Sentence Suggestion of Japanese Functional Expressions for Chinese-speaking Learners

Jun Liu[1], Hiroyuki Shindo[1,2] and Yuji Matsumoto[1,2]
[1]Nara Institute of Science and Technology
[2]RIKEN Center for Advanced Intelligence Project (AIP)
{liu.jun.lc3, shindo, matsu}@is.naist.jp

Abstract

We present a computer-assisted learning system, *Jastudy*[1], which is particularly designed for Chinese-speaking learners of Japanese as a second language (JSL) to learn Japanese functional expressions with suggestion of appropriate example sentences. The system automatically recognizes Japanese functional expressions using a free Japanese morphological analyzer MeCab, which is retrained on a Conditional Random Fields (CRF) model. In order to select appropriate example sentences, we apply Support Vector Machines for Ranking (SVMrank) to estimate the complexity of the example sentences using Japanese-Chinese homographs as an important feature. In addition, we cluster the example sentences that contain Japanese functional expressions to discriminate different meanings and usages, based on part-of-speech, conjugation forms and semantic attributes, using the k-means clustering algorithm. Experimental results demonstrate the effectiveness of our approach.

1 Introduction

In the process of Japanese learning, learners must study many vocabulary words as well as various functional expressions. Since a large number of Chinese characters (Kanji characters in Japanese) are commonly used both in Chinese and Japanese, one of the most difficult and challenging problem for Chinese-speaking learners of Japanese as a second language (JSL) is the acquisition of Japanese functional expressions (Dongli Han, and Xin Song. 2011). Japanese has various types of compound functional expressions that consist of more than one word including both content words and functional words, such as “ざるをえない (have to)”, “ことができる (be able to)”. Due to various meanings and usages of Japanese functional expressions, it is fairly difficult for JSL learners to learn them.

In recent years, certain online Japanese learning systems are developed to support JSL learners, such as Reading Tutor[2], Asunaro[3], Rikai[4], and WWWJDIC[5]. Some of these systems are particularly designed to enable JSL learners to read and write Japanese texts by offering the word information with their corresponding difficulty information or translation information (Ohno et al., 2013; Toyoda 2016). However, learners’ native language background has not been taken into account in these systems. Moreover, these systems provide learners with limited information about the various types of Japanese functional expressions, which learners actually intend to learn as a part of the procedure for learning Japanese. Therefore, developing a learning system that can assist JSL learners to learn Japanese functional expressions is crucial in Japanese education.

In this paper, we present *Jastudy*, a computer-assisted learning system, aiming at helping Chinese-speaking JSL learners with their study of Japanese functional expressions. We train a CRF model and use a Japanese morphological analyzer MeCab [6] to detect Japanese functional expressions. To select the appropriate example sentences, we take Japanese-Chinese homographs as an important feature to estimate the complexity of example sentences using SVMrank[7]. In addition, in order to suggest example sentences that contain the target Japanese functional expression with the same meaning and usage, we cluster the

[1] http://jastudy.net/jastudy.php

[2] http://language.tiu.ac.jp/
[3] https://hinoki-project.org/asunaro/
[4] http://www.rikai.com/perl/Home.pl
[5] http://nihongo.monash.edu/cgi-bin/wwwjdic?9T
[6] http://taku910.github.io/mecab/
[7] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 56–61
Melbourne, Australia, July 15 - 20, 2018.

example sentences, based on part-of-speech, conjugation forms and semantic attributes of the neighboring words, using the k-means clustering algorithm in Scikit-learn[8].

2 General Method

As shown in Figure 1, our proposed system is mainly composed of three processes: automatic detection of Japanese functional expressions, sentence complexity estimation and sentence clustering. In this section, we explain them in detail.

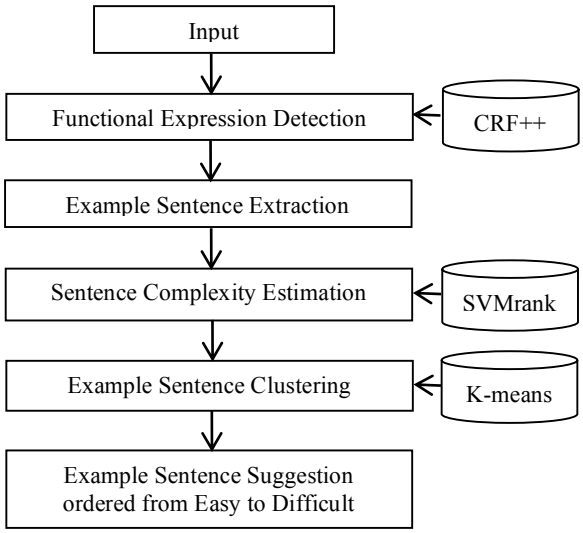


Figure 1: The processing stages of the system

2.1 Detection of Functional Expressions

Several previous researches have been especially paid attention on automatic detection of Japanese functional expressions (Tsuchiya et al., 2006; Shime et al., 2007; Suzuki et al., 2012). However, recognition of Japanese functional expressions is still a difficult problem. For automatic detection of Japanese functional expressions, we apply a Japanese morphological analyzer Mecab, which employs CRF algorithm to build the feature-based statistical model for morphological analysis.

While MeCab provides a pre-trained model using RWCP Text Corpus as well as Kyoto University Corpus (KC), we train a new CRF model using our training corpus, hoping MeCab can detect more Japanese functional expressions. To prepare the training corpus, we firstly referenced certain Japanese grammar dictionaries (Xiaoming Xu and Reika, 2013; Estuko Tomomastu, Jun Miyamoto and Masako Wakuki, 2016) to construct a list of Japanese functional expressions. As a result, we collected approximately 4,600 types of various surface forms in our list. Then we gathered 21,435 sentences from Tatoeba[9] corpus, HiraganaTime[10] corpus, BCCWJ[11] and some grammar dictionaries (Jamashi and Xu, 2001; Xu and Reika, 2013) and segmented each sentence into word level using MeCab. Finally, we manually annotated part-of-speech information for each Japanese functional expression in our training corpus. Figure 2 shows an example sentence after pre-processing.

お	接頭詞, 名詞接続, *, *, *, *, お, オ, オ
風呂	名詞, 一般, *, *, *, *, 風呂, フロ, フロ
に	助詞, 格助詞, 一般, *, *, *, に, ニ, ニ
入っ	動詞, 自立, *, *, 五段・ラ行, 連用タ接続, 入る, ハイッ, ハイッ
てから	**助詞, 接続助詞, 機能表現, *, *, *, てから, テカラ, テカラ**
寝	動詞, 自立, *, *, 一段, 連用形, 寝る, ネ, ネ
ます	助動詞, *, *, *, 特殊・マス, 基本形, ます, マス, マス
。	記号, 句点, *, *, *, *, 。, 。, 。

Figure 2: An example sentence (I will go to sleep after I take a bath.) after pre-processing. In the sentence, the Japanese functional expression and its part-of-speech information are in bold.

2.2 Sentence Complexity Estimation

There are a large number of Japanese words written with Chinese characters. Most of the words share identical or similar meaning with the Chinese words. We define these words as Japanese-Chinese homographs in our study. For Chinese-speaking learners, it is easy to understand their meanings even though they have never learned Japanese. Therefore, Japanese-Chinese homographs should be considered as an important feature in estimating sentence complexity.

In order to construct a list of Japanese-Chinese homographs, we firstly extracted Japanese words written only with Chinese characters from two Japanese dictionaries: IPA (mecab-ipadic-2.7.0-20070801)[12] and UniDic (unidic-mecab 2.1.2)[13]. These two dictionaries are used as the standard dictionaries for the Japanese morphological analyzer MeCab, with appropriate part-of-speech information for each expression. We then extracted the Chinese translations of these Japanese words from two online dictionary websites: Wiktionary[14]

[8] http://scikitlearn.org/stable/modules/clustering.html#clustering

[9] https://tatoeba.org/eng/

[10] http://www.hiraganatiomes.com/

[11] http://pj.ninjal.ac.jp/corpus_center/bccwj/en/

[12] https://sourceforge.net/projects/mecab/files/mecab-ipadic/2.7.0-20070801/mecab-ipadic-2.7.0-20070801.tar.gz/download

[13] http://osdn.net/project/unidic/

[14] http://ja.wictionary.org/wiki/

and Weblio[15]. We compared the character forms of Japanese words with their Chinese translations to identify whether the Japanese word is a Japanese-Chinese homograph or not. Since Japanese words use both the simplified Chinese characters and the traditional Chinese characters, we first replaced all the traditional Chinese characters with the corresponding simplified Chinese characters. If the character form of a Japanese word is the same as the character form of the Chinese translation, the Japanese word is recognized as a Japanese-Chinese homograph, as illustrated in Table 1.

Considering unknown words in the above online dictionaries, we also referenced an online Chinese encyclopedia: Baike Baidu[16] and a Japanese dictionary: Kojien fifth Edition (Shinmura, 1998). If a Japanese word and its corresponding Chinese translation share an identical or a similar meaning, the Japanese word is also identified as a Japanese-Chinese homograph. Ultimately, we created a list of Japanese-Chinese homographs that consists of approximately 14,000 words.

Original Japanese word	Simplified Chinese characters	Chinese translation	Japanese-Chinese homographs
社会 (society)	社会	社会	Yes
緊張 (nervous)	紧张	紧张	Yes
手紙 (letter)	手纸	信件	No

Table 1: Examples of Identification of Japanese–Chinese homographs

To estimate sentence complexity, we follow the standard of the JLPT (Japanese Language Proficiency Test). The JLPT consists of five levels, ranging from N5 (the least difficult level) to N1 (the most difficult level)[17]. We employ the following 12 features as the baseline feature set:

- Numbers of N0–N5 Japanese words in a sentence (Here, N0 implies unknown words in the vocabulary list of JLPT.)
- Numbers of N1–N5 Japanese functional expressions in a sentence
- Length of a sentence

Different from the standard of the JLPT, the words in the list of Japanese–Chinese homographs (JCHs) were categorized separately as a new feature. Ultimately, we combine the following new features with the baseline features (all 17 features), forming our feature set.

- Numbers of JCHs in a sentence
- Numbers of verbs in a sentence
- Numbers of syntactic dependencies in a sentence
- Average length of syntactic dependencies
- Maximum number of child phrases

The last three features are to measure syntactic complexity of a sentence. We used a well-known Japanese dependency structure analyzer CaboCha[18] to divide an example sentence into base phrases (called bunsetsu) and to obtain its syntactic dependency structure. For example, the example sentence “彼は人生に満足して死んだ。(He died content with his life.)” is divided into four phrases: “彼は”, “人生に”, “満足して”, “死んだ”. In this sentence, the first, and the third phrases depend on the fourth, and the second phrase depends on the third. The numbers of syntactic dependencies in this sentence is 3. The length of syntactic dependencies is the numbers of phrases between arbitrary phrase and its dependent. In this sentence, the average length of syntactic dependencies is 1.7 (the length of syntactic dependency between the first and the fourth is 3, the length of syntactic dependency between the second and the third is 1, and the length of syntactic dependency between the third and the fourth is 1). The fourth phrase has two child’s phrases while the third has only one child phrase, so the maximum number of child phrases in this sentence is 2.

2.3 Sentence Clustering

Some Japanese functional expressions have two or more meanings and usages. For example, the following two example sentences contain the identical Japanese functional expression “そうだ”, but have different meanings. However, we can distinguish the meaning of “そうだ” through part-of-speech and conjugation forms of the words that appear just before “そうだ”.

雨が降りそうだ。 (**It looks like** it will rain.)
雨が降るそうだ。 (**It’s heard that** it will rain.)

To obtain example sentences for each of distinct usages of a functional expression, we apply a

[15] http://cjjc.weblio.jp
[16] https://baike.baidu.com
[17] http://jlpt.jp/e/about/levelsummary.html
[18] https://taku910.github/cabocha/

clustering algorithm with a small number of known examples (those appear in dictionaries) and a large number of untagged example sentences. For the features of sentence clustering, we utilize the following features: part-of-speech, conjugation form, and semantic attribute of the word that appear just before or after the target Japanese functional expression.

3 Experiments and Results

3.1 Automatically Detecting Japanese Functional Expressions

This experiment evaluates automatic detection of Japanese functional expressions.

We apply CRF++[19], which is an open source implementation of CRF for segmenting sequential data. We utilized nine features including surface forms and their part-of-speech in our training. The training corpus mentioned in Section 2.1 was used in the CRF++. The CRF++ learned the training corpus and outputted a model file as the learning result. We then applied MeCab, trained on our training corpus, to automatically recognize the Japanese functional expressions.

For the test data, we randomly extracted 200 example sentences from Tatoeba, HiraganaTimes and BCCWJ. Table 2 shows some examples of detected Japanese functional expressions by our system. The final evaluation results are shown in Table 3. We obtained 86.5% accuracy, indicating our approach has certain validity.

Correctly detected Japanese functional expressions
Input: 今、雪が降っている。 (It is snowing now.) Output: 今 、 雪 が 降っ **ている** 。 Input: この箱を開けてください。 (Please open this box.) Output: この 箱 を 開け **てください** 。
Incorrectly detected Japanese functional expressions
Input: 彼女は火にあたってからだを暖めた。 (She warmed herself by the fire.) Output: 彼女 は 火 **にあたって** からだ を 暖め た 。

Table 2: Detection of Japanese functional expressions. In the sentences, Japanese functional expressions are in bold and underlined.

Correctly recognized	173 (86.5%)
Incorrectly recognized	27(13.5%)
Total	200 (100%)

Table 3: Experimental results on detection of Japanese functional expressions

[19] https://taku910.github.io/crfpp/

3.2 Estimating Sentence Complexity

This experiment evaluates sentence complexity estimation, using an online machine learning tool SVMrank.

We first collected 5,000 example sentences from Tatoeba, HiraganaTimes, BCCWJ and randomly paired them and constructed 2,500 sentence pairs. Then 15 native Chinese-speaking JSL learners, all of whom have been learning Japanese for about one year, were invited to read the pairs of example sentences and asked to choose the one which is easier to understand. We asked three learners to compare each pair and the final decision was made by majority voting. We finally applied a set of five-fold cross-validations with each combination of 4,000 sentences as the training data and 1,000 sentences as the test data.

The experimental results using baseline features and our method using all of the proposed features are presented in Tables 4 and 5. Compared with the results using the baseline features, our method enhances the average accuracy by 3.3%, partially demonstrating the effectiveness of our features.

Features	Cross-validations	Accuracy
Baseline Features	1	83.2%
	2	84%
	3	80.4%
	4	82%
	5	81.8%
Average		82.3%

Table 4: Experimental results using baseline features.

Features	Cross-validations	Accuracy
Proposed Features	1	87.6%
	2	86.4%
	3	84.6%
	4	83.8%
	5	85.4%
Average		85.6%

Table 5: Experimental results using our proposed features

3.3 Clustering Example Sentences

This experiment evaluates the performance of sentence clustering, using the k-means clustering algorithm in Scikit-learn.

Here in our study, we took five different types of Japanese functional expressions as the examples. For the test data, we collected 10 example sentences, which were used for the reference, from Japanese functional expression dictionaries

and 20 example sentences from Tatoeba, HiraganaTimes, and BCCWJ for each type of Japanese functional expressions, respectively. We conducted our experiments with the number of clusters ranging from four to six. The clustering result was evaluated based on whether the test data that was clustered into one cluster share the same usage of a Japanese functional expression. The experimental results are shown in Table 6. The average results of accuracies for the number of clusters ranging from four to six are 89%, 93%, 92%, indicating the usefulness of the sentence clustering method for classifying sentences in the same usage.

Functional Expressions	Numbers of Clusters	Accuracy
そうだ (it looks like / it's heard that)	4	97%
	5	97%
	6	97%
とともに (together with / at the same time)	4	87%
	5	97%
	6	87%
ため（に） (because / in order to)	4	83%
	5	83%
	6	90%
に対して (to / every / in contrast to)	4	87%
	5	93%
	6	93%
次第（だ） (as soon as / depends on)	4	93%
	5	93%
	6	93%
Average	4	89%
	5	93%
	6	92%

Table 6: Experimental results of sentence clustering

4 Featured functions of the Demo

In our proposed demo, we have implemented the following main functions.

1. The function to detect Japanese functional expressions. Given a sentence, *Jastudy* automatically segments the input sentence into individual words using MeCab. Difficult Japanese functional expressions (N2 and above) in the input sentence are simplified with easier Japanese functional expressions (N3 and below) or with phrases and shown in the output sentence, using a "Simple Japanese Replacement List" (Jun Liu and Yuji Matsumoto, 2016). An example is shown in Figure 3. Moreover, *Jastudy* represents detailed information about the surface-form, part-of-speech of each word in the input sentence and the output sentence, respectively.

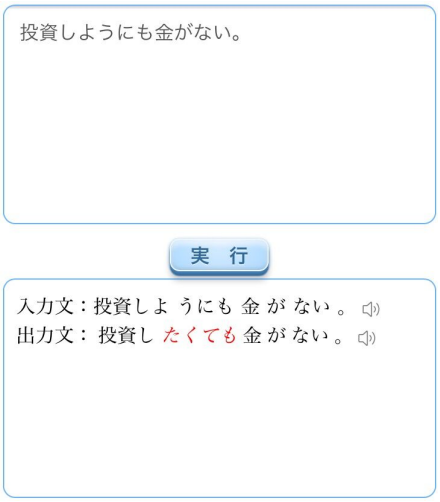
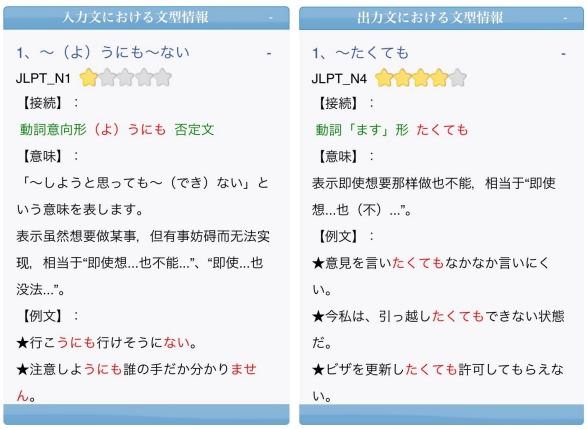


Figure 3: "投資しようにも金がない。(I have no money to invest.)" is typed in the system.

2. The function to provide JSL learners with the detail information about the meaning, usage and example sentences of the Japanese functional expression which appears in the input sentence and the output sentence, respectively. An example is shown in Figure 4. Learners can also choose the Japanese functional expressions they want to learn, based on their Japanese abilities.



(a) (b)

Figure 4: Detailed information of Japanese functional expressions appeared in the input sentence (a) and the output sentence (b).

3. The function to suggest comprehensive example sentences. The JSL learners can search more example sentences through the following three aspects: 1) only keyword, 2) only usage, 3) both keyword and usage. For example, the learner inputs the Japanese functional expression "そうだ" as a keyword and selects its meaning and usage "it looks like" from drop-down list, a list of

example sentences that contain the functional expression sharing the same meaning are retrieved to form the corpus, as shown in Figure 5. The only sentences whose complexity is equal to or below the learner's level are retrieved.

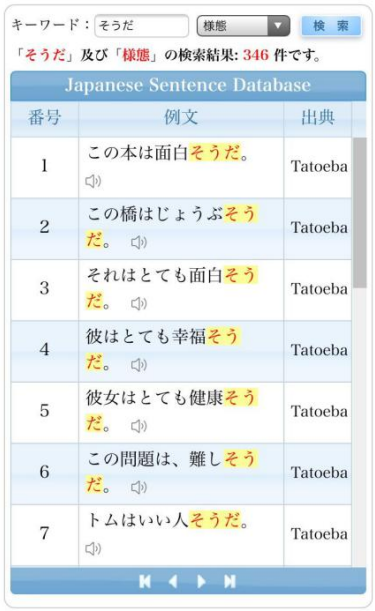


Figure 5: Example sentences suggested by the system, given "そうだ" with its meaning as "様態(it looks like)"

5 Conclusion and Future Work

In this paper, we presented a computer-assisted leaning system of Japanese language for Chinese-speaking learners with their study of Japanese functional expressions. The system detects Japanese functional expressions using MeCab that employs the CRF model we trained. We apply SVMrank to estimate sentence complexity using the Japanese-Chinese homographs as an important feature to suggest example sentences that are easy to understand for Chinese-speaking JSL learners. Moreover, we cluster example sentences containing the Japanese functional expressions with the same meanings and usages. The experimental results indicate effectiveness of our method.

We plan to examine the run-time effectiveness of the system for JSL learners. This will be our future task for improving the performance of our system.

Acknowledgments

We would like to thank anonymous reviewers for their detailed comments and advice.

References

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

Jun Liu, Yuji matsumoto. 2016. Simplification of Example Sentences for Learners of Japanese Functional Expressions. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 1–5.

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*, pages 556–562.

Takao Shime, Masatoshi Tsuchiya, Suguru Matsuyoshi, Takehito Utsuro, Satoshi Sato. 2007. Automatic Detection of Japanese Compound Functional Expressions and its Application to Statistical Dependency Analysis, *Journal of Natural Language Processing*, Vol (14). No.5: 167-196.

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

Takafumi Suzuki, Yusuke Abe, Itsuki Toyota, Takehito Utsuro, Suguru Matsuyoshi, Masatoshi Tsuchiya. 2012. Detecting Japanese Compound Functional Expressions using Canonical/Derivational Relation, In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*.

Estuko Tomomastu, Jun Miyamoto and Masako Wakuki. 2016. Japanese Expressions Dictionary. *Aruku Press*.

Etsuko Toyoda. 2016. Evaluation of computerized reading-assistance systems for reading Japanese texts – from a linguistic point of view. *Australasian Journal of Educational Technology*, 32(5): 94-97.

Masatoshi Tsuchiya, Takao Shime, Toshihiro Takagi, Takehito Utsuro, Kiyotaka Uchimoto, Suguru Matsuyoshi, Satoshi Sato, Seiichi Nakagawa. 2006. Chunking Japanese compound functional expressions by machine learning, In *Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context*, pages 25-32.

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