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

Terumasa EHARA
Yamanashi Eiwa College

ABSTRACT

In this article, we describe system architecture, preparation of training data and experimental results of the EIWA group in the NTCIR-9 Patent Translation Task. Our system is combining rule-based machine translation and statistical post-editing. Experimental results for Japanese to English (JE) subtask show 0.3169 BLEU score, 7.8161 NIST score, 0.7404 RIBES score, 3.43 adequacy score and 0.6381 pair wise comparison score for acceptability. Experimental results for Chinese to English (CE) subtask show 0.2597 BLEU score, 7.2282 NIST score, 0.7455 RIBES score, and 3.05 adequacy score.

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, Japanese to English, Chinese to English

Team name

EIWA

Subtasks/Languages

JE subtask / Japanese to English
CE subtask / Chinese to English

External Resources Used

Two commercial rule-based machine translation systems (J 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 a machine translation system is combining the rule-based machine translation (RBMT) and the statistical post-editing (SPE) [1][2][3][4].

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 a 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 simple system architecture.

2. JAPANESE TO ENGLISH TRANSLATION

2.1 SYSTEM ARCHITECTURE

Our JE translation 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 commercial base translation software for the RBMT part. This software is specialized to the patent translation.

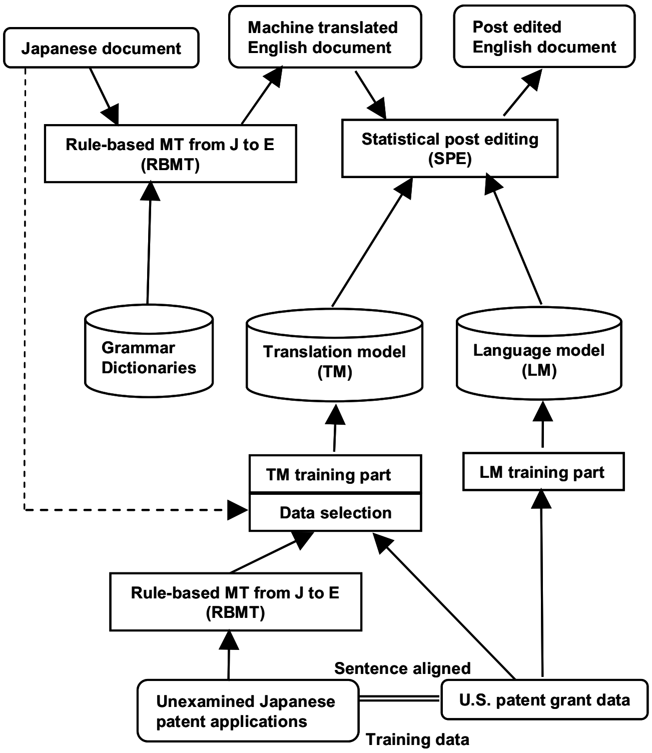
Japanese document
Machine translated English document
Post edited English document
Rule-based MT from J to E (RBMT)
Statistical post editing (SPE)
Grammar Dictionaries
Translation model (TM)
Language model (LM)
TM training part
Data selection
LM training part
Rule-based MT from J to E (RBMT)
Sentence aligned
Unexamined Japanese patent applications
U.S. patent grant data
Training data

Figure 1. JE translation system architecture

The SPE part automatically post-edits the output of the RBMT part to more accurate English document. We use the Moses Rev. 4343 for the SPE part. SPE part needs to include a translation model and a 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 0, because both the source language and the target language of the SPE part is English.

2.2 TRAINING, DEVELOPMENT AND TEST DATA

Training, development and test data used in our experiments are provided from NTCIR-9 Patent Translation Task organizer [5]. Test data include 2,000 Japanese sentences. Development data include 2,000 Japanese and English sentence pairs. We use only initial 300 sentence pairs for the development of JE system.

Training data of JE system consist of two parts. One is the training data for NTCIR-7 task and it includes 1,798,571 Japanese and English sentence pairs. The other is the training data for NTCIR-8 task and it includes 1,387,713 Japanese and English sentence pairs. We use English part of the NTCIR-8's 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 291,475 sentence pairs from the NTCIR-7 and NTCIR-8's 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 which is the same system used in the RBMT phase. The outputted English sentences from the RBMT system and corresponding GOOD reference English sentences in the training data are used as the translation model training. We use Giza-pp v.1.0.3 for the translation model training.

2.3 TRANSLATION MODEL TRAINING

Our JE translation model used 291,475 Japanese English sentence pairs selected from the total training data which include 3,186,284 sentence pairs. The idea for this selection method is to pick up the sentences adapted to input test sentences. Our system does not, then, work in real time, because the training and translation phases must be done at the same time. Construction method of the translation model training is as follows:

(a) Key word extraction: Key words are extracted from test sentences, Japanese part of the development 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 13.1.

(b) Training data selection for the test sentences: 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.

We use the following similarity measure:

$$sim = \frac{2 \times \#(T \cap S)}{\#(T) + \#(S)} \quad (1)$$

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

In this process, we select training sentences for each keyword of the test sentence. Up to ten training sentences which have most similarity to the test sentence are selected. Then the number of training sentences for one test sentence is up to ten times of the number of keywords of the test sentence. The total number of training sentences was 254,362 for 2,000 test sentences. The mean number of training sentences for one test sentence was 127.

(c) Training data selection for the development sentences: Same as the above process (b), we selected 37,113 training sentences for 300 development sentences. We, totally, got 291,475 training sentences.

2.4 EXAMPLE OF THE TRAINING DATA SELECTION

One example of the test sentence is:

この結果、コネクタ本体１４の上方への移動が規制され、コネクタ本体１４が基板１２上に実装される。

The reference English translation of this test sentence which is provided by the task organizer at the time of evaluation results release is:

As a result, upward movement of the connector main body 14 is restricted, and the connector main body 14 is mounted on the substrate 12.

Key words extracted from this test sentence are:

結果, コネクタ, 本体, 上方, 移動, 規制, 基板, 上, 実装

Selected training data for this sentence consists of 45 Japanese and English sentence pairs. The similarity values are spreading from 0.44 to 0.31. The Japanese parts of top three training data are:

端子３０の上方には、端子３０の上方への移動を規制する為の本体４に取付けられた規制部材７０が配置される。
即ち、基板保持部材８０が規制部材８８により上方への移動が規制されるとき、さらに、可動保持部材３６が回動されることによりコネクタ２６がコネクタ７０に対して離隔し始めることとなる。
また、更にこの第２の基板にはコネクタとしてＣＮ１、ＣＮ２、ＣＮ３が実装される。

And corresponding English parts of the training data are:

A restricting member 70, that is attached to the housing 4 in order to restrict the upward movement of the electrical terminal 30, is disposed above the electrical terminal 30.
When the base plate holding member 80 is restricted from movement in an upward direction by the restricting member 88, the connector 26 is caused to shift away from the connector 70 by pivoting the movable holding member 36.
In addition, on the second substrate 2, connectors CN1, CN2 and CN3 are mounted.

Translations of the Japanese parts of above training data by the RBMT are:

Above terminal 30, regulating member 70 attached to main part 4 for regulating movement to the upper part of terminal 30 is arranged.
That is, when movement to the upper part is regulated for substrate attachment component 80 by regulating member 88, when moving holding member 36 rotates, connector 26 will begin to be further isolated to connector 70.
CN1, CN2, and CN3 are mounted in this 2nd substrate as a connector.

2.5 TEST RESULT

Pair wise comparison score of acceptability of our JE system (EIWA) is 0.638 and the system is ranked at the third position in the 14 systems.

The outputs of our system (spe) and the outputs of the RBMT part (rmt) for several test sentences are shown in the Table 4 with Japanese source sentences (src) and English reference translations (ref).

For the first example in the Table 4 which is same to the example described in the section 2.4, the acceptability score for the SPE output is AA. Post-editing part repair the RBMT output phrase "movement to the upper part of connector" to the right expression "upward movement of the connector body". However, for the third example, "identifier storing column" in RBMT output is wrongly post-edited to "identifier column".

EIWA is ranked at the first position for automatic evaluations: BLEU and RIBES. Especially, the scores of BLEU and RIBES of EIWA are both greater than the scores of RBMT1 system[1]. However, the scores of human evaluations: adequacy and acceptability of EIWA are both lesser than the scores of RBMT1 system. Table 2 shows the comparison of scores of acceptability for RBMT1 system and our system.

Table 2. Comparison of acceptability scores of RBMT1 and EIWA

		EIWA	
		AA〜C	F
RBMT1	AA〜C	112	59
	F	36	93

93 sentences of 300 test sentences are scored F for both RBMT1 and our system. If the best translation can be selected from RBMT1 and EIWA, the score of acceptability will be better than the score of the single system.

3. CHINESE TO ENGLISH TRANSLATION

3.1 SYSTEM ARCHITECTURE

Our Chinese to English system architecture is shown in Figure 2, which is similar to Japanese to English system architecture.

The RBMT part translates a Chinese patent document to an English document using rule-based machine translation. We use commercial base translation software for the RBMT part.

The SPE part automatically post-edits the output of the RBMT part to more accurate English document. Translation model and language model are trained by Chinese and English bilingual patent corpus which is provided from NTCIR-9 Patent Translation Task organizer [5]. The Chinese part of the training corpus is machine translated by the same software in the RBMT part of the system before they are used in the translation model training. The distortion limit value for the decoding is set to 0.

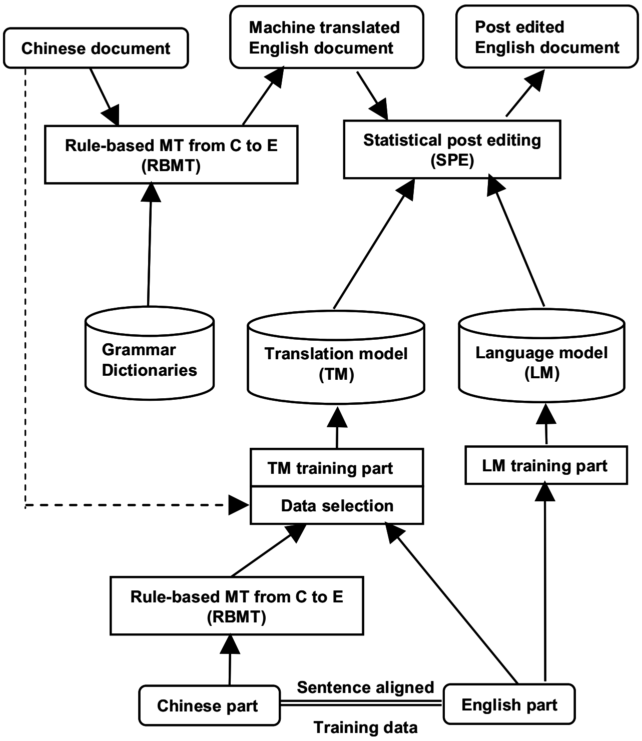

Figure 2. CE System architecture

3.2 TRAINING, DEVELOPMENT AND TEST DATA

Training, development and test data used in our experiments are provided from NTCIR-9 Patent Translation Task organizer [5]. Test data include 2,000 Chinese sentences. Development data include 2,000 Chinese and English sentence pairs. We use only initial 300 Chinese and English sentence pairs for the development. Training data include 1,000,000 Chinese and English sentence pairs.

We use English part of the training data for the language model training.

For the translation model construction, we select 238,787 sentence pairs from the training data. The detail of this selection method is described in the next section. Chinese part of this selected data is translated to English using the rule-based machine translation software which is the same software used in the RBMT part of the system. The outputted English sentences from the RBMT system and corresponding GOOD reference English

[1] RBMT1 is the baseline system ID and RBMT1 does not mean RBMT used in our method. RBMT1 was submitted by the organizers and it was a result translated by a commercial RBMT system.

sentences in the training data are used as the translation model training.

3.3 TRANSLATION MODEL TRAINING

Our translation model used 238,787 Chinese English sentence pairs selected from the total training data which include 1,000,000 sentence pairs. Construction method of the translation model training is similar to the JE system:

(a) Key word extraction: Key words are extracted from test sentences, Chinese part of the development sentences and Chinese part of the training sentences. In this phase, we use simple Chinese word segmenter which is made by ourselves. We extract key words which is not included in a stop-word list which includes 187 frequently occurring words. The mean number of keywords for one test sentence was 19.3.

(b) Training data selection for the test sentences: Algorithm is same as the algorithm of the JE system. The total number of training sentences for 2,000 test sentences was 201,231. The mean number of training sentences for one test sentence was 101.

(c) Training data selection for the development sentences: Same as the above process (b), we selected 37,650 training sentences for 300 development sentences. We, totally, got 238,881 training sentences. We filtered out the data which includes too long (more than 40 words) reference or rmt sentences. Finally, we got 238,787 training sentence pairs.

3.4 EXAMPLE OF THE TRAINING DATA SELECTION

One example of the test sentences is:

图1示出了根据本发明的照明设备1的一个例子。

The reference English translation of this test sentence which is provided by the task organizer at the time of evaluation results release is:

Figure 1 shows an example of a lighting device 1 according to the present invention.

Keywords extracted from this test sentence are:

图1示, 出了根据, 本发明的, 照明, 设备1, 的一个, 例子。

Selected training data for this sentence consists of 74 Chinese and English sentence pairs. The similarity values defined by the equation (1) are spreading from 0.46 to 0.11. The Chinese parts of top three training data are:

图1示出了根据本发明的原理制造的系统10的一个实施例。
图1示出了根据本发明的示例性实施方案的网络环境的框图表示。
图1示出了根据本发明的设备1如何被安排在电视机4中的电视接收机天线2与电视接收机3之间的。

And corresponding English parts of the training data are:

FIG. 1 shows one embodiment of a system 10 made in accordance with the principles of the present invention.
Fig. 1 displays a block diagram representation of a network environment in accordance with an exemplary embodiment of the present invention.
Fig. 1 shows how an arrangement 1 according to the invention is arranged between a TV-receiver antenna 2 and a TV-receiver 3 in a TV-set 4.

Translations of the Chinese parts by the RBMT are:

Figure 1 showed has acted according to this invention a principle manufacture system 10 implementation example.
Figure 1 showed has acted according to this invention the demonstration implementation plan network environment diagram expression.
How did Figure 1 show has acted according to this invention the equipment 1 to arrange in the television 4 television receiver antennas 2 and the television receiver 3 between.

3.5 TEST RESULT

Adequacy score of our CE system (EIWA) is 3.05 and the system is ranked at the 16th position in the 23 systems.

The outputs of our system (spe) and the outputs of the RBMT part (rmt) for several test sentences are shown in the Table 5 with Chinese source sentences (src) and English reference translations (ref).

For the first example in the Table 5 which is same to the example described in the section 3.4, the adequacy score for the SPE output is 4. Post-editing part repair the RBMT output phrase "showed has acted" to the more appropriate expression "illustrates".

Table 3 shows automatic and human evaluation scores for RBMT1, RBMT2 and our system (EIWA). We can see statistical post-editing make improvement of RBMTs for both automatic and human evaluation scores.

Table 3. Evaluation scores for RBMT systems and our system

	BLEU	NIST	RIBES	adequacy
RBMT1	0.108	4.546	0.670	2.277
RBMT2	0.128	5.174	0.694	2.663
EIWA	0.260	7.228	0.745	3.047

4. CONCLUSION

For Japanese to English translation, adding statistical post-editing part to rule-based machine translation, we can improve automatic evaluations: BLEU and RIBES. However, for human evaluations: adequacy and acceptability, the scores of our system are less than the scores of the rule-based system.

For Chinese to English translation, we can improve both automatic evaluation scores and human evaluation scores.

One of the main remaining issues of our technique 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.

5. REFERENCES

[1] Ehara, T. 2010. Machine translation for patent documents combining rule-based translation and statistical post-editing. Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access (May 2010).

[2] 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).

[3] 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).

[4] Ehara, T. 2008. Improving the translation accuracy using phrase-based statistical post-editing. Japio Year Book (Nov. 2008), 262-265, (in Japanese).

[5] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. NTCIR-9, 2011

Table 4. JE Translation Examples

src	この結果、コネクタ本体１４の上方への移動が規制され、コネクタ本体１４が基板１２上に実装される。
ref	As a result, upward movement of the connector main body 14 is restricted, and the connector main body 14 is mounted on the substrate 12.
rmt	As a result , movement to the upper part of connector main part 14 is regulated , and connector main part 14 is mounted on substrate 12 .
spe	As a result, the upward movement of the connector body 14 is regulated, and the connector body 14 is mounted on the substrate 12.

src	検出した信号をセンサ本体５５から速度計に伝え、艇体１１の推進速度を速度計に表示する。
ref	A detected signal is relayed from the sensor main body 55 to the speedometer which then reads the propulsion speed of the hull 11.
rmt	The detected signal is told to a speedometer from sensor body 55 , and the promotion speed of 艇 object 11 is displayed on a speedometer .
spe	The detected signal is transmitted to the speedometer from the sensor body 55, and the speed of the boat body 11 is to be displayed on the speedometer.

src	該テーブル１３０１の構造は、アプリケーション・ファイルシステム間関連テーブル１４０１と同様にアプリケーション識別子格納欄１７０１、ファイルシステム識別子格納欄１７０２、有効期間開始時刻格納欄１７０３、有効期間終了時刻格納欄１７０４からなる。
ref	The table 1301 has a configuration similar to that of the application-file system relation table 1401, and is composed of an application identifier storing field 1701, a file system identifier storing field 1702, an effective period start time storing field 1703, and an effective period end time storing field 1704.
rmt	The structure of this table 1301 consists of application identifier storing column 1701 , file system identifier storing column 1702 , term-of-validity start time storing column 1703 , and term-of-validity finish time storing column 1704 like application-file system related table 1401 .
spe	The structure of the table 1301 is composed of the application identifier column 1701, file system identifier column 1702, the valid period start time column 1703 and the valid period end time column 1704 such as application-file system related table 1401.

src	したがって、第２所定ビット数は、第１予定ビット数（例、１６ビット）よりも少ないビット数（例えば１０ビット）でよい。
ref	Thus, the second predetermined number of bits can be less (e.g. 10 bits) than the first predetermined number of bits (e.g. 16 bits).
rmt	Therefore , the number of bits (for example , 10 bits) smaller than the 1st schedule number of bits (an example , 16 bits) may be sufficient as the 2nd predetermined number of bits .
spe	Therefore, the number of bits (for example, 10 bits) smaller than the first predetermined number of bits (for example, 16 bits) may be used as the second predetermined number of bits.

Table 5. CE Translation Examples

src	图 1 示出了根据本发明的照明设备 1 的一个例子。
ref	Figure 1 shows an example of a lighting device 1 according to the present invention.
rmt	Figure 1 showed has acted according to this invention a lighting attachment 1 example .
spe	Figure 1 illustrates , in accordance with the present invention a luminaire 1 example .

src	该涂层 6 可改变从固态光源 4 发射的光的颜色。
ref	This coating 6 can change the color of the light emitted from the solid-state light source 4.
rmt	This coating 6 may change from the solid state photo source 4 launch light color .
spe	The coating 6 may change from the solid-state light source 4 emission colors of light .

src	实际上，这些方案只对于少量的合作方是可行的，并且通常仅由与它们的几个长期伙伴连接的较大的组织使用。
ref	Implementing a bespoke automated collaborative processing solution using imperative programming requires communication between systems to be in terms of parameters.
rmt	In fact , these plans regarding the few cooperation side are feasible only , and by connects only usually with theirs several long-term partners big organization use .
spe	In fact , these schemes for small collaborator side are viable only and connect to generally only with their several long-term partner large tissue used .

src	另外，复杂的加密计算导致较高的计算开销。
ref	In addition, the complex encryption calculations lead to high computational overheads.
rmt	Moreover , the complex encryption computation causes the high computation expenses .
spe	In addition , the complex encryption calculation results in high computational overhead .