Automatic Selection and Analysis of Synonyms in Japanese Sentences Using Machine Learning

Masaki Murata, Yoshiki Goda, Masato Tokuhisa
Department of Information and Electronics
Tottori University
4-101 Koyama-Minami, Tottori 680-8552, Japan

Abstract—In this study, we conducted automatic selection of an appropriate synonym among multiple candidate synonyms for a certain sentence using machine learning. We experimentally confirmed that the accuracy rate of the proposed method using machine learning (0.86) was higher than that of the baseline method (0.70), which always outputs the synonym that appears most frequently in the training data set. Based on the performance of machine learning, we classified synonym pairs into pairs that tend to require proper synonym selection for a certain sentence and pairs that do not tend to require proper synonym selection, i.e., either synonym from the pair can be used for a certain sentence. We also examined the features used in machine learning for some synonym pairs and clarified the information that is useful for selection and usage of synonyms.

Keywords: Synonym, Sentence, Automatic Selection, Machine Learning, Feature

I. INTRODUCTION

Synonyms are different words with similar meaning. For example, “research” and “study” are a synonym pair. There have been some studies of synonyms. Many researchers extracted synonyms from corpora and dictionaries [1], [2], [3]. Murata et al. conducted automatic selection of an appropriate notational variant among multiple candidate notational variants for a certain sentence using machine learning [4]. Notational variants are different expressions of a certain word. Examples of notational variants are “3,” “III” and “three” in English. Examples of notational variants in Japanese are さん (expressed with Chinese characters; cherry in Japanese) and いちご (expressed with Katakana characters; cherry in Japanese). Although Murata et al. used machine learning, they did not handle synonyms.

In this study, we conduct automatic selection of an appropriate synonym among candidate multiple synonyms for a certain sentence using machine learning. For example, we automatically select of a word “study” among candidate synonyms “research” and “study” for a certain sentence. We refer to automatic selection of an appropriate synonym among candidate multiple synonyms for a certain sentence as automatic selection of synonyms. Our study is useful for the selection of appropriate synonyms in sentence generation [5], [6]. We handle automatic selection of Japanese synonyms.

Although there are many studies on extracting synonyms automatically [1], [2], [3] including using machine learning and there are some studies on analyzing synonyms using similarity (without machine learning) [7], there have been very possibly no studies on mainly handling automatic selection of an appropriate synonym among synonyms with very similar meanings for a certain sentence using supervised machine learning. In this study, our study has an originality that we handled automatic selection of synonyms using machine learning.

For word selection including lexical substitution, there are many studies [8], [9], [10], [11]. They studied word sense disambiguation for word selection in machine translation. They handled selection of an appropriate word among words that have different meaning. In contrast, we handle selection of an appropriate word among words that have very similar meaning. Furthermore, we handle classification of synonym pairs into pairs that tend to require proper synonym selection for a certain sentence and pairs that do not tend to require proper synonym selection, which was not handled by their studies.

We used synonyms in EDR dictionary in this study. Synonyms in the EDR dictionary have very similar meanings. In this study, we handle automatic selection of an appropriate synonym among synonyms with very similar meanings.

Although synonyms have a similar meaning and they seem to have no need for selection and can be used for all cases, some circumstances require the selection of a more appropriate synonym. For example, the two Japanese words, いりよう (clothing) and こはぎ (clothes), have the same meaning as “a thing we wear” has in the EDR dictionary. However, いりよう (clothing) can appear just before hin (item) and こはぎ (clothes) cannot appear just before hin (item), i.e., we can say いりよう hin (clothing item) but we cannot say こはぎ hin (clothes item). When we use hin (item), the pair of いりょう and こはぎ must be appropriately selected for a certain sentence and requires proper synonym selection.

The primary goals of this study are as follows. The first goal is to obtain high performance for automatic selection of synonyms using machine learning. The second goal is to classify synonym pairs into pairs that tend to require proper synonym selection and pairs that do not tend to require proper synonym selection. When the selection of synonyms using machine learning is easy, the pair of synonyms is judged to be a pair that requires proper synonym selection. When the selection of synonyms using machine learning is difficult, the pair of the synonyms is considered not to require proper synonym selection; both synonyms in the pair are likely to be applicable to a sentence. The results of our study will contribute to the selection and usage of synonyms.

II. TASK AND PROPOSED METHOD

A. Task

Two synonyms A and B that have a similar meaning are given. We refer to words A and B as target words in this paper. We collect sentences containing either target word. We then eliminate target words in the collected sentences. Then, the task is to estimate which word existed in the eliminated part of a sentence. When a system selects a word that was
in the original sentence, the system’s selection is considered correct.

B. Proposed method

In this study, we estimate which word among two target words is in the original sentence using machine learning. We use sentences containing either of the two target words as training data. We conduct machine learning by handling a target word included in the original sentence as the category (class) of the sentence. The maximum entropy method [12] is used for supervised machine learning.

The maximum entropy method [12], [13] is used for machine learning because it can produce as good performance as support vector machine, and it can output the importance of features.

One machine learning cycle is used for a pair of target words (synonyms). For n pairs of target words, we use n machine learning cycles.

In our experiments, we use the most frequent two target words. Features are extracted from sentences containing either of the two target words. From the two target words, our machine learning approach selects the target word that is more appropriate in a sentence.

The features used in our method are words in sentences, the three words just before and just after the target word, the words in the bunsetsu (phrase) containing a target word, the content words in the bunsetsu modifying the bunsetsu containing a target word, the content words in the bunsetsu modified by the bunsetsu containing a target word, the POSs (part of speech) of those words, and the first 1, 2, 3, 4, 5, and 7 digits of the category number of those words. We consult Murata et al.’s paper [4] for the features.

The category number in the features is a ten-digit number that is described in Bunri Goi Hyou, a Japanese thesaurus. 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 words adjacent to a target word can be used for synonym selection. The syntactic information of a sentence can also be used in synonym selection.

To classify synonyms into a certain group of synonyms where automatic selection of words is conducted properly by machine learning and other groups, we used “high,” “medium,” and “low” categories on the basis of recall rates of synonym selection. A recall rate is a concept similar to accuracy. A recall rate is the ratio of the number of correct outputs over the number of correct data items.

When the lowest recall rate of the two notational variants of a target word is greater than or equal to 0.8, the word is classified in the “high” category because when both recall rates are high, the estimation is accurate. When the lowest recall rate is greater than or equal to 0.5 and less than 0.8, the word is classified in the “medium” category. When the lowest recall rate is lower than 0.5, the word is classified in the “low” category because when one of the two recall rates is low, the estimation is not accurate.

III. DATA SETS USED IN EXPERIMENTS

In this section, we describe the data sets used in our experiments.

In our experiments on synonym selection, we used synonyms appearing in newspapers. We used Mainichi newspaper articles (all the articles in one year edition, 1991). We extracted pairs of words that satisfy all of the following conditions and used them as pairs of synonyms. Condition 1: The two words in the pair have the same concept identification number in the EDR dictionary. Condition 2: Both words have only one identification number in the EDR dictionary, i.e., the two words have only one meaning. Condition 3: Both words appear 50 times or more in the Mainichi newspaper articles (one year edition, 1991). Condition 4: The representative words of the two words that are given using the morphological analyzer JUMAN are different, i.e., the two words differ and are not notational variants.

Condition 1 is used to extract words with very similar meaning. Condition 2 is used to identify and exclude words with multiple meanings because they are complex and not handled by the proposed method. Condition 3 is used to investigate frequently used words in newspaper articles. Condition 4 is used not to handle notational variants because word selection in notational variants has already been addressed in a previous study [4].

In our data set, 90 word pairs satisfy all four conditions. From these, we randomly extracted 45 pairs to use in our experiments. For each word pair, 100 to 10000 data items exist.

IV. EXPERIMENTS ON SYNONYM SELECTION USING MACHINE LEARNING

A. Experimental methods

We applied our machine learning approach to the 90 word pairs obtained through the process described in Section III.

In our experiments, we used the most frequent two synonyms. Our machine learning approach selects the word from the two synonyms that is more appropriate for a given sentence.

For each pair of words, sentences containing each word were extracted from the Mainichi newspaper articles (one year edition, 1991) and were used as a data set. Sometimes, the same sentences occurred more than once in Mainichi newspaper articles; thus, we eliminated redundant sentences so that each sentence occurred only once. We conducted a ten-fold cross validation for each data set (each word pair). 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 using the maximum entropy method. Then, the estimated category was evaluated using the correct category in the test data set. This process was repeated for all ten parts. Consequently, all ten parts were evaluated.

We used a baseline method to compare the proposed method. The baseline method always outputs the synonym that appears most frequently in the training data set.

B. Experimental results

The results of classifying 45 words into the three categories by machine learning recall rates are shown in Table I. The three categories are “high,” “medium,” and “low,” as discussed in Section II-B. The average accuracy rates for the proposed method and the baseline method are shown in Table II. Values for each category based on recall rates are also described. The results of a comparison of the proposed and baseline methods relative to the accuracies of each synonym pair are shown in Table III. “Even” shows the number of synonym pairs for which the difference is 0.01 or less. “Proposed method
TABLE I: Ratios of words classified into the three categories on the basis of recall rates

<table>
<thead>
<tr>
<th>Category</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.11 (5/45)</td>
</tr>
<tr>
<td>Medium</td>
<td>0.51 (23/45)</td>
</tr>
<tr>
<td>Low</td>
<td>0.38 (17/45)</td>
</tr>
</tbody>
</table>

TABLE II: Average accuracy rates of the proposed method and baseline method

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method</td>
<td>0.95</td>
<td>0.84</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td>Baseline method</td>
<td>0.62</td>
<td>0.69</td>
<td>0.78</td>
<td>0.70</td>
</tr>
</tbody>
</table>

(win)” indicates the number of synonym pairs for which the accuracy rate of the proposed method is higher than that of the baseline method. “Baseline method (win)” indicates the number of synonym pairs for which the accuracy rate of the proposed method is lower than that of the baseline method.

V. DISCUSSION

A. Comparison of the proposed and baseline methods

In Table II, the accuracy rates of the proposed method and the baseline method are 0.86 and 0.70, respectively. The accuracy rate of the proposed method was higher than the baseline method. From Table III, we found that the accuracy rate of the proposed method was higher than that of the baseline method for all “high” and “medium” synonym pairs and was higher for 11 “low” synonym pairs. The synonym pairs where the proposed method won against the baseline method (39 = 5 + 23 + 11) corresponded to 80 percent or more of all 45 synonym pairs. The results indicate that the proposed method and features used in machine learning are useful for synonym selection.

However, in the “low” category, there were four pairs of “even” and two pairs for which the accuracy rate of the baseline method was higher than that of the proposed method. This was likely caused by significant differences in the frequencies of the two words in a synonym pair. This is a case where the frequency of a word in the two words is extremely higher than the frequency of the other word. The baseline method always outputs the synonym that appears most frequently in the training data set among the synonyms. Therefore, when the frequencies of the two words differ significantly, the accuracy rates of the baseline method are likely to be higher. Thus, there would be a small number of cases for which the baseline method was better than the proposed method.

B. Trends based on the degree of recall rates

We examined classified word pairs manually. Most word pairs that were classified as “high” tend to require proper synonym selection. An Example of “high” is the pairing of chokin (“deposit” or “saving”) and chochiku (“storage” or “saving”). Example sentences are below.

Example sentence 1a (Machine learning correctly selected a synonym):

*yubinkyoku* no *chokin* *gaku* *jougen wa*
(post office) (saving deposit) (amount) (maximum)

*hikiagerareru*
(be raised)

(The maximum amount for a saving deposit at a post office will be raised.)

Example sentence 2a (Machine learning correctly selected a synonym):

*iriou* *hin* no *uriage* ga *akii*
(clothing) (item) (sale) (large)

(The sales of clothing items are large.)

Example sentence 1b (Machine learning correctly selected a synonym):

*toashi* ga *yowaku* *chochiku* ga *chisai*
(investment) (weak) (storage) (small)

(Investment is weak and storage is small.)

A pair of chokin (“deposit,” “saving,” or “saved money”) and chochiku (“storage,” “saving,” or “saved thing”) was classified as “high.” These words have the same concept identification “0be6c6” in the EDR dictionary. In the “high” category, machine learning obtained good recall rates (good performance) and word selection was easy and necessary.

The results of machine learning for the pair are shown in Table IV and the three important features with the highest α values are shown in Table V. A normalized α value represents the degree of importance of the corresponding feature used to estimate a synonym learned using the maximum entropy method. Further details can be found in the literature [13]. From “F1: *teigaku* (fixed-amount)” and “*yubin* (post)” in Table V, we found that chochiku is used for a sentence related to postal savings. This is also true of Example sentence 1a. In Example sentence 1a, *youbinkyoku* (post office) was used. From “F1: *touhi* (investment)” and “F1: *beikoku no* (USA),” we found that chochiku was used for a sentence related to investment or countries. We found that chokin and chochiku have their own characteristics, and they tend to require proper synonym selection. This result matches the results of machine learning with high recall rates that determine that word selection is necessary.

Word pairs that were classified as “medium” include two cases. The first case occurs when the meaning of one word of the pair includes the meaning of the other word and is less specific than the meaning of the other word. The second case occurs when only one word of the pair can be used when a certain word appears in the sentence.

An example of the first case is *En* (banquet) and hirouen (reception banquet). Although *en* and hirouen are synonyms in the EDR dictionary, hirouen is a certain special banquet (reception banquet), and the meaning of hirouen is included and is more specific than the meaning of *en* (banquet).

Examples of the second case are shown below.

Example sentence 2b (Machine learning correctly selected a synonym):

*denka seihin* to *iru* *mouhu ga aru*
(electric appliances) (clothes) (blankets) (exist)
A pair of iryou (“clothing material” or “clothing”) and irui (“clothes”) was classified as “medium.” These words have the same concept identification “0e504a” in the EDR dictionary. The results of machine learning for the pair are shown in Table VI. From “F1: hin (item)” in Table VII and Example sentence 2a, we found that iryou (clothing) is used when it appears just before hin (item). However, there were no characteristics other than the above characteristic. In many cases, both iryou and irui can be used. Therefore, we found that iryou and irui do not have their own characteristics, and the tendency to require proper synonym selection was not significantly high.

Most word pairs that were classified as “low” do not tend to require proper synonym selection. Word pairs that were classified as “low” include pairings of a word and its abbreviated word and pairings of a Japanese word and its imported word (a word derived from a foreign word). An example of a pairing of a word and its abbreviated word is *chochiku* (a word) and *cho* (an abbreviated word) (they have the same meaning, i.e., “energy saving”). An example of a pairing of a Japanese word and its imported word is *utai monku* (a Japanese word) and *kycchi hureizu* (its imported word); they have the same meaning, i.e., “catch phrase.” In these cases, the two words in a synonym pair have very similar meaning and do not tend to require proper synonym selection, and can be used for almost all the sentences.

### VI. Conclusion

In this study, we conducted automatic selection of an appropriate synonym among multiple synonyms for a certain sentence using machine learning. In our experiments, we confirmed that the accuracy rate of the proposed method using machine learning (0.86) was higher than that of the baseline method (0.70), which always outputs the synonym that appears most frequently in the training data set. Based on the performance of machine learning, we classified synonym pairs into pairs that tend to require selection for proper usage and pairs that do not tend to require selection. A pairing of *chokin* (“deposit” or “saving”) and *chochiku* (“storage” or “saving”) was classified as tending to require selection to ensure proper usage. We examined features for some synonym pairs and clarified the information that is useful for synonym selection.

### ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23500178.

### REFERENCES