ICIC Express Letters
Part B: Applications
Volume **14**, Number **6**, June **2023**

ICIC International ©2023 ISSN 2185-2766
pp. **657–662**

ACQUISITION OF SYNONYMS FOR COLLOQUIAL EXPRESSIONS IN SCIENTIFIC REPORTS USING *WORD2VEC*

Asako Ohno

Faculty of Engineering
Osaka Sangyo University
3-1-1 Nakagaito, Daito, Osaka 574-8530, Japan
ohno@eic.osaka-sandai.ac.jp

Received November 2022; accepted January 2023

Abstract. *Recent years have seen a decline in Japanese students' reading and writing skills. Although the faculty of engineering departments require students to write experiment reports as part of their coursework, much time is spent on Japanese writing itself rather than the content. Additionally, report instruction is generally inconsistent depending on the faculty member's field of expertise and experience, and there is no systematic instructional method. Therefore, we developed an educational support tool to detect colloquial language, which is one of the most frequently observed items in report instruction by engineering teachers. This tool is designed to help students acquire self-correction ability by having them identify colloquialisms themselves. Because the list of colloquial words used in this tool for science papers is limited to a small vocabulary of detectable words, this study attempts to add synonyms acquired using Word2Vec to our proposed list of colloquial words for science papers to make it capable of handling more colloquial words. We verified the results on actual report documents submitted by students and confirmed that the colloquial word list with added synonyms can be used to detect more colloquial expressions than the conventional one.*
Keywords: Academic writing, Scientific reports, Colloquial words detection, Education support tool, *Word2Vec*

1. **Introduction.** In recent years, there has been a decline in the writing skills of Japanese students. According to the latest (2018) results of the *Programme for International Student Assessment* (*PISA*), conducted by the *Organization for Economic Cooperation and Development* (*OCED*) or 15-year-old students, Japan ranks 15th out of 77 countries in reading comprehension, 6th out of 78 countries in mathematical literacy, and 5th out of 78 countries in scientific literacy. Reading comprehension is at an all-time low [1], which is detrimental as reading and writing skills are essential for university students. In many classes, students are required to write reports to check their understanding of the content, and writing a graduation thesis is a requirement for completing a bachelor's degree. Consequently, "Academic writing" classes are generally offered to improve the reading and writing skills of Japanese students in their first year of college. However, it is difficult for students who have not received sufficient training in reading and writing before entering university to become proficient in academic writing through first-year education alone. This is especially true in the Faculty of Engineering, where the author belongs.

This study aims to improve the writing skills of science students, several of who struggle in this regard. However, scientific reports do not require a high level of skill or the use of metaphors, and by understanding the rules of scientific writing and familiarizing themselves with the format, even students who are poor at writing can fulfill the requirements for writing such reports. In this study, we focused on the detection of colloquial words and rewriting them into formal words, which is one of the most frequently identified

DOI: 10.24507/icicelb.14.06.657

issues by instructors in teaching reports in the Faculty of Engineering. Further, we have developed a prototype version of a colloquial words checker for scientific reports [2]. We add synonyms to the *Scientific Report Colloquial Words list* (*SRCW list*) proposed in the previous study based on *Yamashita's Report Colloquial Words list* (*YRCW list*) [3] using *Word2Vec* [4,5] aiming to make the tool handle more colloquial words.

The remainder of this paper is organized as follows. Section 2 introduces related studies and existing systems developed to improve academic reading/writing skills. Section 3 explains the basic concept of our method. Section 4 reports the procedure for adding synonyms to an existing vocabulary on our *SRCW list*. Then Section 5 concludes the paper.

2. **Related Studies and Existing Systems.** The ability to input (read) and output (write) information from written media is essential in today's information society. Furthermore, it was recently revealed by the 2018 *PISA* results that there is a decline in the reading comprehension skills among the youth in Japan [1]. Many studies have been conducted to improve students' reading and writing skills regardless of their age.

The *Reading Skill Test* (*RST*) [6,7] is a computer-based test (CBT) that assesses "*the ability to read and comprehend questions according to the rules of the Japanese language*" for children in the sixth grade and above as well as adults. Previous RST research has revealed a correlation between reading comprehension and deviation score, i.e., improvement in reading comprehension is essential for academic achievement [7].

Evaluating student submissions often relies on instructors' experiential linguistic knowledge and intuition, making it difficult to generalize the evaluation of learners' writing skills [8]. Lee et al. developed a web-based real-time writing assessment system called *jWriter* [9]; as it is designed for students learning Japanese as a foreign language, it is based on the *I-JAS Corpus* (*International Corpus of Japanese as a Second Language*). The system evaluates sentences based on the difficulty of the words and vocabulary in the input sentences, as well as the number of connecting words, thus allowing Japanese speakers to obtain useful feedback on their own writing. Lee and Hasebe also developed a readability evaluation system called *jReadability* [10]. They visually classified texts extracted from 100 Japanese textbooks into six readability levels. Then, they conducted a discriminant analysis based on the average sentence length, the percentage of Chinese words, Japanese words, verbs, and parts for each group of texts at the six levels. The system calculates readability using a linear regression equation composed of the above variables for the input texts.

A vast vocabulary is considered to be an important requirement for second language acquisition. In other words, measuring the size and quality of a subject's vocabulary may provide an estimate of how proficient he/she is in the language. Hamada et al. [11] developed a method to measure the vocabulary size for Japanese learners of English to assess the learners' vocabulary proficiencies. By using 8,000 words from the new JACET (Japan Association of College English Teachers) Basic Word List, they aimed to eliminate inaccuracy of the content and overestimation of the estimated vocabulary size. It may also be possible to apply this to assessing the reading and writing ability of Japanese students.

As we have seen, most studies have focused on assessing reading and writing skills and it is therefore crucial that these attempts accurately represent and improve students' reading and writing skills. With regard to the specifics of what needs to be improved, we consider the following example: when writing reports, university students are required to explain their content logically. However, in this study, we aim to eliminate more basic expressions that are unlike the ones used in reports; in other words, we focus on colloquial words.

Yamashita et al. [3] developed a web-based system, *Colloquial Words Checker*, to eliminate colloquial words in report documents. First, Yamashita et al. collected colloquial words from 13 Japanese language textbooks and constructed a list of such words. We refer to this list as *Yamashita's Report Colloquial Words list* (*YRCW list*). The tool developed by Yamashita et al. displays the colloquial words in the text input by the user as well as the corresponding written words. This allows the user to obtain information about the colloquial words in his or her own written reports and on ways to rewrite their reports.

3. Colloquial Words Checker for Scientific Reports.

3.1. Basic idea behind the development of the colloquial words list.

This study aims to improve the ability of students to detect expressions that are not typically used in reports, i.e., colloquial language. This ability is easier to acquire via self-learning and entails translating the colloquial words into those that are more appropriate for scientific reports.

The basic analysis performed in this study revealed that many of the colloquial expressions in the *YRCW list* including the corresponding written expressions, are not used in scientific reports [2]. We therefore excluded the highly broken colloquial expressions contained in the *YRCW list* and added the ones extracted from the reports submitted in previous years' classes, to construct the *Scientific Report Colloquial Words list* (*SRCW list*) [3]. The colloquial words on the list were then classified into the following three categories: (Category 1) Appropriate for science reports: the written word that corresponds to the colloquial word is appropriate for science reports, (Category 2) Usable with caution: The written words that are colloquial but can be used in scientific reports (however, they need to be used with caution), and (Category 3) Inappropriate: written words that are colloquial and cannot be used in scientific reports.

The classification was based on the subjective judgments of two teachers: the experiment supervisor and a teacher who teaches Japanese writing to engineering students. See [2] for examples of specific colloquial expressions classified into each category.

3.2. Overview of our tool.

Self-correction of colloquial words requires two abilities: 1) the ability to find colloquial words and 2) the ability to rewrite the identified colloquial words into scientific ones. We developed a prototype of web-based colloquial words checker using *Node.js* to improve this "*self-detection*" ability [2]. Specifically, the procedure employed can be described as follows: students first input a portion of the text of a report and then upload it to the tool by selecting the colloquial words contained in the text on the GUI. The tool then matches the entered colloquial words with the ones on the *SRCW list* and returns the corresponding formal words if there is a match. In other words, the tool performs the detection of colloquial words on its own, and only for those words that it is able to find on its own, the corresponding formal words and usage notes can be visually verified. This allows the user to learn the process of detecting colloquial words and converting them into scientific ones in a step-by-step manner. Further details of this procedure are provided in [2].

4. Acquisition of the Colloquial Vocabulary Using *Word2Vec*.

4.1. *Word2Vec*.

Word2Vec is used in a variety of fields to acquire synonyms. For example, Nguyen et al. [12] used it to determine similar words for opinion mining. In our previous study, we calculated the similarity between the words in the discussion corpus as well as between those in the taxonomy classification table and increased the number of keywords used for classification by adding words with high similarity to the classification table [13].

One-hot vector is a vector whose dimension is equal to the total number of words in the vocabulary, where any one word in the vocabulary is represented by 1 and all the

other words are represented by 0. The number of dimensions of this vector increases with the increase in the number of words in the vocabulary. The distributed representation of words involves expressing the meaning of a single word as a low-dimensional vector of tens to hundreds of dimensions. The basis of the distributed representation is distributional hypothesis: the meaning of a word is determined by its surrounding words. For example, in the sentence, *"I drink soymilk every day"*. If the meaning of the word *"soymilk"* is unknown, it can be inferred from the surrounding words that *"soymilk"* is a type of drink.

Mikolov et al. proposed a method for learning high quality word vectors from a large number of words [4,5]. *Word2Vec* allows us to represent the features of the word using a feature vector as distributed representation. It has the structure of a two-layer neural network consisting of an input layer, a hidden layer, and an output layer, as shown in Figure 1. All the units of the adjacent layers of this network are connected. The occurrence probabilities of the surrounding words are learned by taking a word as input and providing the surrounding words as training data. We obtain the weight matrix $\boldsymbol{W}_{V\times N}$ between the input and hidden layers of the model tentatively in this manner. This weight matrix is the distributed representation of the words, and the goal is to obtain an n-dimensional feature vector that represents features of the input word using the one-hot vector as a lookup table.

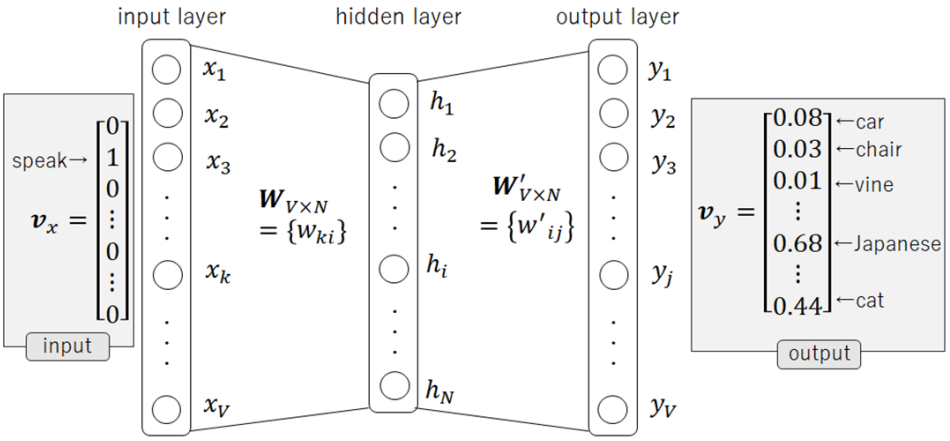

FIGURE 1. Structure of *Word2Vec* (based on the Skip-gram model)

The occurrence probabilities of surrounding words, which is the output of the model, is obtained by calculating the inner product of the weight matrices, which is the distributed representation of the word, and the weight matrix, which is the distributed representation of the surrounding words. The next step is to calculate the cosine similarity between the words. We regard the top 10 similarity words obtained by *Word2Vec* as synonym candidates.

4.2. **Acquisition of synonyms.** We used *gensim*, an open-source Python library for vectorizing the documents, to acquire synonyms using *Word2Vec*. Here, *fastText* [13] and *chiVe* [14] were used as Japanese distributed representations, both of which have a large vocabulary with a vector dimension of 300. *fastText* [13] is a trained *Word2Vec* for Japanese language and has a vocabulary of 2 million words. Its training data were acquired using a web crawler and Wikipedia, and its morphological analyzer is *MeCab*. *chiVe* [14] is the largest Japanese word distribution representation (over 3 million vocabulary) and uses the *Sudachi* morphological analyzer and the National Institute for Japanese Language and Linguistics Web Japanese Corpus (NWJC), which is a large corpus of 25.8 billion words. All of *Sudachi*'s multi-granular segmentation results were used for training *chiVe* to strengthen its vocabulary. In addition, longer entity representations and compound words were made to become more similar to their constituent words.

We obtained synonyms for the 153 words in the *SRCW list* using *fastTest* and *chiVe*. The top 10 words in the calculated similarity per individual colloquial word were treated as candidate synonyms. We determined correctness of the synonyms visually, and the number of correct answers out of the top 10 was calculated as the precision of the synonyms. Precisely, the average precision was calculated by averaging the precision of all 153 results. With *fastText*, we could acquire more synonyms (1,120) than when using *chiVe* (590), whereas *chiVe* had a higher average precision (35%, 208 correct synonyms) than *fastText* (31%, 351 correct synonyms). Overall, a total of 559 synonyms were obtained for the colloquial words in the *SRCW list*.

4.3. **Comparison of colloquial word detection performance.** Among the 559 synonyms acquired, words that could be regarded as the same word, such as the difference between kanji and hiragana, words that were not appropriate for the report, and could be regarded as written words rather than colloquial words were excluded. As a result, as shown in row (a) of Table 1, the total number of colloquial words acquired by *chiVe* was 181* and by *fastText* was 293* (* Some of these words are common to *chiVe*, *fastText*, and *SRCW list*). *All*, which includes the vocabulary contained in all *chiVe*, *fastText*, and *SRCW* without duplication, had a vocabulary count of 549.

Table 1. Comparison of the # of kinds of detected words in 4 sets of colloquial words

Evaluation items	*chiVe**	*fastText**	*SRCW*	*All*
(a) # of colloquial words	181	293	153	**549**
(b) # of kinds of words detected from reports	26	59	52	**107**
Precision (a)/(b)	14%	20%	34%	19%

Next, colloquial words in the four sets of vocabulary listed in Table 1 were retrieved from the entire 120 reports submitted by 20 students as class reports for the first semester of the 2022 experiment. *All* detected 107 types of colloquial words. Although the precision was less than the original *SRCW*, the total vocabulary count increased by 396 and the number of colloquial word detected increased by 55, compared to the original *SRCW*. We have already revealed that *SRCW list* detects more colloquial words than *YRCW list* in [2]. Here, the new *SRCW* (*All*: *chiVe* + *fastText* + *SRCW list*) showed the capability to detect more colloquial words than previous *SRCW list* from an unknown set of reports. We confirmed that the *Word2Vec* vocabulary acquisition improved the colloquial word list.

5. **Conclusions.** This paper reported our first attempt to acquire synonyms using *Word-2Vec* and to increase the vocabulary of a list of colloquial words for scientific reports. We confirmed that the *Word2Vec* vocabulary acquisition improved the colloquial word list. Future tasks include improving the accuracy of synonym acquisition using *Word2Vec* and testing the tool in actual classes.

Acknowledgment. We would like to express our gratitude to Ms. Y. Nakagawa, Lecturer of Osaka Sangyo University, for her cooperation and advice during our study.

REFERENCES

[1] National Institute for Educational Policy Research, Ministry of Education, Culture, Sports, Science and Technology, *Key Features of OECD Programme for International Student Assessment 2018*, 2019, https://www.nier.go.jp/kokusai/pisa/pdf/2018/01_point-eng.pdf, Accessed on April 21, 2023.

[2] A. Ohno et al., Prototype of a colloquial words checker for writing science reports, *Proc. of the 7th Int. Conf. Electron. Software Sci.*, pp.45-50, 2022.

[3] Y. Yamashita et al., Development and evaluation of Japanese colloquial writing checker, *Trans. Japanese Soc. Inf. Syst. Educ.*, vol.38, no.4, pp.369-374, 2021 (in Japanese).

[4] T. Mikolov et al., Distributed representations of words and phrases and their compositionality, *Adv. Neural Inf. Process. Syst.*, vol.26, pp.3111-3119, 2013.
[5] T. Mikolov et al., Efficient estimation of word representations in vector space, *Int. Conf. Learning Representations*, 2013.
[6] N. Yamaji et al., How Japanese undergraduates' skills in academic writing develop in the latter undergraduate years: From analysis of self-correction by graduate students in engineering, *J. Technical Japanese Educ.*, vol.16, pp.45-52, 2014 (in Japanese).
[7] *Reading Skill Test*, https://www.s4e.jp/about-s4e, Accessed on April 21, 2023 (in Japanese).
[8] T. Arai et al., Evaluating reading support systems through reading skill test, *Proc. of the 40th Annual Cognitive Sci. Soc. Meeting*, 2018.
[9] *jWriter: Learner Text Evaluator*, https://jreadability.net/jwriter/en, Accessed on April 21, 2023.
[10] J. H. Lee and Y. Hasebe, Quantitative analysis of JFL learners' writing abilities and the development of a computational system to estimate writing proficiency, *Learner Corpus Studies in Asia and the World*, vol.5, pp.105-120, 2020.
[11] A. Hamada et al., Development of a vocabulary size test for Japanese EFL learners using the new JACET list of 8,000 basic words, *J JACET*, vol.65, pp.23-45, 2021.
[12] T. N. T. Ngoc et al., Language model combined with Word2Vec for product's aspect based extraction, *ICIC Express Letters*, vol.14, no.11, pp.1033-1040, https://doi.org/10.24507/icicel.14.11.1033, 2020.
[13] A. Ohno et al., Quantification of the depth of student learning in group discussions to support active learning using revised taxonomy, *IEEJ Trans. Electron. Inf. Syst.*, vol.142, no.3, pp.382-388, 2022.
[14] Facebook Inc., *fastText*, https://fasttext.cc/, Accessed on April 21, 2023.
[15] *chiVe: Japanese Word Embedding with Sudachi & NWJC*, https://github.com/WorksApplications/chiVe/blob/master/README_en.md, Accessed on April 21, 2023.