9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Evaluating the Quality of Word Embedding Trained on Wikipedia Articles

Keywords: word embedding, social stereotype, occupational gender bias, Wikipedia, social science research

Extended Abstract

Introduction and Problem Statement: It is becoming increasingly common to use word embeddings in social science research. The prior studies demonstrate word embeddings can capture cultural stereotypes [2] and occupational gender disparities [1]. In this paper, we show while word embeddings can be a valuable tool, word embeddings trained on Wikipedia articles should be used with caution.

Wagner and colleagues find Wikipedia's biographical articles about women tend to contain words denoting females while articles about men do not contain words denoting males [4]. For example, it can be written as *female engineer* instead of just *engineer* to emphasize the fact that the engineer is female. Even if people usually associate engineers with men, it may lead the model to associate engineers more closely with women. With women's occupation participation statistics, the present research aims to evaluate the quality of occupational gender bias in the word embedding trained on Japanese Wikipedia articles.

Data and Methodology: We utilize two publicly available word2vec embedding [1][2]. The first model is trained on all Japanese Wikipedia articles which we refer to as Wikipedia embedding. It has 200 dimensions and a vocabulary size of 75,1361 words. For comparison, we also use the model called hottoSNS-w2v which is trained on the Japanese texts from blogs, Twitter, Wikipedia, and web page. It has 200 dimensions and a vocabulary size of 2,067,629 words. Although hottoSNS-w2v is also trained on Wikipedia articles, the size of Wikipedia data used for the training is much smaller than those of other domains of data.

Our approach to quantify gender bias in word embeddings builds on established methodology [1, 2]. We compute per-gender vectors by averaging the vectors of words representing female or male. We make per-gender vectors more robust by averaging vectors of multiple gender-representing words. We collected gender-representing words by referring to the lists of English gender words from [1, 2] and translated them into Japanese. Additional words which are specific to Japanese are added to the list. Then, we construct a gender vector $\vec{g}$ 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 gender bias of occupations by computing how this gender vector relates to occupation terms. The relation between a vector of an occupation term and the gender vector is approximated by cosine similarity. The cosine similarity calculates both direction (whether it has male or female bias), and strength of bias.

The occupations we are examining and the women's occupation participation statistics are collected from the National census conducted by the Japanese government[3]. Please note that

[1]https://github.com/singletongue/WikiEntVec/releases
[2]https://github.com/hottolink/hottoSNS-w2v
[3]https://www.stat.go.jp/english/data/index.html

some of the occupations are grouped in the analysis as women's participation statistics were not available for each occupation but for a group. The gender bias of a group is calculated as the averaged gender bias of occupations in the group.

Result: The table 1 to 4 present the top 5 occupations that are dominated by women or men and their strengths of female bias in each embedding. Interestingly, while both female and male-dominated occupations exhibit strong gender bias in hottoSNS-w2v (Table 3 and 4), the average strengths of gender bias are largely different in Wikipedia embedding (Table 1 and 2).

To evaluate the overall results, we examined the correlation between the strengths of female embedding bias and the percentage of women in each occupation in Japan (Figure 1 and 2). While both of the embedding show a strong correlation, the correlation was relatively weaker in the Wikipedia embedding (Pearson's correlation coefficient of $\rho = 0.78$ with $p < .001$ for Wikipedia embedding; $\rho = 0.92$ with $p < .001$ for hottoSNS-w2v).

Discussion: We demonstrate that Wikipedia embedding exhibits much weaker male bias in occupations that are dominated by men. We also find the correlation between the strengths of female bias and the percentages of women in occupations is relatively weaker in Wikipedia embedding. These results provide evidence that bias in the word embedding trained on Wikipedia articles do not necessarily reflect accurate occupational gender disparities and people's stereotypes. The potential explanation is that femininity is emphasized in male-dominated occupations in Wikipedia. As editors of Wikipedia are mostly male, male is the standard gender in Wikipedia [4]. This explains the reason why biographical articles about men do not contain male-representing words. Similar to our findings, the prior study shows *science* related words are more associated with women while 15% of the scientists with biographies are female in Wikipedia [3]. The implication of our research is that researchers need to pay more attention to social contexts (e.g., gender) of training corpora when using word embeddings in social science research.

The accuracy of the results depends on the occupation terms used in the analysis. One potential problem is some of the occupations have abbreviations and slang. Future research can aim for more fine-grained analysis by improving the quality of the occupation terms.

References

[1] Nikhil Garg et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (2018), E3635–E3644.

[2] Austin C Kozlowski, Matt Taddy, and James A Evans. "The geometry of culture: Analyzing the meanings of class through word embeddings". In: *American Sociological Review* 84.5 (2019), pp. 905–949.

[3] Katja Geertruida Schmahl et al. "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*. 2020, pp. 94–103.

[4] Claudia Wagner et al. "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*. Vol. 9. 1. 2015, pp. 454–463.

Embedding		Most female dominated occupation	Percentage of women	Female bias
Wikipedia embedding	1	助産師(midwife)	100	0.24
	2	歯科衛生士(dental hygienist)	99.8	0.30
	3	保育士(nursery teacher)	96.9	0.30
	4	保健師(public health nurse)	96.8	0.15
	5	栄養士(nutritionist)	95.4	0.20
				Average: 0.24

Table 1: Top 5 female dominated occupations and their strengths of female bias in Wikipedia embedding

Embedding		Most male dominated occupation	Percentage of women	Female bias
Wikipedia embedding	1	船長, 航海士, 機関長, 機関士(captain, navigator, chief engineer, engineer [of a ship])	0.3	-0.07
	2	とび職(scaffold worker)	1.5	-0.05
	3	大工(carpenter)	1.5	-0.06
	4	左官(plasterer)	2.1	-0.05
	5	警備員(security guard)	4.7	-0.02
				Average: -0.05

Table 2: Top 5 male dominated occupations and their strengths of female bias in Wikipedia embedding

Embedding		Most female dominated occupation	Percentage of women	Female bias
hottoSNS-w2v	1	助産師(midwife)	100	0.24
	2	歯科衛生士(dental hygienist)	99.8	0.16
	3	保育士(nursery teacher)	96.9	0.22
	4	保健師(public health nurse)	96.8	0.19
	5	栄養士(nutritionist)	95.4	0.20
				Average: 0.20

Table 3: Top 5 female dominated occupations and their strengths of female bias in hottoSNS-w2v

Embedding		Most male dominated occupation	Percentage of women	Female bias
hottoSNS-w2v	1	船長, 航海士, 機関長, 機関士(captain, navigator, chief engineer, engineer [of a ship])	0.3	-0.18
	2	とび職(scaffold worker)	1.5	-0.25
	3	大工(carpenter)	1.5	-0.26
	4	左官(plasterer)	2.1	-0.22
	5	警備員(security guard)	4.7	-0.12
				Average: -0.21

Table 4: Top 5 male dominated occupations and their strengths of female bias in hottoSNS-w2v

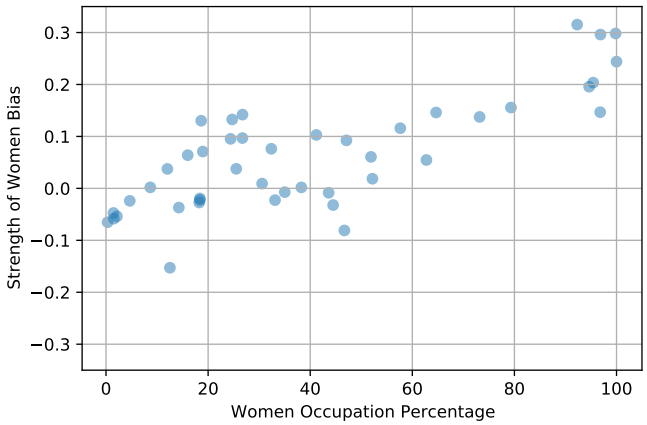

Figure 1: Correlation of percentage of women in each occupation against the strength of female bias for each occupation in Wikipedia embedding. (Pearson's correlation coefficient of $\rho = 0.79$ with $p < .001$)

Figure 2: Correlation of percentage of women in each occupation against the strength of female bias for each occupation in hottoSNS-w2v. (Pearson's correlation coefficient of $\rho = 0.92$ with $p < .001$)