Japanese Reading Difficulty Classification for Non-Native Language Learners

Nicholas Ramkissoon

Abstract

This paper introduces a system for predicting the reading difficulty level of Japanese text for non-native language learners. The system leverages both grammar-based and vocabulary-based features in a text to produce a difficulty classification. Evaluation results from testing on a collection of text from varying difficulty levels and sources show a classification accuracy of 66.67% across 5 difficulty levels with significantly more consistent performance at each level compared to simpler baselines.

1 Introduction

Classifying text based on reading difficulty is a common task in natural language processing. Reading difficulty can mean several different things depending on the task, language, and for whom we are classifying text for. This paper introduces a system for classifying Japanese text into reading difficulty levels for non-native language learners. Therefore, for the scope of this paper, reading difficulty is defined as the readability of a given Japanese text for a non-native student of the language.

Typically, Japanese is taught with an emphasis on learning vocabulary and specific grammatical constructions that students use to form and understand sentences. The vocabulary and grammar gradually becomes more nuanced and complex at higher levels of study. Text that is too easy would provide little useful practice, and text that is too hard, especially text with abundant kanji and vocabulary unknown to the reader, can be impossible to read entirely. Because of this, care must be taken to select appropriately leveled material to study from. Of course, textbooks provide material reviewed by language teaching experts and designed with students in mind. However, if one wished to go beyond textbooks, millions of websites, books, and other media are available, but most of these materials are not labeled by difficulty with respect to language learners. Further, if one wanted to determine the difficulty of the text, they would have to read it anyways, which can be a waste of time because of the already mentioned points about too easy and too difficult texts. As a result, teachers and students face the problem of finding material at an appropriate difficulty level to study from. The system proposed will be a possible solution to this problem by assigning reading difficulty levels to Japanese texts from various sources using both grammatical and vocabulary features.

2 Related Work

There has been several different proposed methods of evaluating Japanese text difficulty in previous work. Earlier methods incorporated more general readability features such as average sentence length and comma to period ratio (Y. Tateisi, Y. Ono, and H. Yamada,

1988). These features are effective in predicting difficulty in a wide range of native-level text including technical materials such as college textbooks and juridical texts. More recent work has focused on classifying Japanese text by reading difficulty from the perspective of non-native language learners. The readability features used in this work better align with what is taught to students in a Japanese language class and textbooks. These features can be broadly categorized into grammar-based and vocabulary-based features.

In a paper tackling the same problem of text difficulty evaluation for language learners, *grammatical templates*, grammar units students learn from class and textbooks, were shown to be important features in predicting Japanese text difficulty. By parsing text and calculating the average number of grammatical templates at different difficulty levels, the difficulty level of material from previous Japanese Language Proficiency Test questions and textbooks can be classified with 87.7% accuracy (Wang and Andersen, 2016). In a later study, Liu and Matsumoto built upon the idea of extracting grammatical features while also considering similarity features between Chinese and Japanese use of kanji in select words to estimate sentence complexity specifically for Chinese-speaking learners of Japanese (Liu and Matsumoto, 2017).

Other work has focused on vocabulary difficulty in text. In a study on lexical simplification for non-native speakers (M. Hading, Y. Matsumoto, and M. Sakamoto, 2016), the methods used for determining vocabulary difficulty involved checking against a list of already-classified words and using a regression model with unigram frequency in several different corpora as features to determine difficulty for any unlisted words. Language learners are likely to know the most common words and phrases in a language, making unigram frequency of a word a good predictor of readability for language learners.

The approach used by the proposed system will combined several of the techniques used by previous work. Specifically, the system will extract usages of specific grammar and vocabulary words which will be used to calculate both grammar-based and vocabulary-based features. In addition, a regression model using both unigram frequency features and features related to the kanji composition of the word will be implemented in order to determine the difficulty of unlisted words.

3 Methods and System

3.1 Difficulty Classification Standard

Because the system will be classifying text based on reading difficulty from the point of view of a non-native speaker, it will follow the standard of the Japanese Language Proficiency Test (JLPT). The JLPT is widely considered to be the standard of Japanese language proficiency examinations as it is often used by Japanese companies for hiring foreigners and by the Japanese government for immigration purposes. The JLPT has 5 different difficulty levels, N5 through N1, N5 being the easiest and N1 being the most advanced. For more context, attaining N5 can be approximated to one semester college course in beginner Japanese, and attaining N1 suggests the ability to consume and understand native-level materials. A more in-depth overview of each of the levels is on the JLPT website[1]. The system takes in Japanese text and produces a level 1 through 5 corresponding to the text's predicted JLPT level.

[1]https://www.jlpt.jp/e

3.2 Data

3.2.1 Corpora

In order to develop and test the system, a corpus consisting of a list of different Japanese text, each labeled with their respective JLPT levels, is required. I compiled a corpus of classified text from several sources.

1. Official JLPT practice problem sets and reading material are available online. The JLPT website provides authoritative examples of Japanese text from each difficulty level.

2. NHK, Japan's national broadcaster, provides simplified versions of their news articles aimed at the N3 level on their website[2].

3. *Genki II* (Eri Banno, Yoko Ikeda, and Yutaka Ohno, 2011) is a standard textbook for the second semester of a beginner Japanese course. It contains example sentences and reading exercises at the N4 level.

4. Unofficial JLPT practice resources and websites[3] provide labeled reading exercises and example sentences.

In related work with similar tasks, manually built corpora consisting of only official JLPT materials and textbook material were used (Wang and Andersen, 2016). However, these corpora only represent a fraction of the study material available for language learners. By including material from multiple other learning resources on the web, the system will be able to better predict text difficulty of material a language learner is likely to encounter. In total, 520 separate texts consisting of 91,206 words were collected and randomly split into training, development, and test corpora (70/10/20 split, respectively).

3.2.2 JLPT Words, Kanji, and Grammar

The system parses sentences for specific words and grammatical constructions that are representative of a specific JLPT level. In order to do this, it leverages a pre-made list of labeled words, kanji and grammatical constructions to check against. The list of classified vocabulary words, and kanji is sourced from three online resources[4]. A word/kanji is included in the final list if it is present in at least two of these resources. A similar method is used by utilizing the same online resources with grammar dictionaries (Makino and Tsutsui, 2016b,a) and *Genki*. In total, the final vocabulary, kanji, and grammar lists consists of 4,859, 1766, and 582 items respectively.

3.3 System Overview

3.3.1 Features

For each input text, the system calculates 10 features:

1. The distributions of N5-N1 words in the text (5 features)

[2]https://www3.nhk.or.jp/news/easy

[3]https://japanesetest4you.com, http://www.tanos.co.uk/jlpt

[4]https://japanesetest4you.com, http://www.tanos.co.uk/jlpt, https://en.wiktionary.org/wiki/Appendix:JLPT

2. The distributions of N5-N1 grammatical constructions in the text (5 features)

Related work has shown that the distribution of grammatical constructions of a text is an effective predictor of its difficulty (Wang and Andersen, 2016; Liu and Matsumoto, 2017), however the difficulty level of the words in the text should also be considered. Table 1 and Table 2 show the average distributions of word and grammar levels respectively in the training corpus. Both tables show similar trends in distributions, N5 vocabulary/grammar makes up a large portion of text of *all* levels, and what differentiates one level, x, from a more difficult level, y, is the proportion of y-level grammar and vocabulary.

	N1 Texts	N2 Texts	N3 Texts	N4 Texts	N5 Texts
N1 Words	7.331%	3.154%	3.703%	0.001%	0.00%
N2 Words	8.242%	7.409%	3.142%	0.516%	0.328%
N3 Words	28.301%	24.745%	17.676%	4.031%	4.926%
N4 Words	13.483%	13.990%	16.386%	9.302%	4.269%
N5 Words	42.641%	50.700%	59.090%	86.046%	90.476%
Total	100%	100%	100%	100%	100%

Table 1: Distribution of Word Levels for each JLPT Level in Training Corpus

	N1 Texts	N2 Texts	N3 Texts	N4 Texts	N5 Texts
N1 Grammar	1.683%	1.171%	0.410%	0.232%	0.502%
N2 Grammar	5.287%	4.054%	2.642%	1.784%	1.758%
N3 Grammar	9.670%	8.443%	5.512%	2.404%	1.005%
N4 Grammar	18.700%	22.521%	17.949%	20.015%	16.834%
N5 Grammar	64.658%	63.808%	73.485%	75.562%	79.899%
Total	100%	100%	100%	100%	100%

Table 2: Distribution of Grammar Construction Levels for each JLPT Level in Training Corpus

3.3.2 Word and Grammar Level Extraction

The system uses the open-source text segmentation and Japanese POS tagger MeCab[4] for processing text at the sentence level. Sentence words are checked against the list of classified vocabulary words (Section 3.2.2). If a words is unlisted, the system relies a regression model that was trained on the listed words to predict the JLPT level. This regression model uses the word's unigram frequency in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and the mode and hardest kanji level in the word (determined using the kanji list described in Section 3.2.2) to predict the words JLPT level. For grammar feature extraction, the system uses the POS tags produced by MeCab and regular expressions to parse for each of the listed grammar items (Section 3.2.2). Table 3 illustrates the process of level extraction on a simple sentence.

After the grammatical and vocabulary compoenents of the text are identified, the word and grammar distribution features are calculated. The example sentence from Table 3 consists of 20% N3 grammar, 25% N3 vocabulary, 80% N5 vocabulary, and 75% N3 vocabulary. From this distribution of grammar and vocabulary, the sentence should be classified as N3.

[4]https://taku910.github.io/mecab

この服のサイズは子供向きです。 (This clothing size in intended for children.)		
Token	Grammar/Vocabulary/Both	JLPT Level
この	Grammar	5
服	Vocabulary	5
の	Grammar	5
サイズ	Vocabulary	5
は	Grammar	5
子供	Vocabulary	5
向き	Both	3
です	Grammar	5
。	N/A	N/A

Table 3: Example Sentence Tokenization into Vocabulary and Grammar Components

3.3.3 Classification Algorithm

The vocabulary/grammar distribution feature data is fed into Support Vector Machines (SVMs) to produce a JLPT level for each text. Specifically, C-Support Vector Classification with RBF kernels is used. Implementation for these algorithms are provided by Python's scikit-learn package[4]. SVMs were selected over other classification techniques such as TF-IDF with cosine similarity because of fast training and testing times once features were calculated and because previously mentioned work achieved good results from using SVMs (Wang and Andersen, 2016). Furthermore, because text of all difficulty levels contain mostly easier grammar and vocabulary, TF-IDF with cosine similarity performed worse than randomly assigning JLPT levels to text (less than 20% accuracy) when using query vectors representing the words/grammar in a specific level; the most cosine similar vector would almost always be the JLPT N5 or N4 vector.

4 Experiments

Vocabulary and grammar distribution features were calculated for training, development, and test corpora. Features from text in training corpora are used to train SVMs which are then used to classify text in the development and test corpora. Results for the test corpus is reported and discussed in the following sections.

4.1 Baselines

Baseline experiments were performed to compare to the main system performance. These expirements were conducted using the same training, development, and test corpora the main system uses. The results of these expirements are included in the subsequent Results and Discussion sections of this paper. The following baselines were implemented:

1. **Hardest Word Baseline** - this baseline simply assigns the JLPT level of the hardest word in a text to that text

2. **Average Sentence Length Baseline** - a regression model was trained to assign a JLPT level based on the average sentence length in an input text, sentence length is equal to number of characters

[4]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

5 Results

5.1 Evaluation Metrics

The evaluation metric for the system will be accuracy. Specifically, overall accuracy is calculated as the total number of correctly classified documents divided by the total number of documents being tested. Additionally, a more lenient accuracy metric, adjacent-level accuracy, will also be reported. Adjacent-level accuracy will define a correct classification as one where the predicted difficulty level is equal to or one level easier than the true level (i.e. a correct prediction for an N3 document can be either an N3 or N4 predicition). This metric essentially collapses the 5 original JLPT levels into 4 levels. Both accuracy and adjacent-level accuracy are reported in Table 4.

Recall, precision, and F-score for each JLPT level is reported for bith the main system and the baselines in Table 5. These metrics are calculated from confusion matrices, and provide insights to the system and baseline performance on a level-by-level basis. In addition to the accuracy metrics, the results of the recall, precision, and F-score metrics and their implications will be discussed further in the Discussion section.

5.2 System and Baseline Results

	Overall Accuracy	Adjacent-level Accuracy
Hardest Word Baseline	53.92%	62.75%
Average Sentence Length Baseline	43.14%	70.59%
System	66.67%	82.35%

Table 4: System and Baseline Accuracy Metrics

	Metric	N1	N2	N3	N4	N5
Hardest Word Baseline	Precision	.5227	.1818	.6316	.7273	.5882
	Recall	.8846	.0909	.6000	.3636	.8333
	F-score	.6571	.1212	.6154	.4848	.6897
Average Sentence Length Baseline	Precision	.7692	.3871	.2500	.4474	.0000
	Recall	.3846	.5455	.2500	.7727	.0000
	F-Score	.5128	.4528	.2500	.5667	.0000
System	Precision	.8571	.5143	.5455	.8421	.8333
	Recall	.4615	.8182	.6000	.7273	.8333
	F-Score	.6000	.6316	.5714	.7805	.8333

Table 5: System and Baseline Precision, Recall, and F-Score for each JLPT Level

6 Discussion

Comparing the accuracy results from the two baselines and main system in Table 4, the main system outperforms the baselines by a significant margin. The overall accuracy of 66.67% is much lower than the state-of-the-art's 87.7% accuracy (Wang and Andersen, 2016), however, the corpus used in that experiment consisted only of previous JLPT questions and textbook materials. Accuracy could have potentially decreased due to the inclusion of unofficial

practice material and new articles which are not designed specifically by JLPT test writers and due to possible differences in classifications for certain vocabulary and grammar. Despite this, given that adjacent-level accuracy is 82.35%, the system can effectively give a ballpark estimate of a text's difficulty level for a large variety of texts.

When looking at the recall, precision, and F-scores for each level, it is clear that the system performs much more consistently across all levels than the baselines. The average sentence length fails to classify any of the N5 texts correctly, and the hardest word baseline performs poorly on N2 texts. F-score and precision are above 0.5 for all levels for the system, indicating that positive predictions for each level are likely to be correct. The system performs best at the N5 and N4 levels and worse at the N3 and N2 levels. N1 documents are frequently labeled as N2, leading to lower N1 recall and N2 precision. N3 texts are often classified as N2. Unfortunately, related work does not report these metrics in their studies so there is not much else to compare against.

7 Conclusion and Future Work

In conclusion, the system is a decent performer because it predicts text difficulty correctly more often than not, and both grammatical and vocabulary features are good indicators of text difficulty. The system fits in with previous word by reinforcing the importance of grammatical features in predicting text difficulty which was established in previous work while also incorporating vocabulary-based features. Future work can greatly expand from just vocabulary and grammar features. Features such as average sentence length, number of clauses per sentence, etc. can be investigated and added to the system if they further improve results. Additionally, although kanji difficulty level is a feature the the regression model for determining word difficulty, the system relied only on a list of approximately 2000 classified kanji. Classification systems specifcally for kanji can be built using features such as number of strokes, unigram frequencies, etc. to provide accurate difficulty classifications for out of vocabulary kanji. In short, there are several different avenues to take and experiement with in order to improve the system.

Apart from improving the system itself, the system or its components can be incorporated into software applications to solve real-world problems. In terms of separate components of the system, they can also be used individually in other applications. For example, during development, over 500 regular expressions were created to identify unique grammatical constructions. These regular expressions can be used to tag a block of text and identify sentences using each grammtical construction. Filtering systems or recommendation systems for example sentences using a specific grammar can be implemented using the regular expressions. With accurate regular expressions and enough text, annotation projects can be undertaken to create a corpus of Japanese text tagged with grammar usages that can be useful in future research on Japanese grammar.

As alluded to in the introduction, practical use cases for the system revolve around language learning. From the perspective of a student, the system can be used as a recommendation system that outputs reading material given a JLPT level by the user. A teacher can also use the system to quickly find material for teaching or review. Building out these systems would require user feedback that can then be used to tune the classification system, and ideally, a larger corpora of text should be used for testing.

Future work can also go beyond Japanese and adapt the methods used in this system to build other text difficulty classification systems for other languages.

References

Eri Banno, Yoko Ikeda, and Yutaka Ohno. *GENKI: An Integrated Course in Elementary Japanese*. The Japan Times and Tsai Fong Books, 2011.

J. Liu and Y. Matsumoto. Sentence complexity estimation for chinese-speaking learners of japanese. *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, page 296–302, 2017.

M. Hading, Y. Matsumoto, and M. Sakamoto. Japanese lexical simplification for non-native speakers. *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, page 92–96, 2016.

Seiichi Makino and Michio Tsutsui. *A Dictionary of Advanced Japanese Grammar*. The Japan Times, 2016a.

Seiichi Makino and Michio Tsutsui. *A Dictionary of Intermediate Japanese Grammar*. The Japan Times, 2016b.

S. Wang and E. Andersen. Grammatical templates: Improving text difficulty evaluation for language learners. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1692–1702, 2016.

Y. Tateisi, Y. Ono, and H. Yamada. A computer readability formula of japanese texts for machine scoring. *Proceedings of the 12th Conference on Computational Linguistics*, pages 649–654, 1988.