International Journal of Innovative
Computing, Information and Control
Volume **9**, Number **10**, October **2013**

 ISSN 1349-4198
pp. **4231–4246**

AUTOMATIC SELECTION AND ANALYSIS OF JAPANESE NOTATIONAL VARIANTS ON THE BASIS OF MACHINE LEARNING

Masaki Murata[1], Masahiro Kojima[2], Takuya Minamiguchi[2]
and Yasuhiko Watanabe[2]

[1]Graduate School of Engineering
Tottori University
4-101 Koyama-Minami, Tottori 680-8550, Japan
murata@ike.tottori-u.ac.jp

[2]Graduate School of Science and Technology
Ryukoku University
Seta, Otsu-shi, Shiga 520-2194, Japan
watanabe@rins.ryukoku.ac.jp

Received October 2012; revised February 2013

Abstract. *Certain words have several notational variants. Thus, when we express such a word, we have to select one of its notational variants. In this study, selecting a variant is termed "notational selection". First, we apply machine learning to determining how difficult it is to perform notational selection on words. In addition, we investigated the reasons it was easy to select the notational variants of certain words. Our experimental results show that when machine learning performs notational selection at a high recall rate, the appropriate notational variant depends on meanings and contexts. Moreover, we show that when machine learning performs notational selection at a low recall rate, any of the notational variants of the word can be used in a sentence. These results are useful to humans performing notational selection. Furthermore, we demonstrate that in certain cases, machine learning can be used to perform notational selection. In experiments conducted with machine learning, the results show we could perform notational selection for 81 out of 939 words with two notations at a recall rate of 80% or higher. We also confirmed that the average of the accuracy rates of our proposed method was higher than that of a baseline method.*
Keywords: Machine learning, Notational selection, Difficulty level, Feature

1. **Introduction.** The subject of notational selection is a part of the information processing field [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. In this study, we use machine learning to automatically select and analyze Japanese notational variants. Our study is related to natural language processing and machine learning [1, 2, 3, 4, 8, 9, 10, 12, 13, 14, 15].

Japanese sentences often contain notational variants of a word. Japanese words can be expressed by Chinese, Hiragana, and/or Katakana characters. For example, the word *sakura* (cherry in Japanese) has three notation variants: "桜" (expressed with Chinese characters), "さくら" (expressed with Hiragana characters), and "サクラ" (expressed with Katakana characters). Moreover, in Japanese, more than one type of character can be used in the same word. For instance, the word *shouyu* has two notation variants. The first is "醤油" (expressed with only Chinese characters), and the second is "しょう油", which is expressed with two types of characters, i.e., Hiragana characters (しょう) and Chinese characters (油). If a word in a sentence has several notational variants, it is often difficult to decide which notational variant to use. For example, because *zeiritsu wo hikiageru* (raise taxes) and *hikiageru* (pull up, raise, withdraw, or go off) have the two notational

variants “引き揚げる” and “引き上げる”, it is difficult to select the appropriate one. A possible solution in these cases is to use a dictionary. However, frequently, the distinction between notational variants is not clear even when we consult a dictionary. Therefore, several studies on notational variant selection have been conducted. Nishikawa et al. investigated the use of word frequencies for notational variant selection [16]. Hiki and Meiseki addressed the problem of selecting notational variants on the basis of newspaper corpora and questionnaires [17]. If we could detect the characteristics of notational variants, we could use them to select the appropriate variant and construct systems that support selecting them.

In this study, we use machine learning to analyze the selection of notational variants. We use machine learning to select the notational variants of numerous words and classify words into easy or difficult to conduct notational selection. Moreover, we investigate the reasons it was easy to select the notational variants of certain words. The results obtained can be used when selecting the notational variant of a word manually.

For certain words, any notational variant of the word can be used to express it. In these cases, machine learning cannot perform notational selection with high accuracy. Therefore, by detecting words that are difficult for machine learning in order to conduct notational selection, we can detect cases where any notational variant of a word can be used. This information can be useful in future studies and systems designed to perform notational selection.

The key contributions of this study are described below.

- By using the recall rates of machine learning, our method can classify words with the notational variants into several categories, such as “high” and “low”. The notational variants of words with “high” recall rates have distinguishing characteristics. For these words, the appropriate notational variant depends on meanings and contexts. Hence, the appropriate notational variant can be selected by using meanings and contexts. The notational variants of words with “low” recall rates have no distinguishing characteristics. Hence, any notational variant of these words can be used in a sentence. In experiments using human subjects, the higher the recall rates of machine learning were likely to be, the higher the accuracy rates of the human subjects.

  Our technique of classifying words with the notational variants into several categories by using machine learning is very original and novel and has not been handled in other literature on notational variants.
- Our proposed method was useful in performing notational selection. Our experiments show that by using our proposed method, we could perform notational selection for 81 out of 939 words with two notations at a recall rate of 80% or higher. We also confirmed that our method was more effective in performing notational selection than was the baseline method that outputted the notation appearing most frequently. In our experiments, the average of the accuracy rates of our proposed method (0.87) was more effective than that of the baseline method (0.84).

  In terms of related studies on notational variants, Kojima et al. automatically extracted variants for information retrieval, Nishikawa et al. used word frequency for notational selection, and Hiki and Meiseki used co-occurring words for notational selection. However, their studies did not use machine learning for notational selection. Our method is original in that machine learning is used for notational selection.

The application of this study is described below.

- The accuracy rates of notational selection with our method are high. Our method would be useful for constructing a system that shows candidate notational variants

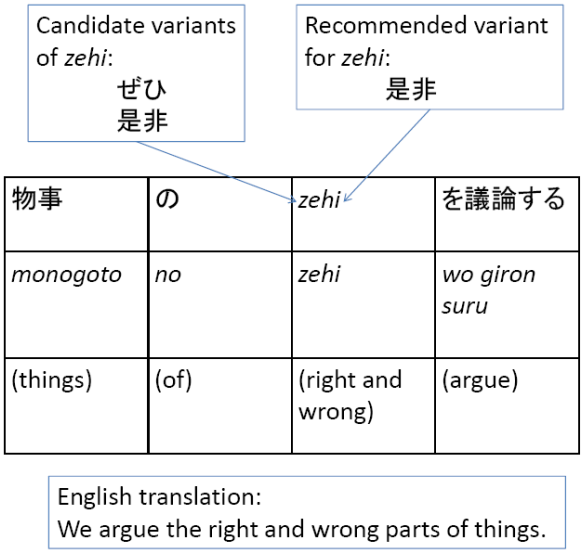

FIGURE 1. Example of a system showing candidate and recommended notational variants

for a word in a sentence. An example of such a system is shown in Figure 1. In this figure, *zehi* has two notational variants, "ぜひ" and "是非". The system recommends "是非" as a proper variant used for the given sentence.

In Section 2, we present background information and introduce the key ideas used in this study. In Section 3, we describe our method, which is based on machine learning. In Section 4, we describe the data sets used in our experiments. In Section 5, we describe experiments on notational selection with machine learning. Our method of notational selection can classify words into categories based on recall rates (accuracies) of machine learning. In Section 6, we examine the classified words with notational variants and clarify the characteristics of the categories based on recall rates. In Section 7, we examine the relationship between categories based on recall rates and human notational selection. In Section 8, we give our concluding remarks.

2. **Background and Key Ideas.** In the field of information retrieval, several studies on notational variants have been conducted. For example, Kojima et al. automatically extracted notational variants for information retrieval [18]. In addition, several studies on notational variant selection have been conducted [16, 17]. Nishikawa et al. used word frequencies in newspapers and academic papers and defined the most frequent variant as the dominant variant, while the other variants were defined as non-dominant [16]. Next, they constructed a system to support humans creating sentences. When a non-dominant variant was used in a sentence, their system produced a message that cautioned the human creating the sentence. Hiki and Meiseki used the co-occurring relationships of nouns in newspaper corpora and questionnaires to analyze the selection of the notational variants of the word *kaeru* (change), which has four notational variants: "変える", "替える", "換える" and "代える" [17].

In this study, we use machine learning to perform notational selection. In machine learning, notational selection is performed by using features (information used for learning). Specifically, the co-occurrences of nouns and verbs can be used as features. Thus, because machine learning uses these co-occurrences as features, it will perform better than that if it only used word frequencies, as seen in the work of Nishikawa et al.

Moreover, we investigated the reasons it was easy to select the notational variants of certain words. The results obtained can be used when selecting the notational variant of a word manually.

For certain words, any notational variant of the word can be used to express it. In these cases, machine learning cannot perform notational selection with high accuracy. Therefore, by detecting words for which it is difficult for machine learning to conduct notational selection, we can detect cases where any notational variant of a word can be used. This information can be useful in future studies and systems designed to perform notational selection.

In this study, we use the accuracy of notational selection and classify words with multiple notational variants into three categories: "high", "medium" and "low". We define these categories in Section 5. Moreover, we examine the reasons it is easy or difficult for machine learning to perform notational selection on words, and we discuss the results. Finally, we examine the characteristics of each category.

3. **Our Method of Notational Selection Using Machine Learning.** In this study, we use machine learning to perform notational selection. Specifically, we use the maximum entropy method [19, 20]. The maximum entropy method is a supervised machine learning method. It can estimate the class of an input data item by using training data items.

A sentence that includes a word with multiple notational variants is input. We estimate which notational variant is the most appropriate for the sentence by using a machine learning method, the maximum entropy method. Notational variants are used as classes in machine learning.

One machine learning cycle is used for a word with notational variants. When we have n words with notational variants, we use n machine learning cycles.

In machine learning, we use features (information used in learning). We describe the features used in our machine learning method below.

In Tables 1 and 2, we list the features used in our experiments. In our experiments, we use the most frequent two notational variants. Features are extracted from sentences containing words with two notational variants. Our machine learning approach selects from the two notational variants the variant that is more appropriate in a sentence.

The category number in Tables 1 and 2 is a ten-digit number that is described in Bunrui Goi Hyou, a Japanese word thesaurus [21, 22]. Words with similar meanings have similar ten-digit numbers. In this study, we use the first five and three digits of the number as features. Therefore, we use the upper concept of each word as features.

The notational variants adjacent to a target word can be used for notational selection. Therefore, during notational selection, we use features F1 to F20, which represent information in the bunsetsu (phrase) that contains a target word with notational variants.

The syntactic information of a sentence can be used in notational selection. Therefore, we use features F21 to F60, which contain useful information when the words in a bunsetsu modify or are modified by the bunsetsu that contains a target word with notational variants.

Furthermore, characters appearing just before or after a target word with notational variants also contain useful information. Therefore, we use features F61 and F62.

In this study, we use a Japanese syntactic parser, KNP, to perform syntactic analysis (parsing) [23].

4. **Data Sets Used in Experiments.** In this section, we describe the data sets used in our experiments.

TABLE 1. Features used in machine learning

ID	Explanation of feature
F1	The first content word in the bunsetsu (phrase) containing a target word for notational selection
F2	The part of speech (POS) of F1
F3	The first five digits of the category number of F1
F4	The first three digits of the category number of the word of F1
F5	The last content word in the bunsetsu containing a target word for notational selection
F6	The POS of F5
F7	The first five digits of the category number of F5
F8	The first three digits of the category number of F5
F9	A content word in the bunsetsu containing a target word for notational selection
F10	The POS of F9
F11	The first five digits of the category number of F9
F12	The first three digits of the category number of F9
F13	The first functional word in the bunsetsu containing a target word for notational selection
F14	The POS of F13
F15	The last functional word in the bunsetsu containing a target word for notational selection
F16	The POS of F15
F17	A functional word in the bunsetsu containing a target word for notational selection
F18	The POS of F17
F19	A symbol in the bunsetsu containing a target word for notational selection
F20	The POS of F19
F21	The first content word in a bunsetsu that modifies the bunsetsu containing a target word for notational selection
F22	The POS of F21
F23	The first five digits of the category number of F21
F24	The first three digits of the category number of F21
F25	The last content word in a bunsetsu that modifies the bunsetsu containing a target word for notational selection
F26	The POS of F25
F27	The first five digits of the category number of F25
F28	The first three digits of the category number of F25
F29	A content word in a bunsetsu that modifies the bunsetsu containing a target word for notational selection
F30	The POS of F29
F31	The first five digits of the category number of F29
F32	The first three digits of the category number of F29

In our experiments on notational selection, we used notational variants appearing in newspapers. Specifically, we used Mainichi newspaper articles (years 2005 – 2007), which contained 3,693,567 sentences. In these sentences, 29,815 words had multiple notational variants as determined by the Juman dictionary [24]. This dictionary lists the notational

Table 2. Features used in machine learning

ID	Explanation of features
F33	The first functional word in a bunsetsu that modifies the bunsetsu containing a target word for notational selection
F34	The POS of F33
F35	The last functional word in a bunsetsu that modifies the bunsetsu containing a target word for notational selection
F36	The POS of F35
F37	A functional word in a bunsetsu that modifies the bunsetsu containing a target word for notational selection
F38	The POS of F37
F39	A symbol in a bunsetsu that modifies the bunsetsu containing a target word for notational selection
F40	The POS of F39
F41	The first content word in the bunsetsu that is modified by the bunsetsu containing a target word for notational selection
F42	The POS of F41
F43	The first five digits of the category number of F41
F44	The first three digits of the category number of F41
F45	The last content word in the bunsetsu that is modified by the bunsetsu containing a target word for notational selection
F46	The POS of F45
F47	The first five digits of the category number of F45
F48	The first three digits of the category number of F45
F49	A content word in the bunsetsu that is modified by the bunsetsu containing a target word for notational selection
F50	The POS of F49
F51	The first five digits of the category number of F49
F52	The first three digits of the category number of F49
F53	The first functional word in the bunsetsu that is modified by the bunsetsu containing a target word for notational selection
F54	The POS of F53
F55	The last functional word in the bunsetsu that is modified by the bunsetsu containing a target word for notational selection
F56	The POS of F55
F57	A functional word in the bunsetsu that is modified by the bunsetsu containing a target word for notational selection
F58	The POS of F57
F59	A symbol in the bunsetsu that is modified by the bunsetsu containing a target word for notational selection
F60	The POS of F59
F61	The 1-gram, 2-gram, 3-gram, 4-gram, and 5-gram characters appearing just before a target word for notational selection
F62	The 1-gram, 2-gram, 3-gram, 4-gram, and 5-gram characters appearing just after a target word for notational selection

variants of words. The notational variants of a word listed in the dictionary are based on the notational variants of the word listed in ordinary Japanese word dictionaries, such

as the Kojien dictionary [25]. In ordinary Japanese word dictionaries, the notational variants of a word are listed under the same entry of the word. The number of words where only one notational variant for a word with multiple notational variants appears in the newspaper sentences is 14,630. The number of words where multiple notational variants for a word appear in the newspaper sentences is 15,185.

From these 15,185 words, we extracted words that satisfy all the following conditions.

Condition 1: The frequency of a word in the newspaper articles is greater than 100.

Condition 2: The morphological analysis of a word performed by using Juman [24] is not ambiguous and does not contain the "@" mark.

Condition 3: The frequency of the second most used notational variant of a word in the newspaper articles is greater than 10.

Condition 1 is used for investigating frequently used words in newspaper articles. Condition 2 is used for extracting notational variants of the same word. If the results of the morphological analysis of a word contain the "@" mark, then we can extract different notational variants of the word. For example, "けいじ" can be a notational variant of "掲示" (placard), "計時" (timekeeper) and "刑事" (detective). When we use Juman to analyze "けいじ", we obtain "掲示", "計時" and "刑事", all of which contain "@". Therefore, the results of the morphological analysis are ambiguous and contain the "@" mark. In this study, we eliminate such ambiguous data items. Condition 3 is applied to ensure that machine learning is executed properly. If the frequency of a notational variant is low, problems occur when machine learning is used in experiments. In our experiments, we use the most frequent two notational variants for notational selection. Therefore, we use Condition 3. In our data set, 1,877 words satisfied all three conditions listed above. Out of these words, we randomly extracted 939 words to use in our experiments.

5. Experiments on Notational Selection Using Machine Learning.

5.1. **Experimental methods.** We applied our machine learning approach to the 939 words obtained through the process described in Section 4.

In our experiments, we used the most frequent two notational variants. Our machine learning approach selects from the two notational variants the variant that is more appropriate in a sentence.

For each word, sentences containing that word were extracted from Mainichi newspaper articles (years 2005 – 2007) and were used as a data set. We conducted a ten-fold cross validation for each data set (for each word). First, we divided the given data set into ten parts. One part was treated as the test data set, and the remaining nine parts were treated as training data sets. The category (class) of each item in the test data set was estimated by learning the training data sets by using the maximum entropy method. Then, the estimated category was evaluated by using the correct category in the test data set. This process was repeated for all ten parts. Consequently, all ten parts are evaluated.

5.2. **Experimental results.** On the basis of the recall rates obtained, we classified words into the three categories: "high", "medium" and "low". A recall rate is a concept similar to accuracy. A recall rate is the ratio of the number of the correct outputs over the number of the correct data items.

When the lowest recall rate of the two notational variants of a target word is higher than or equal to 0.8, the word is classified in the "high" category. This is because when both recall rates are high, the estimation is accurate. When the lowest recall rate of the two notational variants of a target word is higher than or equal to 0.5 and lower than 0.8, the word is classified in the "medium" category. When the lowest recall rate of the two notational variants of a target word is lower than 0.5, the word is classified in the "low"

TABLE 3. Ratios of words classified on the basis of recall rates into the three categories

Category	Ratio
High	0.09 (81/939)
Medium	0.16 (154/939)
Low	0.75 (704/939)

TABLE 4. Averages of accuracy rates in our proposed method based on machine learning and the baseline method

Category	Our proposed method	The baseline method
High	0.95	0.73
Medium	0.87	0.77
Low	0.87	0.87
Total	0.87	0.84

category. This is because when one of the two recall rates is low, the estimation is not accurate.

The results of the classification process described above are shown in Table 3. In this table, the ratio of words is classified into the three categories.

From the results in the table, we see that for 81 out of the 939 words, the recall rates of both of the notational variants obtained were greater than 0.8. Moreover, we discovered that in certain cases, machine learning was effective in performing notational selection.

5.3. **Comparison experiments.** We carried out experiments with a baseline method. This method always outputs the notational variant that most frequently appears in the training data set among the notational variants.[1] We compared the results of our method based on machine learning with those of the baseline method. The compared results are shown in Table 4, which shows the average of the accuracy rates of words in each category ("high", "medium" and "low") and the average of the accuracy rates of all 939 words ("Total" in the table). The accuracy rate is the ratio of the number of the correct ones over the number of all the data items.

In all 939 words ("Total"), the average of the accuracy rates of our proposed method (0.87) was higher than that of the baseline method (0.84). In particular, in the "high" and "medium" categories, the averages of the accuracy rates of our proposed method (0.95 and 0.87) were much higher than those of the baseline method (0.73 and 0.77). We found that our proposed method was more effective in notational selection than was the baseline method.

6. Examinations of Classified Words with Notational Variants.

6.1. **Examinations on words.** Next, we extracted words on the basis of their recall rates and analyzed the characteristics of their notational variants. In this section, we present example sentences, the corresponding results obtained from machine learning, and features of each word that are important in machine learning. First, we present two example sentences for which machine learning selected the correct notational variant, i.e., the notational variant selected by machine learning was the same as that used in the sentence. Next, we present two example sentences for which machine learning did not

[1]In the studies on word sense disambiguation, a baseline method that outputs a sense most frequently appearing in the training data set is often used [26].

select the correct notational variant. For these examples, we identify the three features with the highest normalized α values. A normalized α value represents the degree of importance of the corresponding feature used to estimate a notational variant learned by using the maximum entropy method. A feature, j, with a high normalized α value ($\alpha_{a,j}$) is considered important for the system to categorize a data item with feature j as a. If the normalized α value ($\alpha_{a,j}$) of category a and feature j is equal to x, then when a notational variant is selected by using only feature j, the probability that a notational variant is a is equal to x. For details, please refer to [20, 27].

6.1.1. *Word classified into "high":* *zehi* ("please" or "right and wrong") Notational variants: "ぜひ" (Hiragana characters), "是非" (Chinese characters).

Example sentence 1a (Machine learning correctly selected a notational variant):
１２年後にも ぜひ と 好評 だった
12nengo nimo *zehi* *to* *kouhyou* *datta*
(12 years later) (please) (saying) (received well) (was)
(Saying "Please do so 12 years later" was received well.)

Example sentence 1b (Machine learning correctly selected a notational variant):
物事 の 是非 を知る
monogoto no *zehi* *wo shiru*
(things) (of) (right and wrong) (understand)
(We understand the right and wrong parts of things.)

Example sentence 1c (Machine learning incorrectly selected a notational variant):
ぜひ また 会いたい
zehi *mata* *aitai*
(please) (again) (see us)
(Please see us again.)

Example sentence 1d (Machine learning incorrectly selected a notational variant):
晩婚化 の 是非 を検証する
bankonka *no* *zehi* *wo kenshou suru*
(late marriage) (of) (right and wrong) (investigate)
(We investigate the right and wrong parts of late marriage.)

The word *zehi* is classified into the "high" category. The machine learning results are shown in Table 5. The most important features for machine learning when selecting the notational variants of *zehi* are shown in Table 6. In this table, the precision rate is the ratio of the number of correct outputs over the number of total outputs.

In example sentences 1a and 1b, machine learning selected correct notational variants. Conversely, in example sentences 1c and 1d, machine learning selected incorrect notational variants. For instance, in example sentence 1c, the correct notational variant is "ぜひ", but machine learning selected "是非".

TABLE 5. Machine learning results for *zehi*

Notational variants	Recall rates	Precision rates	Number
"ぜひ"	0.99	0.98	1442
"是非"	0.98	0.99	1642

TABLE 6. Important features for machine learning when selecting the notational variants of *zehi*

“ぜひ”		“是非”	
Features	Normalized α values	Features	Normalized α values
F15: “と” (said)	0.88	F61: “の” (of)	0.96
F61: “を” (objective postpositional particle)	0.76	F62: “この本” (this book)	0.72
F17: “と” (said)	0.74	F15: “も” (also)	0.70

Next, we analyze example sentences for which machine learning correctly selected a notational variant (example sentences 1a and 1b) and identify the features that are important for machine learning (Table 6). We notice that features listed in the table appear in the example sentences that were processed correctly (example sentences 1a and 1b). For instance, in example sentence 1a, feature [F15: “と” (said)] appears in “ぜひと好評”. In example sentence 1b, feature [F61: “の” (of)] appears in “物事の是非”. These features have high normalized α values because there are many cases where [F15: “と” (said)] appears in the sentence that contains “ぜひ” and many cases where [F61: “の” (of)] appears in the sentence that contains “是非”. Machine learning identified these characteristics as features with high normalized α values.

Originally, “ぜひ” and “是非” are notational variants of the same word and have the same meaning. Both “ぜひ” and “是非” can be used to convey “please” and “right and wrong”. However, from example sentences 1a and 1b, we conclude that it is more natural to use “ぜひ” to convey the meaning of “please” and “是非” to convey the meaning of “right and wrong”. This conclusion was reached because “ぜひ” was used often to convey the meaning of “please,” and “是非” was used often to convey the meaning of “right and wrong”.

Because feature [F15: “と” (said)] appears often in sentences containing the meaning of “please”, it is an important expression for detecting “ぜひ”. Similarly, because feature [F61: “の” (of)] appears often in sentences containing the meaning of “right and wrong”, it is an important expression for detecting “是非”. In our proposed method, we can easily detect such important features (clue expressions) by checking the normalized α values.

In ordinary Japanese word dictionaries, such as Kojien, the differences between “ぜひ” and “是非” are not explained. Our proposed method can be used to explain the differences between “ぜひ” and “是非”. Therefore, our method is effective in analyzing how to use notational variants.

Because the recall rates of *zehi* are high, we can accurately select its notational variants. In addition, using the features utilized for notational selection, we can also detect clue expressions such as [F15: “と” (said)] and [F61: “の” (of)]. These results are useful for humans manually selecting notational variants.

Next, we considered using a simple method that always selects the notational variant that is used more often. From Table 5, “是非” is used more often than “ぜひ” ($1642 > 1442$). Hence, this simple method always selects “是非”. In this method, the recall rate for “ぜひ” was 0 and that for “是非” was 1. Although the recall rate for “是非” was high, the recall rate for “ぜひ” was low. Therefore, this simple method that uses only frequencies was not accurate. Conversely, in our method based on machine learning, both recall

rates were very high; the recall rate for "ぜひ" was 0.99, and that for "是非" was 0.98. Hence, we conclude that our method is much better than the simple method that uses only frequencies.

6.1.2. *Word classified into "medium": hikiageru* ("pull up," "withdraw," "go off," or "raise") (*hiki* means "pull" and *ageru* means "up".)

Notational variants: "引き揚げる" (Consists of Chinese and Hiragana characters. "引" and "揚" are Chinese characters. "き", "げ" and "る" are Hiragana characters.), "引き上げる" (Consists of Chinese and Hiragana characters. "上" is a Chinese character.)

Example sentence 2a (Machine learning correctly selected a notational variant):
投機資金を　引き揚げる
toushi shikin wo　*hikiageru*
(investment fund) (withdraw)
(We withdraw funds.)

Example sentence 2b (Machine learning correctly selected a notational variant):
販売計画を　280 万台に　引き上げる
hanbaikeikaku wo 280 man dai ni　*hikiageru*
(sales plan)　(2.8 million cars) (raise)
(We raise the number of cars in our sales plan to 2.8 million.)

Example sentence 2c (Machine learning incorrectly selected a notational variant):
派遣していた　２人を　引き揚げる
haken shiteita hutari wo　*hikiageru*
(dispatched)　(two men) (withdraw or go off)
(We let go of the two men who were dispatched.)

Example sentence 2d (Machine learning incorrectly selected a notational variant):
支払限度額を　引き上げる
shiharai gendogaku wo hikiageru
(payment limit)　(raise)
(We raise the payment limit.)

The word *hikiageru* is classified into the "medium" category. The results obtained from our method based on machine learning are shown in Table 7. The features that are required by machine learning for selecting the notational variants of *hikiageru* are shown in Table 8.

In our data set, "引き揚げる" is often used in sentences containing the meaning "withdraw funds", such as in example sentence 2a. Therefore, feature [F29: 資金 (fund)] received a high normalized α value.

In our data set, "引き上げる" is used often in sentences containing the meaning "raise values to X", such as in example sentence 2b. Therefore, feature [F61: に (to)] received a high normalized α value.

TABLE 7. Machine learning results for *hikiageru*

Notational variants	Recall rates	Precision rates	Number
"引き揚げる"	0.67	0.83	537
"引き上げる"	0.97	0.94	2642

TABLE 8. Important features for machine learning when selecting the notational variants of *hikiageru*

“引き揚げる”		“引き上げる”	
Features	Normalized α values	Features	Normalized α values
F29: 資金 (fund)	0.82	F62: 幅 (range)	0.84
F62: 船 (ship)	0.80	F32: 137 (Semantic concept: “gain”)	0.82
F62: 者 (person)	0.78	F61: に (to)	0.75

From the results presented above, we conclude that “引き揚げる” is used often to convey the meaning of “withdraw” and “引き上げる” is used often to convey the meaning of “raise”. These results are useful for humans manually selecting notational variants.

In ordinary Japanese word dictionaries, such as Kojien, under the entry for the word *hikiageru*, both its notational variants “引き揚げる” and “引き上げる” are described. However, in these dictionaries, the differences between “引き揚げる” and “引き上げる” are not explained. Our proposed method can be used to explain these differences. Therefore, we can conclude that our method is effective in analyzing how to use notational variants.

6.1.3. *Word classified into “low”: moritsukeru* (“arrange”) Notational variants: “盛りつける” (Consists of Chinese and Hiragana characters. “盛” is a Chinese character. “り”, “つ”, “け” and “る” are Hiragana characters.), “盛り付ける” (Consists of Chinese and Hiragana characters. “付” is a Chinese character).

Example sentence 3a (Machine learning correctly selected a notational variant):
総菜を 皿に 盛りつける
souzai wo sara ni moritsukeru
(food) (plate) (arrange)
(We arrange the food on a plate.)

Example sentence 3b (Machine learning correctly selected a notational variant):
彩りよく 盛り付ける
irodoriyoku *moritsukeru*
(in a colorful manner) (arrange)
(We arrange it in a colorful manner.)

Example sentence 3c (Machine learning selected incorrectly a notational variant):
切って 混ぜて 盛りつける
kitte mazete moritsukeru (We cut, mix, and arrange it.)
(cut) (mix) (arrange)

Example sentence 3d (Machine learning selected incorrectly a notational variant):
食卓に 豊かな季節を 盛り付ける
shokutaku ni yutakana kisetsu wo *moritsukeru*
(dining table) (a wealth of seasonal dishes) (arrange)
(We arrange a wealth of seasonal dishes on a dining table.)

The word *moritsukeru* is classified into the “low” category. The results obtained from our proposed method are shown in Table 9. The features that are important to machine learning when selecting the notational variants of *moritsukeru* are presented in Table 10.

TABLE 9. Results of machine learning for *moritsukeru*

Notational variants	Recall rates	Precision rates	Number
“盛りけける”	0.29	0.30	28
“盛り付ける”	0.44	0.43	34

TABLE 10. Important features for machine learning method when selecting the notational variants of *moritsukeru*

“盛りっける”		“盛り付ける”	
Features	Normalized α values	Features	Normalized α values
F50: noun	0.62	F38: postpositional particle	0.58
F42: noun	0.62	F36: postpositional particle	0.58
F46: noun	0.60	F34: postpositional particle	0.58

No distinguished features are listed in Table 10. Moreover, in Table 9, we see that the recall rates were low.

In example sentences 3a to 3d, our proposed method could not identify the difference between the meanings of “盛り付ける” and “盛りっける”. Therefore, machine learning could not identify features that are useful for notational selection and consequently could not obtain high recall rates. This result suggests that either “盛り付ける” or “盛りっける” can be used to convey the meaning of *moritsukeru*. These results are useful for humans selecting manually notational variants.

6.2. **Examinations of words classified by recall rates.** Next, we examined words that were classified on the basis of their recall rates into the “high” category. As discussed in the previous section, when a word has high recall rates, such as *zehi*, its notational selection can be easily performed by using features. Moreover, we confirmed that the features that were important for the machine learning method to select notational variants appeared in sentences. Our machine learning method accurately recognized important features and was effective in performing notational selection by using features. For words with “high” recall rates, the appropriate use of a notational variant in a sentence depends on meanings and contexts. Our machine learning method accurately recognized the meanings and contexts as features. Humans can select appropriate notational variants by considering the features that the machine learning method identified as important. These results are useful for humans selecting notational variants manually.

Next, we examined words that were classified into the “low” category. We concluded that for words with “low” recall rates, either of the two notational variants of the word could be used in a sentence; no distinguishing characteristics were detected for the two notational variants.

Finally, we examined words that were classified into the “medium” category. Words in this category possess characteristics from both “high” and “low” word categories.

In summary, the notational variants of words with “high” recall rates have distinguishing characteristics. For these words, the appropriate notational variant depends on meanings and contexts. Hence, the appropriate notational variant can be selected by using meanings and contexts.

The notational variants of words with “low” recall rates have no distinguishing characteristics. Hence, any notational variant of these words can be used in a sentence.

These results can be useful for humans selecting notational variants and for future studies on notational variant selection.

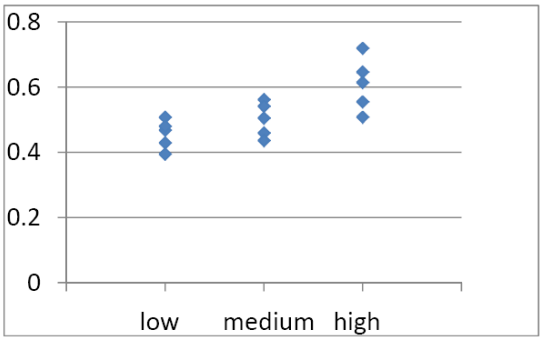

FIGURE 2. Relationship between the categories ("low", "medium" and "high") and the average of the accuracy rates of the ten subjects

7. **Relationship between Categories Based on Recall Rates and Human Selection.** We examined the relationship between categories based on recall rates and human selection. We randomly selected five words among the "low", "medium" and "high" categories used previously and used 15 words in total. Each word had two notational variants ($N1$ and $N2$). Five sentences containing the notational variant $N1$ and five sentences containing the notational variant $N2$ were randomly extracted from Mainichi newspaper articles (years 2005 – 2007). We used ten human subjects in experiments. For each word, ten sentences were given for all subjects. A subject judged which of $N1$ and $N2$ were suitable for each sentence. The accuracy rates of the fifteen words in subjective judgments were calculated. The results are shown in Figure 2. The vertical axis of the figure indicates the accuracy rates of the fifteen words (each accuracy rate is the average of the accuracy rates of the ten subjects), and the horizontal axis indicates the "low", "medium" and "high" categories.

In Figure 2, the accuracy rates in the "high" categories were roughly higher than those in the "medium" categories, and the accuracy rates in the "medium" categories were roughly higher than those in the "low" categories. We used statistical t tests. The average of the five accuracy rates in the "high" category was significantly higher than that in the "medium" category at a significant level of 0.05. The average of the five accuracy rates in the "medium" category was significantly higher than that in the "low" category at a significant level of 0.05. From these results, we found that the higher the recall rates of machine learning were likely to be, the higher the accuracy rates of human subjects. This indicates that our method of using machine learning can roughly estimate the difficulty of manual notational selection.

8. **Conclusions.** In this study, we addressed the problem of notational variants. When we use a word that has multiple notational variants, it is often difficult to select one. In this study, we determined how difficult it is for machine learning to perform notational selection on words. In addition, we investigated the reasons it was easy for machine learning to select the notational variants of certain words. Our proposed method based on machine learning succeeded in clarifying the differences in usages between two different notational variants of a word that were not described in ordinary Japanese word dictionaries. The results of our study are useful for future studies on notational variants. Our proposed machine learning method uses the maximum entropy algorithm.

By conducting experiments with our proposed method, we drew several conclusions. First, the notational variants of words with "high" recall rates have distinguishing characteristics. For these words, the appropriate notational variant depends on meanings and contexts. Hence, the appropriate notational variant can be selected by using meanings

and contexts. Second, the notational variants of words with "low" recall rates have no distinguishing characteristics. Hence, any notational variant of these words can be used in a sentence. Third, we conducted experiments using human subjects. We confirmed that the higher the recall rates of machine learning were likely to be, the higher the accuracy rates of the human subjects.

Next, we confirmed that our proposed method was useful in performing notational selection. Our experiments show that by using our proposed method, we could perform notational selection for 81 out of 939 words with two notations at a recall rate of 80% or higher. We also confirmed that our method was more effective in performing notational selection than was the baseline method that outputted the notation that appeared most frequently. In our experiments, the average of the accuracy rates of our proposed method (0.87) was more effective than that of the baseline method (0.84).

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number 2350 0178.

REFERENCES

[1] M. Murata and H. Isahara, Automatic detection of mis-spelled Japanese expressions using a new method for automatic extraction of negative examples based on positive examples, *IEICE Transactions on Information and Systems*, vol.E85-D, no.9, pp.1416-1424, 2002.

[2] M. Murata, Q. Ma and H. Isahara, Comparison of three machine-learning methods for Thai part-of-speech tagging, *ACM Transactions on Asian Language Information Processing*, vol.1, no.2, pp.145-158, 2002.

[3] M. Murata, M. Utiyama, K. Uchimoto, Q. Ma and H. Isahara, Correction of errors in a verb modality corpus used for machine translation with a machine-learning method, *ACM Transactions on Asian Language Information Processing*, pp.18-37, 2005.

[4] M. Murata, T. Kanamaru, T. Shirado and H. Isahara, Machine-learning-based transformation of passive Japanese sentences into active by separating training data into each input particle, *Coling-ACL 2006*, pp.587-594, 2006.

[5] Q. She, H. Su, L. Dong and J. Chu, Support vector machine with adaptive parameters in image coding, *International Journal of Innovative Computing, Information and Control*, vol.4, no.2, pp.359-368, 2008.

[6] C.-C. Chen and D.-S. Kao, DCT-based zero replacement reversible image watermarking approach, *International Journal of Innovative Computing, Information and Control*, vol.4, no.11, pp.3027-3036, 2008.

[7] X. Zhang, K. Xiao, G. Gao and G. Teng, The improvement of a feature-based image mosaics algorithm, *International Journal of Innovative Computing, Information and Control*, vol.4, no.10, pp.2759-2764, 2008.

[8] W. Zhang, T. Yoshida, T. B. Ho and X. Tang, Augmented mutual information for multi-word extraction, *International Journal of Innovative Computing, Information and Control*, vol.5, no.2, pp.543-554, 2009.

[9] M. Murata, T. Shirado, K. Torisawa, M. Iwatate, K. Ichii, Q. Ma and T. Kanamaru, Extraction and visualization of numerical and named entity information from a very large number of documents using natural language processing, *International Journal of Innovative Computing, Information and Control*, vol.6, no.3(B), pp.1549-1568, 2010.

[10] Q. Ma, S. Sakagami and M. Murata, Extraction of parallel translation expressions for English-writing support systems, *ICIC Express Letters, Part B: Applications*, vol.2, no.1, pp.113-118, 2011.

[11] D. Miao and S. Wang, A quantitative measurement of brain cognitive function based on human voice separation ability, *ICIC Express Letters*, vol.2, no.1, pp.15-21, 2008.

[12] M. Murata, K. Uchimoto, M. Utiyama, Q. Ma, R. Nishimura, Y. Watanabe, K. Doi and K. Torisawa, Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction, *Cognitive Computation*, vol.2, no.4, pp.272-279, 2010.

[13] J. M. Ruiz-Martinez, J. A. Minarro-Gimenez, D. Castellanos-Nieves, F. Garcia-Sanchez and R. Valencia-Garcia, Ontology population: An application for the e-tourism domain, *International Journal of Innovative Computing, Information and Control*, vol.7, no.11, pp.6115-6133, 2011.

[14] Y. Liu, S. Sui, Q. Zhao, Y. Hu and R. Wang, On automatic construction of medical ontology concept's description architecture, *International Journal of Innovative Computing, Information and Control*, vol.8, no.5(B), pp.3601-3616, 2012.
[15] M. Murata and M. Utiyama, Compound word segmentation using dictionary definitions – Extracting and examining of word constituent information –, *ICIC Express Letters, Part B: Applications*, vol.3, no.3, pp.667-672, 2012.
[16] A. Nishikawa, Y. Watanabe, R. Nishimura, M. Murata and Y. Okada, Dominant variant dictionaries for supporting variant selection, *Proc. of IADIS AC 2009 (IADIS International Conference APPLIED COMPUTING)*, pp.265-269, 2009.
[17] M. Hiki and S. Meiseki, Variability on native speakers' use of the Japanese homophonous synonym 'kaeru', *Journal of Humanities and Social Sciences Nagoya City University*, vol.17, pp.187-200, 2004 (in Japanese).
[18] M. Kojima, M. Murata, J. Kazama, K. Kuroda, A. Fujita, E. Aramaki, M. Tsuchida, Y. Watanabe and K. Torisawa, Using various features in machine learning to obtain high levels of performance for recognition of Japanese notational variants, *Proc. of the 24th Pacific Asia Conference on Language, Information and Computation*, pp.653-660, 2010.
[19] E. S. Ristad, Maximum entropy modeling for natural language, *ACL/EACL Tutorial Program*, Madrid, 1997.
[20] M. Murata, K. Uchimoto, M. Utiyama, Q. Ma, R. Nishimura, Y. Watanabe, K. Doi and K. Torisawa, Using the maximum entropy method for natural language processing: Category estimation, feature extraction, and error correction, *Cognitive Computation*, vol.2, no.4, pp.272-279, 2010.
[21] NLRI, *Bunrui Goi Hyou*, Shuuei Publishing, 1964.
[22] M. Murata, K. Kanzaki, K. Uchimoto, Q. Ma and H. Isahara, Meaning sort – Three examples: Dictionary construction, tagged corpus construction, and information presentation system –, *Computational Linguistics and Intelligent Text Processing, the 2nd International Conference, CICLing 2001*, Mexico City, pp.305-318, 2001.
[23] S. Kurohashi and D. Kawahara, *Japanese Dependency/Case Structure Analyzer KNP Version 2.0*, Department of Informatics, Kyoto University, 2005 (in Japanese).
[24] S. Kurohashi and D. Kawahara, *Japanese Morphological Analysis System JUMAN Version 5.1*, Department of Informatics, Kyoto University, 2005.
[25] I. Niimura, *Kojien (Wide Garden of Words)*, Iwanami Publisher, 1998 (in Japanese).
[26] M. Murata, M. Utiyama, K. Uchimoto, Q. Ma and H. Isahara, Japanese word sense disambiguation using the simple bayes and support vector machine methods, *Proc. of SENSEVAL-2*, 2001.
[27] M. Murata, K. Uchimoto, Q. Ma and H. Isahara, A machine-learning approach to estimating the referential properties of Japanese noun phrases, *Computational Linguistics and Intelligent Text Processing, the 2nd International Conference, CICLing 2001*, Mexico City, pp.142-154, 2001.