Can Existing Methods Debias Languages Other than English? First Attempt to Analyze and Mitigate Japanese Word Embeddings

Masashi Takeshita, **Yuki Katsumata**, **Rafal Rzepka**, and **Kenji Araki**

Graduate School of Information Science and Technology
Hokkaido University, Sapporo, Japan
{takeshita.masashi, katsumata, rzepka, araki}@ist.hokudai.ac.jp

Abstract

It is known that word embeddings exhibit biases inherited from the corpus, and those biases reflect social stereotypes. Recently, many studies have been conducted to analyze and mitigate biases in word embeddings. Unsupervised Bias Enumeration (UBE) (Swinger et al., 2019) is one of approach to analyze biases for English, and Hard Debias (Bolukbasi et al., 2016) is the common technique to mitigate gender bias. These methods focused on English, or, in smaller extent, on Indo-European languages. However, it is not clear whether these methods can be generalized to other languages. In this paper, we apply these analyzing and mitigating methods, UBE and Hard Debias, to Japanese word embeddings. Additionally, we examine whether these methods can be used for Japanese. We experimentally show that UBE and Hard Debias cannot be sufficiently adapted to Japanese embeddings.

1 Introduction

Word embeddings are widely used in natural language processing tasks, and they have been reported to inherit social stereotypes, e.g. gender and racial stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017). For example, "programmer" and "homemaker" should be gender neutral by definition, but the analogy of "man is to programmer as woman is to homemaker" holds as observed by Bolukbasi et al. (2016). Such biases cause differences in F1 scores between the pro- and anti-stereotypical conditions. For example in the coreference resolution task, it is difficult to correctly link "physician:she" and "secretary:he" for systems which use gender-biased word embeddings, because "physician:he" and "secretary:she" are strongly related more than "physician:she" and "secretary:he" in the word embeddings (Zhao et al., 2018a). Therefore, in recent years, research has been conducted to mitigate the bias in word embeddings (Bolukbasi et al., 2016; Zhao et al., 2018b; Wang et al., 2020). However, to the authors' best knowledge, most of them have focused on English (Sun et al., 2019; Blodgett et al., 2020), and no study has addressed word embeddings of languages other than Indo-European languages about bias analysis and mitigation.

We hypothesize that it is not obvious that the method developed for English can be easily adapted to other languages for two following reasons. First is due to various grammatical features which do not exist in English. Embeddings can have different characteristics depending on language, for example Spanish words have gender which leads to the grammatical gender bias (Zhou et al., 2019). There is a substantial risk that we cannot adapt the bias mitigation methods meant for English while working on such a language. Secondly, especially when the language family differs, not only the characteristics of a given language but also the cultural background of its users changes, which in turn influences further the bias in the embeddings (Raijmakers, 2020). Therefore, it may not be possible to directly apply bias analysis and mitigation methods developed for English to other languages.

Bias statement Following categorization of Crawford (2017), we focus on representational bias, especially stereotyping one, which means that a system "propagates negative generalisations about particular

social groups" (Blodgett et al., 2020, p.5456). Stereotyping happens in natural language processing tasks when an unfair association of words represents a particular social group with other concepts (not included in its definition), like an analogy of "man is to programmer as woman is to homemaker". If an AI agent has such stereotypes, they can appear in its output as reported in works on dialogue systems (Liu et al., 2019), possibly harming users.

There are several works on stereotypes in word embeddings for English (Bolukbasi et al., 2016; Zhao et al., 2018b; Wang et al., 2020) and some other languages (Sahlgren and Olsson, 2019; Pujari et al., 2019), but to the authors' best knowledge, research regarding Japanese word embeddings does not exist.

In this paper, we analyze the representational bias in Japanese word embeddings, and attempt to mitigate gender bias by using existing methods designed for English. We also show that those methods are difficult to generalize to Japanese.

2 Related work

2.1 Bias in word embeddings and its mitigation for English

This section describes bias analysis and gender bias mitigation for English word embeddings.

2.1.1 Bias analysis

Caliskan et al. (2017) proposed the Word Embedding Association Test (WEAT) to evaluate the inherent social biases in embedding. WEAT measures the difference of semantic similarity with a word embedding between two sets of target words (e.g. "male" and "female" names) and attribute words (e.g. "career" and "family" terms). This metric was used to show that social biases of embeddings are correlated with social stereotypes and the proportion of gender of workers in each occupation.

Swinger et al. (2019) adapted WEAT and proposed Unsupervised Bias Enumeration (UBE) to discover the biases in embedding by unsupervised clustering using first names. They asked crowdworkers to evaluate the results of WEATs which are outputted by UBE and confirm if these results capture social stereotypes, such as gender as well as religion and race.

2.1.2 Approaches to bias mitigation

Bolukbasi et al. (2016) confirmed the existence of the gender bias in English word embeddings, and proposed a method called Hard Debias to mitigate the gender bias. Hard Debias uses words that should be neutral to gender, such as "doctor" and "programmer", and reduces the bias by subtracting the vector components of gender directions from gender neutral words. Gender directions are defined by the first principal component of a word vector of each word consisting of a gender definition word pairs, such as "she" and "he".

However, Gonen and Goldberg (2019) proved experimentally that Hard Debias could not sufficiently remove gender bias and that it can be recovered from embeddings after mitigation.

In the work of Mu and Viswanath (2018), the most statistically dominant principal components are encoding the frequency of words. Their method improves performance of embedding by subtracting the common mean vector from each word vector and removing the dominant principal components. Wang et al. (2020) proposed Double-Hard Debias which was inspired by work of Mu and Viswanath (2018). They improved Hard Debias by deciding the dominant principal component of gender bias before performing Hard Debias. Experiments on English embeddings, including the neighborhood metric (Gonen and Goldberg, 2019), showed improved results.

All of the above-mentioned research examples work on English language. Next, we present studies on the bias inherent in non-English embeddings.

2.2 Word embedding biases in languages other than English

There are two major directions of research on non-English word embedding bias. The first is a bias study of multilingual embeddings, which compares what biases exist in embeddings available in both English and other languages, e.g. Spanish and French, and how they differ depending on language (Zhou et al., 2019; Zhao et al., 2020). The second direction is to address biases in monolingual embeddings of languages other than English (Zhou et al., 2019; Sahlgren and Olsson, 2019; Pujari et al., 2019;

Raijmakers, 2020). For example, the gender bias has been found and mitigated in Swedish (Sahlgren and Olsson, 2019) and Hindi (Pujari et al., 2019). Both used Hard Debias for gender bias mitigation – Sahlgren and Olsson (2019) could not mitigate the gender bias but Pujari et al. (2019) were able to achieve that goal. However, (Sahlgren and Olsson, 2019) analyzed their results only partially. The problem in the Pujari et al. (2019) method is that they used Support Vector Machine (SVM) trained on gender-biased embeddings during Hard Debias evaluation for Hindi. Raijmakers (2020) proposed a WEAT-extended method to investigate gender bias in monolingual embeddings of 26 languages, including Japanese, but did not attempt to mitigate any of them. This work also lacks a detailed analysis, as it only investigates the overall gender bias of embeddings and does not assess whether gender neutral words have gender bias.

In this paper, we examine biases in Japanese monolingual embeddings and attempt to mitigate gender bias as a case study.

3 Specificity of Japanese language

Japanese and Western languages use different types of characters. There are three types of characters in Japanese language: phonetic *hiragana*[*], *katakana*, and ideographic *kanji*. Embeddings of *kanji* may capture not only the meaning of the word but also the meaning of the characters. For example word 数学 ("maths") consists of two ideograms: 数 ("number") and 学 ("learning"). *Katakana* often represents a foreign word プログラマ ("programmer"), while words written in rounded shape of *hiragana* like ふわふわ (*fuwafuwa*, "fluffy") are often associated with a feminine image (Iwahara et al., 2003).

4 Experiments

In this section we explain word embeddings we used, describe UBE (Swinger et al., 2019) used in the bias analysis experiment, and two other methods (Hard Debias (Bolukbasi et al., 2016), Double-Hard Debias (Wang et al., 2020)) used in the bias mitigation experiment. Finally, we explain our evaluation methodology.

4.1 Word embeddings for experiments

As the target of our analysis we use two publicly available embeddings: word2vec (Mikolov et al., 2013) trained on Japanese Wikipedia (Suzuki et al., 2018)[†] and fastText (Bojanowski et al., 2016) trained on the Wikipedia text and Common Crawl. Number of dimensions in these embeddings is 200 and 300, respectively.

We use 50,000 most frequent words (Bolukbasi et al., 2016) and also limit the words to be assessed for bias to nouns, verbs, adjectives, adjectival verbs and adverbs in their dictionary forms using morphological analyzer Juman++ (Morita et al., 2015).

4.2 Bias analysis experiment

Unsupervised Bias Enumeration (UBE) In this subsection, we introduce procedural steps of UBE which is a method to detect various biases in embeddings using names.

As the first step, we filter out possibly problematic first names. In many languages there are polysemous first names such as "May" in English (name of a month). Also in Japanese first names that have other meaning, such as *Hoshi* (star), can be found. We filter them out because of the ambiguity they tend to bring. Identically to Caliskan et al. (2017), we remove 20% of names with the lowest mean of cosine similarity between a name and all other names. Then, after filtering, the names are clustered with k-means++ (Arthur and Vassilvitskii, 2006) included in scikit-learn library (Pedregosa et al., 2011). Female and male first names data is borrowed from JMnedict[‡]. Names being used for both genders are treated as neutral. The number of clusters was experimentally set to 10 in both embeddings (word2vec and fastText). JMnedict also includes foreign surnames. Initially, we were going to exclude them, but we thought that we might be able to find social stereotypes regarding foreigners, so we eventually included their names in the dataset. The results of the filtering are shown in Table 1.

[*] An italic represents romanization of Japanese words.

[†] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/ (2017.2.2 version)

[‡] https://www.edrdg.org/enamdict/enamdict_doc.html

Embeddings	Neutral first name	Female first name	Male first name	Foreign surname	Total
word2vec	302	7,750	2,714	717	11,483
fastText	319	8,439	2,585	558	11,901

Table 1: The number of names after filtering

Secondly, we cluster the words which are included in the most frequent M tokens into clusters of m words. In work of Swinger et al. (2019), occupation and food-related clusters were generated for English. We set m to 64 as in their setup, but increase M from 30,000 to 50,000 in order to match the bias mitigation experiment of (Bolukbasi et al., 2016).

Thirdly, each m cluster is further divided into Voronoi sets with a high degree of dot product between a word vector and the vector mean of each name cluster. In this step, all word vectors and name vectors are normalized to size 1. After that, the most relevant words were chosen as t in each Voronoi set, and we set $t = 3$, following Swinger et al. (2019). However, if the number of elements in each Voronoi set generated after Voronoi partitioning is smaller than t, all elements are used.

Finally, in the fourth step, we compute the WEAT score and p-value. First, we calculate the WEAT score for each cluster of names and the t words included in Voronoi sets and chosen in order of relevance. Next, we calculate the p-value. Following Swinger et al. (2019), we use "rotational null hypothesis" for p-value. We multiply each name vector by an uniform Haar random orthogonal matrix and perform the above-described third step identically to how the WEAT score is computed. This is done $R = 10,000$ times, and the percentage of times the score is higher than the original score becomes the p-value. Finally, the statistically significant WEATs are outputted. For determining the critical p-value, we follow Swinger et al. (2019), who utilized method of Benjamini and Hochberg (1995) to guarantee an α bound on false discovery rate. The α is set to 0.05 as in Swinger et al. (2019).

Our hypothesis is that the use of first names in Japanese does not reflect social stereotypes. As mentioned in Section 3, *kanji* ideograms have their own specific meanings, and Japanese first names are sometimes given with the intention of expressing the meaning of the *kanji*. In the case of a name consisting of a single *kanji* character, its meaning may have a significant impact on the information conveyed by embeddings. Additionally, as mentioned in Section 3, since the usage of embeddings may differ depending on a character type, we assume that such types may have an influence on Japanese embeddings. Therefore, we presume that embeddings of names are unlikely to reflect social stereotypes and that clusters are formed by the character type and the meaning of ideograms.

4.3 Gender bias mitigation

We target bias mitigation for gender bias in Japanese embeddings by using Hard Debias (Bolukbasi et al., 2016) and Double-Hard Debias (Wang et al., 2020).

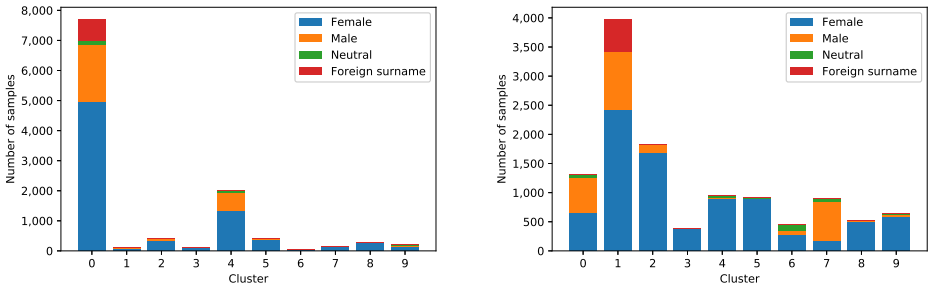
4.3.1 Mitigating methods

Hard Debias Hard Debias is a method for bias mitigation by removing gender direction from gender neutral words. Gender direction is defined in advance as the first principal component of gender definition word pairs. Original Hard Debias (Bolukbasi et al., 2016) normalizes a word vector to size 1, but we do not so, because its length can contain important information as pointed out by Ethayarajh et al. (2019).

Double-Hard Debias Double-Hard Debias follows Mu and Viswanath (2018), before doing Hard Debias, first centralizing the entire embedding and then removing the dominant principal component of the gender bias. Hard Debias was improved by performing these steps.

4.3.2 Gender Definition and Specific Words

Bolukbasi et al. (2016) defines gender specific words in advance, then uses them in the training data and extends the gender specific words with SVM. However, it has been pointed out that searching for gender specific words using embeddings of the bias mitigation targets poses the problem of not being able to



(a) word2vec (b) fastText

Figure 1: Results of clustering names by first names and foreign surnames with $n = 10$

properly classify truly gender specific words (Ethayarajh et al., 2019; Kumar et al., 2020). For that reason we collect gender specific words using Knowledge Based Classifier (KBC) proposed by Kumar et al. (2020).

The KBC has been implemented as follows. First we translate the definition words used by Bolukbasi et al. (2016) and use them as gender definition words for Japanese. However, since "herself" and other words they utilized do not exist in Japanese embeddings, we instead use, for example, synonyms of "mother" to match the number of pairs[§]. Then, for any word w, we check whether the definition of w contains gender definition words or not by using Wordnet (Bond et al., 2012) and ConceptNet (Speer and Havasi, 2013). If the gender word is present in a definition or node, w is treated as a gender specific word, and if not, it is labelled as a gender neutral word. However, our preliminary experiments showed that some relationships in ConceptNet contained gender bias themselves, so we chose edges for which effects of gender bias were not significant: IsA, PartOf, HasA, Synonym, Antonym, DefinedAs, and MannerOf.

4.4 Evaluation methods

Experiment 1: bias analysis We select the top 12 WEATs with the highest WEAT scores among the output WEATs in the bias analysis experiment and check whether these WEATs reflect social stereotypes. Five illustrative names for each name cluster were used for the evaluation. They are selected using a simple greedy heuristic presented in the original paper (Swinger et al., 2019). To evaluate whether the output WEATs reflected social stereotypes, we asked seven native Japanese speakers (5 males and 2 females, 19-29 years old) to associate statistically significant cluster of words with one most stereotypically related cluster of names. If WEATs represent a social stereotype, there should be high agreement between WEATs and annotators. Pairs of names/words clusters selected by more than 50% annotators were treated as correct associations (annotation guideline follows Swinger et al. (2019) but no rewards were given to annotators).

Experiment 2: mitigating gender bias We evaluate gender bias of the Japanese word embeddings using the neighborhood metric (Gonen and Goldberg, 2019).

The neighborhood metric is a measure of bias, which clusters $n \times k$ words with the largest bias in embedding before mitigation into k clusters by using k-means++, and then evaluates bias level providing the percentage of words belonging to each cluster that is consistent with the original bias. Higher percentage indicates that the word embedding includes a bias. We use the difference in cosine similarity between the word vectors of "woman" and "man" and between "she" and "he" as the magnitude of the gender bias. After compressing the data into two dimensions using tSNE (van der Maaten and Hinton, 2008), we perform further clustering also using k-means++. For this experiment, we set $k = 2$ to evaluate the gender bias related to females and males. We conduct experiments setting n to 100, 500, and $1,000$, following Wang et al. (2020).

[§]Definitional word pairs we used are: ["woman", "man"], ["female", "male" (gender)], ["female", "male" (sex)], ["girl", "boy"], ["little girl", "little boy"], ["mother", "father"], ["mother parent", "father parent"], ["daughter", "son"], ["she", "he"],["*Hanako*", "*Taro*"]

w2v F0	w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9
Hiroji	*Shinzaemon*	*Kyoko*	***Kasumi***	*Kotaro*	***Yomogi***	***Yu***	*Rie* (h)	*Yukino* (h)	***Etsu***
Akiko	*Ikurumi*	*Mai*	***Suzu***	*Akari*	***Mari***	***Syu***	*Chika* (h)	*Juri* (h)	***Ryo***
Asuka	*Noriaki*	*Sachiko*	***Mine***	*Tomihisa*	***Satsuki***	*Shichiro*	*Akio* (h)	*Yae* (h)	***Itsuki***
Shigetaka	*Toriha*	*Sekiko*	***Usagi***	*Zyotaro*	*Sachi*	***Sada***	*Ura* (h)	*Ao* (h)	***Atsushi***
Sachino	*Ayame*	*Kazuki*	***Midori***	*Kiyono* (h)	*Kuon*	***Hisao***	*Kaoru* (h)	*Atsumi* (h)	***Kou***
+7,712	*+97*	*+417*	*+113*	*+1,993*	*+419*	*+52*	*+134*	*+290*	*+206*
64% F	*72% F*	*77% F*	*92% F*	*67% F*	*85% F*	*54% F*	*96% F*	*98% F*	*64% F*

Table 2: Clustering results of the first names and the foreign surnames using word2vec (w2v) with $n = 10$ and the illustrative names of each cluster. (h) indicates a *hiragana* word, and **bold** font represents single kanji names. "% F" in the last row indicates female name ratio in the cluster.

ft F0	*ft F1*	*ft F2*	*ft F3*	*ft F4*	*ft F5*	*ft F6*	*ft F7*	*ft F8*	*ft F9*
Sachio	*Yumie*	*Fuyu*	*Mitsuki*	*Kaede*	*Mana*	***Hiro***	*Masato*	*Ayano*	*Miyoko*
Katsuyo	*Kikue*	*Akiho*	*Yoshino*	*Teruka*	*Kaori*	***Akira***	*Eiichi*	*Matsue*	*Harue*
Takashige	*Mitsuki* (k)	*Raiko*	*Arisu*	*Kikyou*	*Ena*	***Kei***	*Yoshihiro*	*Nao*	*Kazuko*
Yoshimi	Jewison (k)	*Takie*	*Yuuki*	*Midori*	*Yuki*	***Akane***	*Kenji*	*Hiroyasu*	*Akie*
Sukeichi	*Yurie*	*Ruuku*	*Ebiko*	*Tsukuyo*	*Nana*	***Ken***	*Kano*	*Chiho*	*Katsuko*
+1,301	*+3,978*	*+1827*	*+377*	*+940*	*+913*	*+456*	*+901*	*+522*	*+636*
50% F	*61% F*	*92% F*	*98% F*	*95% F*	*96% F*	*61% F*	*19% F*	*93% F*	*91% F*

Table 3: Clustering results for the first names and the foreign surnames using fastText (ft) with $n = 10$ and the illustrative names of each cluster. (k) indicates a *katakana* word, and **bold** font represents single kanji names. "% F" in the last row indicates female name ratio in the cluster.

5 Results

5.1 Experiment 1: bias analysis

The results of clustering names are shown in Figure 1a for word2vec, Figure 1b for fastText, and in Tables 2, 3, correspondingly. There are several possible readings of *kanji* ideograms for a single Japanese name, but we use only one reading in the tables. Figures 1a and 1b show the overall results of clustering names. Tables 2 and 3 list the illustrative names of each cluster.

In work of Swinger et al. (2019), distinct clusters are generated for both genders. However, as shown in the Figure 1a and Table 2, in the case of Japanese, no clusters of male names are formed from word2vec embedding and most of the names are clustered in cluster 0. Names in *hiragana* gathered in clusters 7 and 8. On the other hand, as shown in Figure 1b and Table 3, male names are grouped in cluster 7 when fastText is used. In both word embeddings, clusters of single *kanji* ideograms (3, 5, 6 and 9 on word2vec and 6 on fastText) and female names ending with "*-ko*" (cluster 2 on word2vec, cluster 9 on fastText) were formed. We can observe that each cluster captures some distinctive characteristics, but all of them are formed rather by the character type or number of characters, not by features that reflect social stereotypes.

The top 12 WEATs outputted by UBE are shown in Tables 4, 5. Table 4 illustrates the results of UBE on word2vec and Table 5 on fastText. The fastText lexicon contains a number of uninterpretable parts of words that could not be removed by the morphological analyzer Juman++, and we enclosed them in quotes. The colored background indicates cases where the annotators agreed with WEATs that the words reflect social stereotypes of the names. As far as Tables 4 and 5 are concerned, we can observe that most WEATs fail to capture social stereotypes (15% agreement for word2vec, 24% for fastText).

5.2 Experiment 2: mitigating gender bias

The results of experiments using the neighborhood metric are shown in Tables 6 and 7. The tSNE visualization is shown in Figures 2 and 3.

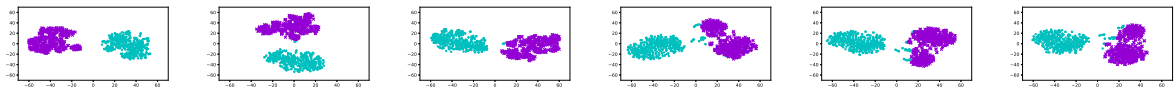
Regardless of which pair ("she/he" or "women/men") is used to evaluate the size of the gender bias in Japanese word2vec embedding, neither Hard Debias nor Double-Hard Debias come close to sufficient

w2v F0	w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9
investigate, grow old, warp	escape safely, betray, patrol	meet, be irritated, be enthusiastic	disperse, lithography, burn		offer devoutly, deify, funeral	in time, good offices, assault	'boru' (h), keep (h), some (h)	'nosu' (h), 'noku', pain (h)	stupid, salvation of country, yin-yang
	Chikugo, *Kofu*, *Komoro* (places)	*Keisuke*, *Hiro*, *Sekine*	astringent, white horse, Mt.Fuji	country club (k), *Kissho-ji* (place), Japanese old ordinary high school	*Asama* (place), imperial capital, Mt.Yae	*Nagai*, *Akamatsu*, *Nabeshima*		*Tochigi*(h), *Saitama* (h), *Nanba* (places) (h)	*shogi*, northern seas, Konan (place in China)
	Chikuzen (place), *Shimofusa* (place), Edo shogunate	*Chube*	tray, folding screen, the Healing Buddha		devine sprit, deify, dedication	family of shogunate, *Yoshinori*, *Harunobu*			trick, Toi (ancient China class), Buddhist priest
		pleasure, *Zyunichi*, boy friend	beautiful, many, *kirakira* (glitter) (k)				'yo' (h), 'sun' (h), irresponsible	adult (h), guy (h), No.1 (h)	pain, intelligent person, captivation
		buddy, *Shinji*, transfer student	frog, spider, fang		emperor, sanctuary, superiority	stratagem, revenge, assassin		rat (h), life (h), original title	die out, hollow, thief
Venezuela (k), Slovakia (k), Croatia (k)	dukedom, Ruthenia (k), Netherland (k)	American, Britisher, Japanese diaspora			Arabia (k), Hindu (k), Jerusalem				Guangzhou , Yunnan, Fujian (places in China)
	castle town, villa, the main enclosure of a castle	apartment (k), one house, *manshon* (k) (rich apartment)	giant tree, fountain, stone pillar		tomb, mosque (k), royal palace				study (room), warehouse
	direct line of descent, relative, collateral line	childhood friend, childhood friend (only *kanji*), same age				eldest son, successor, father and son			princess, empress dowager, mother-in-law
general education college, graduate school of letters, department of sociology		classmate, upperclass-man, pupil				study of Chinese classics, assistant professor, school principal			academy, degree, pass an examination
	topography, ancient manuscript, genealogy	conversation, story of one's experience, recollection	scroll, collection of haikus, iconography		inscription, series of publication, hymn	history book, historical material, transcription			word, anthology, national history
		fullname (k), family name, initial (k)	seal, character used as a phonetic symbol, Greek		free translation, Greek, original meaning			*hiragana* (h), word (h), written in English	translation into classical Chinese, classical Chinese, translitera-tion
		confess, go around together, meet	dance, light up, plant		divine, praise, protect	ask, look after, beg			do, bestow, destroy

Table 4: The top 12 highest-scoring WEATs output (statistically significant) by UBE on word2vec. 'w2v F' indicate the cluster in Table 2. (h) indicates a *hiragana* word, (k) stands for a *katakana*. All other words are written in *kanji* ideograms except ones in quotation – they are uninterpretable parts of words (noise). Orange cells indicate the clusters of names and words selected by more than 50% annotators matches the generated WEAT.

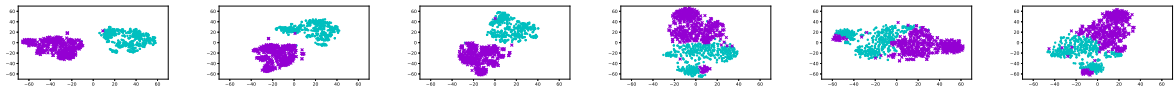
ft F0	ft F1	ft F2	ft F3	ft F4	ft F5	ft F6	ft F7	ft F8	ft F9
director, investigate, assistant professor	sweet novel comic (k), comedian, *Kaiseisha* (company)	go (h), get up (h), feel (h)	Iceland (k), Toulouse (k), America (k)	orange (k), leaf, the Milky Way	bikini model (k), girl, idol (k)		*Yukio*, *Yuji* (h), factory	Nuremberg (k) (place), *Hitachi-naka* (h) (place), Okhotsk (k)	career woman (k), wife, lady (k)
		dry, be dazzled, mold	enough, very (h), excellent	somehow, all year round, *hirahira* (h) (fluttering)	erotic (k), cute, look like a grown-up	split, I (*ware*), too much	leading person, go through (h), plan		
	Joseph (k), Norman (k), Harry (k)			aurora (k), Laguna (k), acacia (k)	Erina (k), Emily (k), Lilly (k)		*Hiroshi* (k), *Kenji* (k), Ministry of Transport	*Yawatahama*, *Dazaifu*, *Wakayama* (places)	*Toru* (k), Susan (k), *Takeshi* (k)
			stone wall, imperial guards, *Chika-matsu*	enjoyment of the moon, wild cherry tree, Japanese apricot with red blossoms		Horse (old orthography), stipend, vivid		equator, *Okinawa* (place), *Kyushu* (place)	
			Ito (h), *Hida* (h) (place), 'koji' (h)	kid (h), crab (h), burnt (h)	'koru' (h), 'ri' (h), 'puri' (h)	connection, detail, cut off	'rero' (h), line (h), feeling (h)		
						to (old unit of volume), disaster, I (*onore*)	hero, expert, primeval man		royal princess, imperial princess, princess
			very, a bit, first	various, other place, this way			tax included, immediately after, pipe (h)		
successor, third son, eldest son							sworn friend, brother, family		mother, ex-wife, married couple
			particular, multiple, diverseness			little, slant, error		the whole country, rising, neighborhood	
family name, brief history, pen name						meaning, abbreviation, character (letter)	omitted letter, name, one's title		old name, favorite phrase, speech
raise, explain, be granted		distribute, compare, return					prompt one to do, neglect, receive		
						price, side, measure			

Table 5: The top 12 highest-scoring WEATs output (statistically significant) by UBE on fastText. 'ft F' indicate the cluster in Table 3. (h) indicates a *hiragana* word, (k) stands for a *katakana*. All other words are written in *kanji* ideograms except ones in quotation – they are uninterpretable parts of words (parser noise). Orange cells indicate the clusters of names and words selected by more than 50% annotators matches the generated WEAT.



(a) word2vec (b) Hard debiased word2vec (c) Double-Hard debiased word2vec (d) fastText (e) Hard debiased fastText (f) Double-Hard debiased fastText

Figure 2: tSNE visualisation of the top 500 words in the case of "she" and "he". Graphs (a-c) show the results for word2vec. Graphs (d-f) show the results for fastText.



(a) word2vec (b) Hard debiased word2vec (c) Double-Hard debiased word2vec (d) fastText (e) Hard debiased fastText (f) Double-Hard debiased fastText

Figure 3: tSNE visualisation of the top 500 words in the case of "women" and "men". Graphs (a-c) show the results for word2vec, (d-f) for fastText. Clusters in 3d-3f are not separated, so gender bias is not visible.

mitigation of the gender bias. Also when fastText is used, neither of the bias mitigation methods is able to effectively mitigate the gender bias in the "she/he" case. However, when "woman/man" were used, gender bias could not be confirmed even before mitigating bias using the neighborhood metric.

6 Discussion

6.1 Experiment 1: bias analysis

Based on our experimental results, it is difficult to say that WEATs reflect social stereotypes. This supports our hypothesis that Japanese first name embeddings do not reflect social stereotypes. However, Ethayarajh et al. (2019) noticed that WEAT systematically overestimates the bias. We need to examine their findings in the future.

As mentioned in Section 5.2, each cluster of names is formed by the character type, which also supports our hypothesis that clusters are formed by the surface characteristics of Japanese language, not by the meaning. However, clusters are not formed by the meaning of *kanji* included in the names themselves. Particularly, our hypothesis that the clustering would be affected by a single *kanji* character was not supported by the experimental results. Rather than single ideograms, the single *kanji* character names are grouped, and we were able to confirm that clusters were not formed by the meaning of these characters. We also confirmed that clusters of three or more character names were created ("ft F1" cluster in Table 3). Foreign surnames also did not form their own clusters, but were grouped into the element-richest clusters. Therefore, our experimental results show that name embeddings form concentric circles of names merely from superficial information of character type and number of characters rather than meaning, gender or nationality.

Based on the above considerations, it can be said that Japanese first name embeddings do not contain much of social stereotypes, and the similarity between name and word vectors are affected by character types of a word rather than the meaning of the word itself. We think that the fact that WEATs failed to reflect social stereotypes is because the main information conveyed by name and word embeddings is mostly superficial. Swinger et al. (2019) express their concern about the difficulty of applying UBE with respect to groups that cannot be significantly distinguished by name. The results of our experiment support that speculation.

6.2 Experiment 2: mitigating gender bias

Gonen and Goldberg (2019) showed experimentally that Hard Debias fails to mitigate the gender bias when the neighborhood metric is used. We replicated this phenomenon in Japanese word embeddings. According to Wang et al. (2020), Double-Hard Debias can mitigate gender bias with the neighborhood

Embedding	Method	Top 100	Top 500	Top 1000
word2vec	Original	1.00 (1.00)	1.000 (0.994)	1.000 (0.999)
	Hard Debias	1.00 (1.00)	0.995 (0.992)	0.993 (0.988)
	Double-Hard Debias	1.00 (1.00)	**0.960 (0.978)**	**0.933 (0.967)**
fastText	Original	1.0 (1.00)	0.753 (0.972)	0.594 (0.959)
	Hard Debias	**0.99** (1.00)	**0.607** (0.974)	0.593 (0.976)
	Double-Hard Debias	**0.99** (1.00)	**0.607 (0.973)**	**0.592 (0.958)**

Table 6: Experimental results on the neighborhood metric in the case of "she" and "he". The accuracy of the metric after dimensionality reduction with tSNE is shown in parentheses.

Embedding	Method	Top 100	Top 500	Top 1000
word2vec	Original	1.00 (0.99)	1.000 (0.982)	0.996 (0.945)
	Hard Debias	1.00 (1.00)	0.993 (0.965)	0.993 (0.971)
	Double-Hard Debias	1.00 (**0.98**)	**0.966 (0.937)**	**0.916 (0.940)**
fastText	Original	**0.64 (0.51)**	0.585 (0.622)	0.645 (**0.643**)
	Hard Debias	0.64 (0.68)	**0.583** (**0.598**)	0.647 (0.652)
	Double-Hard Debias	**0.64 (0.51)**	**0.583** (0.647)	**0.573** (0.662)

Table 7: Experimental results on the neighborhood metric in the case of "women" and "men". The accuracy of the metric after dimensionality reduction by tSNE is shown in parentheses.

metric when targeting English GloVe and word2vec (results of the latter only shown in their Appendix). However, the bias could not be sufficiently mitigated in Japanese embeddings by using their method[¶].

One of the reasons might be related to the way how the gender definition words are predefined in those methods. Ethayarajh et al. (2019) comment on the results of Gonen and Goldberg (2019) stating that Hard Debias removes only the components of the predefined gender direction, and that it is impossible to remove other undefined components of the gender direction. Their conclusion is that even if one mitigates the bias with non-exhaustive gender definition word pairs, potential gender directions remain (Ethayarajh et al., 2019, p.1699).

We think this is true even if we remove the dominant principal components and make the embedding space isotropic, so the same criticism applies to Double-Hard Debias. In our opinion, the experimental results presented in this paper indicate that the list of gender definition word pairs we used was not sufficient to mitigate the gender bias. This poses the following problem. The number and types of words for gender naturally vary from language to language. Depending on the language, the exhaustive set of gender definition word pairs will differ. Also, the gender direction affecting the downstream task is not guaranteed to be identifiable or known a priori by simply using gender definition words translated from English. Therefore, it will be generally difficult to provide a comprehensive set of gender definition word pairs, suitable for downstream tasks, especially working with languages of a small NLP research population and limited resources.

7 Conclusion

In this paper, we analyzed the representational bias of Japanese word embeddings and attempted to mitigate the gender bias in these embeddings with previous methods developed for English. The experimental results showed that Japanese first name embeddings do not include social stereotypes and that the similarity of word vectors is influenced by the superficial information of character type. And, the existing gender

[¶]Unfortunately, there is no pre-trained GloVe model available for Japanese, so we were not able to investigate the influence of the embedding type.

bias mitigation methods did not sufficiently mitigate the gender bias in Japanese embeddings. These results suggest that it is difficult to generalize the previous methods for English to Japanese. This, in turn, may be suggesting that it could be difficult to apply those methods not only to Japanese, therefore it is important to consider whether and how they can be used to analyze and mitigate bias in other languages.

In the future, we will develop methods for bias analysis and of bias mitigation specifically dedicated to Japanese language. We will also examine the generalizability of other existing methods, and try to answer remaining question: what are the meta-conditions for a method to be independent of a language.

References

David Arthur and Sergei Vassilvitskii. 2006. k-means++: The advantages of careful seeding. Technical Report 2006-13, Stanford InfoLab, June.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th global WordNet conference (GWC 2012)*, pages 56–63. Citeseer.

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

Kate Crawford. 2017. The trouble with bias. Keynote at Neural Information Processing Systems (NIPS'17).

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy, July. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. 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)*, pages 609–614, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Akihiko Iwahara, Takeshi Hatta, and Aiko Maehara. 2003. The effects of a sense of compatibility between type of script and word in written japanese. *Reading and Writing*, 16(4):377–397.

Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings.

Haochen Liu, Jamell Dacon, Wenqi Fan, H. Liu, Zhiwei Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv*, abs/1910.10486.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal, September. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective post-processing for word representations. In *6th International Conference on Learning Representations, ICLR 2018*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2019, page 450–456, New York, NY, USA. Association for Computing Machinery.

Thijs Raijmakers. 2020. Gender bias in word embeddings of different languages.

Magnus Sahlgren and Fredrik Olsson. 2019. Gender bias in pretrained Swedish embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43, Turku, Finland, September–October. Linköping University Electronic Press.

Robyn Speer and Catherine Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*, pages 161–176. Springer.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 2018. A joint neural model for fine-grained named entity classification of wikipedia articles. *IEICE Transactions on Information and Systems*, E101.D(1):73–81.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 305–311, New York, NY, USA. Association for Computing Machinery.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Online, July. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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, New Orleans, Louisiana, June. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jieyu Zhao, Subhabrata Mukherjee, saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China, November. Association for Computational Linguistics.