Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT

Jin'ichi Murakami
Department of Information and Knowledge Engineering
Faculty of Engineering
4-101 koyamachou Tottori City
Tottori 680-8552, Japan
murakami@ike.tottori-u.ac.jp

Isamu Fujiwara
Department of Information and Knowledge Engineering
Faculty of Engineering
4-101 koyamachou Tottori City
Tottori 680-8552, Japan
s072046@ike.tottori-u.ac.jp

Masato Tokuhisa
Department of Information and Knowledge Engineering
Faculty of Engineering
4-101 koyamachou Tottori City
Tottori 680-8552, Japan
tokuhisa@ike.tottori-u.ac.jp

ABSTRACT

Pattern-based machine translation is a very traditional machine translation method that uses translation patterns and translation word (phrase) dictionaries. The characteristic of this translation method is that high-quality translation results can be obtained if the input sentence matches the translation pattern and this translation pattern is correct. However, translation patterns and translation word dictionaries are usually made manually. Therefore, there are many costs in making a pattern-based machine translation system.

We propose making translation patterns and translation word dictionaries automatically by using statistical machine translation methods. Using these methods, we decreased the costs in making a pattern-based machine translation system.

We demonstrate the effectiveness of the proposed method in a Japanese-English machine translation patent task (NTCIR-10). We obtained good results.

Team Name

TOTTORI

Subtasks

Japanese to English

Keywords

Pattern-Based Machine Translation, Statistical Machine Translation, Pattern-Based Statistical Machine Translation GIZA++, IBM model 1-5, Moses

1. INTRODUCTION

Recently, phrase-based statistical machine translation, which we describe as "phrase-based SMT," has been very popular. However, there are many serious problems. One is the translation quality. For Japanese-English translation, a rule-based machine translation system is better than phrase-based SMT [4].

There are about 3,000,000 Japanese-English parallel translating patents sentences[4]. Nevertheless, the performance of phrase-based SMT is lower than rule-based machine translation. Commercial machine translation systems are classed as rule-based machine translation systems. We considered that this poor performance is the fundamental problem of phrase-based SMT and especially caused by the reordering model.

There are three models for phrase-based SMT: translation, language, and reordering models. These models each have problems. The translation model is the probability of a source phrase matching a target phrase. This model is calculated by using Och's heuristic and IBM model 1-5. However, this model produces strange grammar phrases. The language model normally uses N-gram, which is very reasonable for stochastic language model. However, the N-gram model has local information and does not have global information. Also, the reliability of high order N-gram (for exmaple 5-gram) is low because there are many parameters. Therefore, an oracle number of monolingual sentences is needed. To overcome these problems, smoothing techniques like delete interpolation or Kneser-Ney are used. However, these techniques sometimes decrease the translation performance. Finally, we consider that the reordering model have the most important problems. Normally, the N-gram model has local, not global, information. To surmount this problem, the reordering model is used. However, this model is not so effective for Japanese-English translation. In our opinion, word reordering is also local, not global, information. And as a more serious problem, the word reordering may be deterministic and not statistical.

To overcome these problems with the pattern-based machine translation and statitical machine translation system, we propose a pattern-based statistical machine translation system. The conventional pattern-based machine translation is a kind of rule-based machine translation and uses translation patterns and translation word dictionaries. Translation patterns provide word order. This means that the reordering problem is no longer problem for pattern-based machine translation. Therefore, the output is grammatical and tends to be a good translation. However, this system is costly because the translation patterns and word dictionaries are made manually. On the other hand, the statistical machine translation is low in cost because it uses only source and target sentence pairs and does not have to be made manually. Using these tools, we can implement automatic pattern-based statistical machine translation. GIZA++[3] can get the source and target word pairs automatically from the source and target sentence pairs. Also, we can make Japanese-English translation patterns by using the automatically obtained word pairs.

Finally, we investigated the output sentences of the proposed method and surveyed the rule-based machine translation to make a comparison.

2. PATTERN-BASED MACHINE TRANSLATION

2.1 Outline of Pattern-Based Machine Translation

Pattern-based machine translation [7] is a very traditional machine translation method that was proposed in the 1960s. This form of machine translation uses source language and target language translation patterns and translation word (phrase) dictionaries. It has certain advantages. In pattern-based machine translation, a translation pattern provides a word order. Therefore, if the input sentence matches a translation pattern, the translated sentence will be of high quality. However, this form of machine translation has disadvantages as well. It cannot translate input sentences that do not match any of the stored translation patterns. This means that to match many sentences, we either have to make many patterns or generalize these patterns.

2.2 Japanese-English Pattern-Based Machine Translation

The conventional Japanese-English pattern-based translation method is as follows [5].

Step 1 Prepare Japanese-English translation patterns and Japanese-English word pairs.

Step 2 Input an Japanese sentence.

Step 3 Search for a Japanese translation pattern that matches the input of Step 2.

Step 4 Output an English translation pattern corresponding to the English translation pattern made in Step 3.

Step 5 Generate a English sentence by using the Japanese-English word pairs and English translation patterns.

Table 1 shows an example of Japanese-English pattern translation, and Table 2 shows examples of Japanese-English word pairs.

Table 1: Example of Japanese-English Pattern Translation

Input sentence	図 9 はそのときの特性図である。
Japanese pattern	X1 X2 はそのときの X3 図である。
English pattern	X1 X2 shows a X3 of a time.
Output sentence	Fig. 9 shows a characteristic of a time.

Table 2: Example of Japanese-English Word Dictionary

図	Fig.
9	9
特性	characteristic

2.3 Program of Pattern-Based Machine Translation

For traditional pattern-based machine translation, source language and target language translation patterns and translation word (phrase) dictionaries are made manually. Therefore, the costs are very high. Hence, the amount of research on pattern-based machine translation has declined.

3. PHRASE-BASED STATISTICAL MACHINE TRANSLATION

3.1 Outline of Statistical Machine Translation

Statistical machine translation (SMT) was proposed in the 1990s [1]. This translation method uses source and target sentence pairs and has a translation model and language model. A decoder uses these models to output a target sentence with the maximum probability.

The following is an example of Japanese-English SMT [9].

$$J = argmax_w P(e|j) \quad (1)$$
$$\simeq argmax_e P(j|e)P(e) \quad (2)$$

Here, $P(j|e)$ means the Japanese-English translation model, and $P(e)$ means the English language model. The translation model has the probabilities of Japanese words translated into English words. These probabilities are calculated from Japanese and English sentence pairs. The language model has the probabilities of target word strings.

The decoder selects the target sentence by referring to the translation model and language model probabilities. Statistical machine translation was initially word-based. Recently, it has become phrase-based.

3.2 Program of phrase-based SMT (Word reordering)

As stated in the introduction, phrase-based SMT has been very popular. However, there are many serious problems. One is the translation performance. For Japanese-English translation, the rule-based machine translation system is better than the phrase-based SMT[4]. There are about 3,000,000 Japanese-English parallel sentences used with translating patents[4]. Even when using these parallel sentences, the performance of phrase-based SMT is lower than that of a rule-based machine translation system.

We considered that this is caused by the reordering model. Normally, the N-gram model is used for the language model. However, this model has local, not global, information. To surmount this problem, a reordering model is normally used. However, this model is not so effective for Japanese-English translation. In our opinion, word reordering is also local, not global, information. There are many reordering models. For example, “msd-bidirectional-fe” is normally used. However, we think that word reordering is related to grammar, especially case grammar [2]. Therefore, we believe it is not a statistical phenomenon.

4. PROPOSED METHOD

Conventional pattern-based machine translation costs a lot because its translation patterns are made manually. In return, the output of such translation is grammatical and tends to be a good translation. In comparison, statistical machine translation is low in cost because it uses only source and target sentence pairs. However, such translation often outputs ungrammatical translations. In our opinion, this is caused by the reordering model. We believe it is not a statistical phenomenon.

To overcome the above mentioned problems, we proposed pattern-based statistical machine translation. We focused on the corresponding word pairs between the source language and the target language that can be automatically found with GIZA++. GIZA++ gets the source and target word pairs automatically from source and target sentence pairs. The Japanese-English translation patterns can then be made from these Japanese-English word pairs. The steps of the proposed method are described below.

4.1 Make the Japanese-English Word Dictionary

Translating only one way from Japanese to English will result in an unreliable dictionary. Thus, we used both Japanese-English word pairs and English-Japanese word pairs to make the dictionary.

The word dictionary was made as follows.

Step 1 Make Japanese-English word pairs and English-Japanese word pairs using GIZA++.

Step 2 Multiply the translation probabilities of the Japanese-English word pairs and English-Japanese word pairs. Select the word pairs with probabilities higher than a threshold (α) and put them in the dictionary.

GIZA++[3] gets the source language and the target language word pairs by using the maximum likelihood correspondence from the source sentences and the target sentences. It also assigns a translation probability. In this experiment, we used GIZA++ to obtain Japanese-English word pairs and English-Japanese word pairs. Examples of using GIZA++ are shown in Tables 3 and 4.

Table 3: Example of Japanese-English Word Pairs found by GIZA++

図	Fig.	0.37
特性	characteristic	0.49

Table 4: Example of English-Japanese Word Pairs found by GIZA++

FIG	図	0.22
characteristic	特性	0.71

Table 5 shows an example of the Japanese-English word dictionary.

Table 5: Example entries of the Japanese-English Word Dictionary

図	Fig.	0.08
特性	characteristic	0.34

4.2 Make the Japanese-English Translation Patterns

We made the Japanese-English translation patterns by using the Japanese-English word dictionary and Japanese-English sentence pairs. The translation patterns are created with the following steps.

Step 1 Compare each Japanese word of a Japanese-English sentence pair with an Japanese word of the Japanese-English word dictionary

Step 2 Compare an English word of the Japanese-English word dictionary with each word of the English sentence of the translation pairs

Step 3 Match up the Japanese-English word pairs and replace each pair with a variable, such as X1, X2, X3.

Step 4 Repeat steps 1 to 3 for all sentence pairs.

Figure 1 shows an example of making a Japanese-English translation pattern.

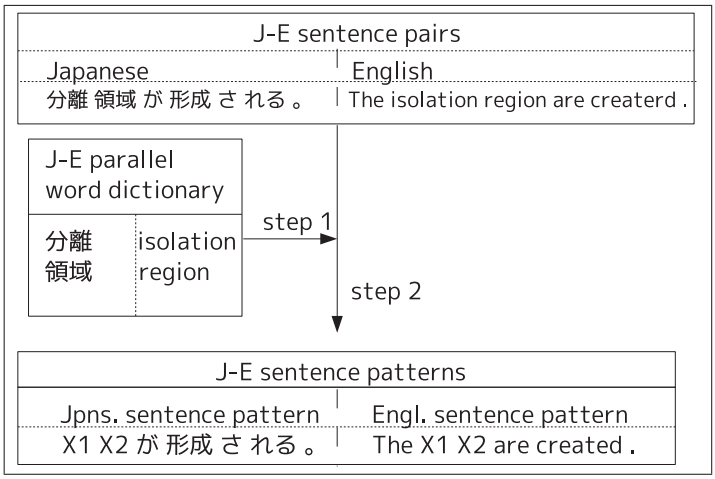

Figure 1: Making a Japanese-English Translation Pattern

4.3 Generate the English Translation Sentence

We generate English translation sentences by using the Japanese-English word dictionary and the Japanese-English translation patterns. The English translation sentences are made as follows.

Step 1 Select Japanese translation patterns corresponding to the input Japanese sentence.

Step 2 Find the variables in the Japanese translation patterns and obtain the Japanese words corresponding to the variables.

Step 3 Obtain the English translation patterns corresponding to the Japanese translation patterns.

Step 4 Find the variables in English translation patterns and search for the English words corresponding to the variables.

Step 5 Replace variables in English translation patterns with the English word (Step 4).

Step 6 If the result of step 5 is that multiple English translation sentences are output, select only one sentence by referring to a English word tri-gram. These tri-grams are calculated from the Japanese-English sentence pairs.

Figure 2 shows an example of generating a English translation sentence.

4.4 Notes

- The English word tri-gram is calculated with the base 2 logarithms.
- If the probability of the word tri-gram data is 0.0, we set -1000.0 as a penalty.
- The following cases are not outputted as English translation sentences.

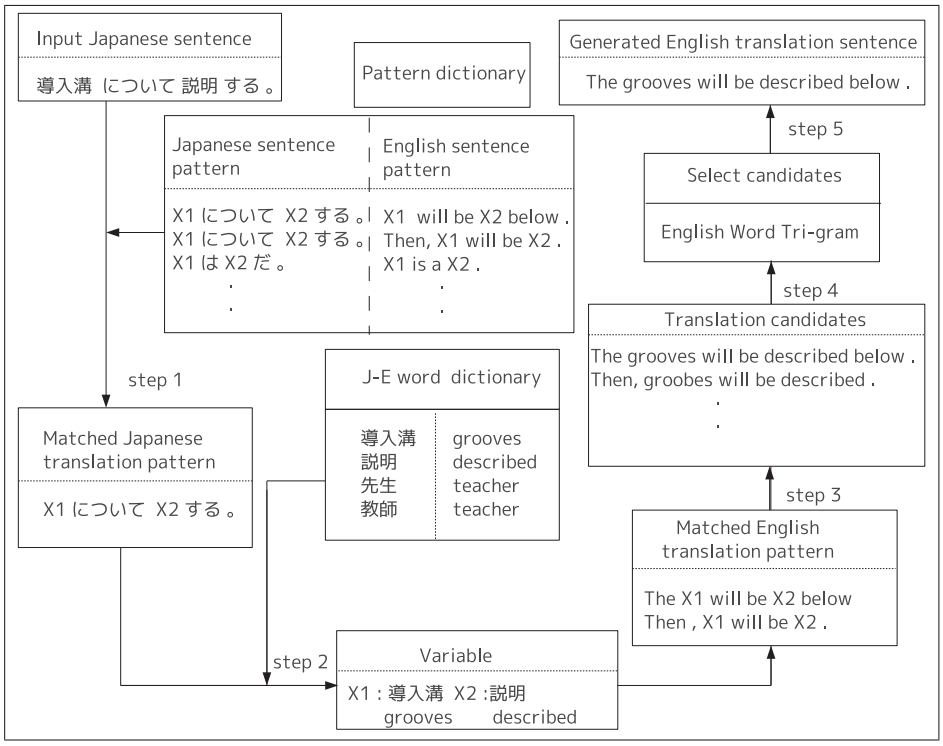

Figure 2: Generating a English Translation Sentence

- The input Japanese sentence does not match any of the Japanese translation patterns.
- For Step 2 or Step 4 of Section 4.3, the Japanese word or the English word could not be found in the Japanese-English word dictionary.

5. EXPERIMENTAL CONDITIONS

1. Database

   We used NTCIR-10 Japanese-English patent sentence pairs. We used 3,000,000 Japanese-English patent sentence pairs for the training. We used Mecab [8] as the morphological analyzer and the standing tokenizer of Moses [6].

2. Threshold

   (a) Making Japanese-English patterns

   We used a word dictionary with α = 0.1 to make the Japanese-English translation patterns. (α is used in Step 2 of Section 4.1.) As a result, we obtained 31,843 Japanese-English word pairs (word dictionary) and 3,158,406 Japanese-English translation patterns from the training.

   (b) Generating Japanese translation sentences

   We used a word dictionary with α = 0.01 to generate the Japanese translation sentences. As a result, we obtained 125,194 Japanese-English word pairs (word dictionary).

3. Tri-gram Data

   We used about 3,000,000 Japanese-English sentence pairs to calculate the English word tri-gram for the language model.

4. Phrase-Based Statistical Machine Translation (Moses)

   We used Moses [6] as the phrase-based SMT for comparison.

5. Rule-Based Machine Translation System

   For comparison, we used the art of trial rule-based machine translation system as a rule-base machine translation.

6. EXPERIMENTAL RESULTS

We classified the output English translation sentences into four types (A rank - D rank). We used the sum of English word tri-gram scores (β) as a classifier. The four value types are shown below.

Table 6: Four Value Types for the Sum of English Word Tri-gram Scores

A rank	$-1000.0 < \beta \leq 0.0$
B rank	$-2000.0 < \beta \leq -1000.0$
C rank	$-3000.0 < \beta \leq -2000.0$
D rank	$-3000.0 \leq \beta$

β: sum of English word tri-gram scores

6.1 Example of Patent Task

Here, we show the patent translation results. In Table 7 - Table 10, "input" means the input Japanese sentence. "Japanese pattern" means the Japanese translation pattern matching the input. "English pattern" means English translation patterns corresponding to a Japanese pattern. "Proposed" means the translation sentence obtained by the proposed method. "β" means the sum of

English word tri-gram scores. ”Reference” means a correct sentence. ”Moses” is the output of phrase-based SMT and means Moses. ”RBMT” is the output of rule-based machine translation.

6.1.1 Example of A-rank translation

We obtained 22 sentences in the A rank. The A-rank results were excellent translation results. The quality of these results was comparable with human results. An example of the A-rank translations is shown below.

Table 7: A-rank Example

Input	これにより、分離領域４６０が形成される。
Japanese pattern	これにより N00 N01 N02 N03 N04 N05 が形成される。
English pattern	As a result N00 the N01 N02 N03 N04 N05 is created.
Proposed	As a result, the isolation region 460 is created. ($\beta = -58.7402$)
Reference	By this, the isolation region 460 is formed.
Moses	As a result, the isolation region 460 is formed.
RBMT	Thereby, the separation domain 460 is formed.

In Table 7, the results of the proposed method were the same quality as the reference sentence. This shows that the proposed method worked correctly in this case.

6.1.2 Example of B-rank translation

We obtained 13 sentences in the B rank. The B ranks had the same as the A rank. We show an example below.

Table 8: B-rank Example

Input	図１３は消去動作時におけるメモリセルアレイ１０の回路図である。
Japanese pattern	N00 N01 N02 N03 N04 N05 N06 N07 N08 N09 N10 N11 N12 N13 N14 図である。
English pattern	N00 N01 N02 a N14 diagram N06 a N07 N08 N09 N05 N11 N12 in N13 N03 N04 N10.
Proposed	Fig. 13 is a circuit diagram of erasing operation time in memory cell array 10. ($\beta = -1065.68$)
Reference	Fig. 13 is a circuit diagram of the memory cell array 10 in an erase operation.
Moses	Fig. 13 is a circuit diagram of the memory cell array 10 in an erase operation.
RBMT	Fig. 13 is a circuit diagram of the memory cell array 10 at the time of elimination operation.

In Table 8, the result of the proposed method resembles that of the reference sentence.

6.1.3 C-rank translation

We obtained 8 sentences in the C rank. Some of the C-rank translations were better, and some were worse than the baselines. Below is an example.

Table 9: C-rank Example

Input	スロットルボディ３３は、空気フィルタ３２の下流側の吸気管３１に配置される。
Japanese pattern	N00 N01 N02 N03 N04 N05 N06 N07 N08 N09 N10 N11 N12 N13 N14 N15 N16 N17 に N18 される。
English pattern	An N07 N01 N06 N13 N00 signal N02 N03 N04 N05 N08 N18 to the N14 N11 N12 N17 of the N09 N10 N15 N16.
Proposed	The throttle body 33 is a layout to side tube 31 intake the air filter 32 the location. ($\beta = -2128.98$)
Reference	The throttle body 33 is disposed in the intake pipe 31 on the downstream side of the air filter 32.
Moses	The throttle body 33 is arranged in the intake pipe 31 downstream of the air filter 32.
RBMT	The throttle body 33 is arranged at the induction pipe 31 by the side of the lower stream of the air filter 32.

In Table 9, the result of the proposed method resembled that of the reference sentence.

6.1.4 Example of D-rank translation

We obtained 89 sentences in the D rank. The D-rank translations were not always excellent.

1. Example of D-rank translation

Below is an example of a D-rank translation.

Table 10: Example of D-rank translation

Input	断熱材６５を取付けるか、取付けないかは任意である。
Japanese pattern	N00 N01 N02 N03 N04 N05 N06 N07 N08 N09 N10 N11 N12 である。
English pattern	The N06 N05 the N00 N01 N02 N03 N04 N07 N08 N09 N10 N11 N12.
Proposed	The one attaching the insulation member 65 to and installed there it is desired. ($\beta = -4079.35$)
Reference	Whether to mount the heat insulation material 65 is arbitrary.
Moses	Whether or not a heat insulating material 65 is mounted to the mounting is arbitrary.
RBMT	It is arbitrary whether the thermal insulation 65 is attached or it is not attached.

In Table 10, the proposed translation method was unsuitable.

6.2 Human Evaluation Results

There are many human evaluation methods. Amongst them, we chose the ABX test for reliability. We carried out the ABX test on the proposed method and rule-based machine translation. This involves a count of the sentences by using the following criteria.

- Proposed ◯: The proposed translation method was better than the rule-based translation method.
- RBMT ◯: The proposed translation method was worse than the rule-based translation method.
- No difference: There was no difference in translation quality between the proposed method and the rule-based translation method.
- Same: Both outputs were completely the same.

6.2.1 Human Evaluation

We survayed all sentences for A-rank, B-rank, and C-rank. Also we selected 20 sentences at random for D-rank. The results of the ABX evaluation are listed in Table 11.

Table 11: Results of Human Evaluation

Rank	Proposed ◯	RBMT ◯	No difference	Same
A rank	8	5	10	0
B rank	3	2	8	0
C rank	2	4	2	0
D rank	1	10	9	0

Table 11 indicates that the proposed method was superior to the rule-based translation for the A and B ranks. In comparison, its results were split for the C rank, and it was inferior to the rule-based translation for the D rank. This shows the effectiveness of the proposed method for the A and B ranks.

7. DISCUSSION

7.1 With Rule-Based Machine Translation

In our experiments, the translation performance of A and B ranks had excellent quality. The quality of these ranks was comparable with that of the human translation results. However, there were many non-translated sentences that did not match translation patterns. Also, the results of D ranks were not so good. So we used a trial rule-based machine translation to translate non-translated sentences and D-rank sentences. Finally we submitted these results to an NTCIR-10 organizer for our translation results(TORRI).

According to the NTCIR-10 organizer, the results of our system were as follows. Our system was in 5th place amongst 19 systems for the adequacy score, in 2nd place amongst 9 systems for the acceptability score, in 8th place amongst 30 systems for RIBES, and in 22nd place amongst 30 systems for BLEU.

This means that our system was good for human evaluation. However, the results of the automatic evaluation were not so good.

7.2 Examination of Word-based Statistical Machine Translation Decoder

The first generation of the statistical machine translation was word-based, and its performance was low. More recently, phrase-based statistical machine translation has gotten better results.

In the proposed method, we think that the Japanese-English word dictionary and Japanese-English translation patterns were equivalent to the translation model of SMT. We also think that the word tri-gram was equivalent to the language model of SMT. Consequently, we think that the proposed method was equivalent to a word-based SMT decoder.

7.3 Increasing the Number of A-rank Translations

In this experiment, we obtained only 22 sentences in the A rank out of 2300 sentences. This means that we must increase the number of A-rank translations. Moreover, the proposed method was word-based in order to make translation patterns. In the future, we will make a program that is phrase-based.

8. CONCLUSION

The characteristic of the pattern-based machine translation method is that high-quality translation results can be obtained if the input sentence matches the translation pattern. However, translation patterns are made by hand in pattern-based machine translation. In comparison, phrase-based SMT is very popular and low in cost. However, the rule-based machine translation system is better than the phrase-based statistical machine for Japanese-English translation. In our opinion, this is caused by the reordering model, and we believe it is not a statistical phenomenon.

To overcome these problems, we proposed pattern-based statistical machine translations. In such translation, the reordering problem is no longer a problem because translation patterns are used. By using phrase-based SMT tools, we can implement this translation method automatically. In the NTCIR-10 patent task, we obtained high-quality translation sentences under certain conditions. The proposed method was especially effective in the human evaluation in the A-rank and B-rank classification.

In the future, we will make translation patterns with phrase-based Japanese-English word pairs in order to increase the number of translations in the A rank.

9. REFERENCES

[1] P. F.Brown, S. A. Pietra, V. J. Pietra, and R. L.Mercer, editors. *The mathematics of statistical machine translation: Parameter Estimation*. Proceedings of the ACL 2007 Demo and Poster Sessions, 263-311, 1993.

[2] Fillmore and C. J, editors. *The Case for Case*. In Bach and Harms (Ed.): Universals in Linguistic Theory, 1968.

[3] GIZA++, editor. *Training of statistical translation models*, http://www.fjoch.com/GIZA++, .

[4] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou, editors. *Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop*. Proceedings of NTCIR-9 Workshop Meeting, 559-578, 2012.

[5] S. Ikehara, M. Saraki, M. Miyazaki, N. Ikeda, Y. Nitta, S. Shirai, and K. Shibata, editors. *Analogical Mapping Method for MT based on Semantic Typology*. EiC, 7-12, 2002.

[6] P. Koehn, M. Federico, B. Cowan, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, editors. *Moses: Open Source Toolkit for Statistical MachineTranslation*. Proceedings of the ACL 2007 Demo and Poster Sessions, 177-180, 2007.

[7] H. Maruyama, editor. *Pattern-Based Translation: Context-Free Transducer and Its Applications to Practical NLP*. In Proc. of Natural Language Pacific Rim Symposium, 232-237, 12 1993.

[8] Mecab, editor. *Japanese morphological analyzer*, http://mecab.sourceforge.net/, .

[9] R. Zens, F. J. Och, and H. Ney, editors. *Phrase-based Statistical Machine Traslation*. KI, 35-56, 2002.