Evaluation of Gender Bias in Japanese Word Embedding Model

Shintaro Sakai Yasuhiro Suzuki

Nagoya University

With substantial progress in Natural Language Processing(NLP), word embedding made it possible to represent semantic relationships in a low-dimensional vector space. Although word embeddings are widely used in various applications, they have been reported to inherit and amplify social biases such as gender and racial biases. While a growing number of studies have been conducted to analyze biases in word embeddings, research examining bias in Japanese word embedding model is scarce. The present research aims to fill this gap by empirically evaluating gender bias in occupation with Japanese word embedding trained on Wikipedia articles. Our experiments show some of the occupations such as nurse and dental hygienist show strong women biases. We also found these strength of biases are highly correlated with percentages of women in each occupation in Japan($\rho = 0.78$ with $p < .001$). Our research shows Japanese word embedding model also inherits gender bias in occupations and they accurately reflect real-world gender disparities in occupations.

1. Introduction

With substantial progress in NLP, prediction-based word embeddings made it possible to represent semantic relationships in a dense and low-dimensional space [Mikolov 13a, Pennington 14, Joulin 17]. In word embedding, each word in the corpus is represented as a unique vector. Words are assigned a position in a vector space based on their distributions of co-occurring words in the corpus. More specifically, words sharing similar contexts are positioned close together in a vector space, while words sharing different contexts are located farther apart. Thus, words that are semantically close tend to have similar vectors. Calculating the distance of the positions between words provides a measurement of how similar the words are. Distance between words is approximated by the cosine similarity between the vectors. The cosine similarity computes the similarity between words by measuring the angle between vectors and it ranges from -1 (= opposite meaning) to 1 (= same meaning). In addition to distance, the direction of vectors is a useful property in word embeddings. The famous example is the analogy of *man* is to *woman* as *king* is to *queen*. Word embedding is able to approximate $\vec{queen}$ by calculating $\vec{king}$ - $\vec{man}$ + $\vec{woman}$.

Although word embeddings are widely used in various applications, they have been reported to inherit and amplify social bias such as gender and racial bias [Bolukbasi 16, Caliskan 17]. Due to their widespread and sensitive applications [Bolukbasi 16, Zhao 18, Wang 20], it is critical to quantify bias in word embeddings. However, to my best knowledge, research examining bias in Japanese word embeddings is scarce. Thus, the present research aims to quantitatively uncover social bias that Japanese word embedding models incorporate. While there are several types of bias to be examined, this study focuses on gender biases in occupations. It is important to assess whether findings in English word embeddings are applicable to other languages so that we can expand and refine our understanding. Although effective mitigation should be tailored to downstream applications, this study also demonstrates word embedding itself is a useful means to understand cultural associations [Caliskan 17, Kozlowski 19]. To verify that the embedding bias reflects real-world occupational gender associations, it also assessed whether gender biases in word embedding correlates with the percentage of women in occupations. While it is possible to examine cultural associations with interviews or surveys, these method typically take cost and time and the amount of data one can collect is limited.

Contact: Shintaro Sakai, Graduate School of Informatics, Nagoya University, sakai.shintaro.h1@s.mail.nagoya-u.ac.jp

2. Related Work

There are several studies that examine occupational gender biases in word embeddings.

The pioneering work by [Bolukbasi 16] observed the word embedding models trained on Google News articles exhibit occupational gender bias. They conducted an online survey to judge whether bias in word embedding reflect actual stereotypes by humans. They found both stereotypes are strongly correlated with each other. Caliskan and colleagues proposed the Word Embedding Association Test (WEAT) to evaluate the social biases in embedding. They found female names were more associated with family words while male names were with career words. Gender biases of word embeddings measured by WEAT are highly correlated with the results from Implicit Attitude Tests [Caliskan 17]. By using word embeddings trained with Google News, Garg and colleagues found that occupations such as nurse, librarian, and housekeeper are more associated with women while carpenter, mechanic, and engineer are more associated with men. These gender biases are highly correlated with empirical percentages of women in each occupation and people ' s stereotypes measured on an online survey [Garg 18]. With word embedding models trained with Google Ngram corpus, another research by [Kozlowski 19] found *nurse* is more associated with female while *engineer* and *journalist* are more associated with male. However, by evaluating associations for decades, they also concluded that gender biases become smaller over time. Like other studies, gen-

der biases in word embedding are strongly correlated with people's stereotypes measured on an online survey. There is one study examining biases in Japanese word embedding models [Takeshita 20]. Unlike a similar analysis conducted in English word embeddings [Swinger 19], the study found Japanese first names do not reflect social stereotypes.

3. Methodology

In this section, I briefly describe the data and approach used to capture gender bias in word embeddings.

3.1 Data

I utilize a publicly available pre-trained word embedding model trained on all Japanese Wikipedia articles [Suzuki 18]. I use pre-trained model as it is possibly used for real applications. The word embedding was trained using skip-gram word2vec algorithm proposed by [Mikolov 13b, Mikolov 13c]. I pick the latest of the available models trained in 2019. It has a 200 dimensions and vocabulary size of 75,1361 words. I chose the model trained on Wikipedia articles because of Wikipedia's scale and information richness. It is also a good benchmark for future studies. Pretrained embeddings are available at this page so that one can reproduce my research[*1].

3.2 Quantification of Gender Bias

My approach builds on established methodology by [Kozlowski 19, Bolukbasi 16, Garg 18]. First, I compute per-gender vectors (female and male vectors) by averaging the vectors of words representing female or male. Female and male word sets described later are used to construct per-gender vectors. I make per-gender vectors more robust by averaging vectors of multiple gender-representing words. Then, I construct a gender vector (namely gender dimension) $\vec{g}$ by subtracting the averaged male vector from the averaged female vector. In this case, $\vec{g}$ points to female while $-\vec{g}$ points to male. After that, one can measure the bias of occupations by computing how this gender dimension relates to occupation terms. As already noted, the relation between an occupation term and the gender dimension is approximated by cosine similarity. The cosine similarity calculates both direction (e.g., whether 看護師（nurse） has male or female bias), and magnitude of bias. For example, if the cosine similarity between the vector of 看護師（nurse） and $\vec{g}$ is negative, it suggests the embedding more closely associates 看護師（nurse）with women.

3.3 Occupation and Gender Terms

The occupations I am examining are collected from the National census（国勢調査） conducted in 2020[*2]. Only the occupation terms that are available in the embedding model were chosen. Occupation terms such as 家政婦（housekeeper） were excluded from the analysis as the orthography is already gendered. The final list contains 48 occupation terms. Please note that some of the occupations are grouped together for the analysis(e.g., 裁判官（Judge）, 検察官（Prosecutor）, 弁護士（Lawyer） were grouped together) as gender participation statistic was not available for each occupation. The gender bias of a group is averaged gender biases of occupations in the group.

For terms representing gender, I collect a set of antonym word pairs denoting femininity and masculinity. I referred to the lists of English gender words from [Caliskan 17, Garg 18, Kozlowski 19, Arseniev 22] and translated them into Japanese. To increase the robustness of gender dimension, multiple word pairs are chosen. Words that do not exist in the embedding model were excluded from the list. Additional words which are specific to Japanese are added to the list. For example, while the English word *sister* refers to both older and younger sisters, Japanese has two different words for each expression such as 姉（older sister）and 妹（younger sister）.

Table 1 shows full lists of gender-representing terms.

男性を表す単語（Words representing male）	女性を表す単語（Words representing female）
男, 彼, 男の子, 男子, 男性, 少年, 男らしい, 父, 父親, お父さん, 父さん, パパ, 息子, 祖父, おじいちゃん, 夫, 兄, 弟, お兄ちゃん, 義父, 甥, 伯父, 叔父	女, 彼女, 女の子, 女子, 女性, 少女, 女らしい, 母, 母親, お母さん, 母さん, ママ, 娘, 祖母, おばあちゃん, 妻, 姉, 妹, お姉ちゃん, 義母, 姪, 伯母, 叔母

Table 1: The lists of words representing each. gender. The averaged vectors are used to construct per-gender vectors.

4. Result

4.1 Occupational Gender Bias

The table 2 and 3 present the top 5 occupations that are biased toward men and women. Female biases appeared in occupations that are traditionally dominated by women such as 看護師（nurse）, 保育士（nursery teacher）, and 助産師（midwife）. It is also understandable that 美容師（hairdresser）are more female-biased than 理容師（barber）. In contrast, although table 3 itself is convincing, some of the occupations that are traditionally dominated by men were not listed amongst top 5 such as 自衛官（Self-Defense Official）, 警察官（police officer） and 海上保安官（coast guard officer）. In fact, 警察官（police Officer） and 海上保安官（coast guard officer）exhibit weak female bias (Female bias = 0.04). Interestingly, when comparing the strength of biases of table 2 with table 3, absolute values are much larger for female-biased occupations. Another interesting finding is, while 幼稚園教員（kindergarten teacher）exhibit strong female bias (Female bias = 0.20), 小学校教員（elementary school teacher）, 中学校教員（middle school teacher） and 高等学校教員（high school teacher） showed weak gender biases(Female bias = 0.05, -0.03, -0.01).

4.2 Validation

To evaluate the quality of gender biases found in the embedding, I examined the correlation between embedding bias and the percentage of women in each occupation. It has been suggested that implicit gender-occupation biases are linked to gender gaps in occupational participation

∗1 https://github.com/singletongue/WikiEntVec/releases
∗2 https://www.e-stat.go.jp/dbview?sid=0003464389

	Occupation	Female bias
1	看護師（nurse）	0.32
2	歯科衛生士（dental hygienist）	0.30
3	保育士（nursery teacher）	0.30
4	美容師（hairdresser）	0.27
5	助産師（midwife）	0.24

Table 2: Top 5 occupations associated with women

	Occupation	Female Bias
1	技術者（engineer）	-0.14
2	著述家（writer）	-0.08
3	農業従事者（farmer）	-0.07
4	船長（captain），航海士（navigator），機関長（chief engineer），機関士（engineer）	-0.07
5	大工（carpenter）	-0.06

Table 3: Top 5 occupations associated with men

[Nosek 09]. I use National census data in 2020 [*2] to extract the percentage of women in each occupation. Figure 1 plots the percentage of women in each occupation against the strength of female bias for each occupation term in the embedding model. I found the strength of female biases is strongly correlated with the percentage of women in each occupation (Pearson's correlation coefficient of $\rho = 0.78$ with $p < .001$). Overall, the result suggests the correlation of women biases in embedding model with gender participation statistics accurately reflect real-world gender disparities in occupations.

5. Discussion and Conclusion

The present research aims to quantify occupational gender bias in the word embedding trained on Japanese Wikipedia articles. I demonstrated some of the occupations are highly associated with women and the strength of female biases in the embedding model are strongly correlated with the actual gender participation statistics. The result is consistent with prior work conducted with the English word embeddings by [Caliskan 17, Garg 18, Friedman 19]. To my best knowledge, this is the first study examining occupational gender bias in the Japanese word embedding.

Although the majority of the results are consistent with prior work conducted in English, it is interesting that there are more female biased occupations than male biased occupations. 30 out of 48 were female biased occupations. Furthermore, the strengths of female bias are larger than that of male bias. Is this an artifact of the algorithm or training data? Here, I discuss two potential reasons. First, the accuracy of the results depends on the occupation terms used in the analysis. Although I collected the occupation terms from the National census data, some of the terms are uncommon or vague. For example, 著述家 (writer) is not usually used on a daily basis and 調査員 (investigator) is

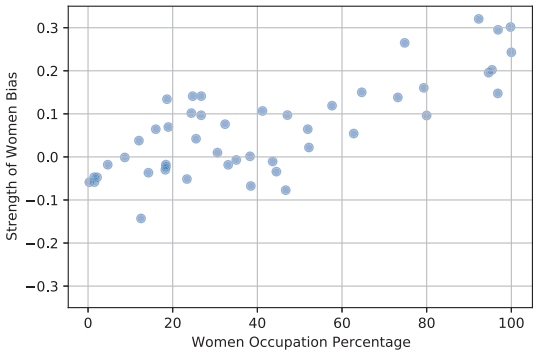

Figure 1: Correlation of percentage of women in each occupation against the strength of female bias for each occupation. (Pearson's correlation coefficient of $\rho = 0.78$ with $p < .001$)

vague as there are several types of investigator. Another concern is some of the terms have several ways to call such as abbreviations and slang. For example, 警察官（police officer）can be called 警察 and 警官 for abbreviations and 刑事 and お巡りさん for synonyms. While 警察官（police officer）does not show strong male bias, other terms such as 警察, 警官, 刑事, and お巡りさん might exhibit strong male bias. Future research can aim for more fine-grained analysis by improving the quality of the word lists. Another potential reason is the fact that Wikipedia's biographical articles about women tend to contain words denoting female [Wagner 15]. For example, it can be written as *female engineer* instead of *engineer* to emphasize the fact that an engineer is female. It leads the model to associate engineer more closely with women. While the study by [Wagner 15] did not examine Japanese Wikipedia, male bias might become weaker in some of the occupations because femininity is emphasized in Japanese Wikipedia. Similar to our finding, Schmahl and colleagues found *science* related words are more associated with women while 15% of the scientists with biographies are female [Schmahl 20].

References

[Arseniev 22] Arseniev-Koehler, A., and Foster, J. G.: Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. Sociological Methods Research, 51(4), pp. 1484-1539(2022)

[Bolukbasi 16] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., and Kalai, A. T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29. (2016)

[Caliskan 17] Caliskan, A., Bryson, J. J., and Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), pp 183-186(2017)

[Friedman 19] Friedman, S., Schmer-Galunder, S., Chen, A., and Rye, J.: Relating Word Embedding Gender Biases to Gender Gaps: A Cross-Cultural Analysis. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pp. 18–24(2019)

[Garg 18] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences, 115(16), pp. E3635-E3644(2018)

[Joulin 17] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T.: Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 427–431(2017)

[Kozlowski 19] Kozlowski, A. C., Taddy, M., and Evans, J. A.: The geometry of culture: Analyzing the meanings of class through word embeddings. American Sociological Review, 84(5), pp. 905-949(2019)

[Mikolov 13a] Mikolov, T., Yih, W. T., and Zweig, G.: Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, pp. 746-751(2013)

[Mikolov 13b] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space. In Proceedings of the 2013 International Conference on Learning Representations, pp 1–12(2013)

[Mikolov 13c] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, pp. 3111–3119(2013).

[Nosek 09] Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., et al.: National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences, 106(26), pp. 10593-10597(2009).

[Pennington 14] Pennington, J., Socher, R., and Manning, C. D.: Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543(2014)

[Schmahl 20] Schmahl, K. G., Viering, T. J., Makrodimitris, S., Jahfari, A. N., Tax, D., and Loog, M.: Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. In Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pp. 94-103(2020)

[Suzuki 18] Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., and Inui, K.: A joint neural model for fine-grained named entity classification of wikipedia articles. IEICE Transactions on Information and Systems, 101(1), pp. 73-81(2018)

[Swinger 19] Swinger, N., De-Arteaga, M., Heffernan, N. T., Leiserson, M. D., and Tauman Kalai, A.: What are the biases in my word embedding? In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES ’19, pp. 305–311(2019)

[Takeshita 20] Takeshita, M., Katsumata, Y., Rzepka, R., and Araki, K.: Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, pp. 44-55(2020)

[Wagner 15] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M.: It’s a man’s Wikipedia? Assessing gender inequality in an online encyclopedia. In Proceedings of the international AAAI conference on web and social media, pp. 454-463(2015).

[Wang 20] Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., and Xiong, C.: Double-hard debias: tailoring word embeddings for gender bias mitigation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5443–5453(2020)

[Zhao 18] Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K. W.: Learning gender-neutral word embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4847–4853(2018)