Machine translation system for patent documents combining rule-based translation and statistical post-editing applied to the NTCIR-10 PatentMT Task

Terumasa EHARA
Yamanashi Eiwa College

ABSTRACT
In this article, we describe system architecture, preparation of training data and discussion on experimental results of the EIWA group in the NTCIR-10 Patent Translation Task. Our system is combining rule-based machine translation and statistical post-editing. The thing about our new system compared with NTCIR-9 PatentMT task is to implement automatic selecting method from multiple translations: rule-based MT output and statistical post-editing output.

Categories and Subject Descriptors
I.2.7 [**Natural Language Processing**]: Machine translation

General Terms
Experimentation

Keywords
Patent translation, Machine translation, Hybrid system, Rule-based machine translation, Statistical post-editing, Automatic selecting from multiple translations, Japanese to English, English to Japanese, Chinese to English

Team name
EIWA

Subtasks/Languages
JE subtask / Japanese to English
EJ subtask / English to Japanese
CE subtask / Chinese to English

External Resources Used
Two rule-based commercial machine translation systems (J from/to E and C to E), Srilm ver.1.5.5, Giza-pp v.1.0.3, Moses Rev. 4343

1. INTRODUCTION
One of the architectures of combining rule-based technique and statistical technique in the machine translation field is combining rule-based machine translation (RBMT) with statistical post-editing (SPE) [1] [2] [3].
This architecture can use both advantages of rule-based method and statistical method. The former advantage is to use sophisticated translation rules accumulated in a long history of the machine translation technology. The latter advantage is to use powerful computational power and data power. These advantages may give a good effect for a hybrid translation, especially between structurally different languages like Japanese and English.
However, NTCIR-9 JE subtask results showed that the simple RBMT system (RBMT1) exceeded our hybrid system (EIWA) [4]. We compared adequacy and acceptability scores of RBMT1 and EIWA shown in Table 1. Sign test result for adequacy shows that RBMT1 is significantly higher than EIWA with 5% significance level. On the other hand, although acceptability of RBMT1 is higher than EIWA, but the sign test result shows that it is not significant with 5% level.

Table 1. Comparison of human judgment results of RBMT1 and EIWA in NTCIR-9 JE subtask

(a) Adequacy

Won system	Counts
RBMT1	97
Tie	134
EIWA	69

(b) Acceptability

Won system	Counts
RBMT1	86
Tie	147
EIWA	67

If we can select better output from RBMT and SPE outputs, we can make a more accurate system. To implement this idea, we must make an automatic translation selecting method from multiple translations.

2. AUTOMATIC SELECTING METHOD FROM MULTIPLE TRANSLATIONS
Several methods are proposed to automatically select a better translation from multiple translations [5] [6] [7]. We use an inverse translation method [8] for this task.

The procedure of our method is described in Figure 1. Source sentence is machine translated by several translation systems 1 to n. The n translated sentences are inversely translated to the source language expressions by a backward machine translation system. These inversely translated sentences 1 to n are compared with the original source sentence and the system calculates evaluation scores for each inversely translated sentences. The system selects translated sentence i as an output, where i is the sentence number that inversely translated sentence i has the best score.

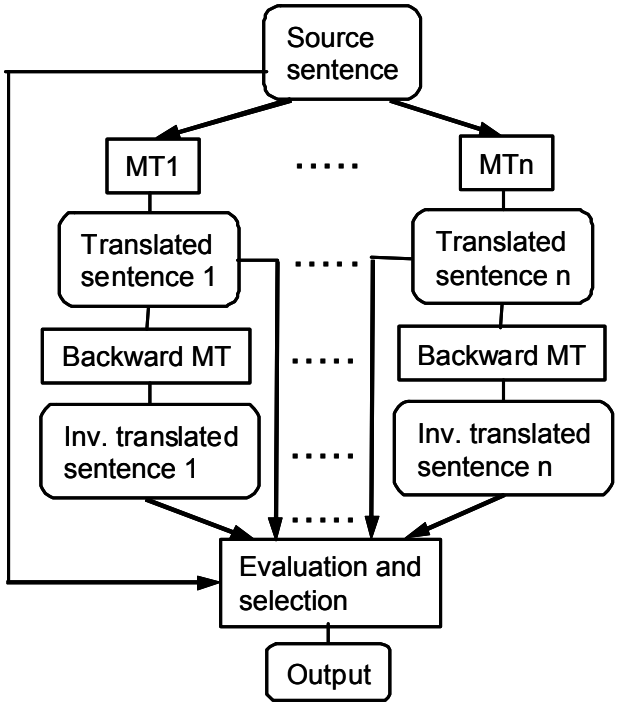


Figure 1. Translation selection method using inverse translation

In this evaluation part, we use evaluation criterion IMPACT [9] which has high correlation with human judgment [10]. We made a preliminary experiment comparing three criteria: sentence level BLUE, RIBES and IMPACT using NTCIR-9's data. From the result, we select IMPACT as the best evaluation criterion. Using this method, our translation system's architecture is described in Figure 2. Training part and bottom half of translation part is same as described in the previous paper [2]. The new part is to add the "evaluation and translation selection" phase to evaluate RBMT output and SPE output with a source sentence using backward MT and IMPACT score calculation tool. Here, backward MT tools are rule-based commercial MT systems from the venders that are same venders providing the forward RBMT systems.

There are two problems in this evaluation method. Firstly, if a translated sentence includes some source words, an inversely translated sentence also includes these source words. Then IMPACT score for the inversely translated sentence is rather high. So, if the case of including source word in a RBMT output, the system forces to select a SPE output. Secondly, an inversely translated sentence of a RBMT output has a tendency to have a high IMPACT score, because a SPE output is obtained by two step translations, while a RBMT output is obtained by one step translation. So we use "bonus score" for SPE outputs. Only the case of the following condition is satisfied, a RBMT output is selected as the system output[1].

$$impact(rbmt) > impact(spe) + bonus$$

where $impact(rbmt)$ means IMPACT score of a RBMT output and $impact(spe)$ means IMPACT score of a SPE output. Bonus scores for each subtask are chosen as Table 2 by the preliminary experiments using the NTCIR-9 results. In EJ case, we don't use any RBMT output, because they are rather worse compared with SPE output. It means the bonus score for EJ is infinite.

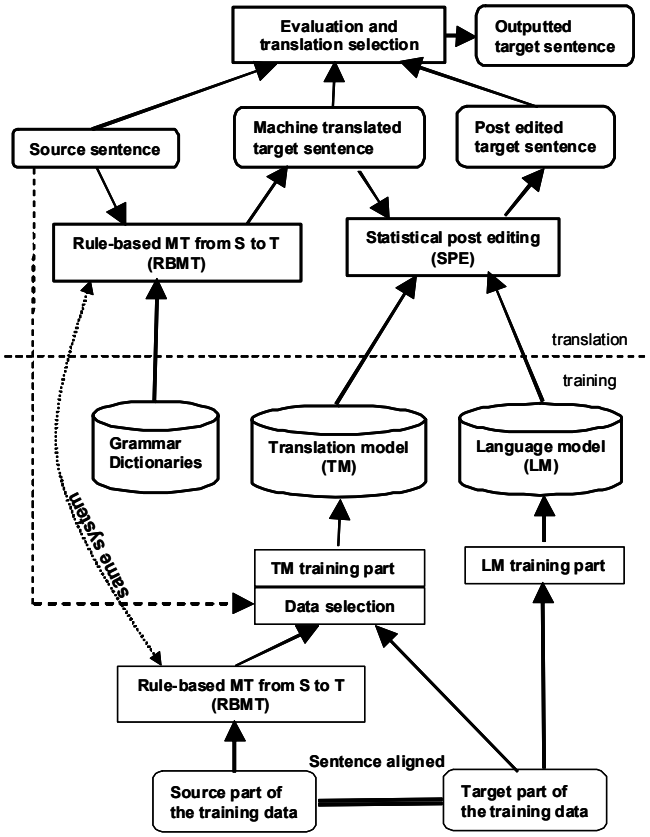


Figure 2. EIWA's translation system architecture

Table 2. Bonus scores for SPE output

Subtask	Bonus
JE	0.1
EJ	∞
CE	0.2

We tested our method using NTCIR-9's data. The result of JE subtask is shown in Table 3 and Figure 3.

Both adequacy and acceptability scores of our method are higher than RBMT1. But, Sign test shows that these differences are not significant with 5% level.

[1] We use bonus score as additive constant. Other calculating method such as scaling factor can be considered, but we don't try them. It is an issue of the future It would be better than MT1.investigation.

Table 3. Comparison results of RBMT1 and our method using human judgment scores of NTCIR-9 JE subtask

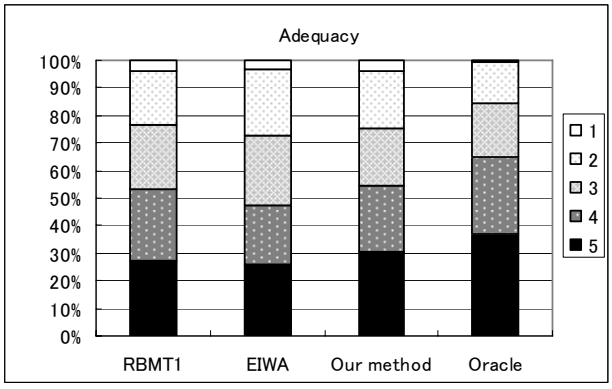
(a) Adequacy

Won system	Counts
RBMT1	34
Tie	228
Our method	38

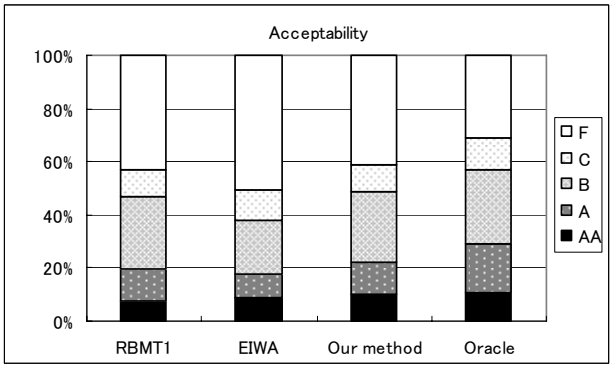
(b) Acceptability

Won system	Counts
RBMT1	26
Tie	235
Our method	39

In Figure 3, "oracle" means the method which can perfectly select a better translation from RBMT1 and EIWA outputs. Our method has 3.563 for average adequacy score comparing 3.530 of RBMT1 and 3.430 of EIWA. Our method has 59% coverage as C or higher rank in acceptability score comparing 57% of RBMT1 and 49% of EIWA. If we can use the oracle system, average adequacy score will reach 3.853 and coverage as C or higher rank in acceptability score will reach 69%.



(a) Adequacy



(b) Acceptability

Figure 3. Test results using NTCIR-9's JE subtask data

3. TRAINING, DEVELOPMENT AND TEST DATA

Training, development and test data used in our experiments are provided by the NTCIR-10 Patent Translation Task organizer [11]. Our system does not use all of the training data to make a translation model for SPE. We only use data which are fitted to the test data. This data selection method for CE and JE subtask is same as described in the previous paper [2]. For EJ subtask we use English stop word list to delete non key words in key word extraction phase. This stop word list includes 31 words.

As the results, we get the training data size for translation model training shown in Table 4[2].

Table 4. Training data size for translation model training

Subtask	Phase and Eval.	Test/dev sentences	Training sentences
JE	Development	2,000	253,333
	Test (IE PEE)	2,543	357,443
EJ	Development	2,000	181,000
	Test (IE)	2,300	205,460
	Test ChE	2,000	183,663
CE	Development	2,000	115,528
	Test (IE)	2,300	99,732
	Test (ME PEE)	2,282	126,321

4. TEST RESULTS

Human judgment results for EIWA's output are summarized in Table5. Here "Accept." means the rate of C or higher ranked in the acceptability judgment. In JE subtask our adequacy score is lower than RBMT1's score: 3.57 [11].

Table 5. Summary of the EIWA's experimental results

Subtask	Adequacy	Accept.
JE	3.53	0.44
EJ	3.42	0.59
CE	2.80	---

Results of translation selection method for JE and CE subtask is given in Table 6. For JE subtask, SPE outputs covers 87% of total 300 system outputs and RBMT outputs covers 13%. Adequacy and Accept. scores when the system selects RBMT output are higher than the case when SPE output is selected. So our strategy is effective in the JE subtask.

For CE subtask, only 6% of total system outputs arc come from RBMT outputs. Adequacy of RBMT is lower than SPE's. So translation selection does not give a high score in the CE subtask.

Table 6. Translation selection results

Subtask	Coverage		Adequacy		Accept.	
	SPE	RBMT	SPE	RBMT	SPE	RBMT
JE	0.870	0.130	3.487	3.821	0.421	0.564
CE	0.940	0.060	2.798	2.778	---	---

[2] For ChE of JE and CE subtask, we use RBMT and SPE outputs of NTCIR-9. So we need not make any training data for these evaluations.

5. DISCUSSION

Here, we compare RBMT1's results and our results for JE subtask. Table 7 shows the comparison of adequacy scores of RBMT1 and EIWA. EIWA's total results are divided into two cases. These are SPE output case and RBMT output case. Adequacy scores of RBMT1 are higher than EIWA's. However, sign test shows these differences are not significant with 5% level. Considering NTCIR-9's result that adequacy score of RBMT1 is significantly high compared with EIWA's score, our translation selection method is somewhat effective.

Table 7. Comparison of adequacy scores of RBMT1 and EIWA for JE subtask

Won system	Total	SPE	RBMT
RBMT1	62	56	6
Tie	182	151	31
EIWA	56	54	2

Why all of the RBMT outputs are not tied in the Table 7 ? The reason is that RBMT1 system of NTCIR-10 may be a new version of our RBMT system. Actually, there are several differences between RBMT1 results and our RBMT outputs. An example is as follows:

Id: 20041019_2004304769=20051014_11249308-10158
Scr.: また、作用角調整機構５３は上記バルブ特性調整機構を構成している。
Ref.: The operational angle adjustment mechanism 53 constitutes the valve characteristic adjustment mechanism.
RBMT1: Working-angle adjustment mechanism 53 constitutes the above-mentioned valve characteristic adjustment mechanism.
Adequacy of RBMT1: 5
EIWA (RBMT output case): Angle-of-action adjustment mechanism 53 constitutes the above-mentioned valve characteristic adjustment mechanism.
Adequacy of EIWA: 4

The task organizer does not provide a rule based translation results for CE subtask. Then we can't compare our results with rule based system's results in CE subtask.

For EJ subtask, we see RBMT4 results are similar to our RBMT outputs. However, any human judgment score for RBMT4 are not provided. Then we can't compare RBMT4 result and our result.

6. CONCLUSION

System architecture, preparation of training data and discussion on experimental results of the EIWA group is described. Our basic idea is to combine rule-based MT (RBMT) and statistical post editing (SPE). The new thing in this experiment is to add automatic translation selection from RBMT output and SPE output. Using JE subtask data, we can conclude our new method is somewhat effective.

One of the main remaining issues with our system is to improve the parsing accuracy in the RBMT part. Syntactically collapsed outputs from the RBMT part can't be recovered by our SPE part.

7. REFERENCES

[1] Ehara, Terumasa 2010. Machine translation for patent documents combining rule-based translation and statistical post-editing. *Proceedings of NTCIR-8 Workshop Meeting* (June 2010), 384-386.

[2] Ehara, Terumasa 2011. Machine translation system for patent documents combining rule-based translation and statistical post-editing applied to the PatentMT Task. *Proceedings of NTCIR-9 Workshop Meeting* (Dec. 2011), 623-628.

[3] Ehara, Terumasa 2005. Automatic selecting system of best translation from multiple translations using re-translation. *Japio Year Book* (Nov. 2005), 254-257, (in Japanese).

[4] Goto, Isao; Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou 2011. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. *Proceedings of NTCIR-9 Workshop Meeting* (Dec. 2011), 559-578.

[5] Akiba, Yasuhiro; Taro Watanabe and Eiichiro Sumita 2002. Using language and translation models to select the best among outputs from multiple MT systems. *Proceedings of the 19th international conference on Computational linguistics* (2002).

[6] Ehara, Terumasa 2011. Japanese to English machine translation system combining rule-based machine translation and statistical post editing (3). *2010 annual report of AAMT/Japio Special Interest Group on Patent Translation* (Mar. 2011), (in Japanese).

[7] Suzuki, Hirokazu 2011. Automatic Post-Editing based on SMT and its selective application by Sentence-Level Automatic Quality Evaluation. *Proceedings of the Machine Translation Summit XIII* (2011), 156-163.

[8] Yokoyama, Shoichi; Akira Kumano, Masaki Matsudaira, Yoshiko Shirokizawa, Mutsumi Kawagoe, Shuji Kodama, Hideki Kashioka, Terumasa Ehara, Shinichiro Miyazawa and Yasuo Nakajima 1999. Quantitative evaluation of machine translation using two-way MT. *Proceedings of the Machine Translation Summit VII* (1999), 568-573.

[9] Echizen-ya, Hiroshi and Kenji Araki 2010. Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (2010), 108-117.

[10] Echizen-ya, Hiroshi; Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro and Noriko Kando 2009. Meta-evaluation of automatic evaluation methods for machine translation using patent translation data in NTCIR-7. *Proceedings of the Third Workshop on Patent Translation, MT Summit XII* (2009), 9-16.

[11] Goto, Isao; Ka Po Chow, Bin Lu, Eiichiro Sumita and Benjamin K. Tsou 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. *Proceedings of NTCIR-10 Workshop Meeting* (June 2013).