The attempt to simplify Japanese words

Hiroshima Prefectural Hiroshima Kokutaiji High School

01 Introduction

Aim: to convert Japanese words (hereafter referred to as 'words') into simpler words.

Background:

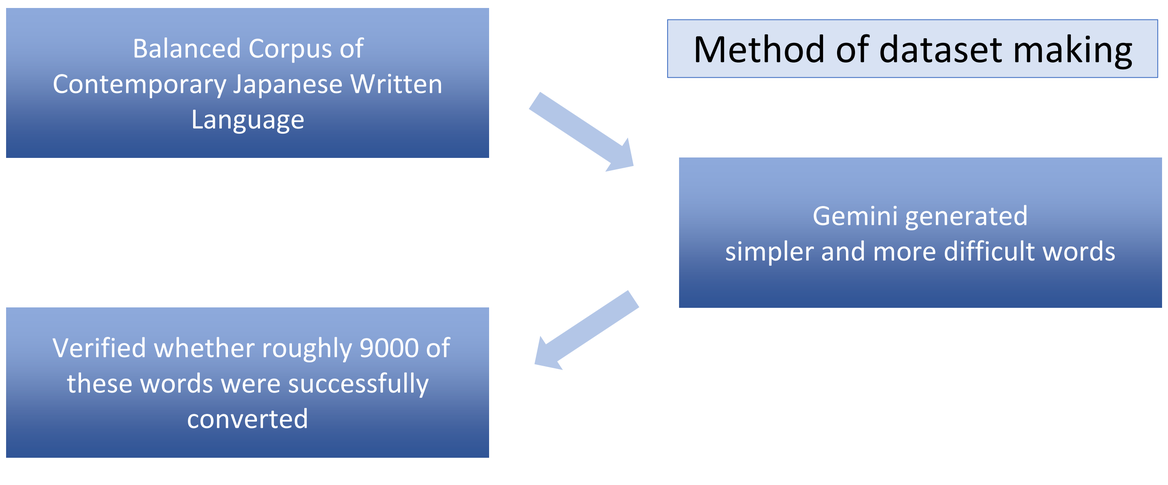
We see many situations in which we encounter difficulties in the Japanese language. For example, the difficulty level of learning Japanese is considered high compared to other languages.
In addition, there are many situations where people experience difficulty in expressing themselves when searching on the Internet.
We thought that by expressing words more simply, it would be easier to learn the language and search the Internet more efficiently.

02 Previous Research

Prior to this research, we consulted some relevant literature, including previous studies. According to the Kyoto University article, “Estimating Word Difficulty Using Wikipedia”[1], the method of estimating word difficulty based on the frequency of words used in Wikipedia commentaries was concluded to be not very good.
In addition, according to an article by the same university, “An Attempt to Estimate Japanese Difficulty Levels”[2], it was concluded that the method of estimating word difficulty levels through machine learning using the first occurrence of document difficulty levels and the first occurrence of test of Kanji skills difficulty levels as reference data, was the most accurate method for estimating difficulty levels. However, the estimation using web appearance frequencies and document difficulty averages did not work effectively.
From these findings, we decided to take other approaches because the approach of assessing word difficulty and transforming by comparing them is strongly influenced by the data used as a reference.

03 Research Method

In this study, we created a model to simplify more difficult Japanese words. This was done by converting words into vectors using Word2vec[3] (a set of models used to generate word embeddings). Thus, it was necessary to prepare a dataset containing the post-transformed and pre-transformed words.The model was then trained. Therefore, a dataset was generated using Gemini[4] (Google AI), and its reliability was checked and used for training.



04 Results and Considerations

We ordered Gemini to generate a dataset with simpler and more difficult words, resulting in a total of about 100,000 words from the “Balanced Corpus of Contemporary Japanese Written Language”[5]. Approximately 9,000 of these words were randomly extracted to check whether the words were converted successfully or not. The results are as follows.

According to the table shown, when using Gemini to convert words, the accuracy of simplifying words was as high as 79.36% and the accuracy of making words more difficult was also high at 74.34%. It was also found that there were few words predicted by AI to be easy, which were difficult for humans, and inversely, few words predicted by AI to be difficult, which were easy for humans. So, we think the dataset is generally useful.

ANN (Artificial Neural Network)[6], a computational model which imitates the functioning of biological neurons, was created based on this and the accuracies of the AI model and samples were output. The coefficient of determination was -0.0444, which was not very flattering. Examples of the results are as follows:

< **Success Samples** >

同等(same)　⇒　同じ(same)　　規約(rule)　⇒　ルール(rule)

< **Failure Samples** >

仏国(France)　⇒　法律(law)　　調教(breaking)　⇒　車(car)
アンモナイト(ammonite)　⇒　鳥(bird)

AI predicted / Human judgement	Easy words	Moderate words	Difficult words
Easy words	**79.36%**	**15.92%**	**5.47%**
Moderate words	**16.70%**	**62.61%**	**20.19%**
Difficult words	**3.94%**	**21.48%**	**74.34%**

05 Outlook

Future prospects include increasing the accuracy of the conversion of difficult to simple words. Specifically, the following improvements can be made: increasing the accuracy of the dataset by changing the prompts to Gemini, examining and selecting the entire dataset, increasing the amount of data, creating a structure which understands the context, and identifying words that cannot be converted. In the future, these will be achieved and the accuracy of the conversion will be improved.

References

1. 「ウィキペディアを利用した単語の難易度推定」
(“Word Difficulty Estimation Using Wikipedia”)
山河 絵利奈(京都大学工学部情報学科) 田島 敬史(京都大学大学院情報学研究科)
Erina Yamakawa (Department of Informatics and Mathematical Science, Faculty of Engineering, Kyoto University), Keishi Tajima (Graduate School of Informatics, Kyoto University)
(https://proceedings-of-deim.github.io/DEIM2021/papers/F25-2.pdf) (2024, December, 11)
2. 「日本語単語の難易度推定の試み」
(“An Attempt at Estimating the Difficulty of Japanese Words”)
水谷 勇介 河原 大輔 黒橋 禎夫(京都大学 大学院情報学研究科)
Yusuke Mizutani, Daisuke Kawahara, Sadao Kurohashi (Graduate School of Informatics, Kyoto University)
(https://www.anlp.jp/proceedings/annual_meeting/2018/pdf_dir/B4-3.pdf) (2024, December , 11)
3. https://code.google.com/archive/p/word2vec/ (2024, December, 11)
4. https://gemini.google.com/ (2024, December, 11)
5. 『現代日本語書き言葉均衡コーパス』短単位語彙表(Version 1.0)
“Balanced Corpus of Contemporary Japanese Written Language” Short Unit Lexicon (Version 1.0).
(https://repository.ninjal.ac.jp/records/3234) (2024, December, 11)
6. https://www.sciencedirect.com/topics/neuroscience/artificial-neural-network (2024, December, 11)

Acknowledgments

We thank our advisor, Mr. Kitayama, and many other people for their cooperation in carrying out this research.