SYNERGY PUBLISHING PTE. LTD.

Modern Electronic Technology

http://www.advmodoncolres.sg/index.php/amor/index

Japanese-Chinese Machine Translation of Japanese Determiners Based on Templates

Ling Wang Zhongjian Wang*

Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China

ARTICLE INFO

Article history
Received: 19 March 2022
Revised: 26 March 2022
Accepted: 9 April 2022
Published Online: 16 April 2022

Keywords:
Japanese determiners
Similarity calculation
Machine translation
Translation templates

ABSTRACT

The machine translation of Japanese sentences with determiners, like "shika...nai", "tyoutto...dakedeha", "tada...dake" and so on, are more special and regular on sentences structure. The research collects and classifies the Japanese sentences which contain the determiners. The classification is carried out by according to the characteristics of Japanese sentences and translation habit of Chinese sentences. Through further abstraction and simplification, translation templates are extracted by gathering grammar rules information, studying syntax and analysis the collocation mode of sentences. Those determiners express confirmed meaning, and the corresponding translation Chinese sentences have the same characteristic. By analyzing the sentence characteristics with determiners and formalizing the sentences structure, the translation templates are abstracted. By investigating the structure characteristic of original sentences with translation templates, the similarity algorithm was defined. The threshold value of the similarity calculation was obtained by preliminary experiments, and the experiments of Japanese-Chinese translation are carried out by a small corpus. The experimental results for several kinds of Japanese sentences with determiners show the translation accuracy rate is 68.6%, template coverage rate reach 83.3%. At last, through the analysis for the translation errors, following conclusion is drawn: the results of morphological analysis are erroneous, because the error of word segmentation the part of speech tagging also are erroneous, result in the grammar structure cannot match with templates; the original sentences are long and especially complex sentences; the templates are too complicated; the similarity calculation method needs to discuss further, and so on.

1. Introduction

The research of the machine translation has a long history [1], although the great progress has been achieved, because the complexity of natural language there are still a lot of difficult problems. There are various machine translation techniques, such as rules based, examples based, statistical machine translation. Usually, the different machine translation techniques have different advantages; they are used according to the different problem domain [2]. Some machine translation methods are as follows.

Example-based machine translation methods are trained from bilingual parallel corpora, which contain sentence pairs. Sentence pairs contain sentences in one language with their translations into another language. Statistical machine translation methods are applied to generate results using bilingual corpora. There are three different models of statistical machine translation, statistical word based, statistical phrase based and statistical syntax based models.

**Corresponding Author:*
Zhongjian Wang,
Guangzhou College of Technology and Business, Foshan, Guangzhou, 510800, China;
Email: 3591256684@qq.com

 DOI: https://doi.org/10.26549/met.v6i1.11422

The other method is rule based machine translation. This method needs a large amount linguistic rules, which are building considering both the languages source language and target language. The rule based machine translation method is based on linking the structure of the given input sentence with the structure of the target output sentence, preserving their unique meaning. The method uses large collection of manually developed rules used for mapping source language into target language text. These can be edited to improve translations [3].

Paper [6] proposes a method for translating English sentences to Malayalam by rule based method. The main process is mediated by bilingual dictionaries and rules for converting source language structures into target language structures. The rules used in this approach are prepared based on the parts of speech tag and dependency information obtained from the parser. In their method, the transfer link rules are used for generating target structure. Morphological rules are used for assigning morphological features.

Any kind of translation method has advantages and shortages, the rule based machine translation are also no exception. There are also a lot of researches about the machine translation method based on templates.

Such as paper [5] proposed a method based on templates, the method is used to improve the performance of patent machine translation. They created more than 600 templates manually and integrated it to a rule-based MT system. Their evaluation experiment shows that the translation quality of 40% test text is improved.

Template-based machine translation is widely used in the machine translation of limited field or specific issue, such as automatic patent translation. Because in those translation problem, there are enormous amount of jargons, those jargons have uneven word frequency distribution and data sparseness, there are serious problems when applying statistical method to automatic template acquisition.

In this paper, we will discuss the translation of Japanese sentences, which contains determiners. Japanese sentences which contain the determiners are a kind of frequently used sentence. Those sentences have special determiners, rigorous grammar, and strong structural characteristics. So it is very suitable for template-based translation method which having particularly effective in the phrase-level alignment of parallel corpora.

2. The Translation Process of Japanese Sentences with Determiners

Template-based machine translation is a kind of intuitive representation method. It does not require massive knowledge of linguistics, and labor cost is low and easy to get translation templates by using corpus.

First of all, it needs to be explained, the determiners in this paper are "…sika…nai", "…nikagitte…", "tada…… dake", "…tekosohajimete…", "…kagiri…", "…nomida", "…nitodomaru", etc. In this paper, we deal with the sentences which express that that's all there is. Those sentences use the especially auxiliary word to express restriction of scope, estimation of quantity etc. For those sentences, we collect data, classify the sentences, and use the tool of Japanese morphological to carry out morphological analysis, and use the morphological analysis results to do grammar analysis and sentences structure analysis, then extract translation templates.

2.1 Classification of Japanese Sentences with Determiners

We collect and classify the Japanese sentences which contain the determiners. The classification is carried out by according to the characteristics of Japanese sentences and translation habit of Chinese sentences. The parts of sentences with determiners are listed as following in Table 1.

The Table 1 indicated nine kinds of sentences and each kind has its corresponding Chinese translation keyword.

The " しか…ない " is used to indicate that there is nothing else. For example " これしかない。(There's nothing but this.)".

The phrase "… ただ … だけ …" express various degree of amounts. For example, sentences like " 私はひとつだけ食べました。(I only ate one.)", " この乗車券は発売当日のみ有効です。(This boarding ticket is only valid on the date on which it was purchased.)". A particle that is essentially identical both grammatically and in meaning to " だけ " is " のみ ". " だけ " is used in regular conversations and " のみ " is usually only used in a written context.

2.2 Morphological Analysis of Japanese Sentences with Determiner

The Japanese text is same as Chinese text; there are no spaces between words. The Japanese text is comprised of three main written characters: Hiragana, Katakana, and Kanji. The Japanese morphological analysis includes following works, such as segmenting text into words, part-of-speech tagging, get dictionary forms for inflected verbs and adjectives and extracting readings for kanji. In this paper, WinCha [7] is used to analyze Japanese text. In fact, the morphological analysis results by WinCh include more information, but we only use the results of word segmentation and part of speech tagging.

DOI: https://doi.org/10.26549/met.v6i1.11422

Table 1. The sentences with determiners

No.	The sentences	The determiners	The Chinese translation
1	これは僕しか知らない話だ。	…しか…ない…(…しかない)	只有…；仅仅…
2	その日に限って、帰りが早かった。	…に限って…	唯独…；只有…；仅…；只限于……
3	彼よりほか知っている者はいない。	…よりほかはない…	只有…；只好…
4	ちょっと読んでだけでは、分かりません。	…ちょっと…だけでは…	只是…；光…
5	彼女はただ笑うだけで、答えません。	…ただ…だけ…	只是…；仅仅…
6	自分でやってこそはじめて分かる。	…てこそはじめて…	只有（唯有）…才能…
7	生命のつづくかぎり祖国のために尽くす。	…かぎり…	只要…就…；除非…就…
8	あとは返事を待つのみだ。	…のみだ	只有…；唯有…
9	単に希望を述べたにとどまる。	…にとどまる	只是…（而已）

We sum up nine kinds of the Japanese sentences as following.

- The particle “だけ” is used to express that there is all there. “ただ” is used with “だけ”, to emphasize expression meaning. “ただ” is used with “だけ”, just apple (and nothing else) to express more stronger meaning than only “だけ”.
- “しか…ない” is used to indicate that there is nothing else. The sentence express always negative. For example, “これしかない。 There is nothing but this.” “しか” has an embedded negative meaning while “だけ” doesn’t have any particular nuance.
- “に限って” are consist of “限っ” and “て”, to express specifically things and specifically scope.
- “よりほかはない” are consist of “より”, “ほか”, “は” and “ない”, to express the meaning that affirm this and negate other else.
- “ちょっと…だけでは” are consist of “ちょっと”, “だけ”, “で” and “は”, to express the meaning that some things happen in a particular situation. sometimes only “ちょっと…だけ” is used.
- “てこそはじめて” express the meaning when only a certain kind of situation occurs, the other kind of situation may appear.
- “かぎり” express the limit and range of things, structure adverbial phrase. Represents the maximum extent of competence, degree, knowledge and so on. Example, 力のかぎり戦ったのだから、負けても悔しく思いません。（Has gone all out to fight, so even if it is lost, I will not regret it.）
- “のみだ” is consist of “のみ” and “だ”. A particle that is essentially identical both grammatically and in meaning to “だけ” is “のみ”. However, unlike “だけ”, which is used in regular conversations, “のみ” is usually only used in a written context.
- “にとどまる” express the limit, the scope or the degree are limit in the scope that describe by the words before “に”. such as an example, “単に希望を述べたにとどまる。/I just expressed my hope.”

2.3 The Grammar Analysis and the Template Abstract

To get abstract form of Japanese-Chinese translation pair, morphological analysis of Japanese sentences and the Chinese target sentences are carried out meanwhile. By comparing the part of speech tagging of bilingual sentences, we extract grammar structure expressed by POS, key words and part of variables. We summarize the sentences’ grammar structure as following in Table 2. Each kind of the sentence is illustrated by part of speech tagging of corresponding words.

Through further abstraction and simplification, translation templates are extracted by gathering grammar rules information, studying syntax and analysis the collocation mode of sentences. The Japanese sentence patterns are a lot and its form of expressions are abundant. Especially, the sentences of including determiners are variety, various styles. In this paper emphasizes describes on translation of determiners, especially the determiners for list in Table 1.

According to the results of sentence morphological analysis, we summarize and formalize the grammar structure of sentences, thus extract the translation templates. Some examples are shown in Figure 1.

Example 1:

これはあの店でしか売っていない。
これ/名詞-代名詞-一般/は/助詞-係助詞/あの/連体詞/店/名詞-一般/で/助詞-格助詞-一般/しか/助詞-係助詞/
売っ/動詞-自立/て/助詞-接続助詞/い/動詞-非自立/ない/助動詞/。/記号-句点

Original sentence template: N１＋は＋N２＋で＋しか＋V＋ない
Extract translation template: N１+只有+N2+V】

Figure 1. A translation example

 DOI: https://doi.org/10.26549/met.v6i1.11422

We must consider much more factors when extract the translation templates because the words sequence of sentences is too long and consist of various words with different part of speech tagging. To get a translation template that can represent feature of sentence exactly, we take words as a basic unit to abstract the translation templates. The part of translation templates are shown in Table 2.

Table 2. The extraction of translation templates

Japanese Determiners	Chinese translation	Abstracted translation templates
しか…ない、しかない	只有...，仅有....	N+は+NP+しか…N+が+ない/N 除了 NP 之外没有 N N が A1+A2+しか+V+ない/因为+N＋A１，只能+V+A2
に限って、に限り、に限る、に限らず	唯有...；只有...；仅...；只限于...	N1+に+限って,N2+が+A/只限于+N1+N2+A N1+に+限って,N2+は+V/只限于+N1+V+N2
よりほかはない、よりほかには…ない	只有...;只好...	N１＋よりほか+VP＋N２＋は＋ない/除+N１＋之外没有 N２＋VP N１＋V１＋には＋N２＋を＋V２＋ほかはない/V１＋N1＋只有+N２
のみだ、	只有...唯有...	N１＋のみならず、N２＋も＋N/不仅+N1,而且 N2 N１＋は＋N２＋A+のみならず、N3 も＋A/N1+N2+不仅+A，而且＋N3+A
にとどまる、にとどまらず	只是...(而已)	N１＋は＋N２＋にとどまる/只限于，N１＋只限于+N２ N１＋を＋V＋にとどまる/只限于+V＋N１

3. System Implementation and Translation Experiments

We developed the system and carried out the evaluation experiments by collection sentences from Japanese textbook.

3.1 Translation Process

The translation process is shown in Figure 2.

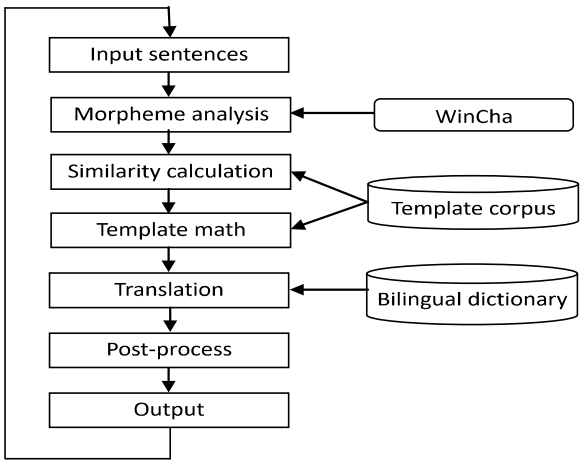

Figure 2. The translation Process

The translation process is described as follows, for the input Japanese sentences, morphological analysis is carried out. Here the WinCha is used. We only take the part of speech tagging as basic unit to formalize the grammar structure of sentence, remain the determiners and particles, got a formalizing sentence grammar structure. Then we compare it with translation templates; calculate similarity of the formalizing sentence grammar structure with translation template. When the similarity calculation value is greater or equal to the threshold value that is determined by preliminary experiments, translation template are selected. The translation is carried out by using the bilingual dictionary. Here the first entry in bilingual dictionary is used. Figure 3 is an example of translation.

3.2 Experiments and Evaluation

To evaluate the proposed method, it is necessary to carry out experiments. Through we collect 480 Japanese sentences with determiners from elementary Japanese textbook, use they as experimental data. We use 100 sentences to do preliminary experiment, to get the threshold. The similarity of source sentence with template is calculated by formula (1) as following.

$$TemSim[\%]=\begin{cases}0 & DW(T,S)=0\\ \dfrac{\alpha\cdot KW(T,S)+\beta\cdot RW(T,S)}{KW(T,S)+RW(T,S)}\times DW(T,S)\times 100 & DW(T,S)\neq 0\end{cases} \quad (1)$$

where ***KW*** is the number of part of speech tagging of the being original sentence to compare with the template; ***RW*** is the number of particle of the being original sentence to compare with the template; ***DW*** is a fixed value. When the original sentence and template contain same determiners ***DW*** is equal to 1, else 0. ***TemSim*** is the similarity of the being original sentence to match with the template. Here

DOI: https://doi.org/10.26549/met.v6i1.11422

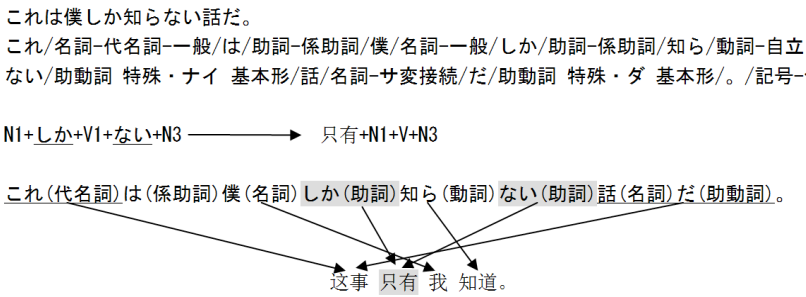

Figure 3. An example of translation

α and β are weighted parameter. The optimum coefficients are decided by the preliminary experiments using Greedy method, α=2 and β=5.

About the evaluation of experimental results, we judge the readability of translation sentence according to word order and expression meaning correctness. For each translation sentence, a score is given between 1 and 0. Here 1 is correct and 0 is error. The translation correct rate is calculated by formula (2).

$$TCR[\%] = \sum_{i=1}^{sum} \frac{S_i}{sum} \times 100 \quad (2)$$

Here TCR is the translation correct rate, Si is the score of the *i*th translation sentence, and sum is the total number of using in the evaluation experiment.

The coverage rate of translation templates (CRT) is calculated by formula (3), equal to the quotient of the template number which is used in translation experiment and the total number of translation template.

$$CRT[\%] = \frac{used\ template\ number}{the\ total\ template\ umber} \times 100 \quad (3)$$

By judging the score of each translation sentence, using formula (2) to get the TCR is 68.6%, and by formula (3) CRT is 83%. The reason that TCR is lower will be illustrated in the next section.

3.3 Analysis of Erroneous Results

We got a lower translation correct rate. Through analysis the result of translation, the summing up is as following.

1) The results of morphological analysis are erroneous. Such as Figure 6 and Figure 7. Because the error of word segmentation the part of speech tagging also are erroneous, result in the grammar structure cannot match with templates.

2) The original sentences are long and especially complex sentences. The result of morphological analysis contains various part of speech tagging, clause structure.

3) The templates are too complicated.

4) The similarity calculation method needs to discuss further.

Word boundary ambiguity cannot be ignored when dealing with Japanese sentences. Because word boundaries are not clear in Japanese, some mistakes occur when morphological analysis is carried out. For example, Figure 4 and Figure 5 show.

しばらく	シバラク	しばらく	副詞-助詞類接続	
し	シ	する	動詞-自立	サ変・スル 連用形
か	カ	か	助詞-副助詞／並立助詞／終助詞	
滞在	タイザイ	滞在	名詞-サ変接続	
でき	デキ	できる	動詞-自立	一段 未然形
ない	ナイ	ない	助動詞	特殊・ナイ 基本形
。	。	。	記号-句点	

Figure 4. A mistake of "sika" morphological analysis

表通り	オモテドオリ	表通り	名詞-一般
の	ノ	の	助詞-連体化
みか	ミカ	みか	名詞-固有名詞-人名-名
路地	ロジ	路地	名詞-一般
の	ノ	の	助詞-連体化
隅々	スミズミ	隅々	名詞-一般
まで	マデ	まで	助詞-副助詞

Figure 5. A mistake of "nomi" morphological analysis

According to error analysis of experiment result, the translation error rate that occurs owing to the mistakes of morphological analysis is 15%. The reasons of translation error are shown as Table 3.

Table 3. The ratio of the various translation error

Error reason	Word segment & POS error	Template error	Computing error of similarity	Others
Proportion [%]	15	10.5	5.8	0.7

The template matching error includes wrong template extraction and the inaccuracy of similarity calculation is 10.5% and 5.8% respectively. This shows the need for some longer sentences the proposed approach should be further research, and the similarity calculation also needs further improvement. The others error 0.7% is the transla-

DOI: https://doi.org/10.26549/met.v6i1.11422

tion errors except for the first three items listed in Table 3.

4. Conclusions

The advantage of template-based approach is that it can carry out research in the lack of resources, but the shortcoming is expenditure of manual work in the extraction of translation templates. So the method is suitable to use the translation of the Japanese sentences those have some special vocabularies and special language phenomenon. Error analysis pointed out the existing problems of the proposed approach and disposing the details, in the future study we plan for further search on extraction of template and the improvement of similarity calculation method.

References

[1] Chen, J.R., 2013. New Approach to Translation Technologies (In Chinese). Journal of Southwest Jiaotong Universit (Social Science Edition). 6(14), 109-113.

[2] Chen, Y., Zhang, P.H., Ren, L.H., 2013. A Review on Machine Translation (In Chinese). Value Engineering. (1), 174-176.

[3] Feng, Zh.W., 2010. Machine Translation: from rule based technology to statistic based technology. CTPF, China Translation Profession Forum.

[4] Rajan, R., Sivan, R., Ravindran, R., et al., 2009. Rule Based Machine Translation from English to Malayalam. Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT '09. International Conference on Date of Conference.

[5] Zhang, D.M., Liu, X.D., Ji, Y.H., 2013. Chinese-English patent machine translation based on templates(In Chinese). Application Research of Computers. 7(30), 2044-2046.

[6] Wang, Zh.J., 2012. Machine Translation of Japanese-Chinese for Conjunctive particles. Applied Mechanics and Materials.

[7] ChaSen's Wiki. http://chasen.naist.jp/hiki/ChaSen/.

DOI: https://doi.org/10.26549/met.v6i1.11422