ChGaR: A Performance-Guaranteed Chinese Gender Debiasing Rewriter

Danying Xu, Jinghan Zhang, Yingtong Lu, Beilun Wang, Meng Wang
Southeast University

Abstract

Gender bias is the tendency to prefer one gender over another, which is learned in various NLP models. Chinese is an isolating language containing a large number of characters that makes gender bias hide deeply in implicit meanings. Past work targeted at Chinese either focused on model debiasing, but performed poorly in downstream tasks because of biased corpora; or concentrated on locally data debiasing via converting a portion of certain words, which are unable to solve implicit bias problems. Our work proposes a novel Chinese Gender Debiasing Rewriter (ChGaR) framework. We first transform the gender-related context to opposite ones via a neural machine translation model and then aggregate both two data sets. The bias in ChGaR-augmented data reduced 45.95% in average in the evaluation of WEAT, WinoBias, and EEC tests, respectively. In addition, we achieve the close performance in two downstream tasks compared to the original dataset.

1 Introduction

Powerful deep learning models are able to learn inappropriate gender biases from human languages, even perpetuate and accentuate them. For example, Amazon AI resume screening model has been proved to prefer male applicants(Hu et al., 2022)(Dastin, 2018), which is caused by human bias in training data. For Chinese women with a labor force participation rate of more than 60%, if this kind of unfair resume screening system is used, the negative social impact will be especially significant.

As an isolating language (William and Thomson, 1990), Chinese has no explicit word-form change, so that its gender characteristics and sexism are deeply hidden. Another feature is that Chinese words are mainly compound words. Additionally, it has a low ratio of morpheme-per-word, so there are loads of characters in Chinese, each of which has multiple meanings. To debias Chinese text, we need to replace meaningful characters, so that compound words are innovated, which means a hard work from scratch for the machine.

Many approaches have been proposed in the past to avoid structural discrimination against protected groups. Two main approaches for bias mitigation can be identified: one is to adjust hidden output in the model, another is to augment data in the training set.

It is polular to make post-processing for word representations such as hard debiasing (Bolukbasi et al., 2016,Mu et al., 2017), but past studies lack description of model performance after adding debiasing layers. Liang et al. (2020)'s work, extracting the dimensions of the gender subspace and zero them out, mentions that their model performs poorly no matter which zeroing scheme is used. There are two reasons for the poor performance. On one hand, since the test set is also biased, the unbiased model and the biased data are mismatched, thus leads to obvious accuracy reduction. On the other hand, model debiasing is equivalent to adding a new task of unsupervised distinguishing whether the learned knowledge has gender bias or not to the original model, which increases the difficulty of training. In this circumstance, model debiasing methods are hard to popularize in real-world applications, as some people may be reluctant to degrade model performance for fairness.

In terms of data debiasing, the resampling approach can be regarded as a baseline to alleviate the amount imbalance (Burnaev et al., 2015)(Chawla et al., 2002). It not only has the risk of overfitting but also doesn't change the contextual scene of gender words. For example, women still appear more in the family context, but less or not in the workplace. Only rewriting gender words in sentences can biases be balanced in context, such as having women mentioned more as sci-

entists and men more involved in household matters.

Vanmassenhove et al. (2021) converted gender-related pronouns ('she','he') into neutral ones ('they'), which is not appropriate for Chinese because group word 'they' is written as 'many he' ("") in Chinese. The work of Xuewen et al. (2021) for Chinese rewrites the pronouns ('he' to 'she', 'she' to 'he') and take translation as the downstream task. However, pronouns are only a small portion of gendered words, so that pronouns rewriting is inadequate. Jain et al. (2021) proposes a rewriter for Spanish, but as a inflectional language, Spanish gender words have word-form transformation, so that this work only targets at short sentences and requires only one gendered word in one sentence, which is seriously limited. As far as we know, there's no rewriting work with full coverage of gender words has been done for Chinese or other isolating languages.

We propose ChGaR, a performance-gauranteed Chinese rewriter framework capable of gender debiasing. It consists of three parts: pattern transformation, neural machine translation (NMT) and data aggregation. Supported by the theoretical support of Chinese linguistics, our framework can directly rewrite Chinese sentences and remove bias by increasing the amount and diversity of data. We evaluate ChGaR using three methods: direct evaluation via Word Embedding Association Test (WEAT, Caliskan et al., 2017) and indirect evaluation, WinoBias(Zhao et al., 2018)(He and Choi, 2021)(Rudinger et al., 2018) in coreference resolution and Equity Evaluation Corpus (EEC, Kiritchenko and Mohammad, 2018) in sentiment analysis.

The main contributions of our work are as follows:

1. We propose the first effective comprehensive data debiasing approach, which has great generalization when applying on other corpus.

2. We prove in downstream work that model performance wouldn't be affected when trained with unbiased dataset produced by our rewriter.

3. We provide three gender bias evaluation metrics for Chinese that can be used in future work.

2 ChGaR

ChGaR can be divided into three segments: pattern transformation, NMT and data aggregation. Pattern transformation provides balanced data for training the NMT which can complete the gender transformations. Data aggregation provides some ways to combine the original data with aggregated data from NMT. We use two translation models (the difference is whether attention mechanism exists) considering character level and word level to instantiate our framework presented in Figure 1.

2.1 Pattern Transformation

Unlike most of the inflectional languages such as English and German, Chinese is an isolating language that are usually expressed in the form of compound words and phrases(Chen, 2005). We summarize four aspects of gender-convertible opposites from Chinese language-level research.

Pronoun "()", which is usually used to refer to a male, is used to refer to any person (people) of indeterminate gender or a person in general (people). This results in far more pronouns referring to males than females in the text data.

Derogatory words about women Inequality between the two genders is common in Chinese. For example, "[1]" can be used for any adult male in any context while the word "[1]" is a small social term used exclusively for unmarried women while still has a pejorative emotional connotation in some cases. Meanwhile, some curses and dirty words in Chinese are often aimed at women, and derogatory words for women are more figurative.

Treating women as exceptions Some words to express identity, occupation, and status are assumed to identify men by default. For example, "[2]" refers to men by default, and gender is required as a definite article when describing "[2]".

Sexes-ordered words In Chinese, when a word contains both genders, the language order usually follows the principle of male first and female second, such as "[3]", "[4]". This linguistic order enables women to fully embody the higher social status of men in Chinese tradition.

We design two lists and use the rule-based method to convert gender opposite words for the above four situations. However, some words that do not contain gender meanings but contain gender words were incorrectly converted, such as the word "[5]" is converted by mistake to "[5]". In order to

[1]English:"Sir"; "Miss"

[2]English:"driver"; "women-driver"

[3]English: parents that in the order of father and mother.

[4]English: men farming and women weaving

[5]English: "Company". Its literal translation into Chinese is "Male Management", thus is wrong translated to "Female

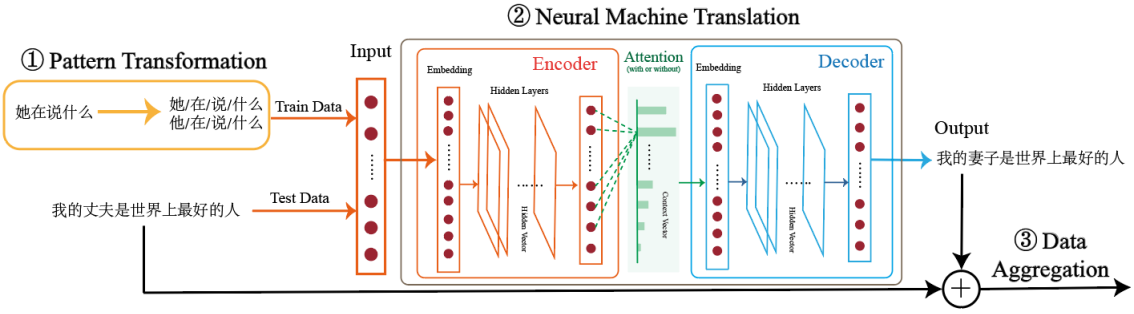


Figure 1: The pipeline of ChGaR Framework. It contains three segments: pattern transformation, NMT and data aggregation. ""means "What does she say", which is converted to "What does he say" through pattern transformation. The test data "" ("My husband is the best person in the world") can be converted to "" ("My wife is the best person in the world") and output the aggregated data through this end-to-end ChGaR framework.

solve the problem that patterns only convert opposite words without considering Chinese compound words and context, we believe that it is necessary to build a neural translation model based on Chinese gender debiasing rewriter.

2.2 Neural Machine Translation

Considering that Chinese texts are different from English texts divided by spaces, which may cause some words to be inaccurate in translation if only the sentences are divided by characters when generating vectors, we use the methods of single-character and pos-tagging respectively for sentences splitting to construct the word-num mapping dictionary.

We choose Seq2Seq model for neural machine translation and use GRU as the encoder and decoder in it to achieve higher training efficiency. The "teacher_forcing_ratio" is set to 0.5 in the Seq2Seq model to avoid the subsequent training results being affected by errors during the training process.

To address the problem that the Seq2Seq model cannot effectively focus on the input target when used alone, we also use the Seq2Seq with attention model(Tang et al., 2016). The attention mechanism can encode the encoder into different contextual variables based according to each time step of the sequence. When decoding, the decoding output is combined with each different context variable, which makes the conversion of gender words in the data more accurate.

2.3 Data Aggregation

Several methods can be used in data augmentation. One is to integrate all the data before and after rewriter processing to expand the scale of the dataset. The second is to extract from the data containing gender words before and after the rewriter conversion with proportion 1:1. They are then fused with gender-free data, keeping the dataset of the same size.

3 Results & Analysis

Both debiasing effect and performance are evaluated on the rewriter. Due to the lack of evaluation methods designed for Chinese, we culturally-adaptively rewrite the following approaches to Chinese.

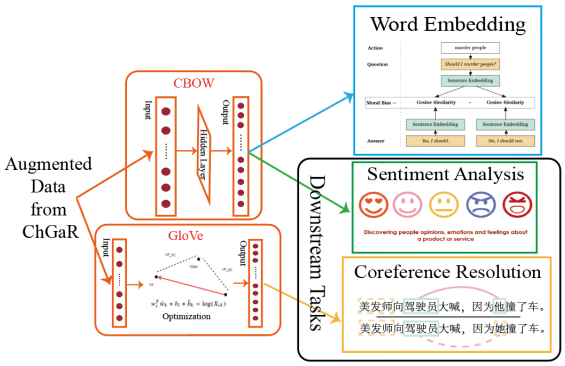


Figure 2: Three evaluations for ChGaR: word embedding (evalluate directly) and two downstream tasks coreference resolution and sentiment analysis (evaluate indirectly).

3.1 Datasets

ChGaR We use the Fudan University Chinese text classification dataset that contains 9833 documents from 20 categories. We split the document content by sentence, intercept 2,223,107 sentences of length 3-50 after long-tail analysis (see Appendix) and set "sentences with gender words:

Management" which has no meaning in Chinese.

sentences without gender words = 1:1". We end up with a dataset of 148,002 Chinese sentences in total.

Coreference Resolution We use the WinoBias proposed by (Zhao et al., 2018) contains a total of 3,167 augmented synthetic sentences under 2 sentence types. We rewrite it all into Chinese for use.

Sentiment Analysis See the source of dataset in Ethics Statement which has 159,081 sentences with "train:validation:test = 7:2:1".

3.2 Word Embedding

Word2vec(Mikolov et al., 2013) is an approach for training digital representation for words due to context, which causes that words semantically close to each other would also be neighbors in high-dimensional embedding space. It can reflect how people with different genders are treated differently when they are mentioned in texts. We mine this difference in word embeddings via a classic method WEAT(Caliskan et al., 2017), whose basic idea is to compare the distances of vectors. Larger distance means more evident gender bias.

The dataset scores are shown in Figure 3 with lists we choose (A, B, $X(123)$, $Y(123)$). Detailed description see Appendix. The average bias reduces 43.6%.

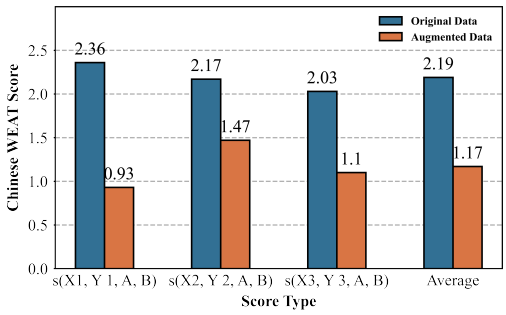


Figure 3: The word embedding evaluation scores (WEAT) experimented on original dataset and ChGaR augmented dataset.

3.3 Coreference Resolution

In the coreference resolution task, when a sentence contains a pronoun and two biased occupation words with different genders, the results may change with the change of gender pronoun. We used e2e-coref model for the processed winobias data and demonstrated the presence of gender bias in the model. We retrained the model using the debiased data and used the test set for validation.

The test set was divided into two parts, one of which was generated after gender rewriting in the other part. The pronouns and the referred occupations were used as the labeled clusters. The model should yield the same results for both parts of the data without gender bias. We used the Euclidean distance to measure the bias, which decreased 38.24%, from 0.34 for the original model to 0.21 for the retrained model. In addition, we compared the initial clusters with the resultant clusters and calculated that the accuracy improved from 77.66% for the initial model to 79.0% for the retrained model, indicating that the retrained model guaranteed the original performance while reducing gender bias.

3.4 Sentiment Analysis

Sentiment and emotion analysis systems trained by biased data consider utterances from one gender differently simply because of their gender, for example, customer support systems may prioritize a call from an angry male over a call from the equally angry female. This will be reflected in the difference in sentiment scores given by the system in the face of sentences that differ only in gender. EEC proposed in (Kiritchenko and Mohammad, 2018) is a designed evaluation dataset containing eleven templates, and one can calculate the gender-paired two sample t-test to determine whether the mean difference between the two sets of scores is significant. Therefore, we trained TextCNN-based(Kim, 2014) sentiment analysis models on original data and ChGaR-augmented data, respectively. We apply EEC to our two models, and the difference significance reduced 50.02%.

4 Conclusion

This paper proposes an end-to-end Chinese data debiasing rewriter framework ChGaR. After the training data is rewritten by pattern, it can be more balanced. We use the NMT method for model training, which can more intelligently learn the gender opposite words that need to be converted, to avoid the rewriting errors caused by the rule-based approach. After one direct evaluation and two evaluation methods via downstream tasks, we show that ChGaR has certain advantages.

In future work, we plan to find NMT models with better performance to expand the advantages

of ChGaR and make up for the insufficiency of related work on Chinese gender debiasing rewriters.

Limitations

There are some limitations in our work listed as follows.

1. The dataset used to train the rewriter (from Fudan University) is selected from news and professional literature. There are many sentences that do not conform to normal grammar and are not easy to remove (such as professional symbols, etc.), which may disturb the model during training.

2. The coreference resolution dataset WinoBias needs to be translated from English. After a certain number of sentences are translated, the format will change, which may disturb the model. And due to the small amount of data, the model may be under-trained or over-fit.

3. The rewriter is limited to isolated languages such as Chinese. Similar rewriters these years all require special patterns for each language, just like the earlier dictionary-based translations. We look forward to general pattern alternatives being proposed.

4. The translation model is lightweight due to the limited GPU. The effect has not been verified on a large model, but it should be better in theory.

Ethics Statement

In order to avoid inaccurate extraction and use of Chinese gender words, we refer to the language characteristics and gender word analysis of many Chinese language and literature majors(Chen, 2005)(Xie, 2004)(Li, 2011).

We used three public datasets for our experiments. The Chinese text classification dataset of Fudan University was provided by Li Ronglu, Fudan University, from the Natural Language Processing Group of the International Database Center, Department of Computer Information and Technology, Fudan University. The coreference resolution dataset was created by the (Zhao et al., 2018) paper and distributed under the MIT license. EEC test comes from templates designed by Kiritchenko and Mohammad (2018). The sentiment analysis dataset refers to the "simplify_4_mood" dataset from a repository on github[6].

[6]https://github.com/SophonPlus/ChineseNlpCorpus

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*, volume 29. Curran Associates, Inc.

E Burnaev, P Erofeev, and A Papanov. 2015. Influence of resampling on accuracy of imbalanced classification. In *Eighth International Conference on Machine Vision*, volume 9875, page 987521.

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

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Congyun Chen. 2005. Sexism in chinese. *Journal of Nantong University: Philosophy and Social Sciences Edition*, 21(4):105–107.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

S Hu, JA Al-Ani, KD Hughes, N Denier, A Konnikov, L Ding, J Xie, Y Hu, M Tarafdar, B Jiang, et al. 2022. Balancing gender bias in job advertisements with text-level bias mitigation. *Front. Big Data 5: 805713. doi: 10.3389/fdata*.

Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating gender augmented data for nlp. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Haixia Li. 2011. Modern chinese common gender words "male/female" synonym series. *Theory Research*, (13):232–233.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. All-but-the-top: Simple and effective postprocessing for word representations. *arXiv preprint arXiv:1702.01417*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Yujin Tang, Jianfeng Xu, Kazunori Matsumoto, and Chihiro Ono. 2016. Sequence-to-sequence model with attention for time series classification. In *2016 IEEE 16th International Conference on Data Mining Workshops*, pages 503–510. IEEE Computer Society.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948.

Thomson William and William Thomson. 1990. Typology and universals.

Lixia Xie. 2004. A comparative study of gender words in english and chinese languages. *Yun Meng*, 25(3):111–113.

Shi Xuewen, Huang Heyan, Jian Ping, and Tang Yikun. 2021. The method for reducing gender bias of pronouns in uyghur to chinese neural machine translation. *Journal of Xiamen University(Natural Science)*, 60(4):693–700.

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 (Short Papers)*, pages 15–20.

A Appendix

There are some more figures during experiments we would like to attach.

The long-tail relationship for dataset used in ChGaR are shown in figure 4, 5. We select sentences with a sentence length of 3 to 50 words according to long-tail statistics.

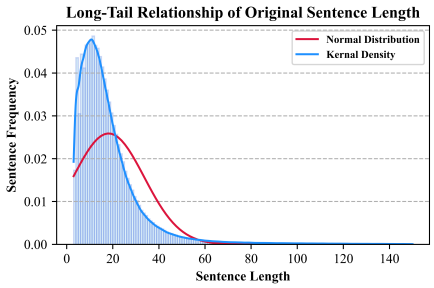


Figure 4: Before Data Processing

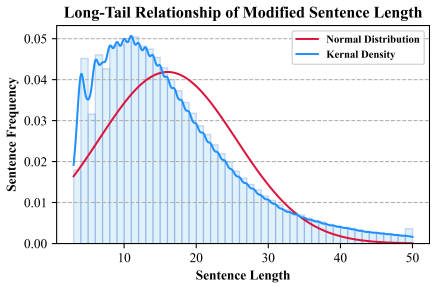


Figure 5: After Data Processing

Figure 6, 7, 8 show the train and validation loss process during NMT training.

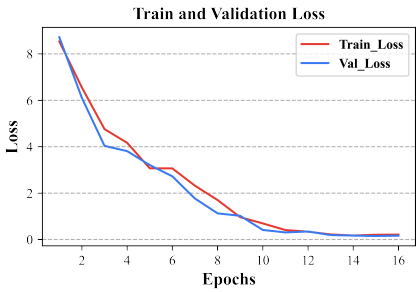


Figure 6: Train and Validation Loss of Epochs in Seq2Seq Model on Character Level

For word embedding evaluation (WEAT), A, B, X3 and Y3 are lists for gender words with respect to women, men, feminine adjectives and masculine adjectives. X1, X2, Y1, Y2 are lists for different categories with respect to family, career, arts and science.

To be detailed, the average of the difference of the two results on female wordlist called A and on

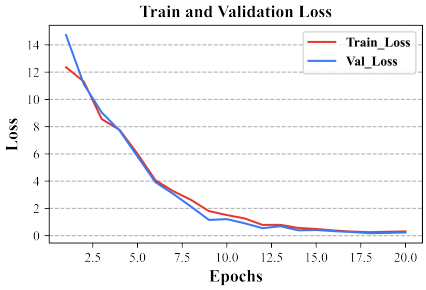


Figure 7: Train and Validation Loss of Epochs in Seq2Seq Model on Word Level

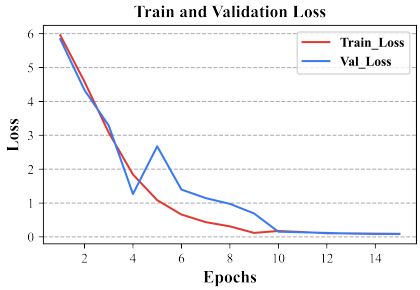


Figure 8: Train and Validation Loss of Epochs in Seq2Seq with Attention Model on Character Level

male one called B is noted as $s(w, A, B)$:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b}) \quad (1)$$

where w is a word belongs to female-related wordlist X or male-related Y. After adding up all of the differences of distances and normalization, we will get the final WEAT score $s(X, Y, A, B)$:

$$\begin{aligned} s(X, A, B) &= \text{mean}_{w \in X} s(w, A, B) \\ s(Y, A, B) &= \text{mean}_{w \in Y} s(w, A, B) \\ s(X, Y, A, B) &= \frac{s(X, A, B) - s(Y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}. \end{aligned} \quad (2)$$

Figure 9 shows the training and validation accuracy under original dataset and augmented dataset in sentiment analysis. The similar accuracy reveals that ChGaR is guranteed for the downwtram tasks after data augmentation.

Figure 10 is the results for sentiment analysis that shows the gender difference is lower after the data augmentation through ChGaR. The results are compared between "ori_0" and "aug_0", respectively.

Our experiments are based on GeForce RTX 2080 Ti GPU. The code language is Python, and the deep learning part is written based on Pytorch.

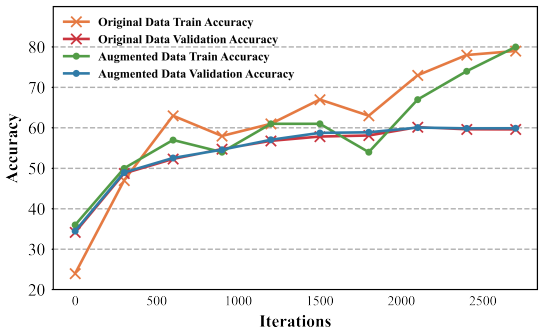
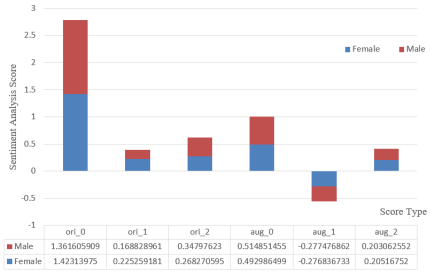


Figure 9: The sentiment analysis evaluation model (ECC) accuracy during training and validation.



	ori_0	ori_1	ori_2	aug_0	aug_1	aug_2
Male	1.361605909	0.168828961	0.34797623	0.514851455	-0.277476862	0.203062552
Female	1.42313975	0.225259181	0.268270595	0.492986499	-0.276836733	0.20516752

Figure 10: The results for sentiment analysis.

The software uses jupyter notebook and pycharm. The training of NMT cost a week.