Investigating Gender Bias in Pre-trained Korean Word Embeddings

Chani Jung [1] **Sojeong Song** [2] **Gyojin Han** [2]

Abstract

Various previous studies on word embeddings have shown that word embeddings have social bias such as gender bias. Such social bias can influence in a way that harms fairness in the down-stream NLP tasks that use word embeddings. Therefore, various debiasing methods are being studied to avoid such threats. Nevertheless, studies on the gender bias of Korean word embeddings have not been actively conducted. In this study, we provide Korean word list dataset including gender-definition words, non-gender-definition words, and profession words to check the existence of gender bias in Korean word embeddings. Additionally, we check the gender bias of Korean word embeddings through bias-by-projection and five GBWR tasks and check whether Half-Sibling Regression (HSR), a powerful debiasing method proposed by Zekun Yang and Juan Feng, is also effective in debiasing the gender bias of Korean word embeddings. Experiments on our dataset demonstrate that the gender bias exists in Korean embeddings and can be mitigated through HSR.

1. Introduction

Word embeddings trained with large text corpus often demonstrate discriminative social biases, such as gender, racial or ethnic biases. The down-stream NLP tasks which make use of the word embeddings are highly likely to reflect or even amplify the bias in word embeddings. Therefore, the word embeddings need to be debiased before they are used in downstream tasks.

There have been several works that attempted to measure and mitigate social biases in English word embeddings. Specifically, gender bias has been investigated with the most effort among several categories of social biases. Nonetheless, research on gender bias in Korean word embeddings has not progressed much in spite of increased demand for Korean NLP systems.

We aim to investigate gender bias in Korean word embeddings by measuring and mitigating the bias in KLUE-BERT embeddings. First of all, we constructed the Korean word list dataset which is required for measuring and reducing gender bias in Korean word embeddings. We measured projection of the embeddings on gender direction [1] and the relation among gender-biased words [6] to estimate the amount of gender bias included in word embeddings. Then, we reduced the bias by Half-Sibling Regression (HSR), which removes the statistical dependency between gender-definition and non-gender-definition word embeddings.

As a result, we identified the gender bias existing in Korean word embeddings, which showed a similar tendency with that of English word embeddings. HSR was also effective in reducing the bias in Korean word embeddings. In addition, we could verify the necessity and validity of our Korean word list dataset by the experimental results.[1]

2. Approach

We check whether the gender bias exists in Korean word embeddings through the bias-by-projection and the five Gender-Biased Word Relation (GBWR) tasks. Additionally, we check whether Half-Sibling Regression (HSR), a post-processing algorithm that uses causal inference method to learn and subtract spurious gender information contained in non-gender-definition, is also effective in debiasing Korean word embeddings gender bias. The rest of this section describes bias-by-projection, GBWR tasks, and Half-Sibling Regression algorithm.

2.1. Bias-by-projection

Bias-by-projection is the dot prouct between the target word and the gender direction $\overrightarrow{he} - \overrightarrow{she}$ [6]. The best corresponding words for 'he' and 'she' in Korean are '그' and '그녀'. However, the word '그' is used not only to refer to a man but also to a woman, and is used for purposes other than pronouns that refer to a person. Therefore, we selected '

[1]School of Computing, KAIST, South Korea [2]Electrical Engineering, KAIST, South Korea. Correspondence to: Chani Jung <1016chani@kaist.ac.kr>, Sojeong Song <akqjq4985@kaist.ac.kr>, Gyojin Han <hangj0820@kaist.ac.kr>.

[1]Codes and datasets are available at https://github.com/akqjq4985/CS575_Team4.git

Table 1. Comparison of word list sizes

	Gender-Definition (pairs)	Non-Gender-Definition		Profession
Ours	33	Gender-neutral	KLUE-BERT vocabulary	109
		878	32K	
Original (Yang et al.)	223	1.9M		320

남성 (male)' and '여성 (female)' as words for calculating gender direction in Korean.

2.2. Gender-Biased Word Relation Tasks

Gender-Biased Word Relation (GBWR) tasks examines whether gender bias exists in word embedding relation. The five GBWR tasks were proposed by Gonen and Goldberg [2] and we use them to measure the gender bias of Korean word embeddings.

Clustering. We take the top K male-biased words and the top K female-biased words according to the bias-by-projection value of original embedding and clusters them into two clusters using k-means [4], where K is the parameter of the number of the biased words.

Correlation. We measure the correlation between bias-by-projection and bias-by-neighbors. The bias-by-neighbors of a target word refers to the percentage of male/female socially-biased words among the k-nearest neighbors of it.

Profession. For this task, we use our proposed Korean Profession word list. For each profession word, we first calculate the top K nearest neighbors and count the number of male-biased neighbors. Then, we compute the Pearson correlation coefficient between the number of male-biased neighbors and the original bias-by-projection values.

Association. We choose clear male/female names from the names in the KLUE-BERT vocabulary list. Then, we measure the association between male/female names and three groups of words including family and career words, arts and mathematics words, and arts and science words.

Classification. In this task, we sample the top 400 male-biased words and 400 female-biased words (total 800 words). Then, we trained an SVM classifier with a subset of 600 words to predict the gender of the rest 200 words.

2.3. Half-Sibling Regression

Zekun Yang and Juan Feng [6] propose that the debiased non-gender-definition word embeddings are learned by substracting the approximated gender information from the original non-gender-definition word embeddings, where the approximated gender infromation is obtained by predicting the original non-gender-definition word embeddings using gender-definition word embeddings. To calculate gender information, Zekun Yang and Juan Feng use Ridge Regression. Through learning the approximated gender information in gender-biased word embeddings, both gender bias in word embedding relation and gender bias associated with gender direction are learned and subtracted.

3. Experiments

3.1. KLUE-BERT word embeddings

We used KLUE-BERT embedding as the target Korean word embedding to evaluate and debias. KLUE-BERT is a Korean pre-trained language model suggested by Park et al. [5]. The authors pre-trained the model with 473M Korean sentences from five publicly available corpora - MODU, CC-100-Kor, NAMUWIKI, NEWSCRAWL, PETITION. They designed and used morpheme-based subword tokenization, which first tokenizes raw text into morphemes and apply Byte Pair Encoding (BPE) to get the final vocabulary. They claimed that this worked better than using BPE solely, because Korean is an agglutinative language. To get KLUE-BERT word embeddings, we passed a single-word sentence through KLUE-BERT, and averaged the second-to-last hidden layer embedding of each token to get a single 768 dimensional embedding.

3.2. Korean Word Lists

For measuring and reducing gender bias of word embeddings, there were three kinds of word lists required. Since no such Korean word lists were available, we built our own datasets by translating English word lists and manually rectifying them.

One required word list was gender-definition word list. We translated the corresponding English dataset suggested by Zhao et al. [7] after excluding the words that do not exist in Korean language.

For non-gender definition word list, we used the dataset translated from the English gender-neutral word list built by Kaneko et al. [3]. Since this gender-neutral word list has much smaller size compared to the dataset used in the original paper, we also experimented with another word list for comparison to check the validity and effectiveness of

the gender neutral word list. It was KLUE-BERT vocabulary from which gender-definition words were excluded. This was the imitation of the original paper that used GloVe vocabulary excepting gender-definition words. Note that, unlike GloVe, KLUE-BERT vocabulary includes many subwords and single letters that don't have meaning when they are used on their own.

For the profession word list, we translated the stereotypical profession word list built by [3]. We compared the sizes of our word lists and those used in the original paper [6] in Table 1.

4. Experimental Results

In order to figure out whether the HSR debiasing method is effective on Korean dataset, we replicate the HSR algorithm. When implementing the algorithm, a hyper-parameter, which is Ridge Regression constant α, has to be tuned. In the referenced paper, 60 is chosen for the value. However, we found that the debiasing effects are largely different according to the constant value so it should be tuned depending on the dataset. For example, when we choose 60 as in the previous paper, for our Korean dataset, the bias associated male is enhanced. We fix α = 0.01 by performing many experiments to reduce both male- and female-bias, for example, bias-by-projection of '주부' is reduced from -0.048 to -0.017, and the one of '교수' is also reduced from 0.04 to -0.01.

4.1. Bias-by-projection

The Table 2 shows the average absolute bias-by-projection of the embedding of samples from each dataset, and this refers the gender bias associated with gender direction in Korean language model. The sampling rate is fixed to 2% of the whole data as in the previous paper. For our dataset, top 10 male- and female-biased words are sampled, and top 320 male- and female-words are chosen for KLUE-BERT vocabulary set. As seen in the table, there are gender bias both in our dataset and KLUE-BERT vocabulary set, but the bias in both datasets is significantly well-removed by HSR algorithm.

Table 2. Bias-by-projection of our dataset and KLUE vocabulary

Dataset	Before HSR	After HSR
Ours	0.03	0.0113
KLUE	0.027	0.0117

4.2. Gender-Biased Word Relation Tasks

Clustering. The Figure 1 shows 2D visualization of true labels and clustering results and the Table 3 shows the precision values of clustering results by using K-Meas. K is set to 10 for our dataset and 640 for KLUE-BERT vocabulary set. The precision values decrease after HSR debiasing in both datasets. Especially gender bias in KLUE-BERT vocabulary set is significantly removed by HSR algorithm.

Table 3. Precision values of clustering experiments.

Dataset	Before HSR	After HSR
Ours	1.00	0.95
KLUE	0.8820	0.6117

Correlation. Bias-by-neighbors refers to the percentage of male/female biased words among the k-nearest neighbors of the target word, where k = 20 for our dataset and k = 640 for KLUE-BERT vocabulary set. The higher Pearson correlation coefficient between k-nearest neighbors and bias-by-projection means that larger gender bias exists. It is observed that gender bias in our dataset removed while the coefficient slightly increases for KLUE-BERT vocabulary set. It seems that HSR algorithm is not effective on correlation task for KLUE-BERT vocabulary, but this is because gender-definition words could not be completely extracted from KLUE-BERT vocabulary.

Table 4. Pearson correlation coefficient of correlation experiments.

Dataset	Before HSR	After HSR
Ours	0.6542	0.6011
KLUE	0.6908	0.6948

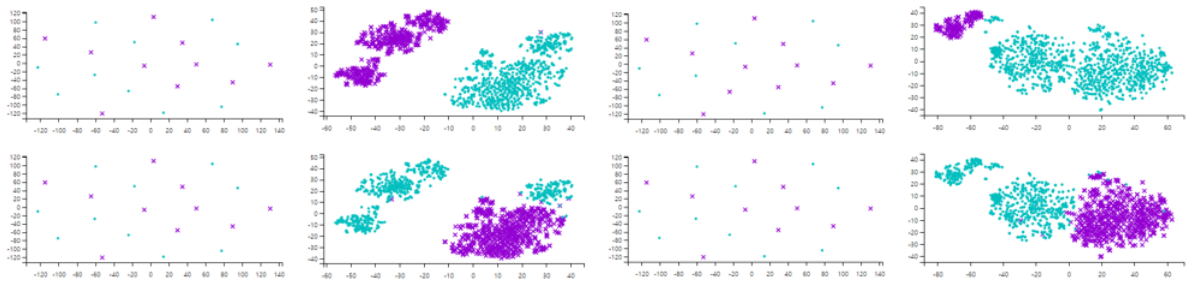
Profession. For this task, we first determined the number of nearest neighbors, K, for this task. When K was in the range of 100 to 200, the correlation values were stable. This study included the experimental results when $K = 100$. As shown in Figure 2, the number of the male biased neighbors and the bias-by-projection values have a positive correlation. Additionally, the experiments using our dataset shows similar results to the experiments using the KLUE-BERT vocabulary list. Next, the experimental results after debiasing through HSR were confirmed.

Table 5. Correlation values of profession experiments.

Dataset	Before HSR	After HSR
Ours	0.705	0.499
KLUE	0.651	0.345

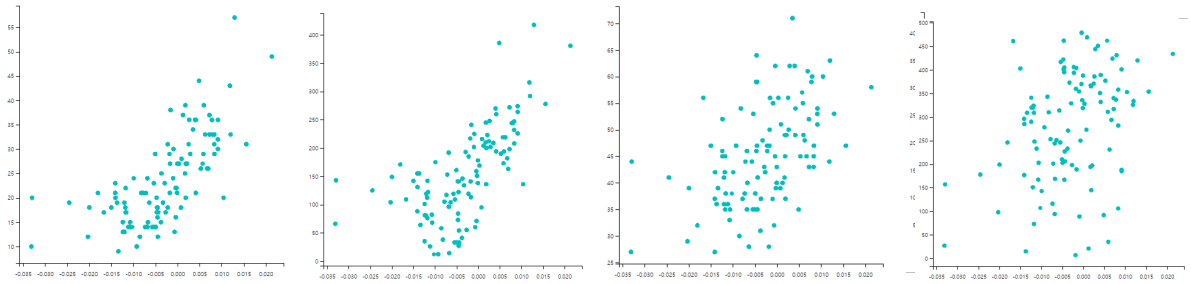
Both the Figure 2 and the table 5 show that the correlation is greatly reduced through HSR.

Association. We used typical male and female names in-



(a) Ours before HSR (b) KLUE before HSR (c) Ours after HSR (d) KLUE after HSR

Figure 1. **Experiment results of clustering task.** These are 2D visualization results of the t-SNE method for each case. Upper figures are the clustering results of K-Means and lower figures show the true labels. Top 10 male/female biased words are used for our dataset and top 640 male/female biased words are used for KLUE-BERT vocabulary set. The number of male-biased neighbors corresponding to the bias-by-projection values.



(a) Ours before HSR (b) KLUE before HSR (c) Ours after HSR (d) KLUE after HSR

Figure 2. **Experiment results of profession task.** The number of male-biased neighbors corresponding to the bias-by-projection values.

cluded in the KLUE-BERT vocabulary list. In addition, words belonging to the arts, mathematics, and science categories are also used in this experiments. The words we used in our experiments were uploaded to our personal website. In association experiments, p-values lower than 0.5 indicate gender bias towards typical male and female names.

Table 6. P-values of association experiments.

Experiments	Before HSR	After HSR
Arts/Mathematics	0.0771	0.1382
Arts/Science	0.0038	0.0275

In this task, we confirmed that female names are close to arts and male names are close to mathematics and science. As shown in Table 6, HSR is effective in reducing the gender bias between male/female names and the group of words.

Classification. We experimented with whether a classifier can learn to predict the direction of gender bias of test data based only on their representations. The experimental results can be checked in Table 7.

Table 7. Test accuracy of classification experiments.

Dataset	Before HSR	After HSR
Ours	0.860	0.805
KLUE	1.000	0.995

Prediction was more difficult in our dataset, and HSR increased the difficulty of prediction.

5. Discussion

5.1. Contributions

We built Korean gender-related word lists including gender-definition, non-gender-definition, and profession word list. We made these word lists publicly available to help the future research on gender bias in Korean NLP systems.

We measured the gender bias in KLUE-BERT embeddings by the two metrics - bias by projection and word vector relation. As a result, we found out that there exists gender bias in KLUE-BERT embeddings, and it can be successfully reduced by HSR.

In particular, we compared the experimental results using

two different non-gender-definition word lists - gender-neutral word list and the KLUE-BERT vocabulary excluding gender-definition words. Despite the small size of the gender-neutral word list, there was no big difference in the results of the two datasets. However, when we used KLUE-BERT vocabulary, HSR was not effective in reducing the correlation between bias-by-neighbors and bias-by-projection. We conjecture that this would be because we could not remove all the gender-definition words in KLUE-BERT vocabulary by excluding the words in our gender-definition word list. Therefore, we conclude that the gender-neutral word list would be more appropriate to be used in the future research.

5.2. Limitations

We found out several limitations of the bias-measuring and mitigating methods of the original paper [6]. In GBWR tasks, the result highly depended on the parameters of sampling rate of words - for example, the sampling number of gender-biased words in Clustering task and the number of nearest neighbors in Correlation task. Also, K-means, used in GBWR clustering task, is simple to implement but sometimes failed to build reasonable clusters from the data. Therefore, we could try other clustering methods such as hierarchical and density-based clustering in future work.

There were also limitations of our own work. In applying HSR, we tuned the Ridge Regression constant α through a few rounds of experiments varying the constant. However, it could also be set as a trainable parameter so that we could learn its optimal value. Also, KLUE-BERT is designed to output sentence embedding, so it yields different embeddings for the same word depending on its surrounding words. Therefore, the word embeddings that we extracted by passing single-word sentences might have missed some meanings of the word.

References

[1] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in neural information processing systems* 29 (2016).

[2] Hila Gonen and Yoav Goldberg. “Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them”. 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)*. 2019, pp. 609–614.

[3] Masahiro Kaneko and Danushka Bollegala. “Gender-preserving Debiasing for Pre-trained Word Embeddings”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1641–1650.

[4] Stuart Lloyd. “Least squares quantization in PCM”. In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[5] Sungjoon Park et al. “KLUE: Korean Language Understanding Evaluation”. In: *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*. Advances in Neural Information Processing Systems. 2021.

[6] Zekun Yang and Juan Feng. “A causal inference method for reducing gender bias in word embedding relations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 9434–9441.

[7] Jieyu Zhao et al. “Learning Gender-Neutral Word Embeddings”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.