Applying Word Embeddings to Measure the Semantic Adaptation of English Loanwords in Japanese and Korean

Akihiko Yamada
(Seoul National University)
Hyopil Shin
(Seoul National University)

Akihiko Yamada and Shin, Hyopil. (2017). Applying Word Embeddings to Measure the Semantic Adaptation of English Loanwords in Japanese and Korean. *Language Research*, 53.3, 473-500.

With the internationalization of English, many English words are used as loanwords around the world. These English loanwords can bring about confusion, however, as the semantic usage of a loanword often differs from region to region. Therefore, it is important to examine how the semantic adaptation of a given English loanword has occurred when it is introduced into the lexicon of a country. The word vector model has been used to study semantic changes. In this paper, using these techniques, we investigate the semantic adaptation of English loanwords in Japanese and Korean. In addition, we investigate the correlation between the number of meanings of original English words and cosine similarity. As a result, we have brought a new insight into the computational contrastive semantics.

Keywords: loanword, semantic difference, distributional semantic model, word embedding, Word2vec

1. Introduction

In recent decades, English has become an international language. English is spoken as the native language in several countries and taught as a second language in many more. Over the course of English's rise as an international language, many English words have had influence

on the native languages of countries where English is not the mother tongue. Foreign words are often incorporated into a language in order to express a specific concept that cannot be expressed using the words of the mother tongue alone. For example, consider the word *resident*. *Resident* means '*a person staying in a specific area*' and '*a person who is training to be a doctor*' in English. However, *resident* as a loanword is mostly used with the second meaning in Japanese and Korean, because these two languages each have a native word for the first meaning of *resident*. This example shows that some of the original meanings of a loan word are not used in foreign countries (Kay 1995, Okawa 2008, and SM Cheon 2008). Furthermore, loanwords are often used figuratively. In Japanese and Korean, the word *corner* indicates not only '*a positional area*', but also '*a section provided for a specific purpose*'. The word *stand* is also frequently used to mean '*a desk lamp*' in Japan and Korea. Due to this phenomenon, the same English word is often used differently depending on the language. This semantic difference can pose a challenge in computational tasks such as machine translation and information retrieval. In addition, the semantic difference of loanwords can also pose a challenge to language learners. For these reasons, the task of investigating the nature of a semantic adaptation when a word enters from a foreign language is an important one.

In order to deal with the challenges posed by loanwords, it is first necessary to develop a methodology for detecting the meaning difference of loanwords. To this end, we review the previous studies of computational models for word meaning change. Kulkarni et al. (2014) propose a new computational approach for tracing change of meaning and usage of words from a historical perspective. They construct a property time series of word usage and apply statistically sound change point detection algorithms to show the semantic change. The result shows interesting patterns of language change. Hamilton et al. (2016) compare three major computational methods, PPMI, SVD, word2vec, and develop a powerful methodology for quantifying historical semantic change. They also tackle linguistic complications related to historical semantic change-the relationships between semantic change and word frequency and between semantic change and polysemy. As a result, they propose two quantitative laws of semantic

change. Takamura et al. (2017) apply a word vector space model for semantic changes in Japanese loanwords. They train a word vector space model with English and Japanese text data and map Japanese loanword vectors onto the English vector space. After that, a Japanese loanword's vector is compared with an original English word vector according to their cosine similarity. This method is evaluated by several tests and is verified as a reliable method for studying semantic change in loanwords. As demonstrated in these previous studies, the word vector space model is considered one of the most powerful methods for detecting differences in word meaning. Based on these previous studies, it is highly probable that the word vector space model is also powerful for detecting English loanwords as well as their semantic adaptation.

In fact, Fenogenova et al. (2017) try to apply the word vector space model to detect English loanwords in Russian data. Their detection method is based on the idea that the original Latin word is similar to its Cyrillic analogue in terms of scripting, phonetics and semantics. They also assume that English loanwords and their original English words should be close in their meanings; their vector value is also similar. On this assumption, they develop a filtering system for detecting real loanwords from several loanword candidates in Russian data. As a result, they improve the accuracy of detecting English loanwords. However, their method only manipulates the loanwords that have the same meaning as the original English word. Thus, in this paper, we apply the word vector model to the task of detecting English loanwords whose semantic usage is different from its source English word and for measuring the degree of its semantic adaptation.

In addition to this methodological purpose, we verify the relationship between polysemy and meaning adaptation. As mentioned earlier, the main purpose of using loanwords is introducing a new concept. Thus, loanwords will tend to have only a part of the meaning that the word originally had. Given this supposition, it can be predicted that if an original word has several meanings (polysemy), the meaning between loanwords and the original English word will be much different. Hamilton et al. (2016) study the relationship between polysemy and the meaning change

of a word, but they study only from the perspective of historical meaning change and do not investigate the relationship from the point of view of meaning change in loanwords. In order to verify this prediction, we examine the relationship between the number of original meanings of the English word and the degree of semantic adaptation using the word vector model.

2. Word Embedding

For processing natural language with a computer, it is necessary to represent words with numeric values. One of the methods of converting word meaning to a numeric value is to use a distributional model. This method is based on the hypothesis: word meaning depends on the context in which the word appears. This hypothesis was formed based on the work of Wittgenstein (1997). He wrote that "the meaning of a word is its use in the language". And this concept is applied to a practical language model: the distributional model. The distributional model predicts a semantic similarity based on the distributional hypothesis (Harris 1954): if two words tend to occur in similar contexts, we can assume that they are similar in meaning.

Distributional models are typically carried through high-dimensional vector space. In these models, the representation for a word is a point in a high-dimensional space. The dimensions stand for context items (for example, co-occurring words), and the coordinates depend on the co-occurrence counts.

There are several methods for converting a word to a vector. One uses a neural network, a mathematical model inspired by neural cells and their connections in a human brain. Neural networks are composed of an output layer, one or more hidden layers, and an input layer. Weight "W" indicates the strength of the connections of neurons from one layer to those in another. Figure 1 provides an outline of a neural network. Neurons in a human brain communicate through electrical signals. The extent of the information transmitted depends on the strength of the bond

between the synapses. This mathematical model represents the strength of a bond between artificial neurons with the weight "W". When a data set is input to the input layer X1, it is then multiplied by the weight W1 and the resulting value Y1 is output in the hidden layer. Next, W2 is multiplied by the value of the Y1, and then the value Z1 is output in the output layer. Finally, the Z1 value is compared with the actual value in the data set. Ideally, this process will repeat until "W" converges to a value that produces a Z1 that best approximates the observations found in the data set, i.e., the final output value is as close as possible to the value of the actual observed quantity the model is trying to predict.

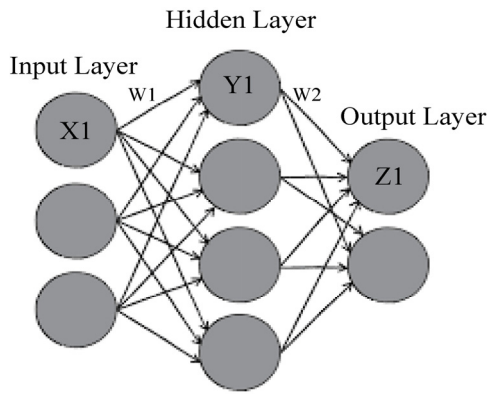

Figure 1. The basic neural network model.

Currently, in the field of natural language processing and artificial intelligence research, the word embedding using this neural network is mainstream. Benjio (2003) represents the distributional feature of word meaning as a computational statistical language model through the successful application of neural networks. Following this lead, many computational linguistic researchers have started to use a distributional model for tasks such as machine translation, information retrieval and sentiment analysis. Mikolov et al. (2013a) developed a new calculation method, the Skip-gram, and applied it to the neural network model. As a result, a word can be vectorized more effectively and more accurately. This neural network model is known as Word2Vec.

Word2Vec is an effective tool for representing word meaning as a vector, and there are several reasons for why Word2Vec is a good tool for language studies. Firstly, the model can rapidly process a large amount of data. Thus, a vector can be generated based on rich context information. This this helps to ensure that the vector represents the actual meaning of the word accurately.

The second reason is the ease of the vectorization. In conventional machine learning methods, a data set labeled by human hands is required. Preparing labeled data is costly in terms of both time and money. However, a labeled data set is not necessary with a neural network model. The model calculates and generates the meaningful patterns of words from a large amount of non-labeled data. This makes computational semantic research more efficient and less expensive.

Finally, differences between the meanings of words can be calculated mathematically in this model. Since the meaning of the word is represented by a vector, the computer calculates the difference of meanings using a simple vector calculation. Normally, the cosine similarity is used as the index of the difference of meaning in this model. Cosine similarity is equal to the cosine value of two word vectors. If the meanings of the two words are similar, the word vector values are also similar and the angle between the two word vectors is near zero. When the cosine value of the angle is almost 1, the cosine similarity is also near 1. Conversely, the angle between the words which measures the difference between the meanings of the words approaches -1 is it grows larger.

Another interesting aspect of this model is the handling of the addition and subtraction of meaning. The output vector encodes the explicit number of linguistic regularities or patterns. Surprisingly, a lot of these patterns can be expressed as a linear translation. This allows for the addition and subtraction of meaning. Concretely, a vector calculation like "vec ("*Tokyo*") - vec ("*Japan*") + vec ("*Korea*")" is closer to the value of vec ("*Seoul*") than the to the value of the vectors of the other words. As mentioned above, a neural network model of semantic meaning is utilized in many area of computational linguistic research. With this as our context, we use Word2Vec in this paper.

3. Methodology

We use the word vector model for detecting English loanwords that have different meanings from their source words and for measuring the degree of their semantic adaptation. For this purpose, the Word2vec skip-gram model (Mikolov et al. 2013a) is chosen to generate the word vector space model with reference to Hamilton et al.'s work (2016) and Takamura et al.'s work (2017). We chose English, Japanese and Korean, because English loanwords that are semantically distinct from their source words are abundant in both Japanese and Korean. At first, we create word embedding for the three languages: English, Japanese and Korean. Next, we calculate the cosine similarity and dissimilarity between the original English words and their Japanese or Korean loanword counterparts. For this purpose, the two language's words should be represented in the same vector space with the same coordinates. For mapping the embeddings into the same vector space, we choose one of the simplest methods developed by Mikolov et al. (2013b). The method is represented by the equation below. By calculating the equation using seed words, the transformation matrix W is obtained. To make the bilingual seed word pairs, we used the most frequent nouns from monolingual source data sets, and translated those words using Google Translate like Mikolov et al. (2013b). By multiplying the value of an English loanword vector in Japanese or Korean by the transformation matrix W, it becomes possible to compare the loanword vectors in the English word vector space.

$$\min_{W} \sum_{i=1}^{n} \| Wx_i - z_i \|^2$$

After this transformation, we can get the N-nearest neighbors of the English loanword in the English vector space and can calculate the cosine similarity between the English loanwords and the original English words. If the value of cosine similarity is low, it shows that the English loanword meaning is very different from the original word, and thus we can detect the English loanwords that are used with significantly different meanings

in Japanese and Korean. In the next section, we present our data set and experiment for English loanword detection in Japanese and Korean.

4. Data and Experiment

The data set used for training Word2vec was obtained from Wikipedia dump data[1] in May of 2017 for English, Japanese and Korean. The text data was extracted by a Wikipedia extractor[2] from each Wikipedia dump data set. The English Wikipedia data is 13.6 GB, the Japanese Wikipedia data is 2.5 GB and the Korean Wikipedia data is 606 MB. In the case of English text data, non-alphabetic symbols are removed and all alphabetic characters are lowered. For Japanese data, word segmentation is done using the Japanese morphological analyzer MeCab (Kudo et al., 2004). For Korean text data, we apply the open-source Korean text tokenizer Twitter[3] for Korean Text. These preprocessed data are used for training Word2vec (dimensions = 200, min count = 20, window size = 15) in the Gensim[4] Python package.

For calculating the transformation matrix, a bilingual word list -- English-Japanese word list and English-Korean word list -- is necessary. In this experiment, a bilingual list is prepared according with the method of Mikolov et al. (2013b). Mikolov et al. (2013b) selects the high frequency words from English corpus and translates them into another language with Google translator. It may be easy to make a bilingual list in Spanish or French for calculating transformation matrix but difficult in the case of Japanese and Korean because these languages have very complex inflection system. This complex inflection system makes one-to-one mapping of English words to Japanese (or Korean) words in the language database difficult. This is a source of various difficulties when training the trans-

1) https://dumps.wikimedia.org/enwiki/
https://dumps.wikimedia.org/jawiki/
https://dumps.wikimedia.org/kowiki/.
2) http://medialab.di.unipi.it/wiki/Wikipedia_Extractor.
3) https://github.com/twitter/twitter-korean-text.
4) https://radimrehurek.com/gensim/index.html.

formation matrix, and as a result, it may be difficult to obtain an accurate transformation matrix. For example, when Google Translate makes a bilingual list of English and Japanese (or Korean) words, Google Translate translates "*eat*" to "*mekta*" and "*beautiful*" to "*alumtawun*". If you calculate the transformation matrix using this bilingual list, "*eat*" is mapped with "*mokta*" and "*beautiful*" is mapped with "*alumtawun*". However, "*mokta*" is actually used in a different form such as "*mokulye*" or "*mokess*" in texts. Similarly, "*alumtawun*" is used in different forms, such as "*alumtawess*" or "*alumtapta*". Therefore, if the bilingual list created by Google Translate is used to get a transformation matrix, other forms of "*mokta*" and "*alumtawun*" are ignored in the process of calculation. As a result, the transformation matrix based on this bilingual list will not be accurate. Thus we chose high frequency English nouns because noun has little inflection in Japanese and Korean. After translating these English nouns into Japanese and Korean, loanwords are removed for training transformation matrix properly. Finally, we compute the transformation matrix with about 5000 word pairs in the list.

After learning Word2vec with using the Wikipedia data and applying transformation, the nearest neighbors are used to check whether the Word2vec and the transformation matrix are trained properly. By way of example, several loanwords which meaning is different from the original English word are selected from previous studies (H-s Min 1998, MS Choi 1996). Table 1 shows the nearest neighbors of Korean loanwords in English vector space.

Table 1. Nearest Neighbors of Korean Loanwords in English Vector Space

Original English Word	Korean Loanword	The nearest neighbors of Korean loanwords in English vector space
consent	khonsenthu	device, grommet, rack, tube, generator
corner	khone	episode, diner, cabbie, show, skit
maker	meyke	brand, producer, toy, designer, maker
propose	phuropocu	girl, kiss, wedding, bride, princess
stand	sutayndu	corner, ramp, lock, ball, cleat
talent	thayllenthu	actress, performer, actor, dancer, entertainer

The target loanwords that we study the semantic differences of in this research are selected from the loanword list distributed by the National Institute of Korean Language.[5] In order to calculate cosine similarity, it is necessary to prepare bilingual loanword pair lists: an English-Korean loanword pair list and an English-Japanese loanword pair list. For obtaining these bilingual loanword lists, the Korean loanwords are translated to English and Japanese with Google Translate. We create the English-Korean loanword and the English-Japanese loanword lists using these translations and then calculate the cosine similarities of the bilingual loanword pairs in these lists. All experimental processes in this study are summarized in Figure 2. In the next section, we present the result of this experiment.

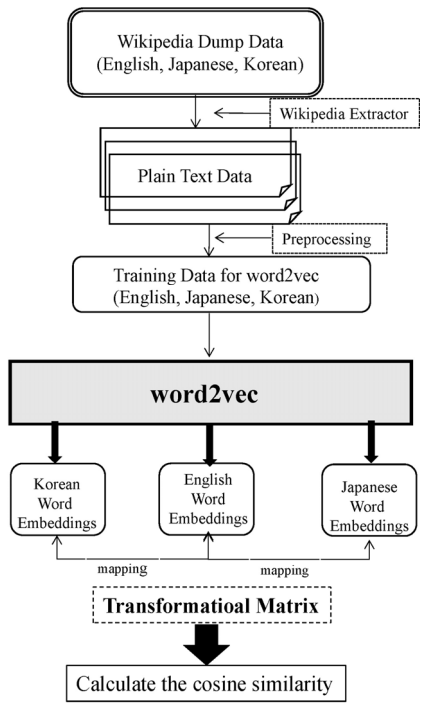

Figure 2. The experimental scheme of measuring the semantic adaptation of English loanwords in Japanese and Korean.

5) http://www.korean.go.kr/front/etcData/etcDataView.do;front=8E3DC144E9BBA954E0BE198B8481950A?mn_id=46&etc_seq=322&pageIndex=1.

5. Result and Discussion

In this section, we demonstrate how accurately the word vector model finds the differences in semantic usage of loanwords in Japanese and Korean. The value of cosine similarity is calculated based on the bilingual list that was explained in detail in section 4. By observing the loanword pairs with low cosine similarities, we can see that there are some outputs that have been calculated with mistakenly paired words — i.e., the loanword is not the loanword of the original English word. These errors are the result of mistranslations made by Google Translate when the bilingual loanword list was generated. The 20 lowest cosine similarity word pairs left following the removal of these errors are represented in tables in each of the following subsections.

Moreover, there remains the possibility that a learning error of word2vec produced these low cosine similarities. Thus, in the case of low cosine similarity word pairs, the frequency of the word in each language corpus is also shown. In addition to word frequency, the N-nearest neighbors of several words in low cosine similarity word pairs are shown for the purpose of checking the accuracy of the word2vec learning process.

In section 5.4, we also discuss the relationship of cosine similarity with the number of original meanings possessed by the English words as described at the beginning of this paper.

5.1. Japanese

Table 2. Twenty English Loanwords with High and Low Cosine Similarities with their Corresponding English Source Words

Similar pairs in Japanese		Dissimilar pairs in Japanese	
English word	cosine	English word	cosine
concert	0.868	spoke	-0.230
ideology	0.866	curl	-0.205
design	0.857	vantage	-0.182
tunnel	0.856	felt	-0.165
robot	0.845	all	-0.089

Similar pairs in Japanese		Dissimilar pairs in Japanese	
English word	cosine	English word	cosine
data	0.842	lighter	-0.088
engine	0.841	cant	-0.083
text	0.840	won	-0.068
model	0.839	dipole	-0.063
message	0.833	plank	-0.054
knife	0.830	centering	-0.033
curriculum	0.828	sage	-0.032
system	0.828	miscast	-0.027
silhouette	0.818	emery	-0.025
campus	0.817	tax	-0.023
hotel	0.817	fume	-0.004
energy	0.816	choline	0.000
radar	0.816	wake	0.003
nationalism	0.816	saving	0.005
approach	0.811	carry	0.006

Table 3. The frequency of words in low cosine similarity pairs in each language corpus

Word	Frequency in English corpus	Frequency of loanwords in Japanese corpus	Word	Frequency in English corpus	Frequency of loanwords in Japanese corpus
spoke	56240	684	centering	2739	108
curl	2792	19455	sage	15149	55
vantage	3048	37	miscast	313	27
felt	134917	1291	emery	4847	99
all	2803188	9670	tax	136584	44
lighter	22285	6520	fume	596	450
cant	11681	1558	choline	616	2301
won	1029101	2542	wake	40052	122
dipole	3149	23	saving	31286	50
plank	4949	635	carry	101652	1628

Table 4. The Nearest Neighbors of the Lowest Cosine Similarity Pairs in English and Japanese (Japanese Words are Translated to English Words as Described in Parenthesis in this Table)

Word	nearest neighbors of original English word	nearest neighbor of Japanese loanword	Word	nearest neighbors of original English word	nearest neighbor of Japanese loanword
spoke	talked speaks speaking speak wrote	akusuru (axle) syarin (wheel) brekikyaripa (brake caliper) bearingu (bearing) spuroket (spurocket)	vantage	jumping-off view stopping-off focal panoramic	bandēji (bandage) hōtai (bandage) tēpingu (taping) tatakitukeru (beat) gāze (gauze)
curl	curls jheri curled rip eyelash	hansu (Hans) adorufu (Adolf) furantsu (Franz) hainrihi (Heinrich) arekusandā (Alexander)	lighter	heavier thicker thinner softer slimmer	furiraitā (free writer) koramunisuto (columnist) jyānarisuto (journalinst) purodyūsā (producer) direkutā (director)
felt	opined remarked commented remarking feels	beruku (Belc) rōto (Roth) apusyutatto (Abstatt) kanpu (Kampf) main (Main)	centering	centered centring focusing centred focusses	kōnākikku (corner kick) hedhingu (heading) kurosubōru (cross-ball) rongupasu (long shot) midorusyūto (middle shot)

Table 3 shows the cosine similarities between an original English word and the corresponding English loanword in Japanese (the Japanese English loanword). For your reference, the frequency of low cosine similarity words in English and Japanese is given in Table 3. Table 4 contains examples of the nearest neighbors of the lowest cosine similarity pairs in English and Japanese. As can be seen from *vantage, dipole* and *saving* in Table 3 and Table 4, we may safely say that word2vec learns the semantic information of the word properly even if the word frequency is only around 20-50.

In Table 2, the left column shows the original English words that have high cosine similarities with their corresponding Japanese English loanwords, and the right column shows the English words that have low cosine similarities. In the left column, almost all words are technical words such as *ideology*, *curriculum*, *engine*, *energy*, *data*, *nationalism* etc. This result corresponds to the findings of Takamura et al. (2017).

In the right column, several words show interesting tendencies of meaning adaptation. For example, consider the word *spoke*. The English *spoke* means *'the thin metal bars which connect the outer ring of a wheel'*. The Japanese English loanword of English *spoke-spōku*-is used in the same sense as English *spoke*. But, English *spoke* has the same spelling as the past form of the English verb *speak*. Thus, English spoke can be used in a wider variety of contexts than the Japanese English loanword *spōku*. As indicated in Table 4, English *spoke* is actually learned as the past form of *speak* in the word2vec model. The low value of the cosine similarity of "*spoke*" indicates this meaning ambiguity.

Next, let us take English *felt* and Japanese *feruto*. As can be seen in Table 4, English *felt* is used as the past form of *feel*. On the other hand, Japanese *feruto* is used as the name of a city. The low cosine similarity is interpreted as the difference of meaning usage of *felt* and *feruto* in table 4. The explanation can be applied to the English word *curl*. The original meaning in English is *'to form a twisted or curved shape'* and Japanese *kāru* is also used with the same meaning. However, as described in Table 4, the word2vec model detects that Japanese *kāru* is also frequently used as a name of a person in Japanese. These examples indicate the low cosine similarities that result from the usage of loanwords for city names or person names.

Finally, according to Table 4, the ambiguous spelling system of Japanese words — a loanword with the same spelling is used to express a variety of different English words — has a close relationship with the value of low cosine similarity. This relationship can be observed in the low cosine similarity value between English *lighter* and the Japanese English loanword *raitā*-the loanword has the meaning of both *writer* and *lighter*. The word2vec model indicates *lighter* means the comparative form of *light* in English

data and *raitā* means several professions related to *writer* in Japanese data. Based on such examples, we can conclude that the ambiguous spelling system of Japanese can be detected in the word2vec model. Similarly, the word2vec model detects the word sense ambiguity of English *vantage* and *centering*. English *vantage* means *'a good position from which you can see something'*, but the Japanese *vantēji* is used as a name for a protective material used in fighting sports. In this example, English *vantage* and Japanese loanword *vantēji* are used with completely different meanings. For this reason, cosine similarity is low between the English *vantage* and the Japanese English loanword *vantēji*. English *centering* and Japanese loanword *sentaringu* are similar: English *centering* means "*to move something to the center position*" while Japanese *sentaringu* means "*the cross ball in a soccer game*" in Japan.

From these examples, it is shown that this word vector model detects the several patterns of meaning adaptation of English loanwords in Japanese. In summary, out of the 20 lowest cosine similarity pairs, 8 words are influenced by the name of a person or city, 7 words are influenced by word sense ambiguity. It is probable that the other 5 words are inaccurately identified in the learning process of the word2vec model.

5.2. Korean

Table 5. Twenty English Loanwords which have High and Low Cosine Similarities with their Corresponding English Source Words

Similar pairs in Korean		Dissimilar pairs in Korean	
English word	cosine	English word	cosine
software	0.849	zone	-0.251
internet	0.838	felt	-0.231
Spain	0.836	spoke	-0.201
ideology	0.835	canter	-0.162
journalist	0.828	catarrh	-0.141
logo	0.824	on	-0.132
berlin	0.822	stole	-0.130
fascism	0.820	stepping	-0.118
energy	0.817	carry	-0.113

Similar pairs in Korean		Dissimilar pairs in Korean	
English word	cosine	English word	cosine
henry	0.811	under	-0.082
message	0.810	won	-0.076
marketing	0.810	combine	-0.068
producer	0.809	sling	-0.066
tunnel	0.808	camel	-0.057
Syria	0.808	polling	-0.055
season	0.805	pinion	-0.054
text	0.805	leak	-0.051
Ukraine	0.803	peace	-0.053
project	0.802	current	-0.046
network	0.802	wife	-0.033

Table 6. The Frequency of Words in Low Cosine Similarity Pairs in each Language Corpus

Word	Frequency in English corpus	Frequency of loanwords in Japanese corpus	Word	Frequency in English corpus	Frequency of loanwords in Japanese corpus
zone	129495	12904	won	1029101	14744
felt	134917	709	combine	24331	163
spoke	56240	27	sling	2212	134
canter	892	33	camel	7904	22
catarrh	108	1031	polling	12545	201
on	16944404	16270	pinion	1018	78
stole	13091	283	leak	9152	880
stepping	9687	24	peace	165246	1025
carry	101652	575	current	410770	21
under	1679197	280	wife	397592	61

Table 7. The Nearest Neighbors of the Lowest Cosine Similarity Pairs in English and Korean (Korean Words are Translated to English Word as Described in Parenthesis in this Table)

Word	nearest neighbors of original English word	nearest neighbor of Korean loanword	Word	nearest neighbors of original English word	nearest neighbor of Korean loanword
zone	zones region zone's area regions	jeimsu (James) ropethu (Robert) thoaesu (Thomas) ayntulu (Andrew) wiliem (William)	won	winning earned competed finished defeated	manwen (ten thousands won) yewen (the name of magazine) maneyn (ten thousands yen) eyn (yen) ekwen (hundred million won)
felt	opined remarked commented remarking feels	beylukhu (Berg) hophu (Hop) haim (Haim) eylunsuthu (Ernest) phulichu (Pretz)	combine	integrate infuse incorporate harmonize utilize	aie (imaginary planet in a game) oykyeyin (alien) kholloni (colony) cikwuin (earth human) sithateyl (fortress)
catarrh	dropsy tuberculosis hemorrhoids toothaches inflammations	alapeymilithu (the United Arab Emirates) alahulli (Al-Ahli) khwuweyithu (Kuwait) sawutialapia (Saudi Arabia) paleyin (Bahrayn)	camel	camels dromedary dromedaries horse mule	pansu (the name of brand) cotan (Jordan) pulwum (Broome) phiswi (fish) khlayp (crab)

Table 5 shows the cosine similarities between an original English word and the corresponding English loanword in Korean (the Korean English loanword). The frequency of low cosine similarity words in English and Japanese are shown in Table 6. Table 7 reveals the examples of nearest neighbors of the lowest cosine similarity pairs in Table 5.

In Table 5, the left column shows the English words that have high cosine similarities with their corresponding English loanwords in Korean and the right column shows the English words with low cosine similarities with their corresponding loanwords. In the left column, almost all words

are technical words such as *software*, *internet, energy, network, message, project, producer* or academic words such as *ideology, fascism,* and *marketing*. This result is almost the same as in the Japanese data set. From this result we can observe the tendency of technical word meanings to remain constant, which was also observed by Nishiyama (1995); this observation appears to also be applicable in the case of semantic adaptation of English loanwords in Korean.

Next, we consider the right column of Table 5. In the case of *zon*e, Korean *jon* (the Korean English loanword of *zone*) is used not only for *zone* but also *John*; a name for a person. This difference in usage is indicated by a low cosine similarity. English words *catarrh, felt, combine* and *camel* are similar cases to *zone.* The English *catarrh* is used as a medical term but the nearest neighbors in word2vec indicate the English loanword *katareu* often means the name of a country (*Qatar*). Additionally, the nearest neighbors of Korean loanword *pheylthu* indicate *pheylthu* means not only the name of a city or area but the past form of *feel*. Because of these differences in semantic usage, the cosine similarity of *catarrh* and *felt* is low. The English words *camel* and *combine* show low cosine similarity for the same reason. These patterns of meaning adaptation are consistent with the Japanese *kāru* and *feluto.*

The English word *won* also shows an interesting result. The English word *won* is mainly used to indicate the past form of *win*. However, Korean English loanword *won* is used as a unit of money. Due to this difference in usage, the cosine similarity with English turns out to be very low. In summary, out of the 20 lowest cosine similarity pairs in Table 5, 10 words are influenced by the name of a person, city or product, 3 words are influenced by word sense ambiguity. The other 7 words are identified as inaccurately paired in the learning process of word2vec.

From these examples, it can be said that in Korean data the word vector model detects the several tendencies of meaning adaptations in English loanwords in Korean, but the accuracy is lower than the results of the English-Japanese comparison in section 5.1. This may be due to the size of data set.

5.3. Comparison of Cosine Similarities of English Loanwords in Japanese and Korean

In this section, we will present a contrastive study of the difference in semantic usage of English loanwords between Japanese and Korean. We calculate the cosine similarity between Japanese and English and the cosine similarity between Korean and English as we did in section 5.2. After calculating these cosine similarities, we compare the vector values of Japanese English loanwords and Korean English loanwords in order to find English words whose semantic usage in Korean and Japanese are highly distinct. The result of this comparison of cosine similarity is shown in Table 4 and Table 5. In both tables, cosine "C_j" means the cosine similarity between an English word and its corresponding Japanese English loanword and cosine "C_k" means the cosine similarity between an English word and its corresponding Korean English loanword. Thus "C_j-C_k" means the difference between the meanings of English loanwords in Japanese and in Korean. We used different data sets to train the Japanese word2vec and the Korean word2vec models, so comparing the values directly may prove challenging. Nevertheless, we were able to detect the tendencies of semantic usage difference between Japanese and Korean through this contrastive study.

Table 8. The Top Twenty English Loanwords whose Cosine Similarity with their Japanese English Loanword Counterparts is Higher than with their Korean English Loanword Counterparts

The top twenty English loanwords whose cosine similarity with their Japanese English loanword counterparts is higher than with their Korean English loanword counterparts

English word	cosine of Japanese loanword (C_j)	cosine of Korean loanword (C_k)	difference of cosine between Japanese and Korean (C_j-C_k)
jeep	0.696	0.050	0.646
cottage	0.719	0.120	0.599
cost	0.752	0.162	0.590
hall	0.773	0.203	0.570
bed	0.735	0.180	0.555

English word	cosine of Japanese loanword (C_j)	cosine of Korean loanword (C_k)	difference of cosine between Japanese and Korean (C_j-C_k)
caption	0.670	0.131	0.539
passport	0.807	0.279	0.528
shocking	0.611	0.099	0.511
crew	0.675	0.173	0.502
antique	0.679	0.180	0.499
wit	0.623	0.126	0.497
gauze	0.705	0.209	0.496
chuck	0.506	0.014	0.492
saccharin	0.621	0.138	0.482
rope	0.773	0.293	0.480
angle	0.496	0.018	0.477
professional	0.490	0.027	0.463
type	0.705	0.254	0.451
total	0.491	0.042	0.449
rescue	0.672	0.223	0.449

Table 8 shows the English words whose cosine similarity with their Japanese English loanword counterpart is higher than their cosine similarity with their Korean English loanword counterpart. In other words, the meaning of the English word in Table 8 is similar to the meaning of the Japanese English loanword and dissimilar with the Korean English loanword. For example, in the case of the English word *cost*, the Japanese English loanword is also used with the meaning of *cost*, but the Korean English loanword *koseuteu* is used not only as the loanword of the English *cost* but also as the loanword of the English *coast* (with no change in spelling). This difference in usage in the loanwords between the two languages is shown as a difference of cosine similarity in Table 8.

Table 9. The Top Twenty English Loanwords whose Cosine Similarity with their Japanese English Loanword Counterparts is Lower than with their Korean English Loanword Counterparts

The top twenty English loanwords whose cosine similarity with their Japanese English loanword counterparts is higher than with their Korean English loanword counterparts.

English word	cosine of Japanese loanword (C_j)	cosine of Korean loanword (C_k)	difference of cosine between Japanese and Korean (C_j-C_k)
label	0.268	0.687	-0.419
fax	0.249	0.643	-0.394
Olympiad	0.275	0.644	-0.368
pierce	0.232	0.598	-0.366
Versailles	0.261	0.600	-0.339
midi	0.229	0.562	-0.334
midfielder	0.390	0.699	-0.309
hood	0.231	0.530	-0.299
victor	0.210	0.507	-0.297
fuse	0.220	0.513	-0.293
walker	0.409	0.674	-0.264
tint	0.195	0.455	-0.260
carrier	0.228	0.487	-0.259
demo	0.282	0.538	-0.256
Colosseum	0.410	0.663	-0.253
editor	0.223	0.461	-0.238
foundation	0.196	0.434	-0.237
break	0.197	0.431	-0.234
close-up	0.365	0.599	-0.233
roleplaying	0.202	0.430	-0.228

Table 9 shows English words whose cosine similarity with their Japanese English loanword counterparts is lower than their cosine similarity with their Korean English loanword counterparts. In other words, the semantic usage of the English words in Table 9 is more similar with the semantic usage of their Korean English loanword counterparts more than with that of their Japanese English loanword counterparts. From a linguistic

viewpoint, several English words show an interesting tendency in their semantic adaptation.

To begin with, we consider the English word *Versailles*. *Versailles* is the name of a city but the Japanese English loanword *verusaiyu* is used as part of a Japanese cartoon's title. Thus the English loanword *verusaiyu* can be used in several contexts in Japanese. The cosine value of *verusailles* reflects this disparity in meaning. Another example is the English word *midi. Midi* is a technical term related to musical devices and systems. In Korean, the Korean loanword *midi* also has this meaning, but the Japanese English loanword *midi* is used not only with this meaning, but also as the name of a music company. As with our previous examples, this difference in usage is reflected in their cosine similarity scores.

Next, the English *hood* is also an interesting example. The English loanword of *hood* in Japanese is *fūdo. Fūdo* is used not only as the loanword of English *hood,* but also as the loanword of English *food*. This means that while *hood* and *food* are completely different words in English, the loanwords have the same spelling in Japanese. As a result, the cosine similarity between the English *hood* and the Japanese English loanword *fūdo* is quite low. These complicated phonetic systems and the spelling rules turn out to have a great deal of influence on the cosine similarity and the concomitant difference in the meanings of English loanwords.

Through our observation of these words and their cosine similarities, we can gain insight into the process of semantic adaptation of English loanwords in Japanese and Korean. To the best of our knowledge, this paper is the first trial of contrastive semantic study between two languages using a word vector space model. Based on our results, the word vector model seems to be a powerful tool for investigating the semantic differences between several languages. We hope that this experiment can offer some inspiration for new horizons of future research in contrastive semantics and contrastive linguistics.

5.4. The Relationship Between the Number of Meanings and Cosine Similarities

In this subsection, we investigate the relationship between the number of meanings of an English word and the degree of difference in its counterpart loanword's semantic usage. In many cases, loanwords tend to have certain specific meanings that cannot be expressed in the foreign language. Considering this tendency, it can be presumed that the difference of the meaning usage between the loanword and the original word will be large if the original word has a large number of meanings. In order to verify this hypothesis, we first calculated the number of meanings in the original English words using WordNet[6]. Wordnet was chosen because it is freely available and has been used in many previous studies of computational semantics. We set the number of synsets in WordNet as the number of meaning of the English word. We draw a plot showing the number of synsets of an original English word and its corresponding cosine similarity between it and its loanword. After that, we calculate the regression line which best summarizes the scatterplot. Figure 3 and 4 are the plots for the Japanese data set and the Korean data set, respectively.

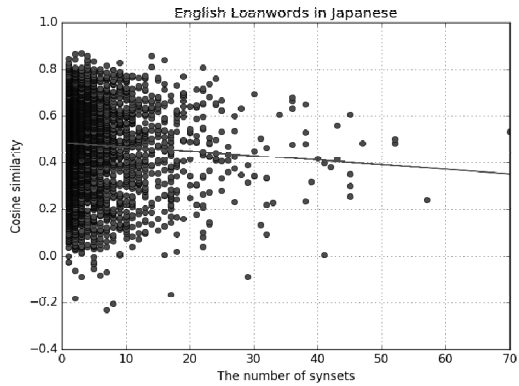

Figure 3. The correlation between number of synsets and cosine similarity between English words and their Japanese English loanword counterparts.

6) https://wordnet.princeton.edu.

Table 10. The Statistical Values of the Relationship between Number of Synsets and Cosine Similarity in English Loanwords in Japanese

variable	Coef	Std Err	t-stat	p-value	CI 2.5%	CI 97.5%
x	-0.0019	0.0005	-4.28	0	-0.0028	-0.0011
intercept	0.4867	0.0031	156.42	0	0.4806	0.4928

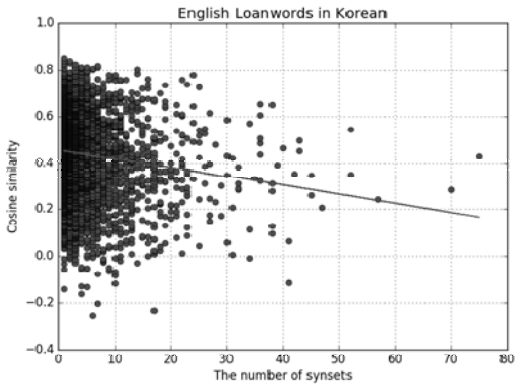

Figure 4. The correlation between number of synsets and cosine similarity between English words and their Korean English loanword counterparts.

Table 11. The Statistical Values of the Relationship between Number of Synsets and Cosine Similarity in English Loanwords in Korean

variable	Coef	Std Err	t-stat	p-value	CI 2.5%	CI 97.5%
x	-0.0039	0.0005	-8.08	0	-0.0048	-0.0029
intercept	0.4584	0.0036	127.37	0	0.4513	0.4654

Firstly, in the case of Japanese data, the slope of the regression line is -0.0019 and the P value is zero (rounded to zero by Pandas, a statistical programming package). As a result, we know that the number of synsets has a significant negative correlation with cosine similarities in Japanese data. In the case of Korean data, we learn the same lesson. The slope of the regression line is -0.0039 and the P value is zero (again rounded to zero by Pandas), which again suggests that the two factors are negatively correlated. Based on these observations, it appears that the semantic usage difference between an original English word and English loanword largely

occurs in cases where an original English word that has many meanings is used as a loanword which has only one certain meaning in order to indicate a specific concept in a foreign country.

6. Conclusion

With the internationalization of English, English is being used throughout the world, and many English words have come into foreign languages. In the case of these English loanwords, the specific meaning that is needed to be used in a given country is selected among the various meanings of the original English word. In addition to this, these English loanwords are also used figuratively or as a product name. Therefore the semantic usage of English loanwords tends to differ from country to country. This tendency can complicate applications in fields such as machine translation and foreign language education.

In order to solve this problem, it is important to detect loanwords which have different semantic usages and to determine to what extent semantic differentiation occurs through the process of adaptation to another language. In previous research, semantic change has been studied from a historical linguistic perspective and semantic difference of English loanwords in Japanese has been studied. However, there are no studies concerning semantic difference of Korean English loanwords and no studies which compare the semantic difference of English loanwords between two languages. Moreover, the previous studies do not extensively consider causes or underlying factors that might explain the difference in semantic usage of English loanwords.

Therefore, in this paper, we conduct a contrastive semantic study of the semantic adaptation of English loanwords in Japanese and Korean by using Word2vec. By analyzing cosine similarities between vectors, we can use Word2vec to detect English loanwords whose semantic usage has been changed and also reveal the degree of this semantic change in Japanese and Korean. Furthermore, to the end of gaining insight into why it is that this difference in semantic usage exists, this model also

provides useful data indicating some possible causes.

Specifically, we interpret the data as suggesting that the number of meanings of the original English word is a factor for differences in semantic usage between languages, and we test whether this hypothesis is valid. We analyze the relationship between the number of WordNet synsets and the value of cosine similarity between English words and their corresponding loanwords. As we expected, the data indicates that there is a negative correlation between the number of synsets and cosine similarity. This result implies that if the original English word has many meanings, the English loanword tends to be used with a different semantic usage in Japanese and Korean.

This study verifies that the word vector model contributes to finding semantic difference of loanwords and to contrastive analysis between two different languages. Concretely, it allows us quantitatively verify the prediction that the number of meanings of an English word is related to the extent of difference of semantic usage in its corresponding loanwords. Based on our work using the word vector model, we believe it could be very useful in future studies of other factors relevant to semantic adaptation, such as figurative usage. Furthermore, it can be expected that the word vector model will contribute to contrastive linguistic research between many languages. We hope this computational method will help to solve problems and provide new insights in several academic fields: a natural language processing, language education and contrastive linguistic analysis.

References

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research* 3, 1137-1155.

Choi, Myung Sook. (1996). English Teaching in primary School by using the Change of the original Meaning of Loan Words. *The Journal of English Education* 16, 43-53.

Fenogenova, Alena, Ilia Karpov, Viktor Kazorin, and Lebedev Innokentii. (2017).

Comparative Analysis of Anglicism Distribution in Russian Social Network Texts. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017"*, 79-88.

Hamilton, William L., 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)*, 1489-1501.

Harris, Zellig S. (1954). Distributional structure. *Word* 10, 146-162.

Kay, Gillian. (1995). English loanwords in Japanese. *World Englishes*, 67-76.

Kudo, Taku, 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)*, 230-237.

Kulkami, Vivek, 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)*, 625-635.

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

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

Min, Hyun-sik. (1998). A Study on the Foreign Words of Korean Language. *Korean Semantics* 2, 91-132.

Nishiyama, Sen. (1995). Speaking English with a Japanese mind. *World Englishes*, 1-6.

Okawa, Daisuke. (2008). A study on degree of recognition about Japanese-style English in Korean. *Journal of North-East Asian cultures* 14, 499-523.

Cheon, Seung Mi. (2008). *A Study of English Loanwords in Korean*. Korean Studies Information

Takamura, Hiroya, Ryo Nagata, and Yoshifumi Kawasaki. (2017). Analyzing Semantic Changes in Japanese Loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 1*, 1195-1204.

Wittgenstein, Ludwig. (1997). Philosophical investigations. Oxford, England: Blackwell Publishers.

Akihiko Yamada
Department of Linguistics
Seoul National University
1 Gwank-ro Gwanak-gu
Seoul, Korea, 08826
Email: a.yamada6022@gmail.com

Hyopil Shin
Department of Linguistics
Seoul National University
1 Gwank-ro Gwanak-gu
Seoul, Korea, 08826
Email: hpshin@snu.ac.kr

Received: October 30, 2017
Revised version received: December 4, 2017
Accepted: December 23, 2017