Sentence Complexity Estimation for Chinese-speaking Learners of Japanese

Jun Liu
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
liu.jun.lc3@is.naist.jp

Yuji Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
matsu@is.naist.jp

Abstract

It is fairly challenging for a foreign language learner to read and understand Japanese texts containing words of high difficulty level or low frequency and complicated linguistic structures. Because a large number of Chinese characters (kanji in Japanese) are commonly used both in Chinese and Japanese, the more confusing problem for Japanese language learners from kanji background countries is the acquisition of various complex Japanese functional expressions. In this study, we propose a method utilizing Japanese kanji characters, particularly Japanese–Chinese homographs with identical or similar meanings, as a critical feature of sentence-complexity estimation for Chinese-speaking learners of Japanese language. Experimental results have partially demonstrated the effectiveness of our method in enhancing the accuracy of sentence-complexity estimation.

1 Introduction

Enhancing reading capability is one of the important purposes in second language teaching and learning. There are various factors that impact learners' reading comprehension. A few of these factors involve the learners' vocabulary knowledge, grammar knowledge, reading strategies, interest, attitude, and motivation (Koda, 2007; Han and Song, 2011; Horiba, 2012; Gilakjani and Sabouri, 2016). Reading comprehension is also influenced by the complexity of the reading material. Texts containing highly demanding vocabularies and highly complex sentence structures are likely to disturb the learners' reading comprehension. Learners of Japanese language from kanji background countries benefit substantially from kanji characters commonly used in both Japanese and Chinese when they read Japanese sentences or documents. However, it is more challenging for them to read and learn various Japanese functional expressions with varied meanings and usages.

The selection of appropriate reading material matching the learners' individual capabilities is highly likely to enable language learners to read in a more focused and selective manner. To support learners in gathering useful information from texts more effectively, certain online public Japanese reading-assistance systems such as Reading Tutor[1], Asunaro[2], Rikai[3], and WWWJDIC[4] are highly effective. These systems are adequately constructed for providing an internet learning environment where learners can make complete use of information from the internet for their Japanese language study, and a few of them are specifically designed to enable language learners to understand Japanese texts by offering words with their corresponding difficulty level information or translation (Toyoda 2016). However, these systems

[1] http://language.tiu.ac.jp/
[2] https://hinoki-project.org/asunaro/
[3] http://www.rikai.com/perl/Home.pl
[4] http://nihongo.monash.edu/cgi-bin/wwwjdic?9T

do not take the learners' native language background into account. Moreover, these systems provide learners with limited information on the grammatical difficulty of all the various types of Japanese functional expressions, which learners actually intend to learn as a part of the procedure for learning Japanese.

In Section 2 of this paper, we introduce some previous works. In Section 3, we describe our method for ranking example sentences of Japanese functional expressions by utilizing Japanese–Chinese homographs with identical or similar meanings, as a critical feature. Section 4 describes the several experiments conducted to examine the effectiveness of our method. Finally, in Section 5, we conclude and describe future work.

2 Previous Research

Text difficulty or text readability evaluation is one of the challenges in natural language processing (NLP) owing to the linguistic complexity generated from both vocabulary and grammar. Researchers have been actively exploring methods to evaluate text difficulty (Gonzalez-Dios et al., 2014; Hancke, Vajjala, and Meurers, 2012; Vajjala and Meurers, 2012; Xia, Kochmar and Briscoe, 2016).

For English texts, there are numerous popular formulas such as Flesch Reading Ease (Flesch 1948) and Flesch-Kincaid Grade Level, all of which are used for several applications such as compilation of reading materials for language learners. Collins–Thompson and Callan (2004) proposed a language modeling method to estimate the readability of English and French texts.

For Japanese texts, Tateishi, Ono, and Yamada (1988a; 1988b) introduced a formula based on six surface characteristics: average number of characters per sentence, average number of Roman letters and symbols, average number of hiragana characters, average number of kanji characters, average number of katakana characters, and ratio of touten (comma) to kuten (period). Formula-based approaches have also been used or teaching Japanese to young native speakers (Shibasaki and Sawai, 2007; Sato, Matsuyoshi, and Kondoh, 2008; Shibasaki and Tamaoka, 2010). To evaluate text difficulty level for foreign language learners of Japanese, Wang and Andersen (2016) introduced an approach for evaluating Japanese text difficulty that focuses on grammar and utilizes grammar templates.

In recent years, a few Japanese text difficulty evaluation systems have been developed to support Japanese language learners (Hasebe and Lee, 2015; Lee and Hasebe, 2016). For example, JReadability[5] can analyze input text and estimate its readability to categorize it as belonging to one of six difficulty levels, on the basis of five characteristics: average length of sentence; percentage of kango (words of Chinese origin), percentage of wago (words of Japanese origin), percentage of verbs, and percentage of particles.

However, JReadability too does not sufficiently consider the various types of Japanese functional expressions with varying difficulty levels. The prediction value calculated by this system is more reliable for long texts (approximately 1000 characters) and not for single sentences.

3 General Method

Japanese and Chinese share a large quantity of homographs that use identical kanji characters (both in simplified Chinese and traditional Chinese). Table 1 presents a few examples of Japanese–Chinese homographs. These words play a significant role while reading Japanese or Chinese texts. According to a report by Wang (2001), approximately 80–95% Japanese–Chinese homographs are used to express identical or similar meanings in both the languages. Foreign language learners from kanji background countries can straightforwardly understand the meaning of these words according to kanji characters. This is occasionally more convenient than grammar for foreign language learners from kanji background countries to learn Japanese.

For Japanese language learners, a vital challenge is to master a large number of complex functional expressions. Hence, providing appropriate example sentences for learners based on their individual Japanese language capabilities are highly likely to aid the enhancement of the efficiency of learning various Japanese functional expressions.

In order to achieve this goal, we utilize Japanese–Chinese homographs as a new feature, which is more or less dissimilar from previous research, to estimate sentence difficulty and select

[5] http://jreadability.net

the most appropriate example sentences as learning content for Japanese functional expressions.

Japanese	Chinese	Meaning
社会(society)	社会(society)	Identical
技術(technology)	技术(technology)	Identical
東西(east and west)	东西(east and west; thing)	Similar
培養(culture)	培养(culture; train)	Similar
手紙(letters)	手纸(toilet paper)	Dissimilar
勉強(study)	勉强(reluctantly)	Dissimilar

Table 1: Examples of Japanese–Chinese homographs.

3.1 Difficulty Level Evaluation Standard

To estimate the difficulties of example sentences, we follow the standard of the Japanese Language Proficiency Test (JLPT). The JLPT consists of five levels: N1, N2, N3, N4, and N5. The least difficult level is N5, and the most difficult level is N1[6]. Since 2010, the JLPT official lists of vocabulary and grammar have not been published in books, we referenced a few books (Xu and Reika, 2013a; Xu and Reika, 2013b) and online learning websites[7,8], all of which provide lists of the JLPT vocabulary and grammar with difficulty levels ranging from N1–N5. Here, we consider levels N3/SP3 and lower as "easy" level, levels N2/SP2 and above as difficult level. A few examples of vocabulary and grammar in JLPT are presented in Table 2.

3.2 List of Japanese–Chinese Homographs

Japanese language learners from kanji background countries can conveniently read and understand majority of the Japanese words written in kanji. However, in the vocabulary list of JLPT, numerous Japanese–Chinese homographs are classified as difficult levels (N2 and above) without consideration of learners' differing mother tongue background. Consequently, we attempt to construct a list of Japanese–Chinese homographs that is likely to be helpful in estimating complexity of example sentences that include Japanese functional expressions.

Japanese vocabulary	Difficulty level
山岳(mountains) 養う(to cultivate) 忙しない(busy)	N1
前提(Presupposition) 迫る(to press) 勇ましい(brave)	N2
愛情(love) 含める(to include) 巨大(huge)	N3
複雑(complex) 捨てる(to throw away) 挨拶(greeting)	N4
学校(school) 明るい(bright) 始まる(begin)	N5
Japanese grammar	Difficulty level
べからざる(must not) がてら(while doing something) を顧みず(regardless of)	SP1
からといって(just because) に加えて(in addition to) に違いない(without a doubt)	SP2
にとって(to) に比べて(compare) わけがない(it is impossible that)	SP3
かもしれない(maybe) ことができる(can) みたいだ(similar to)	SP4
てから(after) 前に(before) ている(am/is/are doing)	SP5

Table 2: Examples of Japanese vocabulary and grammar in JLPT.

To accomplish this task, we first extracted the Japanese words containing only kanji characters from two dictionaries: IPA (mecab-ipadic-2.7.0-20070801)[9] and UniDic (unidic-mecab 2.1.2)[10]. These two dictionaries are used as the standard

[6] http://jlpt.jp/e/about/levelsummary.html
[7] http://www.tanos.co.uk/jlpt/
[8] http://japanesetest4you.com

[9] https://sourceforge.net/projects/mecab/files/mecab-ipadic/2.7.0-20070801/mecab-ipadic-2.7.0-20070801.tar.gz/download
[10] http://osdn.net/project/unidic/

dictionaries for the morphological analyzer MeCab, with appropriate part-of-speech information for each expression. We then extracted the Chinese translation words of these Japanese words from the following online dictionary websites: Wiktionary[11] and Weblio [12]. We compared the character form of the Japanese word with its Chinese translation word to identify whether the Japanese word is a Japanese–Chinese homograph or not. Because Japanese uses both simplified Chinese characters such as "雨(rain), 木(tree), and 本(book)" and traditional Chinese characters such as "車(car), 頭(head), and 雲(cloud)," we replaced all the traditional Chinese characters with the simplified Chinese characters. If the character form of a Japanese word is similar to the character form of the Chinese translation word, the Japanese word is identified as a Japanese–Chinese homograph. Considering unknown words in the above online dictionaries, we also referenced an online Chinese encyclopedia: Baike Baidu [13] and a Japanese dictionary: Kojien fifth Edition (Shinmura, 1998). If a Japanese word and its corresponding Chinese word share an identical or a similar meaning, then, the Japanese word is also identified as a Japanese–Chinese homograph. Finally, we created a list of Japanese–Chinese homographs consisting of approximately 14 000 words.

3.3 Extraction of Japanese Grammar

There are a large number of Japanese functional expressions in Japanese grammar. A problematic feature of Japanese functional expressions is that each functional expression is likely to exhibit numerous surface forms such as "Headword: なければならない(should) and its surface form variations: なければなりません、なければならず、なければならなく、なければならなかっ、なければならぬ. . . ." Based on the grammar list of JLPT, we finally constructed a list of Japanese functional expressions consisting of approximately 680 headwords and 4000 types of their surface form variations, as illustrated in Table 3.

To extract Japanese functional expressions, we use a publicly available morphological analyzer MeCab [14]. We incorporate the list of Japanese functional expressions into the IPA dictionary considering it likely that the morphological analyzer MeCab extracts the usages of functional expressions automatically. Table 4 demonstrates certain extracted examples of Japanese functional expressions.

Headword	Surface Forms	Difficulty Level
をふまえて (in accord with)	をふまえ をふまえた を踏まえて を踏まえ を踏まえた	SP1
にさいして (on the occasion of)	にさいし にさいしまして に際して に際し に際しまして	SP2
ねばならない (should)	ねばなりません ねばならなかっ ねばならなく ねばならぬ ねばならず ねばならん	SP3
ていけない (must not)	ていけなかっ ていけません でいけない でいかなかっ でいけません	SP4
ではない (am/is/are not)	ではありません じゃありません ではなかっ じゃない じゃなかっ	SP5

Table 3: Examples of Japanese functional expressions and surface form variations.

4 Experiments

Because our purpose is to provide the Japanese language learners with straightforward example sentences such that they can understand the meaning and usage of the Japanese functional

[11] http://ja.wictionary.org/wiki/メインページ
[12] http://cjjc.weblio.jp
[13] https://baike.baidu.com
[14] http://taku910.github.io/mecab/

expressions conveniently, it is necessary to solve the problem of displaying the order of the example sentences based on their difficulty. To achieve this goal, we adopt an online machine learning tool, Support Vector Machine for Ranking (SVM^{rank})[15], to estimate the complexity of example sentence.

Input: 彼は学生ではありません。 Output: 彼 **は** 学生 **ではありません** 。 (He is not a student.)
Input: 野菜を食べなければならない。 Output:野菜 **を** 食べ **なければならない** 。 (You must eat vegetables.)
Input: 私は行きたくてたまらない。 Output: 私 **は** 行き **たく** **てたまらない** 。 (I am eager to go.)
Input: 物価は上がる一方だ。 Output: 物価 **は** 上がる **一方だ** 。 (Prices continue to increase.)
Input: 天気いかんにかかわらず来ます。 Output: 天気 **いかんにかかわらず** 来 **ます**。 (Regardless of the weather, I will come.)

Table 4: Extraction of Japanese functional expressions. In the sentences, Japanese functional expressions are in bold and underlined.

4.1 Data Setting

We utilize the Balanced Corpus of Contemporary Written Japanese (BCCWJ) to carry out our experiments:

- BCCWJ [16] is a corpus created for comprehending the breadth of contemporary written Japanese; it contains extensive samples of modern Japanese texts to create as uniquely balanced a corpus as possible. The data comprises 104.3 million words, covering genres including general books and magazines, newspapers, business reports, blogs, internet forums, textbooks, and legal documents.

4.2 Features

Based on the standardization of difficulty level evaluation in JLPT described in Section 3.1, we employ the following 12 features as the baseline readability feature set:

- Number of N0–N5 Japanese words in a sentence (Here, N0 implies unknown words in the vocabulary list of JLPT.)
- Number of SP1–SP5 Japanese functional expressions in a sentence
- Length of a sentence

As a departure from the standardization of difficulty level evaluation in JLPT, we identify the Japanese words in the list of Japanese–Chinese homographs mentioned in Section 3.2 as belonging to the easy level labeled as NJ–C. We assume that if an example sentence contains a higher number of N3–N5 words, SP3–SP5 Japanese functional expressions, and Japanese–Chinese homographs, this example sentence will be more straightforward to read and understand for Chinese-speaking learners. Therefore, we utilize Japanese–Chinese homographs as a new feature in our experiments.

- Number of NJ–C Japanese words in a sentence

Finally, we combine this new feature with the baseline readability features (all 13 features) as we wish to examine whether this new feature will actually help enhance example-sentence-difficulty estimation.

4.3 Example-Sentence-Difficulty Estimation

We first collected 5000 example sentences from the BCCWJ and divided them into 2500 pairs. Then, we invited 15 native Chinese-speaking learners of Japanese language, all of whom have been learning Japanese for ~1 y, to read two example sentences in one pair and select the one that is more straightforward to read and understand. Considering the feasibility of a learner's decision on a particular pair to vary from that of the other learners, we asked every three learners to compare a particular pair. The final decision was made by majority vote. We finally utilized a set of fivefold cross-validations with each combination of 4000 sentences as the training data and 1000 sentences as the test data.

Experimental results using baseline features and our method are presented in Tables 5 and 6, respectively.

[15] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html
[16] http://pj.ninjal.ac.jp/corpus_center/bccwj/en/

Features	Cross-validations	Accuracy
Baseline Features	1	82.4%
	2	82.8%
	3	81.8%
	4	80.8%
	5	81.4%
Average		**81.84%**

Table 5: Experimental results using baseline features.

Features	Cross-validations	Accuracy
Our Method	1	84.4%
	2	86.8%
	3	84.8%
	4	82.8%
	5	83.2%
Average		**84.4%**

Table 6: Experimental results using our method.

According to the experimental results in Tables 5 and 6, our method of incorporating Japanese–Chinese homograph features to baseline readability features effectively estimates the difficulty level of example sentences of Japanese functional expressions, with an average accuracy of 84.4%. In comparison with the experimental results using baseline features, our method enhances the accuracy by 2.56%, partially demonstrating the effectiveness of our method.

5 Conclusion and Future Work

We proposed a method that integrates vocabulary knowledge of Japanese–Chinese homographs that Chinese-speaking learners of Japanese are capable of understanding straightforwardly, with the aim of estimating complexity of example sentences that include Japanese functional expressions. The experimental results demonstrated that this method enhanced the accuracy of estimation of the difficulty levels of example sentences.

However, we did not evaluate the learning effect of using the example sentences of Japanese functional expressions generated by our method. In our future work, we plan to consider other features such as word types and number of verbs to enhance example-sentence-complexity estimation for Chinese-speaking learners of Japanese. Finally, we intend to develop a Computer-aided Language Learning (CALL) system that can recommend learning content to individual learners at appropriate difficulty levels.

Acknowledgments

We wish to thank all those who allocated their time to complete our online survey and the anonymous reviewers for their detailed comments and advice.

References

Abbas Pourhosein Gilakjani and Narjes Banou Sabou. 2016. A Study of Factors Affecting EFL Learners' Reading Comprehension Skill and the Strategies for Improvement. International Journal of English Linguistics, 6(5): pp. 180–187.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? assessing the readability of basque texts. In Proceedings of COLING 2014: Technical Papers, pp. 334–344, Dublin, Ireland, August.

Fudolf Flesch. 1948. A new readability yardstick. Journal of Applied Psychology, 32(3): pp. 221–233.

Dongli Han, and Xin Song. 2011. Japanese Sentence Pattern Learning with the Use of Illustrative Examples Extracted from the Web. IEEJ Transactions on Electrical and Electronic Engineering, 6(5): pp. 490–496.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In Proceedings of COLING 2012, pp. 1063–1080, Mumbai, India, December.

Yoichiro Hasebe and Jae-Ho Lee. 2015. Introducing a Readability Evaluation System for Japanese Language Education. In Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese, pp. 19–22.

Yukie Horiba. 2012. Word knowledge and its relation to text comprehension: a comparative study of Chinese -and Korean-speaking L2 learners and L1 speakers of Japanese. The Modern Language Journal, 96(1): pp. 108–121.

Keiko Koda. 2007. Reading Language Learning: Cross-Linguistic Constraints on Second Language Reading Development. Language Learning, 57(1), pp. 11–44.

Takahiro Ohno, Zyunitiro Edani, Ayato Inoue, and Dongli Han. 2013. A Japanese Learning Support

System Matching Individual Abilities. In Proceeding of the PACLIC 27 Workshop on Computer-Assisted Language Learning, pp. 556–562.

Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In Proceedings of the Sixth International Language Resources and Evaluation, pp. 654–660, Marrakech, Morocco.

Hideko Shibasaki and Yasutaka Sawai. 2007. Study for constructing a readability formula of Japanese texts using a corpus of language school textbooks. IEICE Technical Report NCL2007-32, pp. 19–24.

Hideko Shibasaki and Katsuo Tamaoka. 2010. Constructing a Formula to Predict School Grades 1-9 based on Japanese Language School Textbooks. Japan Journal of Educational Technology 33(4), pp. 449–458.

Izuru Shinmura (Ed. In chief). 1998. Kojien 5th Editon (in Japanese). Tokyo: Iwanani Press.

Yuka Tateisi, Yoshihiko Ono, and Hisao Yamada. 1988a. A computer readability formula of Japanese texts for ma-chine scoring. In Proceedings of the 12th Conference on Computational Linguistics, volume 2, pp. 649–654.

Yuka Tateisi, Yoshihiko Ono, and Hisao Yamada. 1988b. Derivation of readability formula of Japanese texts. IPSJSIG Note 88-DPHI-18-4, Information Processing Society of Japan.

Kevyn Collins-Thompson and Jamie Callan. 2004. A lan-guage modeling approach to predicting reading difficulty. In Proceedings of the HLT/NAACL 2004 Conference, pp. 193–200.

Etsuko Toyoda. 2016. Evaluation of computerised reading-assistance systems for reading Japanese texts – from a linguistic point of view. Australasian Journal of Educational Technology, 32(5). pp. 94–107.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using in-sights from second language acquisition. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 163–173.

Shuhan Wang, Erik Andersen. 2016. Grammatical Templates: Improving Text Difficulty Evaluation for Language Learners. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 1692–1702, Osaka, Japan, December.

Shuyu Wang. 2001. A comparative study of vocabulary in Chinese and Japanese, Sichuan Literature and Art Press.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 12–22.

Xiaoming Xu and Reika. 2013a, Detailed introduction of the New JLPT N1-N5 grammar. East China University of Science and Technology Press.

Xiaoming Xu and Reika. 2013b, Detailed introduction of the New JLPT N1-N5 vocabulary. East China University of Science and Technology Press.