Utilizing Word Embedding Representations in Word Sense Analysis of Japanese Spelling Variants

Tomoki Okugawa
Graduate School of Science and Technology
University of Tsukuba
Tsukuba, Japan
s2220578@u.tsukuba.ac.jp

Takashi Inui
Graduate School of Science and Technology
University of Tsukuba
Tsukuba, Japan
inui@cs.tsukuba.ac.jp

***Abstract*—Spelling variants are an important linguistic phenomenon. In this paper, we focused on Japanese spelling variants caused by two character types, Hiragana and Kanji and attempted to analyze the differences of their word meanings. To realize an in-depth analysis capable of catching word meanings, we utilized word embedding representations. Using 293 variant pairs, we found that spelling variants caused by character types can have different meanings and that words with some semantic relations have a similar usage of the variants in a text.**

***Index Terms*—spelling variant, character type, word embedding**

I. Introduction

Several helpful NLP applications, such as information retrieval and information extraction, have been developed for decades [1], [2]. Although most of these applications require correct POS tagging and syntactic parsing, those fundamental analyses are incomplete because of some obstructions. One obstruction is the existence of spelling variants in the text [3].

Spelling variants are words with the same meanings but with different spells. There are various spelling variants, such as abbreviations(e.g., NLP and natural language processing) and slang (bucks and dollar). In Japanese, spelling variants are also frequently caused by character types because it is usual that four character types, Hiragana, Katakana, Roman alphabet, and Kanji characters, are used together in the same context. For example, the word "apple" is written in Japanese with りんご, リンゴ, Ringo, and 林檎, respectively. And, it is recognized that certain sets of CCT words are used to represent difference word meanings. So, we refer them to CCT words(spelling variants Caused by Character Types) and report the results of our investigations on the differences in word meanings in terms of CCT words.

For the convenience of investigation, in this work, we focused on two character types, Hiragana and Kanji, the main character types in Japanese, and investigated quantitatively whether there are CCT words with different meanings and what kinds of CCT words have different meanings.

Although some previous studies focus on Japanese character types, they are based only on surface aspects of a text, such as which character types are frequently used. An in-depth analysis of the word's meanings has not been conducted yet. To realize an in-depth analysis, we tried using word embeddings, a real-valued vector holding word meanings. Given a CCT word pair such as りんご and 林檎, we first transformed it into the word embedding each other and then compared those embeddings. It assumes that if りんご and 林檎 hold the same meaning, they would have an equivalent embedding vector, and if not, the difference in word meanings would appear by calculating vector subtraction between those embedding vectors. Based on the assumption, given a CCT word pair, we calculated a subtracted vector from each CCT word embedding pair and mainly analyzed.

As a result of analyzing 293 Japanese CCT variant pairs, we acquired the following main findings.

1) CCT word embedding pairs whose part of speech are nouns tend to have low cosine similarity values, while those pairs except nouns tend to have high cosine similarity. It means that CCT noun word pairs may be capable of holding different meanings.
2) From the clustering of subtracted vectors, CCT word pairs consisting of two related words tend to construct clusters. Interestingly, we observed that not only synonym pairs but also antonym pairs such as 大きな (large) and 小さな (small) tend to construct clusters.

II. Related Work

A. Motivation for Changing Character Type

Japanese-origin words tend to be written in Hiragana, Chinese-origin words in Kanji characters, and loan words in Katakana and Roman alphabet. However, notations different from the standard character type are used in some cases. Niwa et al. [4] reported that Hiragana and Katakana are used more frequently in catchphrases than in newspaper headlines and that there are cases in which the expressive power of catchphrases is enhanced by devising character types.

In "Simple Japanese Guidelines for Residential Support" [5], character type conversion from Kanji to Hiragana is actively performed for vocabulary simplification.

B. Survey on Surface Aspects of Character Type Usage

Many conventional works on character types investigated the surface-level word appearance in linguistic literature, such as which character types are frequently used. Chin [6] investigated the frequency of character types in Starbucks' product

979-8-3503-3078-6/23/$31.00

names and advertisements in order to explore the influence of the character types in the actual real-world environment. Kashino and Okumura [7] conducted a character type frequency survey using the "Balanced Corpus of Contemporary Written Japanese" to investigate the use of character types for each corpus genre.

Although their previous works only have surface-level analyses, we conducted semantic-level analyses using word embeddings to investigate the differences in word meanings in terms of CCT.

C. Word Embeddings

Word embeddings are techniques that express the meaning of a single word with a low-dimensional real-valued vector. These techniques are classified into two categories. The static word embeddings represent word meanings with fixed contextual information such as word2vec [8], and the dynamic word embeddings represent word meanings with dynamic contextual information such as BERT [9]. In this work, we adopted word2vec to remove the effects of dynamic contexts from our investigations.

Word2vec realizes a vectorization of words based on the assumption that words used in the same context have similar meanings and can perform addition and subtraction as well as similarity calculations.

III. Preparation

A. Word Embeddings

We used Wikipedia Entity Vectors [10]. It is one of the implementations of word embeddings trained from the Japanese version of Wikipedia articles with the word2vec algorithm, and it contains 751,361 word vectors with a 200-dimensional space.

B. Formal Definition of Subtracted Vector

Given a word w, w_K indicates the Kanji variant of w, and w_H indicates the Hiragana variant of w. And $\boldsymbol{w_K}$ indicates the word embedding corresponding with w_K, and $\boldsymbol{w_H}$ indicates the word embedding corresponding with w_H. Here, we call the embedding pair ($\boldsymbol{w_K}$, $\boldsymbol{w_H}$) CCT embedding pair of w and its subtracted vector is calculated by Equation (1).

$$\boldsymbol{w_{K-H}} = \boldsymbol{w_K} - \boldsymbol{w_H} \tag{1}$$

For example, for the word "美味しい (delicious)", the CCT embedding pair is (美味しい$_{\boldsymbol{K}}$, 美味しい$_{\boldsymbol{H}}$) and its subtracted vector is calculated by Equation (2).

$$\text{美味しい}_{\boldsymbol{K-H}} = \text{美味しい}_{\boldsymbol{K}} - \text{美味しい}_{\boldsymbol{H}} \tag{2}$$

C. Target CCT Embedding Pairs

We selected the target words for our investigation in the following procedures.

We firstly collected words from the staple of Japanese textbooks, "Minna no Nihongo Shokyu I" [11] and "Minna no Nihongo Shokyu II" [12] because they contain basic Japanese words in both Kanji and Hiragana notations. We transformed verb and adjective words into their standard forms for the preprocessing.

Next, we excluded some words that match either of the following conditions.

- words not included in Wikipedia Entity Vectors.
- words not containing the partner word that forms the CCT word pair in the textbooks.

Furthermore, we excluded words with ambiguity caused primarily by the character types. Occasionally, two words that have different meanings each other share the same Hiragana notation in Japanese. We decided to exclude those words because they may harm the investigation.

As a result, we obtained 586 words and those word embeddings, which consist of 293 CCT embedding pairs.

IV. Analyses

A. Analysis 1: Similarity of CCT embedding pairs

At first, we calculated the cosine similarity for each CCT embedding pair and the squared norm for each word embedding consisting of the CCT pair. Fig. 1 shows the scatter plots of the results. Here, we can see that these two measures have a negative correlation relation. So, we focus on our discussion using cosine similarity values.

From Fig. 1, we can see that the similarity values are widely distributed from 0 to 1. Next, Table I and Table II show the samples of the top 10 and bottom 10 of the cosine similarity, respectively. While the samples with high similarity values, such as those in Table I, seem to be normal spelling variants, the opposite side of samples with low similarity values, such as those in Table II, are anomalous. According to the vector information from word2vec, these CCT words have very different meanings between the Kanji and Hiragana variants. We can also see that verbs, adjectives, and adverbs tend to have high similarity values, while nouns tend to have low similarity values. It suggests that noun CCT pairs with low similarity pair are not normal spelling variants but different words with different meanings. In those cases, people might intentionally select a spelling to represent a specific meaning.

B. Analysis 2: Correlations with Character-based Indexes

Next, we investigated the relationship between cosine similarity and some indexes based on characters. Table III shows the correlation coefficients between cosine similarity values and three character-based indexes. Here, given a word w, len_H indicates the number of characters of w_H and len_K indicates the number of characters of w_K. And level_K indicates values defined based on the Kanji distribution table by grade in the elementary school curriculum guidelines [13]. For each Kanji character in w_K, we firstly assigned $n(1{-}6)$ for the Kanji characters taught in nth grade of elementary school in Japan, 7 for the common use Kanji characters taught after elementary school, and 8 for other non-common use Kanji characters. Then, of those values, the maximum value is used as level_K of w.

Table III reveals that there was not much correlation between cosine similarity values and the character-based indexes.

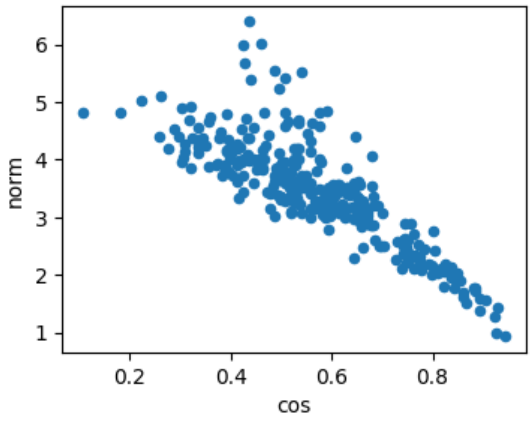

Fig. 1. Scatterplots of the cosine similarity and squared norm

TABLE I
TOP 10 WORDS OF THE COSINE SIMILARITY

$\boldsymbol{w_K}$, $\boldsymbol{w_H}$(English Word)	cosine
殆ど$_K$, 殆ど$_H$(almost)	0.94
大人しい$_K$, 大人しい$_H$(meek)	0.93
例えば$_K$, 例えば$_H$(for example)	0.93
無くなる$_K$, 無くなる$_H$(disappear)	0.92
下さい$_K$, 下さい$_H$(please)	0.91
始める$_K$, 始める$_H$(start)	0.89
捉える$_K$, 捉える$_H$(catch)	0.89
真っ直ぐ$_K$, 真っ直ぐ$_H$(straight)	0.88
皆さん$_K$, 皆さん$_H$(everyone)	0.88
美味しい$_K$, 美味しい$_H$(delicious)	0.88

There seems to be no relationship between the semantics of a CCT variant and the character attributes of that variant.

C. Analysis 3: Clustering of the Subtracted Vectors

We performed a hierarchical clustering to explore words with comparable differences in word meanings in the CCT variants. We applied the WARD algorithm to our 293 subtracted vectors. Fig.2 displays the result. Due to the space limitation, we show only a part of the whole dendrogram.

First, we observed the bottom side of the dendrogram, that is, clusters consisting of two subtracted vectors and found that two subtracted vectors with some semantic relations tend initially to be merged and construct a small cluster. Examples are shown below:

- Synonym
  - 速い (fast) and 早い (quick)
  - 曲がる (bend) and 折れる (break)
- Words with causal or order relationships
  - 合格 (pass) and 卒業 (graduation)
  - 危ない (dangerous) and 壊れる (broken)
- Antonym
  - 大きな (large) and 小さな (small)
  - 登る (climb) and 降りる (climb down)

Interestingly, we observed that not only synonyms are included at the bottom of the dendrogram but also antonyms

TABLE II
BOTTOM 10 WORDS OF THE COSINE SIMILARITY

$\boldsymbol{w_K}$, $\boldsymbol{w_H}$(English Word)	cosine
長さ$_K$, 長さ$_H$(length)	0.11
都合$_K$, 都合$_H$(convenience)	0.18
留守$_K$, 留守$_H$(absence)	0.22
発表$_K$, 発表$_H$(presentation)	0.26
お客様$_K$, お客様$_H$(customer)	0.26
住所$_K$, 住所$_H$(address)	0.28
出発$_K$, 出発$_H$(departure)	0.29
失敗$_K$, 失敗$_H$(failure)	0.3
低い$_K$, 低い$_H$(low)	0.3
近所$_K$, 近所$_H$(neighborhood)	0.3

TABLE III
CORRELATION COEFFICIENTS

	len_H	len_K	level_K
cosine	-0.056	0.189	0.160

are present there. It suggests that words with some semantic relations have a similar usage of CCT variants in a text. However, as can be seen from Fig.2, there is still room for further analysis because there were also clusters for which no clear semantic relationship could be held.

Next, we focused on the remaining part of the dendrogram, that is, clusters consisting of three or more number of subtracted vectors. Then, we found three characteristic clusters.

- Cluster i (total 36 words, e.g. 挨拶 (greeting), 当たる (hit), 可笑しい (funny), 殆ど (almost), 偶に (occasionally))
- Cluster ii (total 38 words, e.g. 駄目 (bad), 始まる (start), 大人しい (meek), 初めて (first time), 実は (actually))
- Cluster iii (total 22 words, e.g. 小学校 (elementary school), 警察 (police), 図書館 (library), 営業 (sales), 法律 (law))

Cluster i and Cluster ii contain a group of words in which the choice of character type can be freely left to the writer. On the other hand, Cluster iii is composed of nouns, most of which are unfamiliar with Hiragana notation in a text. Fig. 3 shows another result of vector visualization in three-dimensional space by principal component analysis. In this figure, we can clearly see that the three clusters are arranged in different positions.

V. CONCLUSION

In this paper, we focused on spelling variants caused by character types and attempted to utilize word embeddings to analyze the differences in word meanings. We analyzed 293 words and found that CCT words can have different meanings, especially CCT nouns, which tend to have low cosine similarity values that suggest holding different meanings. Also, through the clustering of subtracted vectors, CCT word pairs consisting of two semantically-related words tend to construct clusters.

In future work, it is worth conducting similar investigations focusing on other character types, such as Hiragana and Katakana variant pairs. In addition, we would like to attempt to analyze the relationship between the cosine similarity values

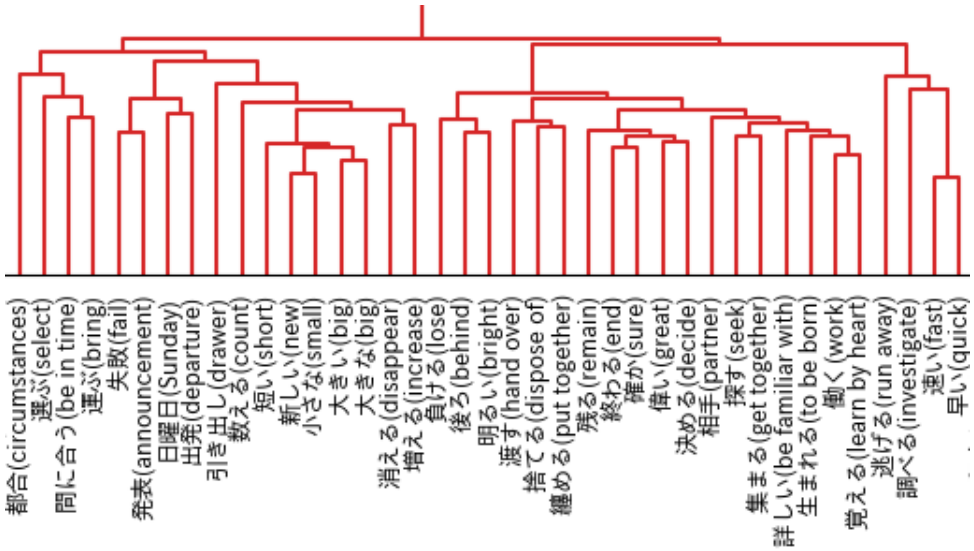

Fig. 2. A part of the clustering result

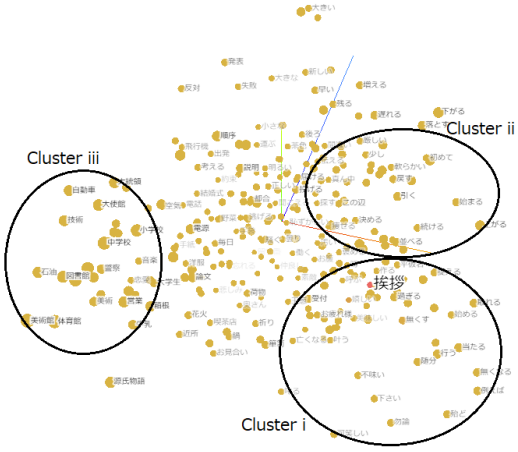

Fig. 3. The result of PCA

and vocabulary-based indices such as word familiarity. In this work, we adopted only word2vec, which constructs static word embeddings. We would also like to conduct a similar analysis using dynamic word embeddings like BERT to further confirm our findings in this work.

REFERENCES

[1] D. Jurafsky and J. H. Martin, "Speech and Language Processing," Pearson Education International, 2008.

[2] J. Eisenstein, "Introduction to Natural Language Processing," The MIT Press, 2019.

[3] K. Yamamoto, "Nihongo no hyokiyure mondai ni kansuru kosatsu to taisho" [On Orthographical Variants Problem and Our Solution], Japio year book, pp. 202–205, 2015 (in Japanese).

[4] A. Niwa, N. Okazaki, K. Nishiguchi, C. Kameyama, and M. Mouri, "Kyatchikopi no jido seisei ni muketa bunseki" [Analysis for Automatic Generation of Catchphrase], In Proceedings of the 25th Annual Meeting of the Association for Natural Language Processing, pp. 558–561, 2019 (in Japanese).

[5] Agency for Cultural Affairs, "Zairyushien no tame no yasashii nihongo gaidorain" [Simple Japanese Guidelines for Residential Support], 2020 (in Japanese).

[6] P. Chan, "Gengokeikan ni okeru gairaigoshiyo –Sutabakkusu no katakanahyokijittai wo chushin ni–" [The Linguistic Landscape of Expression on Katakana Words–Research Based on Expression of Starbuck Company's Katakana Words Expression–], Tohoku University linguistics journal, Vol. 25, pp. 69–84, 2016 (in Japanese).

[7] W. Kashino and M.Okumura, "Wago ya kango no katakana hyoki : 'Gendai nihongo kakikotoba kinko kopasu' no shoseki ni okeru shiyo jittai" [Analysis of Katakana Representation for Japanese Native Words and the Words Imported from Classical Chinese : Using BCCWJ Japanese Corpus], Mathematical linguistics, Vol. 28, No. 4, pp. 153–161, 2012.

[8] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," In Proceedings of Workshop at ICLR, 2013.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding", In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, pp. 4171–4186, 2019.

[10] M. Suzuki, K. Matsuda, S. Sekine, N. Okazaki, and K. Inui, "A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles," IEICE Transactions on Information and Systems, Special Section on Semantic Web and Linked Data, Vol. E101-D, No.1, pp. 73–81, 2018.

[11] 3A corporation, "Minna no nihongo syokyu I dai2han honsatu" [Minna no Nihongo Beginner I 2nd Edition Main Book], 3A corporation, 2012 (in Japanese).

[12] 3A corporation, "Minna no nihongo syokyu II dai2han honsatu" [Minna no Nihongo Beginner II 2nd Edition Main Book], 3A corporation, 2013 (in Japanese).

[13] Ministry of Education, Culture, Sports, Science and Technology, "Shogakko gakushu shido yoryo" [Elementary School Course of Study], 2017 (in Japanese).