2022 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2022, Virtual, December 12-15, 2022

Genre Classification of Modern Japanese Literary Works Based on Word Vectors

Shiori Takenaka†, Jousuke Kuroiwa†, Tomohiro Odaka† and Izumi Suwa‡

†Graduate School of Engineering, University of Fukui
3–9–1 Bunkyo, Fukui, 910-8507, Japan
‡School of Life Science, Women's College of Jin-ai
43–1–1 Amaikecho, Fukui, 910-0124, Japan
Email: jou@u-fukui.ac.jp

Abstract—Word vectors can quantitatively represent the meaning of words in the reduced dimension space. However, the effectiveness of the word vectors of Japanese has not been discussed as much as those of English. Therefore, we investigate the effectiveness of word vectors obtained from the Japanese corpus. The purpose of this paper is to show that word vectors have the ability to represent the characteristic features of each genre of modern Japanese literary works through the classification of works. In classification experiments, we have trained SVM(Support Vector Machine) to perform the categorization giving to two corpora of literary works as inputs. The first categorization is to divide novels and poetry works of a famous modern Japanese novelist. Another is to classify them as novels and essays of the other famous author. The best accuracy of the test data is about 0.97 for each novel and poetry classification, for each novel and essay is about 0.90. This means that the characteristic features of these categories could be represented enough to distinguish between novels and poetry works. The result reflects the similarities between novels and essays compared to ones between novels and poetry works.

1. Introduction

Words that occur in the same context tend to have similar meanings from the view point of the distributional hypothesis [1]. Word2Vec is based on the distributional hypothesis [2]. The projection layer of the Word2Vec can acquire a distributed representation as the word vectors that reflects the meaning of the words. The word vectors have the advantage of being easy to handle on the computer because they are reduced in dimensionality compared to the one-hot vectors, which is the input of Word2Vec. The practicality of word vectors in English sentences has been the focus of much attention and research, but relatively little attention has been paid to Japanese sentences. In addition, it has been shown that the word vectors acquired from Japanese performance is inferior to English in some cases.

ORCID iDs First Author: 0000-0003-4292-1345, Second Author: 0000-0002-6307-5603, Third Author: 0000-0003-4288-5460, Fourth Author: 0000-0002-4625-2911

However, we expect the word vector, which quantitatively represents the meaning of the word in a low-dimensional way, are practical enough for expressing the features of Japanese literary works. Actually, previous studies have followed this assumption and defined the feature value as the center of gravity of the word vector corresponding to all the words used in the sentence of the work. We have confirmed that the writer's features appeared in this feature value.

From this result, we have noticed the possibility that the feature value also would represent the characteristic features of the work's genre. In other words, word vectors acquired from Japanese may have the ability to express genre characteristic features. Therefore, the purpose of the paper is to show that even Japanese literary works by the same writer have different characteristic features depending on the genre of the works using word vectors.

2. Feature Value with Word Vectors

2.1. Continues Bag-of-Words Model

The CBOW(Continues Bag-Of-Words) model, proposed in 2013 by Mikolov et al., is one of the NN models of Word2Vec [2] as shown in Figure 1. This model predicts a word v_i from the context word chain of $v_{i-n}, \ldots, v_{i-1}, v_{i+1}, \ldots, v_{i+n}$ with the window size of n. Vector representations of word meanings can be acquired from the projection layer of the trained CBOW model. Word vector $\boldsymbol{x}_i$ that means i-th word v_i in the corpus can be obtained by the following equation,

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{v}_i \tag{1}$$

where $\boldsymbol{v}_i$ is the one-hot vector of the word of v_i and $\boldsymbol{W}$ is the weight matrix between the input layer and projection layer, and has already been trained as shown in Figure 1. The dimension can be reduced by taking the inner product of $\boldsymbol{v}_i$ and $\boldsymbol{W}$.

2.2. Feature Values

The feature of a work could appear in what vocabulary is used and how often used it in the text. Therefore, we employ the center of gravity of word vectors of all the words

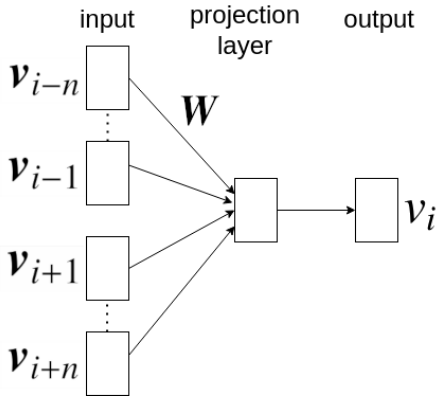

Figure 1: Conceptual diagram of CBOW model

Table 1: Number of works per author for training CBOW

author name	number of works
Sakaguchi Ango	487
Akutagawa Ryunosuke	371
Makino Shinichi	343
Miyazawa Kenji	275
Dazai Osamu	272
Natsume Soseki	109
Kikuchi Kan	75
Kajii Motojiro	47
Arishima Takeo	45
Nakajima Atsushi	34

used in the work as the feature value. Even if the same word appeared again and again, we permit repeated addition in the evaluation of the center of gravity.

The center of gravity of the αth works of the author A is represented by,

$$\boldsymbol{f}^{(A,\alpha)} = \frac{1}{|X^{(A,\alpha)}|} \sum_{i \in X^{(A,\alpha)}} \boldsymbol{x}_i^{(A,\alpha)} \tag{2}$$

where $\boldsymbol{x}_i^{(A,\alpha)}$ denotes the word vector of the one-hot vector of the ith word of the αth works of the author A, $X^{(A,\alpha)}$ describes the set of its vocabulary with the repetition, and $|X^{(A,\alpha)}|$ is the number of the set of $X^{(A,\alpha)}$.

3. Experiment of Genre Classification

3.1. Shaping corpus for learning

In the experiment, we employ 2058 works written by 10 modern Japanese famous novelists published on the web site Aozora Bunko *. The corresponding number of works written by the authors are listed in Table 1. All the works are trained in CBOW as a corpus for acquiring distributed representations of words.

The text files published in the Aozora Bunko include the title of the work, the name of the author, the date of publication, and the other information such as annotations, besides the content of the text. These descriptions not related to the content of the text are deleted as unnecessary for the training. For the same reason, punctuation marks that used frequently, "、", "。", "「", "」", "(" and ")", are also removed. We simultaneously format the corpus as a single-sentence delimiter by delimiting at the place just after marks "。", "!" and "?". A space is inserted before and after each word by the morphological analyzer Mecab since Japanese sentences do not have clear breaks between words. Finally, words in the corpus applied the above processing are transformed into one-hot vectors as input of CBOW model.

*Aozora Bunko has works whose copyrights have expired, digitized into a text file and XHTML(some in HTML) format.

3.2. Classification Method

At first, we have calculated one-hot vector representations of words in 2058 text data applied the processing. We have implemented CBOW model to acquire word vectors with Gensim's Word2Vec. The model has been trained with training parameters, the window size $n = 5$, dimensions 100, number of training epochs 500, and the minimum number of occurrences 1. The window size n specifies the amount of context. All 2058 works of 10 authors are applied to train CBOW model for acquiring the word vectors, and 400 works by two authors are chosen to classify the genre of the works. The first author is Miyazawa, and the other is Akutagawa. From the works of these two authors, we chose 100 works of novel and poetry, respectively, from Miyazawa's works. Likewise, we chose 100 works of novels and essays from Akutagawa's works.

The correct genres are classified based on the NDC(Nippon Decimal Classification) numbers tagged in the Aozora Bunko. In NDC, the classification number 911 is assigned to poetry, 913 is assigned to novels, fiction, and romance, 914 is assigned to essay and prose [3]. For simplicity, this paper refers to genre 913 as "novel", 914 as "essay". In addition, works with the same content but with differences in wording due to the publisher were considered duplicates. We chose the works are no overlap for the classification.

We defined the feature value f of the 400 works by the Equation 2. Each author's works were separated into training data and test data in a ratio of 7:3. Furthermore, the feature $\boldsymbol{f}$ of each work was assigned a label representing the two genres. We inputted training data sets into liner soft-margin SVM algorithm implemented in scikit-learn library for prior learning. After that, we inputted test data into trained SVM to determine the accuracy of its classification. We trained SVM with several patterns by changing the parameter C. C is generally called the regularization parameter in optimization problems solved for $\boldsymbol{w}$ and ζ by SVM(Equation 3). $\boldsymbol{w}$ is parameter corresponding to synaptic weight, $\boldsymbol{\zeta}$ is a non-negative variable that relaxes the con-

Table 2: Miyazawa's Classification results between novel and poetry

C	train_acc	test_acc
10^{-4}	0.51	0.48
10^{-3}	0.99	0.97
10^{-2}	0.98	0.97
1	1.00	0.92

straint.

$$\min_{\boldsymbol{w},\boldsymbol{\zeta}}\left(\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}\zeta_i\right) \tag{3}$$

3.3. Result of Classification

For each author's work, we calculated the accuracy of the training data and the test data when we changed the regularization parameter. Note that the larger C is the less SVM allows for error. The results of classifying Miyazawa's works into novels and poetry works are shown in Table 2, The results of classifying Akutagawa into novels and essays are shown in Table 3.

The result of Miyazawa has the highest accuracy of 97% for the test data. In the case of Akutagawa, the highest accuracy score of the test data is 90% When C is greater than $C = 10^2$. Comparing the results in Tables 2, 3, it is found that the classification of novels and poetry had a higher accuracy than the classification of novels and essays.

4. Discussion

4.1. Evaluation

Table 2 shows the characteristics of the two genres, novel and poetry, were able to emerge in a distinguishable way. In Table 3, the best accuracy on the test data is 90%, which was able to distinguish between novel and essay. On the contrary, about 10% of them are incorrectly classified.

Therefore, we compare the results with the human classification to see whether the 10% misclassification is a large value. The subjects were two people, the one reads books regularly and the other doesn't read regularly. They were asked to classify the genre of the 10 works in Akutagawa as novels or essays. The works used for this classification were selected at random so that there is no overlap in content. As a result, the accuracy of the classification of subjects who read books frequently was 90%, the other was 80%. Taking these results into account, the misclassification of 10% between novels and essays as well as human's classification performance.

4.2. Visualization of Features to See Difference of Two Classifications

The experimental results suggest that poetry are more easily to distinguish from the novel than the essay. To confirm this visually, we used PCA(Principal Component Analysis) to project the feature values f. Figure 2 shows the result of PCA of the works of Miyazawa. Novels are orange dots and poetries is green dots. The cumulative contribution ratio of the first and second principal components was 0.36. Figure 3 shows the result of PCA of the works of Akutagawa. Novels are orange dots and essays in blue dots. The cumulative contribution was 0.40. Compared to Figure 3, dots of different colors, i.e. different genres, tend to be distributed without overlapping in Figure 2.

Table 3: Akutagwa's Classification results between novel and essay

C	train_acc	test_acc
10^{-3}	0.76	0.68
10^{-2}	0.87	0.80
1	0.95	0.83
10	1.00	0.87
10^2	1.00	0.90

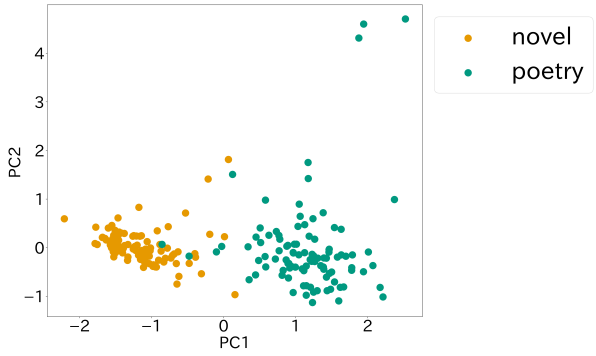

Figure 2: Projected feature values of Miyazawa's works

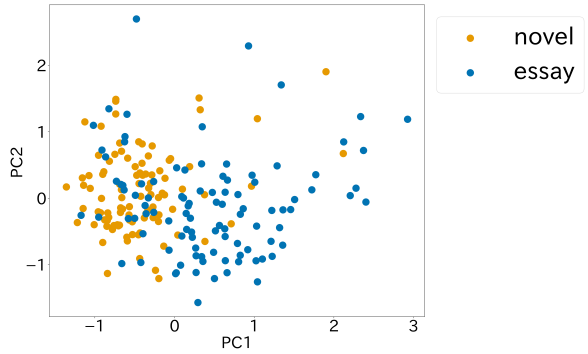

Figure 3: Projected feature values of Akutagawa's works

4.3. Classification Using One-hot Vector

One major advantage of using word embedding is that it can reduce the dimension. In this section, we perform the same classification as Chapter 3 using one-hot-vectors without word embedding. We compare the results with the

Table 4: Miyazawa's Classification results between novel and poetry using one-hot-vector

C	train_acc	test_acc
1	0.51	0.48
10	0.99	0.97
10^2	0.99	0.97
10^3	1.00	0.93

Table 5: Akutagawa's Classification results between novel and poetry using one-hot-vector

C	train_acc	test_acc
10^2	0.89	0.77
10^3	0.99	0.85
10^4	1.00	0.82

previous experimental results in Table 2, 3. We obtained one-hot-vector representations of the vocabulary used in the training and test works. These works are the same as those used in the aforementioned experiment. The dimensional number V of the one-hot-vector was the number of vocabularies in the corpus: $V_{\text{Miyazawa}} = 20743$ for Miyazawa and $V_{\text{Akutagawa}} = 30542$ for Akutagawa. Table 4 shows the classification results for Miyazawa's works, Table 5 shows the classification results for Akutagawa's works. Comparing Table 4, 5 which is the result using one-hot-vector, and Table 4, 5 which is the result using word embedding, the accuracy of the former was found to not inferior or slightly worse than that of the latter for both authors.

Two things can be understood from this result. The first is that the assumption that word types and their frequency of occurrence represent the characteristics of a sentence is more reinforced since the accuracy of the classification results with the one-hot-vector is high enough. Second, the word vectors fully reflect the characteristics of the distributional hypothesis, in which the meaning is determined by the position of the word, compared to the one-hot vector. Rather, the novel and essay of Akutagawa better reflect the features and also allow for classification with a higher degree of accuracy. Therefore, the word vector are effective enough for expressing the feature of Japanese literary works in a low-dimensional way.

4.4. Effects of Punctuation Marks

In the processing of the experiment in section 3.1, several punctuation marks were removed from the corpus. The most frequent symbols "、" and "。" correspond to commas and periods in English. The "「" and "」" work like double quotation marks, the "(" and ")" are used to supplement information. In this section, we examine the effect of these punctuation marks.

We conducted an experiment under the same conditions as above without erasing the punctuation marks. Table 6 shows the classification results of Miyazawa's works, and Table 7 shows the results of Akutagawa's works. Comparing the result of Miyazawa in this section to the result of Miyazawa in Chapter 3, the best accuracy of the test data has not changed, but the dependence of accuracy on the value C has decreased in this result. Because we increased the value of the parameter C, the accuracy value did not change from 0.97. In contrast, the accuracy of Table 7 tends to be lower than that of Table 3. The tendency for poetic works to be more distinguishable than essay works was stronger than when the punctuation marks were not removed. These symbols represent characteristics and contribute to differentiation from other genres because punctuation marks are seldom used in poetry.

Table 6: Miyazawa's Classification results between novel and poetry without erasing two punctuation marks

C	train_acc	test_acc
10^{-4}	0.51	0.48
10^{-3}	0.99	0.97
10^{-2}	0.99	0.97

Table 7: Akutagawa's Classification results between novel and essay without erasing two punctuation marks

C	train_acc	test_acc
10^{-3}	0.73	0.68
10^{-2}	0.87	0.82
1	0.96	0.80

In the case of the essays and the novels, punctuation marks were used frequently in both genres in a similar sense, which may have a negative effect in terms of differentiating features.

From the above, the word vectors acquired from CBOW have a sufficient performance to express the characteristics of genres. If the emphasis is on differentiating the features of a genre, it is better to devise and obtain the features by assuming in advance what symbols and parts of speech will represent the features of the genre.

References

[1] Z. S. Harris, "Distributional structure," *Word*, vol. 10, pp. 146-162, 1954.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[3] K. Mori, "Nippon Decimal Classification New And Revised 10th Ed.," *Japan Library Association*, pp. 412-417, 2014.