RydeeNLP: Optimizing Japanese Learning with Lexical Simplification and Adaptive Translation

Yusuke Satani
Elizabethtown College
Elizabethtown
PA, USA
sataniy@etown.edu

Peilong Li
Elizabethtown College
Elizabethtown
PA, USA
lip@etown.edu

Abstract

In this paper, we present RydeeNLP, an innovative approach to Japanese language learning that leverages lexical simplification and adaptive translation techniques. Our approach introduces a novel difficulty scale encompassing elementary, middle, and high school levels, allowing for more precise and tailored language instruction. By using this scale, we have developed a comprehensive difficulty dictionary that categorizes Japanese words according to their complexity. From this dictionary, we further derived a paraphrase dictionary that maps words of similar meanings but different difficulty levels, providing learners with more nuanced vocabulary options. In addition to these resources, we expanded traditional translation models–often limited to noun replacements–to include verbs and adjectives, thereby offering a more holistic translation experience. We also designed a fine-tuned translation model that adapts output based on user-specified difficulty levels, producing translations that align with the learner's proficiency. The combination of these innovations offers a more effective and customizable solution for Japanese language acquisition compared to previous models.

1 Introduction

The popularity of Japanese language learning has surged in recent years, both in the United States and worldwide. Driven by cultural interests, business needs, and global connectivity, more learners are striving to achieve proficiency in Japanese. Despite this growing interest, learners face significant challenges, particularly when preparing for standardized Japanese tests like the Japanese Language Proficiency Test (JLPT). These tests often emphasize rote memorization and fail to adapt to the varying levels of vocabulary and grammar proficiency among students. Existing research has attempted to address these issues. For example, Kajiwara et al. (2020) explored lexical simplification techniques to make Japanese texts more accessible, while Poncelas and Htun (2022) worked on controlling simplification levels. However, these approaches have limitations, such as restricted vocabulary lists that do not cover the full breadth of the Japanese language, resulting in incomplete or overly simplified learning resources.

In response to these challenges, we propose a novel approach that focuses on enhancing Japanese learning and translation efficiency through a comprehensive lexical simplification model. Our design offers three main contributions: (1) A school-level classifier and expanded dictionaries that consider a broader range of words beyond the limited length list, addressing the vocabulary coverage issue; (2) A fine-tuning translation model designed to adapt to various school levels, delivering clear and understandable sentences tailored to the user's knowledge level; and (3) A word-swapping model that ensures accurate and contextually appropriate vocabulary replacement, even in complex Japanese sentences. These innovations not only address the limitations of previous research but also provide a more tailored and effective solution for learners at different stages of their Japanese language journey.

The remainder of this paper is structured as follows: Section 2 provides an overview of related work, highlighting the existing challenges in Japanese lexical simplification and translation. Section 3 details the datasets we used for the project. Section 4 describes the development of school-level classifier model, difficulty and paraphrase dictionary, and our translation models, including both fine-tuning and word-swapping approaches. In Section 5, we present the results of our experiments, including a comparison of BLEU scores for different models and a discussion of their implications. Section 6 outlines future development directions and potential improvements to enhance the effectiveness of our approach further. And finally, we conclude the paper in Section 7.

2 Background

Our research focuses on developing dictionaries and translation methodologies that build upon prior studies in the field of Japanese language learning. Previous studies have categorized vocabulary using various labels, such as JLPT levels and the Japanese Educational Vocabulary dictionary, which classifies words into six levels based on the input of five Japanese teachers (Sunakawa et al., 2012). However, for our research, we chose to use school or textbook levels–elementary, middle, and high school–as our categorical labels. The public textbook dataset we utilized includes approximately 50,000 words, which is significantly larger than other datasets like the JLPT dataset (Poncelas and Htun, 2022) (15,000 words), and the Japanese Educational Vocabulary dictionary Sunakawa et al. (18,000 words). This extensive dataset provides a broader range of language resources, enhancing the scope of our research compared to previous studies.

From this comprehensive school-level dataset, we developed a classifier capable of predicting the difficulty of words and categorizing them into three school levels. This classifier extends the selection of words beyond those explicitly listed in the textbook dataset, inspired by the methodologies of Hading et al. and Kajiwara et al.. Additionally, we created two types of dictionaries: a difficulty dictionary and a paraphrase dictionary. These efforts are influenced by the research conducted by (Kajiwara et al., 2020) and (Hading et al., 2016).

To construct the difficulty dictionary, we applied our classifier model to predict the school level of words within a large Japanese corpus. Concurrently, we developed a paraphrase dictionary, which groups words with the same meaning but different difficulty levels. According to Kajiwara et al., there are three primary approaches to building a paraphrase dictionary: dictionary-based, parallel corpora, and distributional similarity methods. Our approach combines dictionary-based and distributional similarity methods. By utilizing the thesaurus published by the National Institute for Japanese Language and Linguistics (NINJAL) to include semantically similar words, and integrating our classifier model with the difficulty dictionary, we were able to create an extensive paraphrase dictionary. This comprehensive resource enables the development of a more versatile translation model that goes beyond predefined word lists.

For the translation process, we trained two types of translation models. The first model was developed by fine-tuning an existing English-Japanese translation model, inspired by Poncelas and Htun. The use of tags added to source sentences to control the output of neural machine translation (NMT) models has been explored across different domains (Chu et al., 2017) and languages (Johnson et al., 2017). In our model, tags indicating the school level were added at the beginning of the English input, allowing the model to learn the relationship between words and school levels during the fine-tuning process.

The second model utilizes a pragmatic word swapping approach. This model generates a single Japanese translation according to a user-specified school-level tag, and words beyond the user's specified level are swapped to ensure that all words in the sentence are easier than the chosen difficulty level. Through these translation models, we aim to expand Japanese translation resources and develop a word-level translation model that aligns more closely with users' vocabulary knowledge. The methodologies and resources employed in our research are compared in Table 1.

3 Datasets

The construction of the difficulty dictionary in this study leverages a diverse set of high-quality datasets, carefully curated from multiple authoritative sources. These include a textbook corpus across all subjects, the Balanced Corpus of Contemporary Written Japanese (BCCWJ) for a comprehensive representation of modern written Japanese, and the JA-wiki corpus for extensive lexical coverage derived from online encyclopedic content. Additionally, we incorporated the Asahi Newspaper Word Vector dataset to capture contemporary usage patterns and the Bunrui Goi Hyo Database, a well-regarded Japanese thesaurus, to enhance semantic richness. To ensure the adaptability of our models across different contexts, we also utilized the SNOW T-23 parallel corpus for aligned bilingual data and complemented our resources with a web-scraped dataset to cover emerging trends and colloquialisms. This multifaceted approach ensures a robust and versatile foundation for the development of our lexical simplification tools, enabling more nuanced and context-sensitive applications.

3.1 Existing Datasets

The Textbook Dataset (NINJAL, 2011), provided by the National Institute for Japanese Language and

References	Width of the vocabulary	Word swapping model applied	Fine-tuning model applied
Hading et al.	N/A	✓	×
Kajiwara et al.	67k	✓	×
Poncelas and Htun	22k	✓	✓
This paper	150k	✓	✓

Table 1: Literature Comparison

Linguistics (NINJAL), includes textbooks from the 2005 school year across elementary, middle, and high school levels. This dataset provides detailed word frequency data across various educational levels and subjects, and the words in this dataset includes elementary level: 13k, middle school: 12k, and high school: 23k. For our purposes, words appearing at multiple educational levels were categorized according to the lowest level at which they first appeared, allowing us to establish a baseline vocabulary progression for school-level classifiers.

3.1.1 BCCWJ, JA-wiki, and Asahi Newspaper Word Vector

To supplement the Textbook Dataset, we incorporated additional resources: the Balanced Corpus of Contemporary Written Japanese (NINJAL, 2013b) (BCCWJ), JA-wiki (Wikimedia Foundation, 2024), and Asahi Newspaper Word Vectors (Asahi Shimbun Company and Retrieva, Inc., 2017). These datasets provide a broad spectrum of contemporary written Japanese across different genres, helping to capture a diverse range of vocabulary and usage. Word frequencies from BCCWJ and JA-wiki, along with 300-dimensional vectors from the Asahi Newspaper dataset, were employed as features in the word difficulty prediction model, ensuring a robust representation of Japanese language usage.

3.1.2 Bunrui Goi Hyo Database (Japanese Thesaurus)

The Bunrui Goi Hyo Database (NINJAL, 2004) serves as a comprehensive thesaurus, offering valuable insights into word meanings and synonyms. This information is crucial for building a paraphrase dictionary later.

3.1.3 SNOW T23

The SNOW T23 corpus (Katsuta and Yamamoto, 2018), consisting of 35,000 English-Japanese parallel sentences, provides data on sentence simplification, which helps evaluate our translation and simplification models.

3.2 Scraped Web Dataset

A significant contribution of our research is the creation of a Scraped Web Dataset, specifically curated to capture vocabulary tailored to different educational levels as presented on various online platforms. Unlike existing datasets that are limited to predefined contexts or formats, this dataset dynamically encompasses a wide range of educational materials available on the web, reflecting contemporary language use and emerging trends in Japanese education.

As shown in Figure 1, to construct this dataset, we systematically scraped websites designed for students at various school levels, capturing a diverse collection of words and phrases. By classifying words based on the lowest educational level at which they appear, similar to our methodology for the Textbook Dataset, we ensured consistency while greatly expanding the lexical database. This dataset allows for more granular control over the vocabulary selection process in our difficulty prediction models, ensuring they are relevant, current, and directly applicable to the learners' needs.

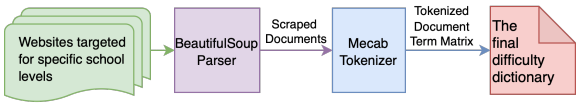

Figure 1: Web Scrapping Process

4 Design

4.1 Overview

In this section, we present the design of our system, which includes five main components: (1) a word difficulty classifier, (2) a difficulty dictionary, and (3) a paraphrase dictionary. These components work together to create (4) a translation model fine-tuned for specific complexity levels and (5) a word-swapping model that adjusts word difficulty according to user specifications. Each component is carefully designed to address the challenges of Japanese language learning and provide tailored resources for learners.

Features
Word Frequency in BCCWJ corpus
Word Frequency in JA-wiki corpus
Part of speech
Goshu (classification of Japanese words by their origin as Japanese, Chinese or Western)
300-dimension Vector Dependency-Based Word Embeddings from Asahi Newspaper Word Vector

Table 2: Features of the School-level Classifier

Model	Hyperparameters	Parameter Map Studied
SVM	gamma	[0.1, 1.0, 10.0, 100.0]
Random Forest	n-estimators	[50, 100, 150]
Random Forest	max-depth	[None, 10, 20]
Random Forest	min-samples-split	[2, 5, 10]
Random Forest	min-samples-leaf	[1, 2, 4]
Random Forest	max-features	['sqrt', 'log2']
MLP,CNN,RNN,LSTM	batch-size	[32,64,128, 256, 512]
MLP,CNN,RNN,LSTM	epochs	[50, 100, 150]
MLP,CNN,RNN	optimizer	['adam', 'rmsprop']
MLP,CNN	l1	[0.001, 0.01, 0.1]
MLP,CNN	l2	[0.001, 0.01, 0.1]
RNN,LSTM	model-lstm-units	[32, 64]
RNN,lSTM	model-dropout-rate	[0.2,0.3]

Table 3: Machine Learning Models and the Parameters Studied

4.2 Word Difficulty Classifier

Our word difficulty classifier is a crucial component designed to categorize words into three school-level labels. The classifier leverages five features, as detailed in Table 2, to accurately predict the difficulty level of a given word.

The classifier employs the MeCab library, a Japanese morphological analysis tool, to obtain detailed information such as part of speech and Goshu. The choice of using the mecab-ipadic-NEologd dictionary allows for a more extensive collection of contemporary words, enhancing the classifier's performance. An example of morphological analysis using MeCab is shown in Figure 2.

Figure 2: Morphological analysis using MeCab

In developing the classifier, we tested various machine learning models, including Support Vector Machine (SVM), Naive Bayes, Random Forest, Multilayer Perceptrons (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory networks (LSTM). Table 3 summarizes the parameters used for these models.

4.3 Difficulty Dictionary

The difficulty dictionary is constructed using a combination of curated and created datasets: a textbook dataset, a web-scraped dataset, and words extracted from BCCWJ, JA-Wiki, and Asahi Newspaper Word Vector. The dictionary categorizes 149,000 entries by school level, significantly surpassing the size of previous dictionaries and providing a more comprehensive resource for assessing word difficulty.

For a word to be included in the classifier, it must appear in these datasets, ensuring consistency and accuracy across our models.

4.4 Paraphrase Dictionary

The paraphrase dictionary is a key component designed to enable nuanced translations by mapping words with similar meanings across different difficulty levels. This allows users to tailor translations according to the desired proficiency level, enhancing the adaptability and educational value of the translations.

Words in the paraphrase dictionary are grouped based on combinations of difficulty levels, such as `(high, elementary)`, `(high, middle)`, and `(middle, elementary)`. This classification helps users select appropriate vocabulary that aligns with specific learning goals.

To construct the paraphrase dictionary, we first identified groups of semantically similar words across different levels using the difficulty dictionary. For words that are not present in the difficulty dictionary but appear in the BCCWJ, JA-wiki, and Asahi Newspaper Word Vector datasets, we employed a classifier to predict their corresponding school grade level.

To determine the most appropriate paraphrases, we calculated the cosine similarity between a higher-level target word and each lower-level word in the group. This metric allowed us to identify pairs of words with the highest semantic similarity, ensuring that the replacements are contextually appropriate and meaningful.

For example, consider a group of words with meanings related to "important" (e.g., [大切, 大事, 重い, 肝要, 肝心, 肝心かなめ, 緊要, 喫緊, 重要, 枢要, 主要], as shown in Table 4). To find the most similar elementary-level word to 肝要 (vital), which is classified at the high school level, we calculated the cosine similarity between 肝要 and four elementary-level words ([大切, 大事, 重い,

重要]). The pair with the highest cosine similarity score was (肝要, 大事) with a score of 0.608, indicating that 大事 (important) is the best match for substituting 肝要 while maintaining the intended meaning (see Table 5). This systematic approach ensures that the paraphrase dictionary is both comprehensive and precise, providing users with reliable word substitutions that are sensitive to varying proficiency levels.

words-group	words
High	肝要　緊要　喫緊　枢要
Middle	肝心　肝心かなめ　主要
Elementary	大切　大事　重い　重要

Table 4: Words and Their Groups Example

Target Word (High)	Elementary words	Cosine Similarity
肝要 **(vital)**	大切 (crucial)	0.55890274
	大事 (important)	**0.60757375**
	重い (important)	0.22134371
	重要 (significant)	0.48181933

Table 5: Cosine Similarity between the Target and Elementary Words

The paraphrase dictionary contains 103k word combinations. The outcome of dictionaries is summarized in Table 6.

Name	Label	Number of Words
Difficulty Dictionary	Elementary	33k
	Middle	16k
	High	100k
Paraphrase Dictionary	High - Elementary	58k
	High - Middle	30k
	Middle - Elementary	15k

Table 6: Word Count Breakdown of Dictionary Dictionary and Paraphrase Dictionary

4.5 Translation Model with Fine-Tuning

Our translation model is designed to produce translations that are tailored to specific complexity levels by incorporating school-level tags into the input sentences. This model is built on the fugumt-en-ja architecture, a transformer-based Sequence-to-Sequence model derived from Marian MT, which has been adapted for English-to-Japanese translation. The fugumt-en-ja model comprises six layers in both the encoder and decoder, providing robust performance for our targeted translation tasks.

To achieve translations suitable for different proficiency levels, we integrate special tokens into the input sentences. This approach, inspired by prior work in domain adaptation and multilingual translation (Chu et al.; Johnson et al.), allows us to control the difficulty level of the output. For Japanese, Poncelas and Htun have demonstrated that adding difficulty tags effectively enhances the precision of translation models by aligning vocabulary complexity with desired learner levels.

The fine-tuning process of the translation model involves the following steps:

Replace Words into Dictionary Form: We use MeCab to process each sentence and extract the dictionary forms of all words. This standardization step is crucial for consistent tagging and processing.

Add School-Level Tags to Each Sentence: Each word's school level is determined by referencing a predefined difficulty dictionary. The overall level of a sentence is set by the highest school level present among its words. We then construct a token in the format Ln based on the sentence's school level n (e.g., $L0$ for elementary, $L1$ for middle school, and $L2$ for high school). By incorporating these tokens, the model learns to associate input tags with corresponding vocabulary levels, enabling controlled output generation during the decoding process.

Expand the English Source-Side Sentence: The English source sentence is expanded by prepending the appropriate school-level token (e.g., $L2$, $w1$, $w2$, ...), ensuring the model aligns the input with the desired complexity level.

Fine-Tune the FuguMT Model: The FuguMT model is fine-tuned using the preprocessed input sentences with embedded school-level tags, optimizing its performance for generating translations that match specified difficulty levels.

To create a balanced training dataset for fine-tuning, we utilized multiple corpora as shown in Table 7, ensuring a diverse and representative sample of text. After classifying sentences by their difficulty levels, we curated datasets to avoid label imbalances, resulting in approximately 0.25 million sentences for each level. This careful balancing ensures that the model learns effectively across all difficulty levels.

4.6 Word Swapping Model

The Word Swapping Model is a second translation model designed to adjust word difficulty levels in translations based on user-specified preferences, using a unified Japanese translation as a starting

Name	Data Size
The Multitarget TED Talks Task (MTTT)	158k
English-Japanese Translation Alignment Data	118k
The Kyoto Free Translation Task	218k
Japanese-English Subtitle Corpus	314k
Tanaka Corpus	148k
Bilingual Corpus of Laws and Regulations	186k
JParaCrawl	200k

Table 7: Datasets used for fine-tuning

point. This model allows for dynamic adaptation of vocabulary to match the desired complexity level, enhancing the educational utility of translations.

A significant challenge in developing this model lies in the complexity of Japanese grammar, particularly in verb conjugations. Japanese verbs undergo various forms of conjugation influenced by their row (gyō, 行) in the syllabary and specific conjugation patterns. Understanding these patterns is essential for accurately modifying words to match different difficulty levels without compromising grammatical correctness.

In Japanese syllabary tables, gyō refers to horizontal rows of kana organized by their initial consonant sounds. Each row is named after its first syllable, as illustrated in Table 8.

	あ (a)	い (i)	う (u)	え (e)	お (o)
あ行 (a-gyō)	あ (a)	い (i)	う (u)	え (e)	お (o)
か行 (ka-gyō)	か (ka)	き (ki)	く (ku)	け (ke)	こ (ko)
さ行 (sa-gyō)	さ (sa)	し (shi)	す (su)	せ (se)	そ (so)
た行 (ta-gyō)	た (ta)	ち (chi)	つ (tsu)	て (te)	と (to)
な行 (na-gyō)	な (na)	に (ni)	ぬ (nu)	ね (ne)	の (no)
は行 (ha-gyō)	は (ha)	ひ (hi)	ふ (fu)	へ (he)	ほ (ho)
ま行 (ma-gyō)	ま (ma)	み (mi)	む (mu)	め (me)	も (mo)
や行 (ya-gyō)	や (ya)	-	ゆ (yu)	-	よ (yo)
ら行 (ra-gyō)	ら (ra)	り (ri)	る (ru)	れ (re)	ろ (ro)
わ行 (wa-gyō)	わ (wa)	-	-	-	を (wo)
ん行 (n-gyō)	ん (n)	-	-	-	-

Table 8: Gojūon (Japanese Syllabary) Table

Japanese verbs are categorized into five main conjugation patterns: (1) **Five-Class Conjugation (五段活用)**, (2) **Upper Ichidan Conjugation (上一段活用)**, (3) **Lower Ichidan Conjugation (下一段活用)**, (4) **K-Verbs Irregular Conjugation (カ行変格活用)**, (5) **S-Verbs Irregular Conjugation (サ行変格活用)**.

Each pattern can transform verbs into six different forms depending on the context, including: (1) **irrealis (未然形)**, (2) **continuative (連用形)**, (3) **conclusive (終止形)**, (4) **attributive (連体形)**, (5) **hypothetical (仮定系)**, (6) **imperative (命令形)**.

Additionally, verbs may undergo special euphonic changes (音便, onbin) in certain forms of the Five-Class Conjugation, such as: **I-Sound Euphony (イ音便)**, **Promotive Euphony (促音便)**, **N-Sound Euphony (撥音便)**.

Another complexity comes from the three types of characters in Japanese: Hiragana (ひらがな), Katakana (カタカナ), and Kanji (漢字). Verbs primarily consist of Kanji and Hiragana. To replace higher-level verbs with simpler ones effectively, we first convert the Hiragana part of the verb to the Roman alphabet, apply the necessary conjugations, and then convert it back to Hiragana. This approach allows for precise management of complex verb conjugations in Japanese. A similar process is employed for adjectives, while noun transformation involves directly swapping one noun for another. Table 9 summarizes this process.

To accurately manage these complexities in word swapping, our model uses the MeCab library to analyze and retrieve detailed grammatical information about each word, focusing on how to modify words while maintaining grammatical accuracy.

Name	POS	Words/Forms that Follow
未然形 Imperfective Form	verb, adjective	～ない (nai), ～う (u),～よう (you)
連用形 Continuative Form	verb, adjective	～ます (masu), ～た (ta)
終止形 Conclusive Form	verb, adjective	Period
連体形 Attributive Form	verb, adjective	Noun
仮定形 Hypothetical Form	verb, adjective	If statement
命令形 Imperative Form	verb	Period

Table 9: Forms Validation of Conjugation Types

The word swapping model follows a structured process to ensure that translations are both natural and contextually appropriate. This process is detailed step-by-step in Algorithm 1. As illustrated in Figure 3, consider the original English sentence, "Treat a sprained foot." Without the word-swapping model, the Japanese translation would be "捻挫した足を治療する," which is classified as high-school level difficulty based on our dataset (Step 1). However, after applying the 8-step word swapping model, the translation is transformed into "くじいた足を治す," which simplifies the language by using more Hiragana and less Kanji, thereby adjusting the difficulty to the elementary school level.

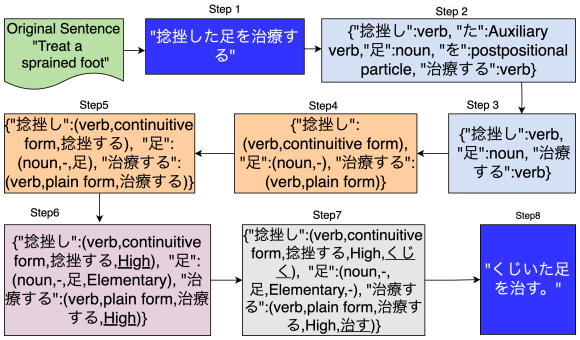

Figure 3: Word Swapping Process

Algorithm 1 Word Swapping Algorithm

Input: Sentence S, Difficulty Level L
Output: Modified Sentence S'
Step 1: Generate Unified Translation
$T \leftarrow FuguMT(S)$
Step 2: Tokenization and POS Tagging
$D \leftarrow MeCab(T)$ {Dictionary D pairs words with their POS}
Step 3: Retrieve Target Words
$W \leftarrow \{w \in D : POS(w) \in \{noun, verb, adjective\}\}$ {Exclude proper nouns}
Step 4: Conjugation Form Recording
for each word $w \in W$ **do**
    $Conj(w) \leftarrow MeCab(w)$
end for
Step 5: Infinitive Form Conversion
for each word $w \in W$ **do**
    $w \leftarrow Infinitive(w)$
end for
Step 6: Identify Upper-Level Words
$W_{upper} \leftarrow \{w \in W : Level(w) > L\}$
Step 7: Word Swapping Based on Difficulty Level
for each word $w \in W_{upper}$ **do**
    $w' \leftarrow ParaphraseDict(w, L)$ {Find simpler word w'}
    if w' **is not found then**
        $w' \leftarrow FindAlternative(w, L)$ {Use Word Vector and Classifier}
    end if
    $w' \leftarrow Conjugate(w', Conj(w))$ {Restore original conjugation}
end for
Step 8: Construct Modified Sentence
$S' \leftarrow Reconstruct(T, W_{upper})$
Return S'

5 Evaluation

5.1 Model Accuracy Performance

The performance of the classifier models was evaluated based on accuracy and the distribution of classifications across different school levels. Figure 4 shows the accuracy results for various models. The Multilayer Perceptrons (MLP) model was selected as the optimal classifier for this study due to its high accuracy and balanced classification distribution.

While the Random Forest classifier achieved the highest accuracy, it tended to classify an excessive number of words as high school level, leading to a significant class imbalance (Figure 5). This imbalance could result in a biased difficulty dictionary, adversely affecting the overall model performance and translation quality. In contrast, the MLP model provided a more balanced distribution across elementary, middle, and high school levels, making it more suitable for generating comprehensive and evenly distributed dictionaries.

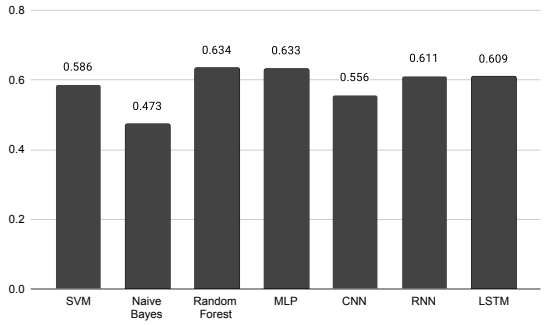

Figure 4: Classifier Accuracy

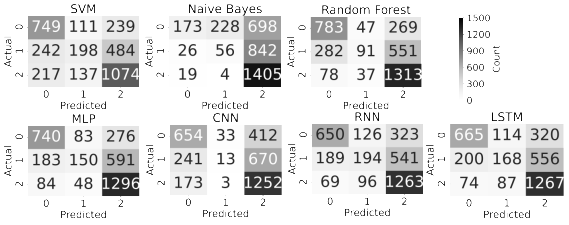

Figure 5: confusion matrix of each model

5.2 Model Latency Test

We also evaluated the latency of each model to understand the computational efficiency of word difficulty evaluation. Figure 6 presents the latency test outcomes, measured in milliseconds per word. Except for the Support Vector Machine (SVM) and Convolutional Neural Network (CNN) models, all models completed the difficulty evaluation in under 1.0 ms per word, demonstrating high speed and suitability for real-time applications. The tests were conducted on a machine configured with an N2-standard-8 instance, 32GB of RAM, and four vCPUs. These results suggest that the MLP model not only offers balanced accuracy but also performs efficiently, making it a strong candidate for practical deployment.

5.3 Translation Models

The quality of the translation models was assessed using the BLEU score, a standard metric for evaluating machine translation quality. The BLEU scores for various models are summarized in Table 11. As a benchmark, we sampled 5000 sentences from the SNOW T23 Parallel Corpus.

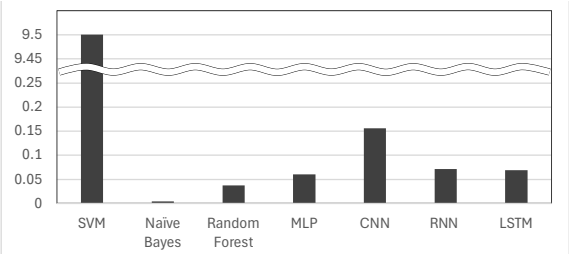

Figure 6: Model Inference Latency (ms/word)

Parameter	Setting
Training dataset in variable lengths	0.2M, 0.4M, 0.75M (full-dataset)
Epochs	1, 2, 4
Learning rate	2e-5, 5e-5

Table 10: Fine-tuning Parameters

The plain Fugu-MT model achieved a BLEU score of 0.191, serving as a baseline for comparison. Fine-tuned models, optimized with various training dataset lengths and epochs (see Table 10), produced lower BLEU scores, with the best-performing models achieving scores of 0.156 for elementary, 0.158 for middle, and 0.159 for high school levels. This decline in BLEU scores suggests that fine-tuning on specific difficulty levels, while improving vocabulary adaptation, may reduce overall translation fluency due to frequent word substitutions.

Interestingly, the word-swapping model consistently outperformed the fine-tuning models across all school levels, indicating that this approach maintains a better balance between preserving translation fluency and adapting vocabulary complexity. Although the BLEU scores of our models are lower than those reported by Poncelas and Htun, this discrepancy is expected because our model considers all words in a sentence, not just those on a limited list like the JLPT. As a result, our model swaps words more frequently, which can naturally lead to a lower BLEU score but provides more comprehensive vocabulary adaptation.

Model	Elementary	Middle	High
Fugu-MT	0.191	-	-
Word-Swapping	0.170	0.176	0.178
Fine-tune[0.2M, 1Epoch]	0.137	0.138	0.139
Fine-tune[0.2M, 2Epoch]	0.137	0.139	0.140
Fine-tune[0.2M, 4Epoch]	0.140	0.144	0.145
Fine-tune[0.4M, 1Epoch]	0.153	0.156	0.158
Fine-tune[0.4M, 2Epoch]	0.155	0.156	0.158
Fine-tune[0.4M, 4Epoch]	0.156	0.158	0.159
Fine-tune[0.75M, 1Epoch]	0.128	0.134	0.137
Fine-tune[0.75M, 2Epoch]	0.132	0.135	0.139
Fine-tune[0.75M, 4Epoch]	0.134	0.138	0.144

Table 11: The BLEU Scores

6 Future Development

To advance our Japanese lexicon simplification and translation methods, several areas need focused development. Enhancing the accuracy of the word-level classifier is a key priority. Refining this classifier with additional training data and advanced techniques could improve its ability to capture nuanced differences in school levels. Improving lower BLEU scores in fine-tuning translation models is also a significant component for the future. By exploring various architectures and hyperparameters, model performance and alignment with desired accuracy may be improved. The word-swapping approach must adopt a more consistent strategy to resolve complicated Japanese grammar, such as prefix issues. The complexity of Japanese grammatical structures and exceptions complicates the production of error-free sentences using the word-swapping model. If the fine-tuning model's performance improves, it is likely to become the more practical choice due to the fine-tuning model's better handling of Japanese grammar.

7 Conclusion

Our approach to simplifying the Japanese lexicon through the creation of difficulty and paraphrase dictionaries, along with a word-level classifier, demonstrates significant potential. The expanded vocabulary coverage should be beneficial for various Japanese translation tasks. The exploration of translation methods—fine-tuning and word-swapping—highlights both benefits and challenges. Although the fine-tuning method currently yields a slightly lower BLEU score, it offers a sophisticated means of learning vocabulary difficulty relationships. Conversely, the word-swapping method, while more direct, presents complexities in ensuring grammatical correctness. Future developments should focus on refining these methods, expanding resources, and exploring hybrid solutions to enhance translation accuracy and usability.

References

2023. フリーのニューラル機械翻訳モデルfugumt. Accessed: 2024-08-19.

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–

1564, Hong Kong, China. Association for Computational Linguistics.

Asahi Shimbun Company and Retrieva, Inc. 2017. 朝日新聞単語ベクトル (asahi newspaper word vector). Accessed on [Insert access date].

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Japan Foundation. 2017. Survey report on japanese-language education abroad 2015.

Muhaimin Hading. 2017. Master's thesis japanese simplification for non-native speakers. Master's thesis, Nara Institute of Science and Technology, Nara, Japan, August.

Muhaimin Hading, Yuji Matsumoto, and Maki Sakamoto. 2016. Japanese lexical simplification for non-native speakers. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 92–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Japan Foundation and Japan Educational Exchanges and Services. 2023. Japanese-language proficiency test: Can-do self-evaluation list. Accessed: 2024-08-19.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Tomoyuki Kajiwara and Mamoru Komachi. 2017. Simple ppdb: Japanese. In *Proceedings of the 23rd Annual Meeting of the Association for Natural Language Processing*, P8-5, pages 529–532.

Tomoyuki Kajiwara, Daiki Nishihara, Tomonori Kodaira, and Mamoru Komachi. 2020. Language resources for japanese lexical simplification. *Journal of Natural Language Processing*, 27(4):801–824.

Akihiro Katsuta and Kazuhide Yamamoto. 2018. Crowdsourced corpus of sentence simplification with core vocabulary (snow t23). In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 461–466. European Language Resources Association (ELRA).

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

NINJAL. 2004. 分類語彙表増補改訂版データベース(ver1.0). Accessed: 2024-08-19.

NINJAL. 2011. 教科書コーパス語彙表. Accessed: 2024-08-29.

NINJAL. 2013a. 『現代日本語書き言葉均衡コーパス』短単位語彙表(ver1.0). Accessed: 2024-08-19.

NINJAL. 2013b. 『現代日本語書き言葉均衡コーパス』長単位語彙表(ver1.0). Accessed: 2024-08-19.

Hitoshi Nishizawa, Dan Isbell, and Yuichi Suzuki. 2022. Review of the japanese-language proficiency test. *Language Testing*, 39:026553222210808.

Alberto Poncelas and Ohnmar Htun. 2022. Controlling Japanese machine translation output by using JLPT vocabulary levels. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 77–85, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Yuriko Sunakawa, Jae-ho Lee, and Mari Takahara. 2012. The construction of a database to support the compilation of japanese learners' dictionaries. *Acta Linguistica Asiatica*, 2(2):97–115.

Wikimedia Foundation. 2024. Japanese wikipedia database dumps. Accessed: 2024-08-19.

理史佐藤 and 玲宮田. 2008. 語彙平易化のための語釈文を用いた類義語抽出. In 言語処理学会第*14*回年次大会発表論文集, pages 1025–1028. 言語処理学会.

大輝柳本, 智之梶原, and 崇二宮. 2023. 単語の難易度埋め込みを用いた日本語のテキスト平易化. In 言語処理学会第*29*回年次大会発表論文集, pages 1007–1011.

智之梶原 and 守小町. 2020. 自動平易化システムの構築と評価データの作成. 自然言語処理, 27(2):189–217.

勇介水谷, 大輔河原, and 禎夫黒橋. 2018. 日本語単語の難易度推定の試み. In 言語処理学会第*24*回年次大会発表論文集, pages 670–673. 言語処理学会.

特定領域研究「日本語コーパス」言語政策班. 2011. 教科書コーパス語彙表(ver1.0). 特定領域研究『日本語コーパス』言語政策班最終成果CD-ROM. Accessed: 2024-08-19.