A Supporting System for Learning and Using Japanese Sentence Patterns

Xin Song　and　Dongli Han

Graduate School of Basic Integrated Sciences,
College of Humanities and Sciences, Nihon University
3-25-40 Sakurajyosui, Setagaya-ku, Tokyo 156-8550, JAPAN

1 Introduction

More and more foreigners began to learn Japanese with the increasing international trade and cultural exchange. However, it is quite difficult for them to create a Japanese document comparing with the case of speaking. Sentences or documents written by foreigners often make natives feel strange somehow although one can get the meaning from the words used.

Two main factors are supposed to be affecting this situation. One is the vocabulary and another is the various sentence patterns in Japanese. There are 3 kinds of characters in Japanese: Hiragana, Katakana, and Kanji, Whose combination has made writing Japanese a bit complicated. Similarly there are many sentence patterns in Japanese to express different meanings in various situations. This is sometimes more difficult than vocabulary for foreigners, especially for those from Chinese-speaking countries. Kanji characters are not that hard for them while the diverse sentence patterns are completely new and quite different from there mother tongue.

In this paper, we first make a survey on existing strategies to support Chinese-speaking foreigners to write Japanese documents, then propose a method for learning and using Japanese sentence patterns in a more practical and more convenient way.

2 Existing strategies and our objective

Generally, we can employ machine translation systems to help create a sentence in the target language while with a significant problem: translation precision. For example we get the Japanese equivalents of a Chinese input sentence: *即使头痛，也不能请假* by Google[1] and Yahoo[2] translation service as follows.

Google：場合でも、頭痛、我々を残していないことができます

Yahoo：仮に頭痛、同様に欠勤不能です

Both the two equivalents are incorrect while the correct translation should be something like *たとえ頭が痛くても休めない*. Another preliminary experiment using the commercial machine translation software: KoryaEiWa! Ippatsu Honyaku with the same input sentence gets the equivalent as *たとえ、頭痛休暇を取ることも出来なくます* which is incorrect too.

Foreigners that are learning Japanese can refer to the translated equivalent to some extent to create a Japanese sentence, but can seldom copy the equivalent in its original form.

Other approaches to help foreigners write Japanese sentences may include researches on phrase parallel translation, and sentence parallel translation. The former one depends mainly on the performance of electronic dictionaries, and is divided into two types. One is the portable electronic dictionary products such as Casio's Ex-word[3]. Another is the dictionary used in the computer or an online dictionary. Some dictionaries have been constructed along this idea [1-2]. We do benefit from various dictionaries when we are writing a document and are wondering which word to use. However we could not create any sentence with the dictionary only if we are not familiar with the grammar or the sentence patterns. The latter one indicates the statistics-based proposals to select an existing Japanese equivalent for an input sentence from a pre-constructed sentence parallel translation database [3]. This method needs exhaustive parallel corpora which is difficult to be prepared.

As described above,neither machine translation nor parallel translation can really satisfy the need of foreigners to generate correct Japanese documents. What they need is something that can teach them which sentence pattern should be

[1] http://translate.google.co.jp/

[2] http://honyaku.yahoo.co.jp/

[3] http://casio.jp/exword/

used in a particular situation and how to express their real thinking correctly in Japanese.

Several studies have made some attempts to teach foreigners some knowledge about Japanese grammar during the procedure of creating documents. For example, Imaeda et al. develop a system that can detect the errors in words and grammars [4], and Yamaguchi et al. propose a method where the system and the writer interact with each other [5]. Another approach has been offered to search Japanese sentence patterns in an input Japanese sentence [6]. These systems could help foreigners who have already learned Japanese and could create a document in Japanese to a certain extent, but will be impractical to Japanese beginners.

In this paper, we consider the above problems, and develop a supporting system for learning Japanese sentence patterns based on the real needs of Chinese-speaking foreigners.

3 Data flow of the system

We set up several goals for our system to support foreigners' document generation.

(1) to find suitable sentence patterns efficiently to create a Japanese sentence.
(2) to support the input with the user's mother tongue (Chinese).
(3) to offer enough and reliable examples of the selected sentence pattern to enhance the learning effect.

For the first goal, we adopt Ajax technology to realize the real-time feature and the quick response of our sentence pattern searching. Then we construct a Chinese-Japanese sentence pattern database according to a dictionary [7]. Using the database and Ajax technology, users could input Chinese sentence and the system will display all possible Japanese sentence patterns partway according to the user's input. Finally, we realize the last goal by gathering sentences containing the selected patterns from the web, calculating their difficulty levels through a machine learning module, and showing them in order of increasing difficulty to the users.

4 Modules in the system

In this section, we describe each module in the system.

4.1 Input and sentence pattern searching

We construct a Chinese-Japanese sentence pattern database in advance according to a Chinese-Japanese grammar reference [7]. The reference contains most common sentence patterns or expressions in Chinese and their equivalents in Japanese. We employ Ajax to realize the real-time sentence pattern searching. Figure 1 is the data flow our system.

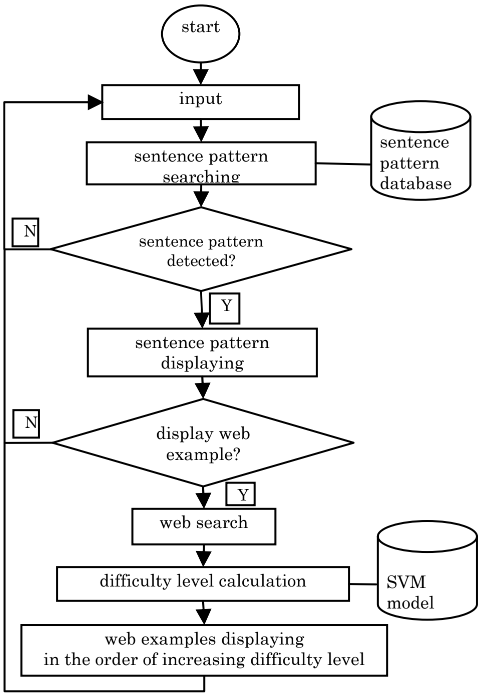

start
input
sentence pattern searching
sentence pattern database
N
sentence pattern detected?
Y
sentence pattern displaying
N
display web example?
Y
web search
difficulty level calculation
SVM model
web examples displaying in the order of increasing difficulty level

Figure 1 Data flow of the system

When the user is inputting some words in Chinese, the system will detect whether the input contains a whole or just a partial sentence pattern by checking against the Chinese entries in the pre-constructed database. If there exists one, its Japanese equivalent will be shown together with the explanation of its usage.

During the process of sentence pattern searching in the database, we adopt a condition called *double-direction inclusion principle* as shown by the following formula.

$$\exists key \quad ((input \subseteq key) \ \| \ (input \supseteq key))$$

Here, a *key* indicates a Chinese sentence pattern in the Chinese-Japanese database, and *input* is the input Chinese text. We search the database to examine whether the user is inputting a Chinese sentence pattern. The search will succeed both in the case the input contains a key and the case a key contains the current input. We try to help users find relevant sentence patterns as more as possible in this manner.

Then the Japanese equivalents of all Chinese sentence patterns that have been found during the above search will be listed up. For example, for the Chinese sentence, *关于外交政策还有许多应该*

讨论的事 (We should argue more about diplomatic policy), we get two sentence patterns, *关于*, and *应该*, meaning *about* and *should*, and find two Japanese equivalents for each Chinese sentence pattern: *について* or *に関して* for *关于*, *はず* or *べき* for *应该* respectively. The user could check all the Japanese equivalents and determine the most proper sentence pattern from all candidates to create a sentence according to the particular situation.

4.2 Web search for example sentence

If the user needs more examples on the usage of the Japanese sentence pattern, a web search is activated according to the user's intention. Here in this section, we describe the procedure of web search for more usage examples.

We use Google API to gather usage examples and in consideration of the unreliability of general web texts, we restrict our web search within the range of Google News. Below we give the processing steps.

Step1: search the selected sentence pattern in Google News
Step2: divide the retrieved snippets into sentences.
Step3: check each sentence obtained in Step2 and select sentences containing the sentence pattern.

In case a Japanese sentence pattern contains multiple words, like 例え…ても… (even if), we will not terminate the check until all parts of the sentence pattern are located in the example sentence. We pass the retrieved example sentences into the next procedure to sort them in an order of increasing difficulty.

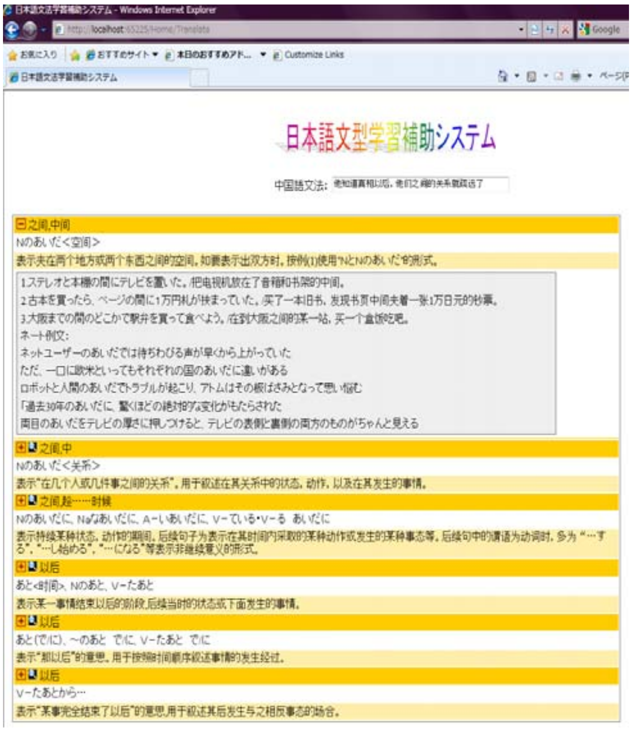
4.3 Example sentence sorting based on SVM

After we gather some example sentences for the sentence pattern from the web, we have to think about the sequence to show the sentences to the user. As mentioned above, our system is developed mainly for the Japanese beginners. That is to say, we should provide them with as simple as possible sentences that they can understand quickly and then focus on the usage of the sentence pattern to enhance the learning effect.

In a study on the standardization of Japanese sentences, Sato et al. defined four factors to examine whether a sentence satisfied a standardized difficulty level [8]. Here in our study, we adopt a machine learning tool, support vector machine[4] (SVM) to decide whether an example sentence is difficult or not. We employ the following features to create the classification model.

- length of the example sentence
- total difficulty level of kanji characters
- total difficulty level of words
- total difficulty level of sentence patterns
- number of verbs
- number of dependency relations

The difficulty level of a kanji character, a word, and a sentence pattern are determined according to the grade they belong to in the Japanese Language Proficiency Test[5]. We gathered 100 example sentences as the training data from the web, and classify them into 2 categories: simple sentences and difficult sentences with handcraft by three Chinese college students. We learn from the training data and create a classification model using SVM.


日本語文法学習補助システム - Windows Internet Explorer
http://localhost:65225/Home/Translate
Google
お気に入り
おすすめサイト
本日のおすすめアド...
Customize Links
日本語文法学習補助システム
ページ(P

日本語文型学習補助システム

中国語文法: 他知道真相以后，他们之 间的关系就疏远了

⊟之间,中间

Nのあいだ<空间>

表示夹在两个地方或两个东西之间的空间。如要表示出双方时，按例(1)使用"NとNのあいだ"的形式。

1.ステレオと本棚の間にテレビを置いた。/把电视机放在了音箱和书架的中间。
2.古本を買ったら、ページの間に1万円札が挟まっていた。/买了一本旧书，发现书页中间夹着一张1万日元的钞票。
3.大阪までの間のどこかで駅弁を買って食べよう。/在到大阪之间的某一站，买一个盒饭吃吧。
ネート例文:
ネットユーザーのあいだでは待ちわびる声が早くから上がっていた
ただ、一口に欧米といってもそれぞれの国のあいだに違いがある
ロボットと人間のあいだでトラブルが起こり、アトムはその板ばさみとなって思い悩む
「過去30年のあいだに、驚くほどの絶対的な変化がもたらされた
両目のあいだをテレビの厚さに押しつけると、テレビの表側と裏側の両方のものがちゃんと見える

⊞之间,中

Nのあいだ<关系>

表示"在几个人或几件事之间的关系"，用于叙述在其关系中的状态，动作，以及在其发生的事情。

⊞之间趁……时候

Nのあいだに、Naなあいだに、A-いあいだに、V-ている•V-る あいだに

表示持续某种状态，动作的期间，后续句子为表示在其时间内采取的某种动作或发生的某种事态等，后续句中的谓语为动词时，多为"…する"，"…し始める"，"…になる"等表示非继续意义的形式。

⊞以后

あと<时间>、Nのあと、V-たあと

表示某一事情结束以后的阶段,后续当时的状态或下面发生的事情。

⊞以后

あと(で/に)、~のあと で/に、V-たあと で/に

表示"那以后"的意思，用于按照时间顺序叙述事情的发生经过。

⊞以后

V-たあとから…

表示"某事完全结束了以后"的意思,用于叙述其后发生与之相反事态的场合。

Figure 2 System interface

Figure 2 is the system interface. When an icon is pressed to display web example sentences, the difficulty level of each sentence retrieved from the web is calculated based on the value of the classification result from SVM, and then the top 5

[4] http://svmlight.joachims.org
[5] http://www.jees.or.jp/jlpt/

examples containing a particular sentence pattern retrieved from Google News are displayed in the order of increasing difficulty level. In this way, the users, i.e., the Japanese beginners could usually refer to as more as possible usage examples from the simplest ones that could help them create an envisioned document.

5 Experiments

In this section, we describe the experiments for evaluating the effect of SVM. We divide the training data into 10 groups, each group with 10 sentences respectively. Then we carry out a set of cross-validations with every combination of three groups of data as the training data and the rest two groups as test data. Table 1 shows the experimental results.

Table 1 Experimental results of SVM

	Precision(%)	recall(%)	F-measure(%)
1	83.3	88.2	85.7
2	81.8	94.7	87.8
3	80.0	83.3	81.6
4	72.2	81.3	76.5
5	76.2	76.2	76.2
6	72.0	78.3	75.0
7	83.3	100.0	90.9
8	81.3	86.7	83.9
9	84.2	94.1	88.9
10	68.8	78.6	73.3
Average	78.3	86.1	82.0

In Table 1, we list the precisions, recalls, and F-values of each cross-validation with the label of 1 to 10, and their average values on the bottom-most line. We calculate the precision, the recall, and the F-measure as shown in the following formulas.

$$precision = \frac{correct\ results\ of\ system}{results\ of\ system}$$

$$recall = \frac{correct\ results\ of\ system}{correct\ results}$$

$$F - measure = \frac{2 * correct\ results\ of\ system}{results\ of\ system + correct\ results}$$

The average results are 78.3%, 86.1%, and 82.0%, indicating the effectiveness of our method using SVM for calculating the difficulty levels of web example sentences.

6 Conclusions

We have built a system for Chinese-speaking foreigners to learn Japanese sentence patterns with the aim of generating proper sentences in Japanese. We adopt SVM to calculate the difficulty levels of web example sentences in order to show them to the user from the simplest ones. The experimental results show the effectiveness of our method. However, we still have some works to do. The scale of the training data for SVM is not large enough to obtain more accurate results, and we have not made any evaluation on the practicality or the operability of the system. These will be our future tasks for improving the performance of our system.

References

[1]Lars Yenken, Zhihui Jin, Kumiko Tanaka-Ishii. “Pinyomi: Dictionary Lookup via Orthographic Associations”. in Proceedings of 10th Conference of the Pacific Association for Computational Linguistics, pp.13-17. (2007)

[2] Tsunakawa, Takashi, Naoaki Okazaki, Xiao Liu, Jun'ichi Tsujii. “A Chinese-Japanese Lexical Machine Translation through a Pivot Language”. ACM Transactions on Asian Language Information Processing, 8(2). pp.9:1-9:21. (2009)

[3]Peng Yang, Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara. “Construction of the J/C Machine Translation System with Valency Patterns”. SIG notes, NL-2008(4), Information Processing Society of Japan, pp.121-126. (2008) (in Japanese)

[4]Koji Imaeda, Atsuo Kawai, Yuji Ishikawa, Ryo Nagata, Fumito Masui. “Error Detection and Correction of Case Particles in Japanese Learner’s Composition”. SIG notes, NL-2003(13), Information Processing Society of Japan, pp.39-46. (2003) (in Japanese)

[5]Masaya Yamaguchi, Masanori Kitamura. “TEachOtherS: A Writing Aid System for Students, in Proceedings of the 11th IASTED International Conference on Computers and Advanced Technology in Education. (2008)

[6]Wenrui Huang. “A Study on the Development of Japanese Sentence Pattern System with Relation Data Model”. Synopsis of master's theses 1999, Tokyo University of Information Sciences, pp.158. (in Japanese)

[7]Group Jamasi, Yiping Xu(translator). “Chubun ban Nihongo Kukei Jiten-Nihongo Bunkei Jiten”. Kurosio Publisher. (2001) (in Chinese and Japanese)

[8]Satoshi Sato, Masatoshi Tsuchiya, Masahiro Murayama, Masahiro Asaoka, Qingqing Wang. “Standardization of Japanese Sentences”. SIG notes, NL-2003(4), Information Processing Society of Japan, pp.133-140. (2003) (in Japanese)