Analyzing Semantic Changes in Japanese Loanwords

Hiroya Takamura
Tokyo Institute of Technology
takamura@pi.titech.ac.jp

Ryo Nagata
Konan University
nagata-acl@hyogo-u.ac.jp

Yoshifumi Kawasaki
Sophia University
kyossii@gmail.com

Abstract

We analyze semantic changes in loanwords from English that are used in Japanese (Japanese loanwords). Specifically, we create word embeddings of English and Japanese and map the Japanese embeddings into the English space so that we can calculate the similarity of each Japanese word and each English word. We then attempt to find loanwords that are semantically different from their original, see if known meaning changes are correctly captured, and show the possibility of using our methodology in language education.

1 Introduction

We often come across advertisements that have extravagant images. In Japan, such images are usually accompanied by the following sentence[1]:

この	画像は	イメージ	です
kono	gazō-wa	imēji	desu
this	image-TOP	image	COP

"*This image is an image.*"

This sentence sounds like a nonsense tautology, but actually means *this image is only for illustrative purposes and may differ from the actual product*. Both *gazō* and *imēji* are Japanese words, each meaning *image*. However, the latter is a loanword from English, i.e., *image*[2]. In the sentence above, *imēji*, the loanword for image, is closer in meaning to the word *impression*, and it makes the sentence roughly mean *this image is just an impression that you might have on this product*. What happens in this seeming tautology is that the loanword changes meaning; i.e., the sense of the loanword deviates from the sense of its original word.

Loanwords from English occupy an important place in the Japanese language. It is reported that approximately 8% of the vocabulary of contemporary Japanese consists of loanwords from English (Barrs, 2013). One noteworthy characteristics of loanwords in Japanese is that their meanings are often different from their original words, as in the above example.Indeed, the meanings of loanwords in any language are not generally the same as those in the language, but according to Kay (1995), Japanese has particularly a strong tendency of changing the meanings of loanwords; Kay argued that in Japan *there is no deep cultural motivation to protect their original meaning*. Daulton (2009) also argued that Japanese loanwords are malleable in terms of meanings. Thus, Japanese loanwords would be an interesting subject to work on in the study of meaning change.

Japanese loanwords from English are also important in language education (Barrs, 2013). Japanese learners of English often make mistakes in using English words that have corresponding loanwords in Japanese but with very different meanings. By contrast, learners are able to make better use of a loanword in conversation if they know that its meaning is the same as that of the original. It is thus important to know which loanwords are semantically different from their original and which are not.

With this background in mind, we work on Japanese loanwords derived from English. Since the word embedding vectors (or simply, embeddings), which have become very popular recently, are powerful tools for dealing with word meanings, we use them to analyze Japanese loanwords. Specifically, we create word embeddings of En-

[1] TOP and COP respectively mean a topic marker and a copula in interlinear glossed text (IGT) representation. The last line is a literal translation of the Japanese sentence.

[2] Note that although *gazō* is also from ancient Chinese, we focus on loanwords from English, which are usually written in *katakana* letters in Japanese.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1195–1204,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

glish and Japanese, and map the Japanese embeddings into the English space so that we can calculate the similarity of each Japanese word and each English word. We then attempt to find loanwords that are semantically different from their original, see if known meaning changes are correctly captured, and show the possibility of using our methodology in language education.

In this paper, we use the term *semantic change* or *meaning change* in a broad sense. Some loanwords are semantically different from the original words because the loanwords or the original words semantically changed after they were introduced into Japanese or because only one of the multiple senses of the original words were introduced. Moreover, some loanwords did not come directly from English, but from words in other languages, which later became English words. Thus, in this paper, the terms semantic change or meaning change cover all of these semantic differences.

2 Related Work

Japanese loanwords have attracted much interest from researchers. Many interesting aspects of Japanese loanwords are summarized in a book written by Irwin (2011). In the field of natural language processing, there have been a number of efforts to capture the behavior of Japanese loanwords including the phonology (Blair and Ingram, 1998; Mao and Hulden, 2016) and segmentation of multi-word loanwords (Breen et al., 2012). The rest of this section explains the computational approaches to semantic changes or variations of words. In particular, there are mainly two different phenomena, namely diachronic change and geographical variation.

Jatowt and Duh (2014) used conventional distributional representations of words, i.e., bag-of-context-words, calculated from Google Book (Michel et al., 2011)[3] to analyze the diachronic meaning changes of words. They also attempted to capture the change in sentiment of words across time. Kulkarni et al. (2014) used distributed representations of words (or word embeddings), instead of the bag-of-context-words used by Jatowt and Duh, to capture meaning changes of words and in addition used the change point detection technique to find the point on the timeline where the meaning change occurred. Hamilton et al. (2016) also used distributed representations for the same purpose and attempted to reveal the statistical laws of meaning change. They compared the following three methods for creating word embedding: positive pointwise mutual information (PPMI), low-dimensional approximation of PPMI obtained through singular value decomposition, and skip-gram with negative sampling. They suggested that the skip-gram with negative sampling is a reasonable choice for studying meaning changes of words. We decided to follow their work and use the skip-gram with negative sampling to create word embeddings.

Bamman et al. (2014) used a similar technique to study differences in word meanings ascribed to geographical factors. They succeeded in correctly recognizing some dialects of English within the United States. Kulkarni et al. (2016) also worked on geographic variations in languages.

With some modification, the methods used in the literature (Kulkarni et al., 2014; Hamilton et al., 2016) can be applied to loanword analysis.

3 Methodology

We use word embeddings to analyze the semantic changes in Japanese loanwords from the corresponding English. Among the methods of analysis, we chose to use the skip-gram with negative sampling for the reason discussed in Section 2 with reference to Hamilton et al.'s work (2016).

First, we create word embeddings for two languages. We then calculate the similarity or dissimilarity between the embedding (or vector) of a word in a language (say, Japanese) and the embedding of a word in another language (say, English). For this purpose, words in the two languages need to be represented in the same vector space with the same coordinates. There are a number of methods for this purpose (Gouws et al., 2015; Zou et al., 2013; Faruqui and Dyer, 2014; Mikolov et al., 2013a). Among them, we choose the simplest and most computationally efficient one proposed by Mikolov et al. (2013a), where it is assumed that embeddings in one language can be mapped into the vector space of another language by means of a linear transformation represented by W. Suppose we are given trained word embeddings of the two languages and a set of seed pairs of embedding vectors $\{(x_i, z_i)|1 \leq i \leq n\}$, each of which is a pair of a vector in one language and a vector in the other language that are translation equiva-

[3] https://books.google.com/ngrams/datasets

lents of each other. The transformation matrix W is obtained by solving the following minimization problem :

$$\min_W \sum_{i=1}^{n} ||Wx_i - z_i||^2, \quad (1)$$

where, in our case, x_i is the embedding of a Japanese seed word and z_i is the embedding of its English counterpart. Thus, the Japanese word embeddings are mapped into the English vector space so that the embeddings of the words in each seed pair should be as close to each other as possible. Although Hamilton et al. (2016) preserved cosine similarities between embedding vectors by adding the orthogonality constraint (i.e., $W^T W = I$, where I is the identity matrix) when they aligned English word embeddings of different time periods, we do not adopt this constraint for two reasons. The first reason is that since we need an inter-language mapping instead of across-time mappings of the same language, the orthogonality constraint would degrade the quality of the mapping; the two spaces might be so different that even the best rotation represented by an orthogonal matrix would leave much error between corresponding words. The second reason is that we do not need to preserve cosine similarities between words in mapping embedding vectors, because we do not use the cosine similarities between mapped embedding vectors of Japanese words.

After mapping the Japanese word embeddings to the English vector space, we calculate the cosine similarity between each Japanese loanword and its original English word. If the cosine similarity is low for a pair of words, the meaning of the Japanese loanword is different from that of its original English word.

4 Empirical Evaluations

Since it is generally difficult to evaluate methods for capturing semantic changes in words, we conduct a number of quantitative and qualitative evaluations from different viewpoints.

4.1 Data and Experimental Settings

The word embeddings of English and Japanese were obtained via the skip-gram with negative sampling (Mikolov et al., 2013b)[4] with different dimensions as shown in the result. The data used for this calculation was taken from Wikipedia dumps[5] as of June 2016 for each language; the text was extracted by using wp2txt (Hasebe, 2006)[6], non-alphabetical symbols were removed, and noisy lines such as the ones corresponding to the infobox were filtered out[7]. We performed word segmentation on the Japanese Wikipedia data by using the Japanese morphological analyzer MeCab (Kudo et al., 2004)[8] with the neologism dictionary, NEologd[9], so that named entities would be recognized correctly.

The list of Japanese loanwords was obtained from Wiktionary[10]. Only one-word entries were used and some errors were corrected, resulting in 1,347 loanwords from English[11].

We extracted seed word pairs from an English-Japanese dictionary, edict (Breen, 2000)[12]; these were used in the minimization problem expressed by Equation (1). Specifically, we extracted one-word English entries that were represented as a single Japanese word. We then excluded the 1,347 loanwords obtained above from the word pairs, which resulted in 41,366 seed word pairs.

4.2 Evaluation through Correlation

To see if the differences in word embeddings are related to the meaning changes of loanwords, we calculate an evaluation measure indicating the global trend. We first extracted one-to-one translation sentence pairs from Japanese-English News Article Alignment Data (JENAAD) (Utiyama and Isahara, 2003). We then use this set of sentence pairs to calculate the Dice coefficient for each pair of a loanword w_{jpn} and its original English word w_{eng}, which is defined as

$$\frac{2 \times P(w_{\text{jpn}}, w_{\text{eng}})}{P(w_{\text{jpn}}) + P(w_{\text{eng}})}, \quad (2)$$

[4] https://code.google.com/archive/p/word2vec/ with options "-window 5 -sample 1e-4 -negative 5 -hs 0 -cbow 0 -iter 3"

[5] https://dumps.wikimedia.org/enwiki/ https://dumps.wikimedia.org/jawiki/

[6] https://github.com/yohasebe/wp2txt

[7] https://en.wikipedia.org/wiki/Help:Infobox

[8] http://taku910.github.io/mecab/

[9] https://github.com/neologd/mecab-ipadic-neologd

[10] https://ja.wiktionary.org/wiki/%E3%82%AB%E3%83%86%E3%82%B4%E3%83%AA%3A%E6%97%A5%E6%9C%AC%E8%AA%9E_%E5%A4%96%E6%9D%A5%E8%AA%9E

[11] Some of these loanwords may have been introduced into Japanese via other languages. However, in this paper, we regard them as from English as long as they are also used in English.

[12] http://www.edrdg.org/jmdict/edict.html

dimension		correlation coefficient	
dim_jpn	dim_eng	Pearson	Spearman
100	100	0.363	0.443
200	100	0.386	0.471
200	200	0.402	0.474
400	200	0.404	0.487
300	300	0.422	0.492
600	300	0.432	0.506

Table 1: Correlation coefficients between the Dice coefficient and the cosine similarity. dim_jpn and dim_eng are respectively the dimensions of the Japanese and English word embeddings; i.e., the dim_jpn-dimensional space is mapped to the dim_eng-dimensional space. All the coefficients are statistically significant (significance level 0.01).

where $P(w_\text{jpn}, w_\text{eng})$ is the probability that this word pair appears in the same sentence pair, and $P(w_\text{jpn})$ and $P(w_\text{eng})$ are the generative probabilities of w_jpn and w_eng. All the probabilities were obtained using the maximum likelihood estimation. The Dice coefficient is a measure of coocurrence and can be used to extract translate equivalents (Smadja et al., 1996). If the Dice coefficient of a word pair is low, the words in the pair are unlikely to be translation equivalents of each other. Therefore, if the meaning of a loanword has changed from the original English word, its Dice coefficient should be low. In other words, the cosine similarity should be correlated to the Dice coefficient if the cosine similarity is a good indicator of meaning change. We thus calculate the Pearson's correlation coefficient between the two. In addition, we calculate the Spearman's rank-order correlation coefficient to examine the relation of the orders given by the Dice coefficient and the cosine similarity.

Note that although we use a parallel corpus for evaluation, it does not mean that we can simply use a parallel corpus for finding meaning changes in loanwords. Parallel corpora are usually much smaller than monolingual corpora and can cover only a small portion of the entire set of loanwords. With the model described in Section 3, we will be able to find meaning changes in loanwords that do not appear in a parallel corpus.

The results for different Japanese and English dimensions, dim_jpn and dim_eng, are shown in Table 1. Pearson's correlation coefficients suggest that the cosine similarity is moderately correlated with the Dice coefficient except for the case dim_eng=100, which shows weak correlation. Spearman's rank-order correlation coefficients also suggest that these two are moderately correlated with each other. The result depends on the dimensions of the word embeddings. Basically, larger dimensions tend to have higher correlation coefficients. In addition, when the dimension is decreased (e.g., $\text{dim}_\text{jpn} = 600$ to $\text{dim}_\text{eng} = 300$), the correlation coefficients tend to be higher, compared with the case where the dimension remains the same (e.g., $\text{dim}_\text{jpn} = 300$ to $\text{dim}_\text{eng} = 300$). This result is consistent with the report by Mikolov et al. (2013a) that *the word vectors trained on the source language should be several times (around 2x-4x) larger than the word vectors trained on the target language*.

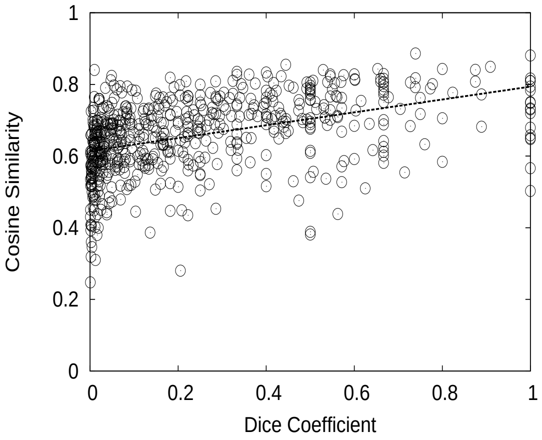

Figure 1: Dice coefficient vs. cosine similarity. Dice coefficients are extracted from a parallel corpus. Cosine similarities are for the embedding vectors of the Japanese loanwords and their English counterparts. The line in the figure is obtained by linear regression.

To examine the relation between the Dice coefficient and the cosine similarity in more detail, we plot these values for the bottom row in Table 1, i.e., where the dimensions for Japanese number 600 and the dimensions for English number 300. The scatter plot that we obtained is shown in Figure 1. The line obtained by linear regression is also drawn in the figure.

4.3 Detailed Evaluation on Known Change

Here, we conduct a detailed evaluation on meaning changes that are already known. We selected the ten Japanese loanwords shown in Ta-

w_{eng}	w_{jpn}	w_a	w_b	$\cos(w_{\text{eng}}, w_a)$ $- \cos(w_{\text{jpn}}, w_a)$	$\cos(w_{\text{jpn}}, w_b)$ $- \cos(w_{\text{eng}}, w_b)$
image	imēji	photo	impression	0.097	0.274
corner	cōnā	crossroad	section	0.099	0.115
digest	daijesuto	dissolve	summary	0.047	0.291
bug	bagu	insect	glitch	0.092	0.200
idol	aidoru	deity	popstar	0.127	0.086
icon	aikon	deity	illustration	−0.035	0.145
cunning	kanningu	shrewd	cheating	0.259	0.273
pension	penshon	annuity	hotel	0.368	0.445
nature	neichā	characteristics	magazine	0.106	0.202
driver	doraibā	chauffeur	screwdriver	−0.063	0.158

Table 2: Differences in cosine similarity. The Japanese loanword from *corner* can mean a small section in a larger building or space. The Japanese loanword from *bug* usually means a bug in a computer program. The Japanese loanword from *cunning* usually means cheating on an exam. The Japanese loanword from *nature* is often used to indicate the scientific journal *Nature*. The Japanese loanword from *driver* can mean both a vehicle driver and a screwdriver (the latter meaning was not one of the original word).

ble 2 that are supposed to have different meanings compared with the original English words. Some of these words were taken from a book about loanwords written by Kojima (1988). The others were collected by the authors. We also added two *pivot* words w_a and w_b for each word[13]. For the first nine words, the meaning of pivot word w_a is supposed to be closer to the English word w_{eng} than to the Japanese loanword w_{jpn}, and the meaning of the pivot word w_b is supposed to be closer to the Japanese loanword w_{jpn} than to the English word w_{eng}. It is thus expected that $\cos(w_{\text{eng}}, w_a) - \cos(w_{\text{jpn}}, w_a) > 0$ and $\cos(w_{\text{jpn}}, w_b) - \cos(w_{\text{eng}}, w_b) > 0$ hold true. The last Japanese loanword in Table 2 is used as both pivot words w_a and w_b, but the original English word is not used as w_b. It is thus expected that $\cos(w_{\text{jpn}}, w_b) - \cos(w_{\text{eng}}, w_b) > 0$ holds true, but $\cos(w_{\text{eng}}, w_a) - \cos(w_{\text{jpn}}, w_a) > 0$ might not be necessarily true. The differences in cosine similarities are shown in Table 2. As expected, almost all the differences are positive, which suggests that the difference of the word embeddings captures the meaning change. However, there was one exception:

$$\cos(\mathit{icon}, \mathit{deity}) - \cos(\mathit{icon}_{\text{jpn}}, \mathit{deity}) = -0.035.$$

The cosine similarity between *icon* and *deity* was 0.266, which is smaller than expected. We randomly sampled 100 lines containing *icon* from English Wikipedia text, which we used for calculating word embeddings, and found that the dominant sense of *icon* in Wikipedia is not *a religious painting or figure*, but *a representative person or thing’* as in the Wikipedia page of a football superstar David Beckham[14] :

> *Beckham became known as a fashion icon, and together with Victoria, the couple became* $\cdots$

Thus, the reason of *icon*’s anomalous behavior is that the distribution over senses in Wikipedia was a lot different from the expected one.

4.4 Nearest Neighbors

We show in Table 3 the English nearest neighbors of the English word w_{eng} and the Japanese loanword w_{jpn} in the 300-dimensional space of English. Japanese loanwords are mapped from the 600-dimensional space of Japanese into the 300-dimensional space of English. The English word *image* is close in meaning to the word *picture*, as suggested by *jpeg* and *close-up*, while its loanword seems to have a more abstract meaning such as *idealizing*. The nearest neighbors of the English word *digest* are influenced by an American fam-

[13] Pivot words are not necessarily synonyms of the corresponding English words. They are the words that we think are useful for capturing how the meanings of the loanwords and the original English are different. We also made sure that pivot words themselves are unambiguous.

[14] `https://en.wikipedia.org/wiki/David_Beckham`

w_{eng}	nearest neighbor of w_{eng}		nearest neighbors of w_{jpn}	
image	file	0.774	idealizing	0.671
	jpeg	0.748	stylization	0.665
	jpg	0.724	inescapably	0.665
	closeup	0.694	evoking	0.664
	close-up	0.658	englishness	0.664
corner	corners	0.727	recapped	0.666
	tiltons	0.646	cliff-hanger	0.644
	goerkes	0.643	“blank”	0.642
	uphams	0.629	announcer’s	0.641
	intersection’s	0.627	sports-themed	0.632
digest	digests	0.609	recaps	0.717
	digest’s	0.594	wrap-up	0.697
	reader’s	0.591	recapped	0.695
	wallace-reader’s	0.573	preview	0.693
	wallace/reader’s	0.556	recap	0.690
bug	bugs	0.672	heartbleed	0.714
	leaf-footed	0.605	workaround	0.695
	motherhead	0.590	workarounds	0.686
	harpactorinae	0.582	glitches	0.684
	thread-legged	0.579	copy-on-write	0.684
icon	icons	0.750	swoosh	0.701
	iverskaya	0.580	viewport	0.694
	nicopeia	0.579	crosshair	0.691
	eleusa	0.570	upper-left	0.684
	derzhavnaya	0.569	wireframe	0.680
nature	teiči	0.649	phytogeography	0.684
	søraust-svalbard	0.643	ethological	0.679
	naturans	0.627	life-history	0.676
	naturata	0.623	paleoclimatology	0.671
	naturing	0.623	archaeoastronomy	0.670
driver	drivers	0.837	driver	0.762
	driver’s	0.703	race-car	0.689
	car	0.685	mechanic	0.649
	co-drivers	0.655	harvick’s	0.645
	owner-driver	0.653	andretti’s	0.642

Table 3: English words that are nearest w_{eng} and w_{jpn}. w_{jpn} is a Japanese loanword and w_{eng} is the original English word. w_{jpn} is mapped into the English vector space. Only words that appear more than 100 times in the Wikipedia corpus are considered as candidates of the nearest neighbors. The value next to each word is the cosine similarity between the nearest neighbor word and w_{eng} or w_{jpn}.

ily magazine *Reader's Digest*[15] by Wallaces, but the terms related to *summary* do not appear in the top-5 list, except for *digest* itself. In contrast, its loanword seems to mean *wrap-up*. We now return to the English word *icon* that was mentioned as an exception in Section 4.3. Besides *icons*, the nearest neighbors of *icon* are *iverskaya*, *nicopeia*, *eleusa*, and *derzhavnaya*. These four words are all related to religious paintings or figures, but they have low cosine similarities. The other parts of the table are also mostly interpretable. The nearest neighbors of w_{eng} *nature* look uninterpretable at a first glance, but they are influenced by the Søraust-Svalbard Nature Reserve in Norway, and Natura naturans, which is a term associated with the philosophy of Baruch Spinoza.

4.5 Ranking of Word Pairs According to Similarity

Here, we investigate the possibility of whether the similarity calculated in the mapped space can be used to detect the loanwords that are very different from or close to the original English words. We show the 20 words with the lowest cosine similarities and the 20 words with highest cosine similarities in Table 4. First, let us take a look at the words on the right, which have high similarities. Most of them are technical terms (e.g., *hexadecane* and *propylene*), and domain-specific terms such as musical instruments (e.g., *piano* and *violin*) and computer-related terms (*computer* and *software*). This result is consistent with the fact that the meanings of technical terms tend not to change, at least for Japanese (Nishiyama, 1995). Next, let us take a look at the words on the left, which have low similarities. Many of them are actually ambiguous, and this ambiguity is often due to the Japanese phonetic system. For example, *lighter* and *writer* are assigned to the same loanword in Japanese, because the Japanese language does not distinguish the consonants l and r. The words *clause*, *close* and *clothe* are also assigned to the same loanword also because of the Japanese phonetic system. Other words are used as parts of named entities, also resulting in low similarity. For example, the Japanese loanword for *refer* is more often used as *Rifaa*, the name of a city in Bahrain, but hardly as *refer*. The loanword for *irregular* is often used as part of a video game title *Irregular Hunter*. However, we can also find words with significant changes in meaning, such as *present*[16] and *coordinate*[17]. Therefore, the result suggests that the similarity calculated by our method has the capability of detecting changes in the meanings of loanwords, but we need to filter out the words that are ambiguous in the Japanese phonetic system.

dissimilar pair		similar pair	
w_{eng}	cosine	w_{eng}	cosine
lac	0.225	piano	0.886
refer	0.245	violin	0.881
police	0.247	cello	0.881
spread	0.251	hexadecane	0.864
mof	0.261	propylene	0.857
pond	0.270	keyboard	0.855
inn	0.274	clarinet	0.851
ism	0.279	cheese	0.849
lighter	0.280	mayonnaise	0.848
root	0.281	software	0.847
tabu	0.284	methanol	0.843
gnu	0.293	hotel	0.843
thyme	0.296	chocolate	0.841
clause	0.310	computer	0.840
board	0.315	engine	0.840
present	0.319	globalization	0.835
coordinate	0.337	tomato	0.833
expanded	0.341	trombone	0.832
irregular	0.342	recipe	0.831
measure	0.346	antimony	0.829

Table 4: Twenty words with the lowest similarities and twenty words with the highest similarities.

We manually evaluated the 100 words that have the lowest similarities to the corresponding loanwords including the 20 words shown in Table 4. Among the 100 words, 21 words are influenced by ambiguity, and 19 are influenced by named entities. Among the remaining 60 words, 57 are judged to be semantically different from their loanwords. For the other three words, the embeddings would not be quite accurate probably due to their infrequency in either the English or the Japanese corpora used for training.

4.6 Evaluation for Educational Use

To see if the obtained word embeddings of English and Japanese can assist in language learn-

[15] http://rd.com

[16] In Japanese, *present* usually means a gift, or to give a gift, but hardly to show or introduce.

[17] In Japanese, this word usually means to match one's clothes attractively.

ing, for purposes such as lexical-choice error correction, we evaluate their usefulness by using the writings of Japanese learners of English. Specifically, we use the Lang-8 English data set (Mizumoto et al., 2011)[18] to calculate the Dice coefficient instead of JENAAD. This dataset consists of sentences originally written by learners, some of which have been corrected by (presumably) native speakers of English. Because we target embeddings of English and Japanese, we only use English sentences written by Japanese among other learners of English. Of those, approximately one million sentences have corresponding corrections. With these sentence pairs, we calculate the Dice coefficient just as in Section 4.2. The coefficient measures how often a word co-occurs in the original sentences and corresponding corrections. If a word is often corrected to another, it tends to appear only in the original sentences and not in the corresponding corrections, and thus, its Dice coefficient becomes small, and vice versa. In other words, the Dice coefficient roughly measures how often a word is corrected in the Lang-8 English data. Considering this, we compare the cosine similarity based on the proposed method with the Dice coefficient by means of the Pearson's correlation to evaluate how effective the cosine similarity is in predicting words in which lexical-choice errors likely occur[19]; the higher the correlation is, the better the cosine similarity is as an indicator of lexical-choice errors. Note that we apply lemmatization to all words both in the original sentences and in the corresponding corrections when calculating the Dice coefficient in order to focus only on lexical-choice errors[20].

It turns out that the value of the Pearson's correlation coefficient shows a milder correlation (ρ=0.302; significant at the significance level α=0.01) in this dataset than in JENAAD. Some loanwords having the almost equivalent senses in English have high values both for the cosine similarity and the Dice coefficient; examples are musical instruments such as *piano* (cos=0.886, Dice=0.951) and *violin* (cos=0.881, Dice=0.914); computer terms *computer* (cos=0.840, Dice=0.865) and *software* (cos=0.847, Dice=0.880) as has discussed in Section 4.5. Moreover, some that do not have equivalent senses show mild correspondences (e.g., *sentence* (cos=0.493, Dice=0.346); *note* (cos=0.470, Dice=0.352)).

By contrast, most of the others show less correspondence. One possible reason is that in the Lang-8 English data, corrections are applied to grammatical errors other than lexical choices, which undesirably decreases the Dice coefficient. Typical examples are errors in number (singular countable nouns are often corrected as corresponding plural nouns; e.g., *book* → *books*) and in inflection (e.g., *book* → *booked*). Therefore, loanwords whose corresponding English words undergo word-form changes less often tend to show strong correspondences as can been seen in the above examples (i.e., *software* and *piano*). This can be regarded as noise in the use of the Lang-8 data set. As mentioned above, we applied lemmatization to reduce the influences by noise. More sophisticated methods such as word alignment might improve the accuracy of the evaluation.

5 Conclusions

We computationally analyzed semantic changes of Japanese loanwords. We used the word embeddings of Japanese and English, and mapped the Japanese embeddings to the space of English, where we calculated the cosine similarity of a Japanese loanword and its original English word. We regarded this value as an indicator of semantic change. We evaluated our methodology in a number of ways.

To detect semantic changes accurately, we have to filter out the words that are ambiguous in the Japanese phonetic system. Such words tend to have low cosine similarities. One direction for future work is application of the technique to similar tasks. For example, we can use our method to analyze semantic changes of cognates. There are also a number of ways to investigate semantic changes of loanwords in more detail. For example, we can examine the relation between the semantic change of a loanword and the time at which the word was introduced in the target language. Hamilton et al. (2016) reported that they

[18] `http://cl.naist.jp/nldata/lang-8/`

[19] Some of the words in the loanword list are too infrequent to calculate the Dice coefficient in the Lang-8 data set. Accordingly, we excluded those appearing fewer than 30 times in it when calculating the Pearson's correlation.

[20] Other grammatical errors including errors in number and inflection often appear in the Lang-8 English data, which are mistakenly included in lexical-choice errors in the calculation of the Dice coefficient. Lemmatization reduces their influences to some extent.

used word embeddings to show the relation between semantic changes and polysemy. It would be interesting to see if similar results are obtained for loanwords.

Acknowledgments

This work was partly supported by Grant-in-Aid for Young Scientists (B) Grant Number JP26750091. Yoshifumi Kawasaki is supported by JSPS KAKENHI Grant Number 15J04335.

References

David Bamman, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834.

Keith Barrs. 2013. Assimilation of English vocabulary into the Japanese language. *Studies in Linguistics and Language Teaching*, 24:1–12.

Alan D. Blair and John Ingram. 1998. Loanword formation: a neural network approach. In *Proceedings of SIGPHON Workshop on the Computation of Phonological Constraints*, pages 45–54.

James Breen, Timothy Baldwin, and Francis Bond. 2012. Segmentation and translation of Japanese multi-word loanwords. In *Proceedings of Australasian Language Technology Association Workshop*, pages 61–69.

Jim W. Breen. 2000. A WWW Japanese dictionary. *Japanese Studies*, pages 313–317.

Frank E. Daulton. 2009. A sociolinguist explanation of Japan's prolific borrowing of English. *The Ryukokou Journal of Humanities and Sciences*, pages 29–36.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the European Chapter of Association for Computational Linguistics*, pages 462–471.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 748–756.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL2016)*, pages 1489–1501. Association for Computational Linguistics.

Yoichiro Hasebe. 2006. Method for using Wikipedia as Japanese corpus (in Japanese). *Doshisha studies in language and culture*, 9(2):373–403.

Mark Irwin. 2011. *Loanwords in Japanese*. John Benjamins Publishing Company.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 229–238. IEEE Press.

Gillian Kay. 1995. English loanwords in Japanese. *World Englishes*, pages 67–76.

Yoshiro Kojima, editor. 1988. *Nihongo no imi eigo no imi (meanings in Japanese, meanings in English (translated by the authors of this paper))*. Nan'un-do.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 230–237.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2014. Statistically significant detection of linguistic change. In *Proceedings of the 24th World Wide Web Conference (WWW)*, pages 625–635.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? quantifying the geographic variation of language in online social media. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, pages 615–618.

Lingshuang Jack Mao and Mans Hulden. 2016. How regular is Japanese loanword adaptation? A computational study. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 847–856.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 147–155.

Sen Nishiyama. 1995. Speaking English with a Japanese mind. *World Englishes*, pages 1–6.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, pages 1–38.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 72–79.

Will Zou, Richard Socher, Daniel Cer, and Christopher Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1393–1398.