Simplification of Example Sentences for Learners of Japanese Functional Expressions

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

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

Abstract

Learning functional expressions is one of the difficulties for language learners, since functional expressions tend to have multiple meanings and complicated usages in various situations. In this paper, we report an experiment of simplifying example sentences of Japanese functional expressions especially for Chinese-speaking learners. For this purpose, we developed "Japanese Functional Expressions List" and "Simple Japanese Replacement List". To evaluate the method, we conduct a small-scale experiment with Chinese-speaking learners on the effectiveness of the simplified example sentences. The experimental results indicate that the simplified sentences are helpful in learning Japanese functional expressions.

1   Introduction

In Japanese grammar, there is a large number of functional expressions consisting of one or more words and behave like a single functional word, such as "たい (want to)、に対して(to)、なければ ならない (must)". Matsuyoshi et al. (2006) developed a Japanese functional expression lexicon consisting of 292 headwords and 13,958 different surface forms. It is crucial to develop a Japanese learning assistant system which supports Japanese language learners to learn such a large number of complicated functional expressions.

Recently, with the help of natural language processing technology, many Japanese learning assistant systems have been constructed. For example, Pereira and Matsumoto (2015) presents a Collocation Assistant for Japanese language learners, which flags possible collocation errors and suggests corrections with example sentences. Han and Song (2011), and Ohno et al. (2013) attempt to develop Japanese learning systems for learning and using Japanese sentence patterns with the use of illustrative examples extracted from the Web.

As mentioned above, some studies have paid attention to assist learners to learn Japanese functional expressions. However, none of the existing studies has aimed at simplifying difficult example sentences for the need of Chinese-speaking learners of Japanese language. In this paper, we describe our proposed method in Section 2. Section 3 explains the result and evaluation of a small-scale experiment for examining the effectiveness of the proposed method. Finally, we conclude in Section 4.

2   Proposed Method

In this section, we propose a method to simplify difficult Japanese sentences that contain Japanese functional expressions for Chinese-speaking learners of Japanese language.

2.1 Making Japanese Functional Expressions List

In order to identify Japanese functional expressions, Tsuchiya et al. (2006) developed an example

database with 337 types of Japanese compound functional expressions. Shime et al. (2007) proposed an approach to detecting 59 types of Japanese functional expressions through a machine learning technique based on a chunking program. However, the major drawback of these studies is in the scale of Japanese functional expressions, which does not reach the need of the five levels (from N1 to N5) of the new Japanese-Language Proficiency Test (JLPT).

According to the requirements of the new JLPT, we manually constructed a list of Japanese functional expressions, which consists of about 680 headwords and 4,000 types of different surface forms. Here, we consider levels 3-5 as easy level, and levels 1-2 as difficult level, respectively. Table 1 shows some examples of Japanese functional expressions and their surface form variations.

Headword	Difficulty Level	Surface Forms
たあとで(after)	N5	たあとで、た後で、だあとで、だ後で
かもしれない(maybe)	N4	かもしれない、かもしれません、かもしれず…
にたいして(to)	N3	にたいして、に対して、にたいしては、に対しては…
ざるをえない(have to)	N2	ざるをえない、ざるを得ない、ざるをえず、ざるを得ず…
やいなや(as soon as)	N1	やいなや、や否や

Table 1: Examples of Japanese functional expressions and surface form variations

2.2 Creating Simple Japanese Replacement List

Text simplification, defined narrowly, is the process of reducing the linguistic complexity of a text, while retaining the original information contents and meaning (Siddharthan, 2014). Watananabe and Kawamura (2013) introduced a Japanese simplification system with the use of a "Simple Japanese Replacement List". Kaneniwa and Kawamura (2013) used the same list to rewrite difficult vocabulary automatically for Japanese learners who have non-kanji backgrounds. Kajiwara and Yamamoto (2015), constructed an evaluation dataset for Japanese lexical simplification. They extracted 2330 sentences from a newswire corpus and simplified only one difficult word using several Japanese lexical paraphrasing databases. Kodaira et al. (2016) built a controlled and balanced dataset for Japanese lexical simplification. They extracted 2010 sentences with only one difficult word in each sentence from a balanced corpus and collected simplification candidates using crowdsourcing techniques.

Different from the previous research mentioned above, we used a large scale Japanese balanced corpus to extract simplification candidates for difficult words and through manual selection we constructed a "Simple Japanese Replacement List" for Chinese-speaking Japanese language learners. For the levels of word difficulty, we consider levels 3-5 of the new JLPT as easy level, and levels 1-2 as difficult level according to the vocabulary list of the new JLPT, which consists of about 16,000 words. Besides, we consider the words which are not included in the vocabulary list of the new JLPT as difficult words. Since a large number of Kanji characters are used both in Japanese and in Chinese, Chinese-speaking learners can easily understand the Japanese words of Chinese origin (Japanese-Chinese homographs), such as "安全(safety)", "学習(study)", "擁立(support)", or easily guess the meaning of many Japanese original words with the help of Kanji characters, such as "壊す(break)", "打つ(hit)", "取り消す(cancel)". Kanji characters in these Japanese words are also included in Chinese dictionary and have the same or similar meaning with Chinese words. Therefore, we consider these words as easy words for Chinese-speaking learners, although some of these words are difficult words in the vocabulary list of the new JLPT.

We obtain a list of similar words associated with each difficult word which we are going to use for replacement, using Word2vec (https://code.google.com/p/word2vec/). As the training data for the Word2vec model, we use the Balanced Corpus of Contemporary Written Japanese (Maekawa, 2008), which consists of about 5,800,000 sentences in various domains. Based on the list of similar words, we choose easy words which are included in the vocabulary list of new JLPT as simplified words. If there is no appropriate easy word in the vocabulary list of the new JLPT, we use Japanese-Chinese homographs whose meaning is the same, or similar to Chinese words based on Japanese dictionaries (Bunrui goihyo zouhokaiteiban, 2004; Kadokawa ruigo shin jiten, 2002; Kojien 5th Editon, 1998). Japanese-Chinese homographs for simplification are specific for Chinese-natives. Therefore, our "Simple Japanese Replacement List" differs from previous research mentioned above on this aspect.

For simplification of Japanese functional expressions, we rewrite some difficult Japanese functional expressions using easy Japanese functional expressions or easy words in the vocabulary list of the new JLPT based on a Japanese sentence pattern dictionary (Group Jamasi, Xu. 2001).

Original words	Difficulty Level	Part of speech	Simplified words	Difficulty Level
合鍵 (duplicate key)	N1	Noun	鍵 (key)	N5
負け戦 (defeat)	N0	Noun	敗戦 (defeat)	N0
出会う(meet)	N0	Verb	会う (meet)	N5
胸苦しい (tough)	N1	Adjective	苦しい (tough)	N3
が早いか (when)	N1	Functional Expression	と (when)	N5
の際には (when)	N2	Functional Expression	の時 (when)	N5/N5

Table 2: Examples of the Simple Japanese Replacement List

Finally, we created a "Simple Japanese Replacement List," which consists of words and functional expressions. Table 2 shows some examples in the list.

3   Experiment and Evaluation

Our aim is to obtain appropriate example sentences that ease understanding of Japanese functional expressions. We conducted a small-scale experiment for evaluating the method for generating simplified example sentences. For the source data, the Balanced Corpus of Contemporary Written Japanese was used. We removed too short or too long sentences by limiting the sentence length between 3 to 25 words and then used the remaining 4,232,120 sentences for the experimental data.

To identify occurrences of Japanese functional expressions in the extracted sentences, we used a publicly available morphological analyzer MeCab (taku910.github.io/mecab/). We add the two lists we created, "Japanese Functional Expressions List" and "Simple Japanese Replacement List", into the IPA (mecab-ipadic-2.7.0-20070801) dictionary used as the standard dictionary for MeCab, with appropriate part-of-speech information for each expression, hoping that the morphological analyzer MeCab extracts the usages of functional expressions automatically. The accuracy is evaluated in the next section. Table 3 shows some example sentences of Japanese functional expressions and their corresponding simplified sentences.

ID	Original sentences	Simplified sentences
1	そしてそこへ、どこからか小鳥がやって来て、その虫をついばみました。	そしてそこへ、どこからか小鳥が**来**て、その虫を**啄**みました。
2	したがって、いろいろな関係部門、団体等と調整しなければならないことは言うまでもない。	**だから**、**色々**な関係部門、団体等と調整しなければならないことは言う**必要がない**。
3	私はそれ以来、数々の失敗を経てつぎのような結論にたどり着きました。	私はそれ以来、**多く**の失敗を経て**次**のような結論に**到達**しました。
4	つめが伸びていると、皮膚を傷つけるおそれがあるので、気をつけてください。	**爪**が伸びていると、皮膚を傷つける**可能性がある**ので、気をつけてください。
5	その土の中から、冬眠を破られた小さな虫が顔をのぞかせました。	その土の中から、冬眠を破られた小さな虫が顔を**見**せました。

Table 3: Examples of original sentences and simplified sentences. The words with underline are functional expressions and the words in bold are simplified words.

3.1   Evaluation of Japanese functional expressions

In this section, we randomly extract 200 sentences from the experimental data to examine whether the identified Japanese functional expressions are correct or not. Table 4 gives the evaluation results of identification of Japanese functional expressions.

Correct rate	Correctly extracted sentences		171 (85.5%)	
Error rate	Incorrectly extracted sentences	10 (5%)		29 (14.5%)
	Japanese functional expressions are not recognized	19 (9.5%)		
Total			200 (100%)	

Table 4: Evaluation results of Japanese functional expression identification

According to Table 4, we obtained 85.5% accuracy for identifying Japanese functional expressions. Cases of failures of functional expression identification can be viewed from the following three causes. First, the lack of discriminative contextual information causes failure. For example, "をめぐる(with related to)" in "決算 を めぐる 政策評価 の 問題", is incorrectly recognized as a literal usage. Here, both literal usage and functional usage of this expression share almost the same contexts and cannot be distinguished only by the surrounding information used in MeCab. Second is the opposite case where a literal usage is recognized as a functional expression. For example, in "青少年 期 から 身 につけて しまう", "につけて(concerning)" was incorrectly extracted. Third, functional expressions are not included in the current "Japanese Functional Expressions List". For example, a colloquial expression "わけじゃない(it does not mean that)" was not recognized as a functional expression.

3.2 Evaluation of Simplified Sentences

In this section, we evaluate the simplified sentences from the following two aspects, fluency and readability. From the 200 example sentences that include functional expressions, we removed 36 sentences which are easy sentences since they include no difficult words. We do not need to simplify such sentences. We then used the remaining 164 sentences for evaluation.

For the evaluation of fluency, we invited three Japanese natives to check the simplified sentences whether they are natural Japanese sentences. Meanwhile, for readability of the simplified sentences, we invited three Chinese-speaking learners who are all beginners of Japanese language. To compare the readability, we provided them with the Chinese translation of the original sentences and simplified sentences using an online translation software "Google translation" (translate.google.cn). We asked them to read and judge which sentence is easier to understand. Tables 5 and 6 show the evaluation results of fluency and readability respectively.

Natural sentences	142 (86.6%)
Unnatural sentences	22 (13.4%)
Total	164 (100%)

Table 5: Evaluation results of fluency of the simplified sentences

Easy to understand	132 (80.5%)
Difficult to understand	32 (19.5%)
Total	164 (100%)

Table 6: Evaluation results of readability of the simplified sentences

According to the evaluation results in Table 6, 80.5% sentences are simplified appropriately and become easier to understand. Two cases are identified in simplification failure. One is lack of appropriate simplified rules. For example, "請求 し うる" contains a functional expression "うる (possible)" with the difficulty level 2. However, no corresponding simplified word for the functional expression "うる" is found in the current "Simple Japanese Replacement List". This case cannot be coped with by lexical simplification. The other is the usage of inappropriate words, which is the main reason for generation of unnatural simplified sentences. For example, "市内の青果物商を片っ端から尋ねて回った。" was rewritten as "市内の野菜と果物商を一つ一つから尋ねて回った。", which is an unnatural sentence that produces unnatural connections for words.

4 Conclusion

In this paper, we presented our attempt to produce simplified example sentences for learning Japanese functional expressions using "Japanese Functional Expressions List" and "Simple Japanese Replacement List". A small-scale experiment was conducted to verify the effectiveness of the proposed method. The experimental results showed that simplified example sentences are helpful in learning Japanese functional expressions. In the future, we plan to estimate the difficulty level of the extracted example sentences automatically and offer better example sentences for Chinese-speaking learners of Japanese language.

Reference

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):490-496.

Group Jamashi, and Yiping Xu. 2001. *Chubunban Nihongo Kukei Jiten-Nihongo Bunkei Jiten (in Chinese and Japanese)*. Tokyo:Kurosio Publishers.

Tomoyuki Kajiwara, Kazuhide Yamamoto. 2015. Evaluation Dataset and System for Japanese Lexical Simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35-40. Beijing, China.

Kumido Kaneniwa, and Yoshiko Kawamura. 2013. Improving the Japanese simplification system for JFL/JSL learners from non-kanji background. *Journal of Japanese language education methods(in Japanese)*, 21(2):10-11.

Tomonori Kodaira, Tomoyuki Kajiwara, Mamoru Komachi. 2016. Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ⁃Student ResearchWorkshop*, pages 1-7, Berlin, Germany.

Kikuo Maekawa. 2008. Balanced Corpus of Contemporary Written Japanese. In *Proceeding of The 6th Workshop on Asian language Resources*, pp.101-102, Stroudsburg, PA, USA.

Suguru Matsuyoshi, Satoshi Sato, and Takehito Utsuro. 2006. Compilation of a dictionary of Japanese functional expressions with hierarchical organization. In *Proc.ICCPOL*, LNAI: Vol.4285, Springer:395-402.

National Institute for Japanese Language and Linguistics. 2004. *Bunrui goihyo zouhokaiteiban (in Japanese)*. Tokyo:Dainihontosho.

Susumu Ohno and Masando Hamanishi. 2002. *Kadokawa ruigo shin jiten (in Japanese)*. Tokyo: Kadokawa shoten.

Takahiro Ohno, Zyunitiro Edani, Ayato Inoue, and Dongli Han. 2013. A Japanese Learning Support System Mathching Individual Abilities. In *Proceeding of the PACLIC 27 Workshop on Computer-Assisted Language Learning,* pages 556- 562.

Lis Pereira and Yuji Matsumoto. 2006. Collocational Aid for Learners of Japanese as a Second Language. In *Proceeding of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA-2)*, pages 20-25, Beijing, China.

Takao Shime, Masatoshi Tsuchiya, Suguru Matsuyoshi, Takehito Utsuro and Satoshi Sato. 2007. Automatic Detection of Japanese Compound Functional Expressions and its Application to Statistical Dependency Analysis. *Journal of Natural Language Processing(in Japanese)*, 14(5):167-197.

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

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*165:2:259-298.

Masatoshi Tsuchiya, Takehito Utsuro, Suguru Matsuyoshi, Satoshi Sato, and Seiichi Nakagawa. 2006. Development and Analysis of An Example Database of Japanese Compound Functional Expressions. *Journal of Information Processing Society of Japan(in Japanese)*, 47(6):1728-1741.

Hyuma Watanabe, and Yoshiko Kawamura. 2013. Basic architecture for a Japanese simplification system. *Journal of Japanese language education methods(in Japanese)*, 20(2):48-49.