu, 2018). As part of the radical system, the gender-related radicals, *i.e.,* "女(female)" and "亻(man)", contains gender information of the corresponding character. Specifically, the radical "女(female)" can denote female and "亻(man)" can denote people, which includes male gender information. For example, characters "姐(sister),妇(wife),妈(mother),姥(grandma)" all have the radical of "女(female)", demonstrating that these are feminine words. Hence, we assume that radical is a natural information source to capture feminine and masculine information, and such information can help the model learn gender definition. Once the model learns what is the definition of gender, it can identify the gender bias that is not actually relevant to gender.

To this end, we propose our Chinese Gender-neutral word Embedding model (CGE) that is based on the classic Word2vec model, where the basic idea is to predict the target word given its context words. CGE has two variations, *i.e.,* Radical-added CGE and Radical-enhanced CGE. Radical-added CGE emphasizes the gender definition information by directly adding the radical embedding to the word embedding. We next propose a Radical-enhanced CGE, where radical embeddings are employed to predict the target word instead of adding to the word embedding. This is a more flexible approach, where the gradients of the embeddings of words and radicals can be different in the training process. Note that the radical can be extracted from the character itself, hence, our model can also learn gender-neutral word embedding in an unsupervised fashion. Experimental results show that

[*] Equal Contribution
[†] Corresponding authors

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 121 - 128
July 15, 2022 ©2022 Association for Computational Linguistics

our methods outperform the supervised models.

2 Related Work

Chinese Word Embedding. Different from the English language where words are usually taken as basic semantic units, Chinese words have complicated composition structures revealing their semantic meanings (Li et al., 2020, 2021). More specifically, a Chinese word is often composed of several characters, and most of the characters themselves can be further divided into components such as radicals. Chen et al. (2015) first presented a character-enhanced word embedding model (CWE). Following this work, Yin et al. (2016) proposed multi-granularity embedding (MGE), which enriches word embeddings by incorporating finer-grained semantics from characters and radicals. Another work (Yu et al., 2017) proposed to jointly embed Chinese words as well as their characters and fine-grained sub-character components. Chen and Hu (2018) used radical escaping mechanisms to extract the intrinsic information in the Chinese corpus. All the above works do not deal with the gender bias phenomena in Chinese word embeddings.

Gender Biased Tasks. Gender biases have been identified in downstream NLP tasks (Hendricks et al., 2018; Holstein et al., 2019). Zhao et al. (2018a) demonstrated that coreference resolution systems carry the risk of relying on societal stereotypes present in training data and introduced a new benchmark, WinoBias, for coreference resolution focused on gender bias. Gender bias also exists in machine translation (Prates et al., 2018), e.g., translating nurses as females and programmers as males, regardless of context. Stanovsky et al. (2019) presented the first challenge set and evaluation protocol for the analysis of gender bias in machine translation. Notable examples also include visual SRL (cooking is stereotypically done by women, construction workers are stereotypically men, (Zhao et al., 2017)), lexical semantics ("man is to computer programmer as woman is to homemaker", (Bolukbasi et al., 2016)) and so on.

Gender-neutral Word Embedding. Previous works demonstrated that word embeddings can encode sexist stereotypes (Caliskan et al., 2017). To reduce the gender stereotypes embedded inside word representations, Bolukbasi et al. (2016) projected gender-neutral words to a subspace, which is orthogonal to the gender dimension defined by a list of gender-definitional words. Concretely, they proposed a hard-debiasing method where the gender direction is computed as the vector difference between the embeddings of the corresponding gender-definitional words, and a soft-debiasing method, which balances the objective of preserving the inner products between the original word embeddings. Zhao et al. (2018a) aimed to preserve gender information in certain dimensions of word vectors while compelling other dimensions to be free of gender influence. Kaneko and Bollegala (2019) debiased pre-trained word embeddings considering four types of information: feminine, masculine, gender-neutral, and stereotypical. Following this work, Kaneko and Bollegala (2021) applied the debiasing technique to pre-trained contextualized embedding model.

Compared with previous works, our work is focused on the Chinese language, and utilizes radicals, a special component of Chinese character.

3 Methodology

We will take CBOW for example and demonstrate our frameworks based on CBOW.

3.1 CBOW

As shown in Figure 1(a), CBOW predicts the target word, given context words in a sliding window. Concretely, given the word sequence $D = (x_1, x_2, ..., x_T)$, the ultimate goal is to maximize the average log probablity:

$$\frac{1}{T}\sum_{t=c}^{T-c} \log P(x_t|x_{t-c}, ..., x_{t+c}), \tag{1}$$

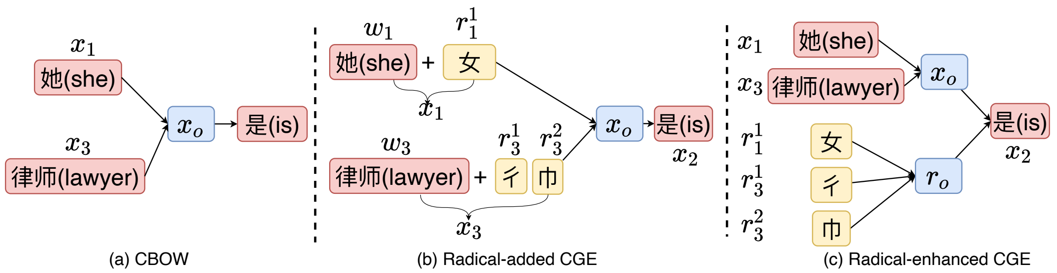
where c is the size of the training context. The prediction probability of x_t based on its context word is defined using softmax function:

$$P\left(x_t|x_{t-c}, ..., x_{t+c}\right) = \frac{\exp\left(\mathbf{x}_o^\top \cdot \mathbf{x}_t\right)}{\sum_{x_{t'}\in W} \exp\left(\mathbf{x}_o^\top \cdot \mathbf{x}_{t'}\right)},$$

where W is the words in the vocabulary. $\mathbf{x}_t$ is the embedding of word x_t, and $\mathbf{x}_o$ is the average of all context word vectors:

$$\mathbf{x}_o = \frac{1}{2c}\sum_{-c\le j\le c, j\ne 0} \mathbf{x}_{t+j}, \tag{2}$$

Since this formulation is impractical because of the training cost, hierarchical softmax and negative sampling are used when training CBOW (Mikolov et al., 2013b).



(a) CBOW (b) Radical-added CGE (c) Radical-enhanced CGE

Figure 1: Illustrations of baseline model and two proposed models. Radical-added CGE directly adds radical embedding to word embedding; Radical-enhanced CGE incorporates radical information to predict the target word.

3.2 Radial-added CGE

Since radical contains rich semantic and gender information, our model considers radical information to improve gender-neutral word embeddings. In Radical-added CGE, we directly add the radical vector representation with word vector, as shown in Figure 1(b).

The pivotal idea of Radical-added CGE is to replace the stored vectors $\mathbf{x}_t$ in CBOW with real-time compositions of $\mathbf{w}_t$ and $\mathbf{r}_t$, but share the same objective in Equation 1. Formally, a context word embedding $\mathbf{x}_t$ is represented as:

$$\mathbf{x}_t = \tfrac{1}{2}\left(\mathbf{w}_t + \tfrac{1}{N_t}\textstyle\sum_{k=1}^{N_t}\mathbf{r}_t^k\right), \tag{3}$$

where N_t is the number of radicals in word x_t, $\mathbf{w}_t$ is the word vector of x_t, and $\mathbf{r}_t^k$ is the radical vector of k-th radical in x_t. Take Figure 1(b) for example, when predicting the word "是(is)", we add the radical vector of "女" to word embedding of "她(she)", and add the average radical vector of "亻,巾" to word embedding of "律师(lawyer)".

3.3 Radical-enhanced CGE

In Radical-added CGE, the context word embedding is the sum of the word vector and radical vector, which ensures that the context word embedding contains the radical information. In this subsection, we propose a more flexible gender-neutral model, *i.e.,* Radical-enhanced CGE, where the radical embedding and the word embedding are separated, where the former is utilized to enhance the latter. The overview of Radical-enhanced CGE is shown in Figure 1(c).

Concretely, the context word embedding $\mathbf{x}_t$ now equals $\mathbf{w}_t$, which means that it does not contain radical embedding. Instead, we use context word vectors as well as context radical vectors to predict target words. Following setting in CBOW, we use $\mathbf{x}_o$ to denote the average of context word vectors, and $\mathbf{r}_o$ to denote the average of context radical vectors:

$$\mathbf{x}_o = \frac{1}{2c}\sum_{-c\leq j\leq c, j\neq 0}\mathbf{x}_{t+j}, \tag{4}$$

$$\mathbf{r}_o = \frac{1}{2c}\sum_{-c\leq j\leq c, j\neq 0}\frac{1}{N_{t+j}}\sum_{k=1}^{N_{t+j}}\mathbf{r}_{t+j}^k, \tag{5}$$

where c is the size of context window.

Next, $\mathbf{x}_o$ is used to calculate the predicted probability $P\left(x_t|x_{t-c},...,x_{t+c}\right)$. Similarly, $\mathbf{r}_o$ is also used to obtain the context radical prediction probability, which is represented as $P\left(x_t|r_{t-c},...,r_{t+c}\right)$:

$$P\left(x_t|x_{t-c},...,x_{t+c}\right) = \frac{\exp\left(\mathbf{x}_o^\top \cdot \mathbf{x}_t\right)}{\sum_{x_{t'}\in W}\exp\left(\mathbf{x}_o^\top \cdot \mathbf{x}_{t'}\right)}, \tag{6}$$

$$P\left(x_t|r_{t-c},...,r_{t+c}\right) = \frac{\exp\left(\mathbf{r}_o^\top \cdot \mathbf{x}_t\right)}{\sum_{x_{t'}\in W}\exp\left(\mathbf{r}_o^\top \cdot \mathbf{x}_{t'}\right)}. \tag{7}$$

Finally, the optimization target is to maximize:

$$\tfrac{1}{T}\textstyle\sum_{t=c}^{T-c}(\log P(x_t|x_{t-c},...,x_{t+c}) + \log P(x_t|r_{t-c},...,r_{t+c})). \tag{8}$$

The intuition behind this model is that the contextual radical embedding $\mathbf{r}_t$ interacts and predicts the target word embedding $\mathbf{x}_t$ so that the gender-related information in radicals is implicitly introduced in the word embeddings. During the back-propagation, the gradients of the embeddings of words and radical components can be different, while they are the same in Radical-enhanced CGE. Thus, the representations of words and radical components are decoupled and can be better trained.

Chinese word pair	English word pair	Category
神父-修女	Father: Nun	Definition
弟弟：妹妹	Headmaster:Headmistress	Definition
狗：猫	Dog: cat	None
书：杂志	Book: Magazine	None
沙发：躺椅	Sofa: Lounge chair	None
杯子：盖子	Cup: Lid	None
医生：护士	Doctor: Nurse	Stereotype
经理：秘书	Manager: Secretary	Stereotype
门卫：收银员	Guard: Cashier	Stereotype
领导：助理	Leader: Assistant	Stereotype

Table 1: Representative cases in `CSemBias` dataset. English words with wavy lines are untranslatable and we replace them with new Chinese words belonging to the same category.

4 Experimental Setup

4.1 Dataset

We adopt the 1GB Chinese Wikipedia Dump[1] as our training corpus. We follow Yu et al. (2017) when pre-processing the dataset, removing pure digits and non-Chinese characters. JIEBA[2] is used for Chinese word segmentation and POS tagging. We add all words in CSemBias in the tokenize vocab dictionary to ensure that the gender-related words are successfully recognized. Along with each character is its radical, and we crawled the radical information of each character from HTTPCN[3]. We obtained 20,879 characters and 218 radicals, of which 214 characters are equal to their radicals.

4.2 Comparisons

We compare our method against several baselines:
GloVe: a global log-bilinear regression model proposed in (Pennington et al., 2014).
Word2vec: introduced by Mikolov et al. (2013a), which either predicts the current word based on the context or predicts surrounding words given the current word. We chose the CBOW model following Chen et al. (2015); Yu et al. (2017).

The above two models denote non-debiased versions of the word embeddings.
Hard-GloVe: we use the implementation of hard-debiasing (Bolukbasi et al., 2016) method to produce a debiased version of GloVe embeddings.
GN-GloVe: preserves gender information in certain dimensions of embeddings (Zhao et al., 2018b).
GP(GloVe) and **GP(GN)**: aims to remove gender biases from pre-trained word embeddings GloVe and GN-GloVe (Kaneko and Bollegala, 2019).

The above three models all rely on additional labeled seed words including feminine, masculine, gender-neutral, and stereotype word lists. We translate their original word lists and adapt them to our Chinese domain. Namely, we add 22 out of 24-word pairs in the test dataset into the supplementary data.

To compare our model with other structure-based Chinese embedding models, we include the performance of other models that also incorporate component information: **CWE** is a character-enhanced word embedding model presented in Chen et al. (2015); **MGE** and **JWE** are multi-granularity embedding model that make full use of word-character-radical composition (Yin et al., 2016; Yu et al., 2017); **RECWE** is a radical enhanced word embedding model (Chen and Hu, 2018). These baselines include radical information in the word embedding construction process, but also take other information sources such as character-level information into consideration, which diminishes the importance and effectiveness of gender-related radicals. The purpose of this comparison is to demonstrate that existing structure-based Chinese word embedding models still suffer from gender bias problems.

4.3 Implementation Details

For all models, we use the same parameter settings. Following Yu et al. (2017), we set the word vector dimension to 200, the window size to 5, the training iteration to 100, the initial learning rate to 0.025, and the subsampling parameter to 10^{-4}. Words with a frequency of less than 5 were ignored during training. We used 10-word negative sampling for optimization. The whole training process takes about six hours.

5 Experimental Result

5.1 Evaluating Debiasing Performance

`CSemBias` Dataset. To evaluate debiasing performance of our model, we come up with a new dataset named `CSemBias` (Chinese SemBias). Concretely, we hire three native Chinese speakers to translate the original English `SemBias` (Zhao et al., 2018b) dataset to the Chinese version. Each instance in `CSemBias` consists of four word pairs: a gender-definition word pair (**Definition**; e.g., "神父-修女(priest-nun)"), a gender-stereotype word pair (**Stereotype**; e.g., "医生-护士(doctor-nurse)")

[1] `http://download.wikipedia.com/zhwiki`
[2] `https://github.com/fxsjy/jieba`
[3] `http://tool.httpcn.com/zi/`

Embeddings	CSemBias-subset			CSemBias		
	Definition ↑	Stereotype ↓	None ↓	Definition ↑	Stereotype ↓	None ↓
GloVe	40.0	37.5	22.5	49.1	31.4	19.5
Word2vec	47.5	30.0	22.5	72.5	17.7	9.8
CWE	45.5	27.5	27.0	57.3	25.2	17.5
JWE	45.0	25.0	30.0	52.3	25.9	21.8
RECWE	50.0	25.0	25.0	60.4	21.4	18.2
MGE	57.5	32.5	10.0	63.6	30.7	5.7
Hard-GloVe	17.5	57.5	25.0	73.6	15.7	10.7
GN-GloVe	17.5	50.0	32.5	92.5	4.5	3.0
GP(GloVe)	15.0	52.5	32.5	71.1	16.4	12.5
GP(GN)	12.5	50.0	37.5	90.4	7.3	**2.3**
Radical-added CGE	**82.5**†∗	**15.0**†∗	**2.5**†∗	**93.4**†∗	**3.9**†∗	2.7†∗
Radical-enhanced CGE	75.0†∗	17.5†∗	7.5†∗	86.8†∗	10.0†∗	3.2†∗

Table 2: Prediction accuracies for gender relational analogies. † and ∗ indicate statistically significant differences against Word2vec and Hard-GloVe respectively.

Model	Wordsim-240	Wordsim-295
GloVe	0.5078	0.4419
Word2vec	0.5009	0.5985
Hard-GloVe	0.5046	0.4378
GN-GloVe	0.5026	0.4400
GP(GloVe)	0.4959	0.4451
GP(GN)	0.4959	0.4451
Radical-added CGE	0.5120	0.5875
Radical-enhanced CGE	0.5067	0.5821

Table 3: Results on word similarity evaluation.

and two other word-pairs that have similar meanings but not a gender relation (**None**; e.g., "狗-猫(dog-cat)", "茶杯-盖子(duc-lid)"). CSemBias contains 20 gender-stereotype word pairs and 22 gender-definitional word pairs, and we use their Cartesian product to generate 440 instances. In the annotation process, for the translatable words, the annotators obtain the same translation results to be included in CSemBias. For untranslatable words, each annotator comes up with a Chinese word belonging to the same category, and they decide the final word together.

Examples are shown in Table 1. Since some of the baselines follow the supervised style, we split the CSemBias into training and test datasets. Among the 22 gender-definitional word pairs, 20-word pairs are used in the training, and the left 2 pairs are used for the test dataset. We name the out-of-domain test dataset as CSemBias-subset.

Debias Evaluation. To study the quality of the gender information present in each model, we follow Jurgens et al. (2012) to use the analogy dataset, CSemBias, with the goal to identify the correct analogy of "he- she" from four pairs of words. We measure relational similarity between (他(he),她(she)) word-pair and a word-pair (a, b) in CSemBias using the cosine similarity between the $\overrightarrow{he} - \overrightarrow{she}$ gender directional vector and $\vec{a} - \vec{b}$ directional vector. We select the word-pair with the highest cosine similarity with $\overrightarrow{he} - \overrightarrow{she}$ as the predicted answer. If the trained embeddings are gender-neutral, the percentage of gender-definitions is expected to be 100%.

From Table 2, we can see our models achieve the best performances on both datasets. In terms of CSemBias-subset, component-based Chinese word embedding models achieve better performance than simple GloVe or Word2vec, which demonstrates that component information is indeed useful in alleviating gender bias. To our surprise, debias models perform poorly on CSemBias-subset, indicating that they do not generalize well to out-of-domain tests. Comparing the performances on CSemBias-subset and CSemBias, we can find that the performance of supervised baseline models highly relies on labeled gender-related word sets. As for our model, both Radical-added CGE and Radical-enhanced CGE achieve comparable and even better performance than the state-of-the-art GN-GloVe model and perform significantly better than Hard-GloVe. Radical-added CGE outperforms Radical-enhanced CGE by a small margin, because it directly stores radical information in word embedding, emphasizing gender information explicitly. Since both of our models are unsupervised, the result means that the radical semantic information in Chinese is especially useful for alleviating gender discrimination, and our models can successfully utilize such information. We use the Clopper-Pearson confidence intervals following Kaneko and Bollegala (2019) to do the significance test.

5.2 Preservation of Word Semantics

Apart from examining the quality of the gender information present in each model, it is also important that other information that is unrelated to gender biases is preserved. Otherwise, the performance of downstream tasks that use these embeddings might be influenced.

Semantic Similarity Measurement. This task evaluates the ability of word embedding by its capacity of uncovering the semantic relatedness of word pairs. We select two different Chinese word similarity datasets, *i.e.,* Wordsim-240 and Wordsim-295 provided by Chen et al. (2015). Wordsim-240 contains 240 pairs of Chinese words and their corresponding human-labeled similarity scores, and the same is true for Wordsim-295. Previous work (Kaneko and Bollegala, 2019) noted that there exist gender-biases even in the English word similarity test dataset. However, we confirm that no stereotype examples exist in Chinese Wordsim-240 and wordsim-295. The similarity embedding score for a word pair is computed as the cosine similarity of their embeddings. We compute the Spearman correlation (Myers et al., 2010) between the human-labeled scores and similarity scores computed by embeddings. Higher correlation denotes better quality. From Table 3, we can see that Radical-added CGE obtains the best performance on Wordsim-240 dataset, outperforming the best baseline Word2vec by 0.0111. A possible reason is that radical information is also useful in semantic similarity tests. Generally, two CGE models perform comparable to Word2vec, indicating that information encoded in Word2vec is preserved while stereotype gender bias is removed.

Analogy Detection. This task examines the quality of word embedding by its ability to discover linguistic regularities between pairs of words. Take the tuple "罗马(Rome):意大利(Italy)-柏林(Berlin):德国(Germany)", the model can answer correctly if the nearest vector representation to $\overrightarrow{Italy} - \overrightarrow{Rome} + \overrightarrow{Berlin}$ among all words except Rome, Italy, and Berlin. More generally, given an analogy tuple "$a : b - c : d$", the model answers the analogy question "$a : b - c :?$" by finding x that:

$$\arg \max_{x \neq a, x \neq b, x \neq c} cos(\vec{b} - \vec{a} + \vec{c}, \vec{x}) \qquad (9)$$

Model	Total	Capital	State	Family
GloVe	0.7846	0.8655	0.9257	0.4926
Word2vec	0.7954	0.8493	0.8857	0.6029
Hard-GloVe	0.7563	0.9099	0.8571	0.3088
GN-GloVe	0.7794	0.9114	0.8857	0.3824
GP(GloVe)	0.7633	0.8715	0.9029	0.4044
GP(GN)	0.7740	0.8996	0.8457	0.4154
Add-CGE	0.7625	0.8400	0.7829	0.4963
Enh-CGE	0.7794	0.8405	0.8914	0.5551

Table 4: Results on word analogy reasoning.

We use the same dataset as in (Yu et al., 2017), which consists of 1,124 tuples of words and each tuple contains 4 words. There are three categories in this dataset, i.e., "Capital" (677 tuples), "State" (175 tuples), and "Family" (272 tuples).

The percentage of correctly solved analogy questions is shown in Table 4. We can see that there is no significant degradation of performance in our model and debias baselines. Specifically, Radical-enhanced CGE performs better than Radical-added CGE. One possible reason is that, in Capital and State related words, the semantic meanings can not be directly revealed by radicals.

6 Conclusion

In this paper, we proposed two methods for unsupervised training in Chinese gender-neutral word embedding by emphasizing gender information stored in Chinese radicals in explicit and implicit ways. Our first model directly incorporates radical embedding in its word embedding, and the second one implicitly utilizes radical information. Experimental results show that our unsupervised method outperforms the supervised debiased word embedding models without sacrificing the functionality of the embedding model.

7 Bias Statement

In this paper, we study stereotypical associations between male and female gender and professional occupations in contextual word embeddings. We regard a system as a biased system if the word embeddings of a specific gender are more related to certain professions. When such representations are used in downstream NLP applications, there is an additional risk of unequal performance across genders (Gonen and Webster, 2020). We believe that the observed correlations between genders and occupations in word embeddings are a symptom of an inadequate training process, and decorrelating genders and occupations would enable systems to counteract rather than reinforce existing gender imbalances.

In this work, we focus on evaluating the binary gender bias performance. However, gender bias can take various formats, and we are looking forward to evaluating the bias in Chinese word embeddings by various methods.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. The work was supported by King Abdullah University of Science and Technology (KAUST) through grant awards Nos. BAS/1/1624-01, FCC/1/1976-18-01, FCC/1/1976-23-01, FCC/1/1976-25-01, FCC/1/1976-26-01.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of bias in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

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

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huanbo Luan. 2015. Joint learning of character and word embeddings. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Zheng Chen and Keqi Hu. 2018. Radical enhanced chinese word embedding. In *Chinese computational linguistics and natural language processing based on naturally annotated big data*, pages 3–11. Springer.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models (extended abstract). In *ECCV*.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miroslav Dudík, and Hanna M. Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *ArXiv*, abs/1812.05239.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. Vmsmo: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369.

Mingzhe Li, Xiuying Chen, Min Yang, Shen Gao, Dongyan Zhao, and Rui Yan. 2021. The style-content duality of attractiveness: Learning to write eye-catching headlines via disentanglement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13252–13260.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Jerome L Myers, Arnold Well, and Robert Frederick Lorch. 2010. *Research design and statistical analysis*. Routledge.

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

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, pages 1–19.

Gabriel Stanovsky, Noah A. Smith, and Luke S. Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *ACL*.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity chinese word embedding. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 981–986.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 286–291.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*.

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