Automatic Easy Japanese Translation for information accessibility of foreigners

Manami MOKU [1] *Kazuhide YAMAMOTO* [1] *Ai MAKABI* [1]

(1) Department of Electrical Engineering, Nagaoka University of Technology,
1603-1, Kamitomioka-cho, Nagaoka-city, Niigata 940-2188, JAPAN
{moku, yamamoto, makabi}@jnlp.org

ABSTRACT

This paper examines the introduction of "Easy Japanese" by extracting important segments for translation. The need for Japanese language has increased dramatically due to the recent influx of non-Japanese-speaking foreigners. Therefore, in order for non-native speakers of Japanese to successfully adapt to society, the so-called Easy Japanese is being developed to aid them in every aspect from basic conversation to translation of official documents. The materials of our project are the official documents since they are generally distributed in public offices, hospitals, and schools, where they include essential information that should be accessed for all residents. Through an analysis of Japanese language dependency as a pre-experiment, this paper introduces a translation by extracting important segments to facilitate the acquisition of Easy Japanese. Upon effective completion, the project will be introduced for use on the Internet and proposed for use by foreigners living in Japan as well as educators.

KEYWORDS : Easy Japanese, Extracting important segments, Translation system, Official documents, Japanese education.

1 Introduction

It is estimated that more than two million foreigners are now living in Japan and roughly a half million of those do not have enough Japanese fluency. Since only Japanese is used in ordinary Japanese society, it has been a problem in Japan in terms of information accessibility to such foreigners.

One solution for this is use of simple and plain expressions for communication to those. Several trials have been attempted to define and spread somewhat simple expressions to the non-Japanese community, mainly by Japanese language teachers. We are joining "Easy Japanese" project (Isao Iori, 2008) since last year. Although it is also a project to teach easy Japanese to foreigners, one goal of this project is to automatically "translate" (or summarize easily) ordinary Japanese sentences into easy one, by use of natural language processing (NLP) techniques. The target material of the project is official documents that are generally distributed in public offices, hospitals, and schools, where they include essential information that should be accessed for all residents.

It is observed that official documents may include some peculiar expressions that make it difficult for foreigners to understand. For example, in case of English, we may see something like: "Please avoid your children's attendance in school with an assessment of the situation by a guardian when the situation is dangerous for children in case of bad weather." Although it is no problem to understand for native speakers, it is far easier for non-native speakers just to say like: "Don't go to school in case of bad weather." We aim to build a system to translate a sentence like the former one into the latter one. We propose in this paper to do that by extracting essential segments and rewriting them into more direct expressions. This paper briefly reports outline of the project, approach of the current translation system, and results of preliminary experiments.

2 Related works

2.1 Easy Japanese

This system of so-called Easy Japanese has been previously researched by those in the translation of news. In one particular study, easy and difficult words from the news were defined (Hideya Mino et al., 2010). In this case, the authors utilized pairs of entities, and the word levels were defined on the basis of a word list from the Japanese Language Proficiency Test (JLPT).[1] This method was general method since there were similar methods.

2.1.1 Easy Japanese system

A previous Easy Japanese system, known as the Plain Japanese (PJ) system,[2] was designed for use in engineering education in Japan. Although such education is generally in Japanese, international students find it difficult not only to learn everyday Japanese but also acquire technical Japanese. In this case, this system used both restricted vocabulary and grammar. Therefore, this method was not suitable for our system since we aim to extract such important contents.

[1] http://www.jlpt.jp/e/index.html : This site is written in English.
JLPT is one of tests for Japanese beginners who learn Japanese. This research use the grade of JLPT, N1~N5.

[2] http://twinning.nagaokaut.ac.jp/PJ/PJ.html : This site is written in Japanese.

2.2 Extraction of important contents

Extracting important contents and sentences (Tsutomu Hirano et al., 2005) was generally used for summarization since the summary maintains natural grammar. However, sometimes, abstract sentences are reconstructed from some natural sentences. In one particular study, important segments were extracted for summaries using Support Vector Machines (SVMs) (Daisuke Suzuki et al., 2006), which was more effective when summarizing documents compared to extracting important sentences. We believe that extracting important segments can be the same as talking with Japanese language beginners. Therefore, we would like to re-introduce an easy process based on Japanese dependency analysis since we do not have more examples of important segment extraction in official documents using SVMs.

3 Data

3.1 Easy Japanese corpus

Easy Japanese overall includes two corpora. The first Easy Japanese pre-corpus was created by two Japanese teachers (Chie Tsutsui, 2010) and included 1,179 sentences from official documents that were rewritten into Easy Japanese. In this case, "easy" implies that Japanese language beginners can easily understand words/sentences, whereas "difficult" indicates that they simply cannot understand the sentences. For this first corpus, the grammar was considered by our project member while the vocabulary was determined on the basis of Japanese Language Proficiency Test (JLPT) levels.

The second Easy Japanese corpus was created by 40 Japanese teachers and it included 42,274 official sentences that were rewritten into Easy Japanese. An example of these language pairs is shown in TABLE 1.

For this paper, Easy Japanese pre-corpus is used for evaluating and extracting important segments. In addition, the Easy Japanese corpus will be used for building the Easy Japanese translation system.

		Kind of corpus		Japanese	English
output	Japanese	Easy Japanese Pre-corpus	Easy Japanese Corpus	予防接種	a vaccination
	Easy Japanese			予防注射	a preventive injection
				病気にならないための注射	an injection which prevents a disease

TABLE 1 - An example of a pair of Japanese and Easy Japanese from each corpora.

4 Pre-experiment for extracting important segments

4.1 Important segment extraction

First, we focused on the predicates of official sentences since the important contents, especially the instructions, were constructed with verbs. In addition, we randomly selected 20 sentences from the Easy Japanese pre-corpus, and the sentences were edited with conjunctions and keywords such as “場合 (in case of)” through morphological analysis by ChaSen.[3] An example is shown in TABLE 2.

Next, the sentences were analyzed through a Japanese dependency analysis by CaboCha,[4] and the output of this process became the candidates for these important sentences. An example is shown in TABLE 3.

		Japanese	English
input		悪天候の際には，大雨警報，暴風警報，大雪警報，暴風雪警報が発令されていなくても，周囲の状況で危険な場合は，保護者の判断で登校を見合わせてください．	Please avoid your children’s attendance in school with an assessment of the situation by a guardian when the situation is dangerous for children and no warning is issued in case of bad weather.
output	I	悪天候の際には，	in case of bad weather
	II	大雨警報，暴風警報，大雪警報，暴風雪警報が発令されていなくても，周囲の状況で危険な場合は，	when the situation is dangerous for children and no warning is issued
	III	保護者の判断で登校を見合わせてください．	Please avoid your children’s attendance in school with an assessment of the situation by a guardian

TABLE 2 - An example of the process for decreasing errors in the Japanese dependency analysis.

		Japanese	English
input		保護者の判断で登校を見合わせてください．	Please avoid your children’s attendance in school with an assessment of the situation by a guardian.
Japanese dependency analysis		**保護者の** –D **判断で** –D **登校を** –D **見合わせてください.**	by a guardian with an assessment of the situation your children’s attendance in school Please avoid
output	I	保護者の判断で見合わせてください．	Please avoid with an assessment of the situation by a guardian.
	II	登校を見合わせてください．	Please avoid your children’s attendance in school.

TABLE 3 - An example of Japanese dependency analysis.

[3] Morphological Analysis, ChaSen, Ver.2.3.3,
Nara Institute of Science and Technology, Computational Linguistics Laboratory
http://chasen.naist.jp/hiki/ChaSen/,

[4] Japanese dependency analysis, CaboCha
Nara Institute of Science and Technology, Taku Kudo
http://code.ge.g.,oogle.com/p/cabocha/

		Japanese	English
output	I	保護者の判断で見合わせてください．	Please avoid with an assessment of the situation by a guardian.
	II	登校を見合わせてください．	Please avoid your children’s attendance to school.

TABLE 4 - An example of output selection.

Finally, we selected the final output from these candidates and focused on postpositional words, especially with regard to particles attached with nouns for easy judgment. In addition, we established an order of priority for the particles. An example is shown in TABLE 4. In the case of example “登校を見合わせてください (Please avoid your children’s attendance to school)”, this phrase was selected as the system’s output.

4.2 Rewriting into direct expressions

The outputs, after extracting the important segments, were shorter than the original sentences. However, it was still difficult for Japanese language beginners to read them. Therefore, we rewrote 165 sentences into direct expressions that could be easily utilized by these beginners, which included pairs of official segments and segments of direct expressions similar to TABLE 1.

5 Evaluating pre-experiments

The Easy Japanese expressions were not only understandable for Japanese language beginners but also native Japanese speakers. Consequently, the outputs were evaluated by one of the authors of this project, who is a native speaker of Japanese.

5.1 Data for evaluation

We randomly extracted 20 sentences from the Easy Japanese pre-corpus and analyzed them for the extraction processes. An example is shown in TABLE 5. The method of evaluation included a two-tiered process that compared the input and output sentences.

		Japanese	English
input		手続きには，診断書はいりません．所定の用紙がありますので，該当するようなけがをした場合は，担任または顧問まですぐにお知らせください．	You don’t need a medical certificate for a processing. Please tell your homeroom teacher or an advisor about your injury with the prescribed form, which follows the rules of our school.
output	I	診断書はいりません．	You don’t need a medical certificate.
	II	所定の用紙があります．	There is a prescribed form.
	III	該当するようなけがをした場合は，	When your injury follows the rules of our school
	IV	お知らせください．	Please tell us about it.

TABLE 5 - An example of evaluation data.

First, the process included extracting important sentences (9.1), which was ineffective according to the results due to the order of priority for the particles. In this case, the particles depend upon each of the verbs. Therefore, it was necessary to consider the particles of each verb because the verbs in data alone were insufficient for obtaining the particles.

Next, the process included rewriting the sentences into direct expressions (9.2), which was also ineffective since the pairs were insufficient for obtaining a significant result. However, we found that the pairs of Japanese and Easy Japanese included many points of similarity. In future research, we will utilize existing pairs of Japanese and Easy Japanese (Manami Moku et al., 2011) or create new pairs from them.

Conclusion and perspectives

When extracting important segments, we considered that predicates included important information and particles were defined by the order of priority. However, the particles relied upon each of the verbs. We believe that our findings will be important for Japanese language beginners, and the Easy Japanese corpus will be utilized for future experiments since the corpus is smaller.

In addition, after rewriting the sentences into direct expressions, we found that the direct expressions had many similarities to Easy Japanese. Furthermore, we will use the pairs of Japanese and Easy Japanese for it.

Finally, in regard to the Easy Japanese system, the system will include three overall steps: (1) Extract important segments; (2) Create tags for representation of intention; and (3) Rewrite Japanese into Easy Japanese. Furthermore, we understand that the direct expressions include many similarities to Easy Japanese. Consequently, we will utilize data comprising pairs of Japanese and easy Japanese sentences for our project, and through the processes, we will create a system that can be used on the Internet by Japanese language beginners.

References

Chie Tsutsui. (2010). Creation of pre-corpus, *The Meeting of Society for Teaching Japanese as a Foreign Language in 2009*, The Spring Meeting in 2009, pages 86 –87

Daisuke Suzuki and Akira Utaumi. (2006). A Method for Extracting Important Segments from Documents Using Support Vector Machines: Toward Automatic Text Summarization, The Japanese Society for Artificial Intelligence, vol.21, no.4, B, pages 330–339

Isao Iori. (2008). Surround Easy Japanese, *The 4th Society to Study for Teaching Japanese as a Foreign Language in Multicultural Symbolical Society*, pages 1–12

Hideya Mino and Hideki Tanaka. (2010). Simplifying noun using Japanese dictionary in news, *The 16th Yearly Meeting of Association for Natural Language Processing*, pages 760–763

Manami Moku and Kazuhide Yamamoto. (2011). Investigation of Paraphrase of Easy Japanese in Official Documents, *The 17th Yearly Meeting of Association for Natural Language Processing*, pages 376–379

Tsutomu Hirano, Hideki Isozaki, Eisaku Maeda and Yuji Matsumoto. (2002). Extracting Important Sentences with Support Vector Machines, *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp.342–348