Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotype: Quantifying Bias in Korean and Japanese Word Embeddings

Youngin Lee
School of Computing, KAIST
antagonist@kaist.ac.kr

Jaewoo Ji
School of Computing, KAIST
jijaewoo@kaist.ac.kr

Minseon Hwang
School of Computing, KAIST
comafj@kaist.ac.kr

Abstract

Language often reflects social bias such as gender and ethnic stereotypes. The original paper builds a framework to quantify this bias using word embeddings. This framework is proven to be robust for an English corpus; in this paper we appy this framework to Korean and Japanese. We calculate the embedding scores for Korean and Japanese embeddings, then perform ordinary least-square regression to compare the score with metrics such as female participation rate in occupations and human labeled stereotype scores. The embedding scores show a strong positive correlation for most of the trials, and we successfully show that the framework is applicable to languages other than English.

1 Introduction

Gender inequality became a prominent topic in Korea, with hate crimes motivated by gender inequality occurring continuously and women's rights movement gaining momentum as a result. It is clear that our language reflects these kinds of social inequality, and the analysis of this reflection lends key insight into this important issue. The evolution of machine learning and natural language processing can provide these analysis with extremely powerful tools, which were previously often done using manual efforts such as human surveys or qualitative analysis on select samples. Unlike the traditional methods of language analysis machine learning can be scalable, and can account for large amounts of data.

In the original paper, Garg et al. (2018) leverages word embeddings, which represents the semantic relationships of words in the form of vectors, to conduct analysis on the bias in natural language. More specifically, the paper constructs a framework to quantify bias between various groups of people by calculating the difference of distance between 'group words', which are words that are used to represent a group of people, and 'neutral words', which are usually assumed to be independent of either group of people. If the differences between the distances for certain 'neutral words' are significant, we can say that the word embedding is biased for the group. Garg et al. (2018) shows that the framework is indeed robust and reasonable by performing ordinary least square regression for the embedding bias score defined by the framework with real life metrics, such as statistics for participation rates of each occupations and human labeled stereotype scores obtained from various studies.

As the original paper works solely with English data, this paper aims to apply this framework to other languages and establish that the framework is indeed robust and reasonable regardless of the target language. We conduct experiments for Korean and Japanese, which we are familiar with, and confirm that there is bias that is positively correlated with real life metrics in both sets of word embeddings. We also qualitatively confirm that the most biased words are aligned with actual gender stereotypes in society.

2 Methodology

Following the original paper, we first compile the lists of group words, which represented the two target groups of people to analyze, and lists of neutral words, which are lists of occupations and characteristic adjectives. The average of word vectors representing the group words are used as the group vector for each group.

As the original paper, we define embedding bias score for group A as follows, where A and B are group vectors for group A and B respectively, while N is the word vector for a neutral word.

$$Score = \|B - N\| - \|A - N\| \quad (1)$$

This definition yields a score that increases if the word favors group A over B. We perform ordinary least square regression to confirm that this bias score is positively correlated with real life metrics.

3 Dataset and Experiments

We design six trials to perform linear regression. For Korean and Japanese, we look for correlations between the embedding bias score and three different metrics. The three metrics are female participation rates for industries, human labeled stereotype scores for occupations, and human labeled stereotype scores for adjectives describing character traits.

We use contemporary word vectors obtained from Wikipedia dumps (Grave et al., 2018). The vectors were learned using fasttext, which shows superior performance for morphological languages (Bojanowski et al., 2017). We choose two of these languages, namely Korean and Japanese, to study the gender stereotypes in the embedding.

The metrics to validate the embedding bias scores were obtained from various sources. The statistics for female participation rate in occupations were obtained from Statistics Korea (이시균 et al., 2015) for the Korean information, and from Statista (Statista Research Department, 2019) for Japanese information. Unlike the U.S. data used in the original paper, only broad statistics for industries as a whole were available for Korea and Japanese. The human labeled score for occupations and adjectives are the same sources used by the original paper. The scores for occupation were gathered on MTurk by Bolukbasi et al. (2016), while the scores for adjectives are derived from a study from 1990 conducted by Williams and Best (1990).

The word list of group words are obtained by translating the original list of group words for male and female in the original paper. Some of the words had multiple translations, and for these cases all of the valid translations were used to obtain the group vector.

The list of neutral words are derived from various sources. For regression against participation rate, we reference the Korean standard classification of occupation (Statistics of Korea, 2017) for the list of jobs. Since the statistics for participation rates were only available for industries and not for individual occupations, the industries are represented by the average of related jobs found in the standard classification.For regression against human labeled scores, both for occupations and adjectives, we translate the list of jobs and adjectives used in the original study. This produces some caveats in our analysis that we will go into deeper in the discussion.

4 Results

4.1 Quantitative Analysis:

Figure 1 are the results of the ordinary least square regression between the embedding bias score and real life metrics for the six trials. Table 1 shows the p-value and the r^2 for all of the trials.

4.2 Qualitative Analysis

Also, following the original paper, we perform qualitative analysis using the word embeddings and the embedding bias scores. We first attempted to find the closest neutral words to each of the group vectors, but we found that there was no significant difference between the top ten closest words for female and male. Our list of words with highest and lowest bias scores were more informative, however. Table 2 and table 3 shows the Korean occupations and adjectives with the highest and lowest bias scores in order of bias score.

5 Discussion

5.1 Quantitative Analysis

The results show a strong positive correlation between the embedding bias and real life metrics for most trials, as shown in Figure 1. With the exception of participation rate for Korea and human stereotype scores for Japanese occupations, the correlations are statistically significant with $p < 0.01$. For the female participation rate for Korea we still see a positive correlation with $p < 0.1$; however, we found no correlation at all for human stereotype scores for Japanese occupation. We also note that for human stereotype scores for Korean occupations, human stereotype scores for Korean and Japanese adjectives, the intercept of the regression is near the origin. For adjectives the intercept is near the $(0, 0)$ point; for Korean occupations it is near $(2, 0)$ as the neutral score for occupations is 2.

Our analysis bares some limitations, as the lack of fine grained statistics and translation caveats might compromise the robustness of our results to a certain degree.

Lack of Fine Grained Statistics: For the analysis on participation rates, for Korea and Japan the

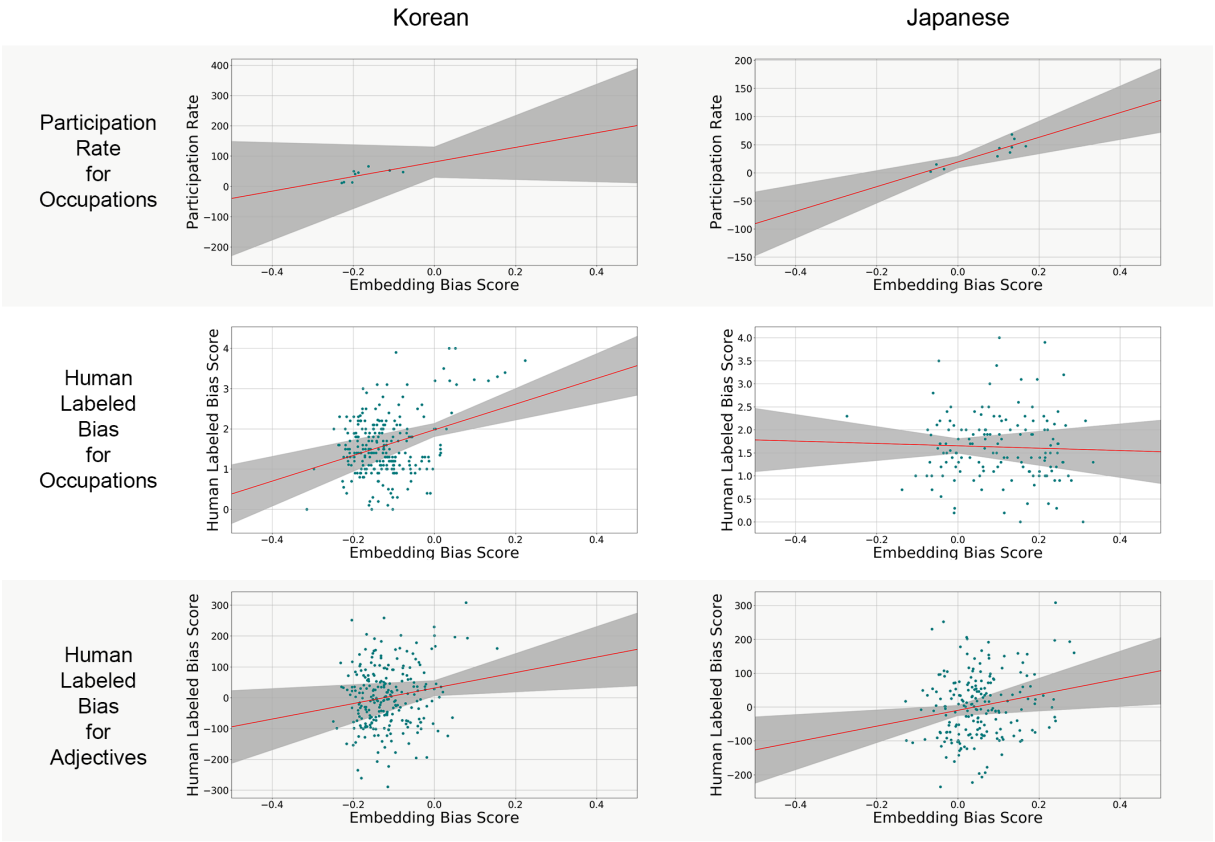

Figure 1: The resulting graphs from the regression. The columns show different languages, while rows show different word lists and metrics. The grey area indicates the 95% confidence interval.

	English(Original)		Korean		Japanese	
Metric	p-value	r^2	p-value	r^2	p-value	r^2
Participation Rate	$< 10^{-10}$	0.499	7.896×10^{-2}	0.3962	6.098×10^{-4}	0.7877
Human(Occupation)	$< 10^{-10}$	0.655	4.051×10^{-8}	0.1186	0.6330	0.0016
Human(Adjectives)	0.0002	0.095	8.559×10^{-3}	-0.0273	5.958×10^{-3}	0.0371

Table 1: p-value and r^2 for all of the trials.

statistics were only available per industry, unlike the original paper which used statistics for individual occupations. Therefore these analysis only have nine data points for Korea and eleven data points for Japan. This might partly explain the weaker correlation for participation rates in Korea.

Translation Caveats: In this analysis we leverage the human labeled stereotype scores used in the original paper, which are originally in English; this posed a few possible problems in our analysis. First, many words were not directly translatable into our target languages. For example, the English word 'businesswoman' translate to '여성 사업자' in Korean, which is not a single word. These words were excluded from the analysis as they had no corresponding word vector. This caveat impacted the analysis on human stereotype scores for Japanese occupation the most heavily, as 48.9% of the original word list was excluded from the analysis as out of vocabulary words. This is much higher than the other trials, of which only 10 20% of the words were excluded.

Second, some of the jobs in the original word list were synonyms with different scores; in order to provide uniformity, we chose the most representative translation for these words as the first word in the NAVER dictionary entry. For the different scores, we took the average of the scores for the synonyms. As the scores did not show a particu-

Highest Bias Score	Lowest Bias Score
수녀	아버지
하녀	지휘관
유모	법학자
창녀	지질학자
비서	부교수
간호사	고문관
주부	은행가
경리	발명가
가정부	군인
여배우	피고용인

Table 2: Jobs with highest and lowest bias score, in order of bias score.

Highest Bias Score	Lowest Bias Score
섹시	적대적
수다적	불성실
여성적	이해
여린	불굴
활기찬	보수적
솔직한	기회주의적
거센	호전적
유치한	열광적
성급한	적응
쾌락주의적	모험적

Table 3: Adjectives with highest and lowest bias score, in order of bias score.

larly high variance the effect of this caveat should be minimal; however, this might indicate a loss of specific nuance in the translation process.

This caveat leads to our third problem, where we want to acknowledge that the translated word list and scores might not correctly reflect the stereotypes of Japan and Korea. The translation poses the problem of lost nuances in the words, as well as the fact that same words might have different connotations in different languages and societies.

Despite these translation caveats our study still produced meaningful results; however, human labeled scored obtained individually for Korean and Japanese might increase the accuracy of the analysis.

5.2 Qualitative Analysis

We attempted to find the closest neutral words to each of the group vectors based on pure distance; however, the results for both group vectors were very similar. On the other hand comparing bias score on both occupation and adjective words shows a more insightful result. The cause of the failures for pure distances might be that some neutral word vectors were much closer to both group vectors than others. In the hopes of confirming this we used PCA to reduce the vectors to 50 dimensions, then used t-SNE to visualized the vectors to a 2-D space. The results are shown in Figure 2. For occupations, we can see that the group vectors are close with each other, therefore sharing the top list of closest neutral words. Although this visualization did not yield the same confirmation for adjectives, it is still possible that the hypothesis is true in higher dimensions.

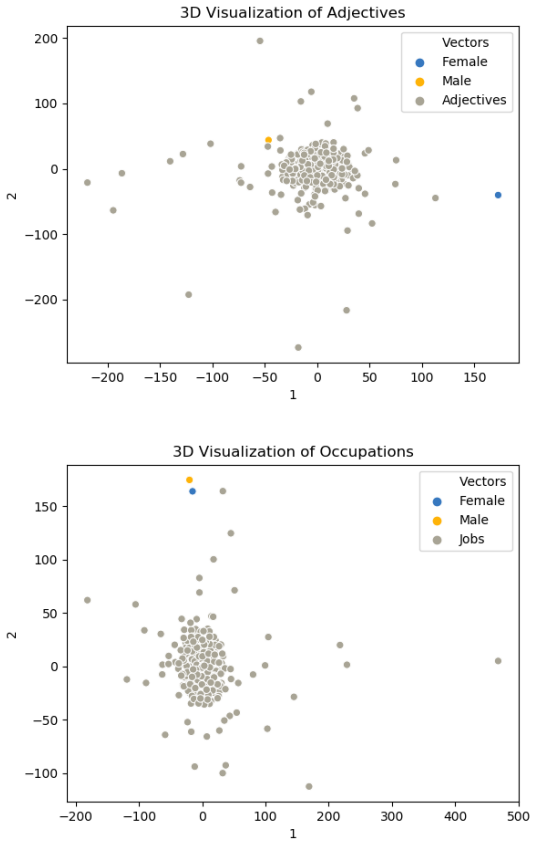

Figure 2: Word vectors visualized on a 2D space using PCA and t-SNE.

6 Conclusion

We successfully confirmed the existence gender bias in Korean and Japanese word vectors by validating the embedding bias score defined in the original paper against real life metrics. Through quantitative analysis using ordinary least square regression, we discovered a significant positive cor-

relation between embedding bias and participation rate, human labeled bias scores for occupations and adjectives. We also qualitatively confirmed that the gender bias score aligns with the general gender stereotypes. However, lack of fine-grained statistics for participation rate and translation caveats might affect robustness of the results. In further research, obtaining human labeled bias scores in Korean and Japanese might yield more accurate results.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Statista Research Department. 2019. Percentage of women employed in japan in 2018, by occupation. https://www.statista.com/statistics/644592/japan-female-employees-share-by-occupation/. Accessed: 2019-12-01.

Statistics of Korea. 2017. Korean standard classification of occupations. http://kssc.kostat.go.kr/ksscNew_web/ekssc/main/main.do. Accessed: 2019-12-01.

John E Williams and Deborah L Best. 1990. *Measuring sex stereotypes: A multination study, Rev*. Sage Publications, Inc.

이시균, 박진희, 성지미, and 김종숙. 2015. *여성 직업별 인력수요 전망*, page 33. 진한M&B.