Machine translation for patent documents combining rule-based translation and statistical post-editing

Terumasa EHARA
Yamanashi Eiwa College

ABSTRACT

In this article, we describe system architecture, training data preparation and experimental results of the EIWA group in the NTCIR-8 Patent Translation Task. Our system is combining rule-based machine translation technique and statistical post-editing technique. Experimental results show 0.344 BLEU score for Japanese to English intrinsic evaluation in the Patent Translation Task.

Categories and Subject Descriptors

[**Natural Language Processing**]: Machine translation

General Terms

Experimentation

Keywords

Patent translation, Machine translation, Hybrid system, Rule-based machine translation, Statistical post-editing

1. INTRODUCTION

One of the architectures of combining rule-based and statistical techniques in machine translation systems is combining the rule-based translation and the statistical post-editing [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. The latter advantage is to use powerful computational power and data power. These advantages may give the good effect for the translation, especially between structurally different languages like Japanese and English.

Recently, more heavy combination of rule-based and statistical techniques is proposed. However, we adopt the light combination because of the easiness of system construction.

2. SYSTEM ARCHITECTURE

Our system architecture is shown in Figure 1. The system consists of two parts: RBMT part and SPE part.

The RBMT part translates a Japanese patent document to an English document using rule-based machine translation. We use a commercial-based translation software for the RBMT part. This software focuses on patent translation.

The SPE part automatically post-edits the output of the RBMT part to more accurate English document. We use the Moses 2007-05-29 version for the SPE part. SPE part needs to include translation model and language model. They are trained from unexamined Japanese patent applications and corresponding U.S. patent grant data. Needless to say, the former data is machine translated by the same software in the RBMT part before they are used in the translation model training. The distortion limit value for the decoding is set to 6 (default value).

3. TRAINING, DEVELOPMENT AND TEST DATA

Training, development and test data used in our experiments are provided from NTCIR-8 Patent Translation Task organizer for the purpose of intrinsic evaluation from Japanese to English translation [4]. Test data include 1,251 Japanese sentences. Development data include 2,000 Japanese and English sentence pairs. Training data consist of two parts. One is the old training data for NTCIR-7 task and it includes 1,798,571 Japanese and English sentence pairs. The other is a new training data and it includes 1,387,713 Japanese and English sentence pairs.

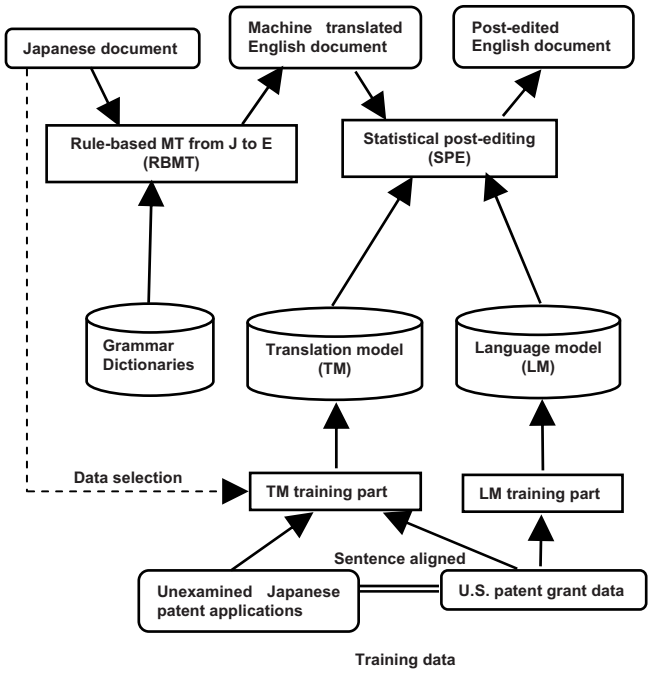

Figure 1. System architecture

We use English part of the new training data for the language model training. Srilm ver.1.5.5 is used for the language model training.

For the translation model construction, we select 152,072 sentence pairs from both the old and new training data. The detail of this selection method is described in the next section. Japanese part of this selected data is translated to English using the rule-based machine translation system. The outputted English

sentences from the RBMT system and corresponding GOOD English sentences in the training data are used as the translation model training. We use Giza-pp v.1.0.1 for the translation model training.

4. TRANSLATION MODEL TRAINING

Our translation model uses 152,072 Japanese English sentence pairs selected from the total training data including 3,186,284 sentence pairs. The idea for this selection method is to pick up the sentences adapted to input sentences. Our system does not, then, work in real time, because the training and translating phases must be done at the same time. Any way, construction method of the translation model training is as follows:

(a) Key word extraction: Key words are extracted from test sentences and Japanese part of the training sentences. In this phase, we use Japanese morphological analyzer, ChaSen and extract the words including Katakana or Kanji as the keywords. The mean number of keywords for one test sentence is 12.6.

(b) Training data selection: For all test sentences, comparing keyword set of the test sentence and keyword sets of the training sentences, we select similar training sentences to the test sentence. In this process, up to the top ten training sentences are selected for each keyword of one test sentence. Then the number of training sentences for one test sentence is up to ten times of the number of keywords of such test sentence. The mean number of training sentences for one test sentence is 125. We use the following similarity measure:

$$sim = \frac{2\times\#(T \cap S)}{\#(T)+\#(S)}$$

where T is a keyword set of the test sentence and S is a keyword set of the training sentence and #(A) means the number of the elements of the set A.

5. EXAMPLE OF THE TRAINING DATA SELECTION

The second test sentence of the data is:

このような構成になる弾性糸のクランプカッター装置１において、弾性糸ＳＹを把持して切断する動作手順を、図１Ａ、図１Ｂ、図１Ｃに示してある。

Key words extracted from this test sentence are:

構成, 弾性, 糸, ク, ランプ, カッター, 装置, 把持, 切断, 動作, 手順, 図, 示し

Selected training data for this sentence consists of 73 Japanese and English sentence pairs. The similarity values are spreading from 0.32 to 0.13. The Japanese parts of the top three training data are:

図２に於いて、切断ドラム３８にはカッター４６が取り付けられる。
まず、図１によってクランプ装置の全体構成を説明する。
次にこのように構成した装置の動作を図１２、１３、１４、１５に示したフローチャートに基づいて説明する。

And corresponding English parts are:

In FIG . 2 , a cutter 46 is attached to the cutting drum 38 .
Firstly , with reference to FIG . 1 , a whole constitution of a clamping apparatus will be explained hereinafter .
Next , the operation of a device constructed in this way is explained using the flowcharts shown in FIGS . 12 , 13 , 14 and 15 .

The translations of Japanese parts by the RBMT are:

Cutter 46 is attached to cutting drum 38 in Drawing 2 .
First , referring to Drawing 1 , the whole clamp device composition is described .
Next , operation of the device constituted in this way is explained based on the flow chart shown in Drawings 12 , 13 , 14 , and 15 .

6. TEST RESULT

The outputs of the system (spe) and the outputs of the RBMT part (rmt) for some test sentences are shown in the Table 2 with Japanese source sentences (src) and English reference translations (ref).

For the second test sentence described at the section 4.1, the modified BLEU score for the RBMT output is 0.2656 and the modified BLEU score for the SPE output is 0.5076. Here, BLEU is modified to perform the sentence level evaluation[1].

The overall modified BLEU and NIST score for the test data using single reference are shown in the Table 1. The official evaluation results by the organizer are shown in [4].

Table 1. Over all BLEU and NIST score for the test data

	BLEU (modified)	NIST
RBMT output	0.1907	6.1466
SPE output	0.3444	7.7538

7. CONCLUSION

Adding statistical post-editing part to rule-based machine translation, we can improve BLEU score from 0.1907 to 0.3444. Mean PER (position-independent word error rate) value for the RBMT outputs compared to the SPE outputs is 0.280. Then we recognize that considerable number of re-writing is performed in the SPE part.

Main remaining issue of the system is to improve the parsing accuracy in the RBMT part. Syntactically collapsed outputs from the RBMT part can't be recovered by the SPE part.

[1] Original BLEU is calculated from the n-gram matches between test sentence and reference sentence(s) from n=1 to fixed N. Usually, N=4 is used. Modified BLEU uses variable N. N is determined as the maximum number of n where n-gram match is not empty and N is up to 4.

8. REFERENCES

[1] Ehara, T. 2005. Extraction of translation knowledge from comparing of rule-based machine translation result and human translation result. Japio Year Book (Oct. 2005), 172-175, (in Japanese).

[2] Ehara, T. 2006. Japanese to English machine translation system for patent documents combining rule-based machine translation and statistical post-editing. Japio Year Book (Nov. 2006), 184-187, (in Japanese).

[3] Ehara, T. 2008. Improving the translation accuracy using phrase-based statistical post-editing. Japio Year Book (Nov. 2008), 262-265, (in Japanese). http://www.japio.or.jp/00yearbook/files/2008book/08_5_04.pdf

[4] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya, Sayori Shimohata. 2010. Overview of the Patent Translation Task at the NTCIR-8 Workshop. Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2010.

Table 2. Translation Examples

src	この第２のスライドブロック５のスライド移動によって、弾性糸ＳＹが、第１のスライドブロック４の下流側面と前記第２のスライドブロック５の上流側面との間で、確実に把持されるとともに、前記弾性糸ＳＹは、第２のスライドブロック５の下流側面と下側の固定ブロック６の上流側面とのスライドによって前記カッター刃１０の作用により切断される。
ref	Due to this slidable movement of the second slide block 5 , the elastic yarn SY is reliably held between the downstream side of the first slide block 4 and the upstream side of the second slide block 5 , and the elastic yarn SY is cut by the operation of the cutter blade 10 due to sliding between the downstream side of the second slide block 5 and the upstream side of the fixed block 6 at the lower side .
rmt	Elastic yarn SY by the slide movement of this 2nd slide block 5 , [between the downstream side of 1st slide block 4 , and the upper side of 2nd above-mentioned slide block 5] While being grasped certainly , the above-mentioned elastic yarn SY is cut by slide with the downstream side of 2nd slide block 5 , and the upper side of lower fixed block 6 by operation of the above-mentioned cutter blade 10 .
spe	The elastic yarn SY by the sliding movement of the slide block 5 , between the downstream side of the first slide block 4 , and the upper side of the second slide block 5 , is gripped , the elastic yarn SY is cut by the slide on the downstream side of the second slide block 5 , and the upper side of the lower fixed block 6 by the operation of the cutter blade 10 .

src	このような構成になる弾性糸のクランプカッター装置１において、弾性糸ＳＹを把持して切断する動作手順を、図１Ａ、図１Ｂ、図１Ｃに示してある。
ref	In the clump cutter apparatus 1 of the elastic yarn configured as described above , the operational procedure to hold and cut the elastic yarn SY is shown in FIG . 1A , FIG . 1B , and FIG . 1C .
rmt	In clamp cutter device 1 of the elastic yarn which becomes such composition , the procedure of operation of grasping and cutting elastic yarn SY is shown in Drawing 1A , Drawing 1B , and Drawing 1C .
spe	The clamp cutter device 1 of the elastic yarn structure , the procedure of the operation of holding and cutting the elastic yarn SY is shown in FIG . 1A , FIG . 1B and FIG . 1C .

src	各紡績ユニット５１は、ドラフト装置５６内の繊維束５５（スライバ５４）に対して、伸縮性の大なる弾性糸ＳＹを供給する弾性糸供給装置２０を備えている。
ref	Each spinning unit 51 is provided with an elastic yarn supplier 20 for supplying the elastic yarn SY having large elasticity to the fiber bundle 55 (the sliver 54) within the draft apparatus 56 .
rmt	Each spinning unit 51 is provided with elastic yarn feed unit 20 which supplies elastic yarn SY which is elasticity , and as for which size becomes to textiles bunch 55 (sliver 54) in drafting apparatus 56 .
spe	The spinning unit 51 is provided with the elastic yarn feeding unit 20 for supplying the elastic yarn SY , elasticity , and larger relative to the fiber bundle 55 (sliver 54 in the draft device 56 .