Overview of the Patent Translation Task at the NTCIR-7 Workshop

Atsushi Fujii[†], Masao Utiyama[‡], Mikio Yamamoto[†], Takehito Utsuro[†]
[†]University of Tsukuba
[‡]National Institute of Information and Communications Technology

Abstract

To aid research and development in machine translation, we have produced a test collection for Japanese/English machine translation and performed the Patent Translation Task at the Seventh NTCIR Workshop. To obtain a parallel corpus, we extracted patent documents for the same or related inventions published in Japan and the United States. Our test collection includes approximately 2 000 000 sentence pairs in Japanese and English, which were extracted automatically from our parallel corpus. These sentence pairs can be used to train and evaluate machine translation systems. Our test collection also includes search topics for cross-lingual patent retrieval, which can be used to evaluate the contribution of machine translation to retrieving patent documents across languages. This paper describes our test collection, methods for evaluating machine translation, and evaluation results for research groups participated in our task. Our research is the first significant exploration into utilizing patent information for the evaluation of machine translations.

Keywords: *Patent information, Machine translation, Patent families, Cross-lingual retrieval, NTCIR*

1 Introduction

Since the Third NTCIR Workshop in 2001[1], which was an evaluation forum for research and development in information retrieval and natural language processing, the Patent Retrieval Task has been performed repeatedly [2, 3, 5, 8]. In the Sixth NTCIR Workshop [5], patent documents published over a 10-year period by the Japanese Patent Office (JPO) and the US Patent & Trademark Office (USPTO) were independently used as target document collections.

Having explored patent retrieval issues for a long time, we decided to address another issue in patent processing. From among a number of research issues related to patent processing [4], we selected Machine Translation (MT) of patent documents, which is useful for a number of applications and services, such as Cross-Lingual Patent Retrieval (CLPR) and filing patent applications in foreign countries.

Reflecting the rapid growth in the use of multilingual corpora, a number of data-driven MT methods have recently been explored, most of which are termed "Statistical Machine Translation (SMT)". While large bilingual corpora for European languages, Arabic, and Chinese are available for research and development purposes, these corpora are rarely associated with Japanese and therefore it is difficult for explore SMT with respect to Japanese.

However, we found that the patent documents used for the NTCIR Workshops can potentially alleviate this data scarcity problem. Higuchi et al. [7] used "patent families" as a parallel corpus for extracting new translations. A patent family is a set of patent documents for the same or related inventions and these documents are usually filed in more than one country in various languages. Following Higuchi et al's method, we can produce a bilingual corpus for Japanese and English. In addition, there are a number of SMT engines (decoders) available to the public, such as Pharaoh and Moses[2], which can be applied to bilingual corpora involving any pair of languages.

Motivated by the above background, we determined to organize a machine translation task for patents in the Seventh NTCIR Workshop (NTCIR-7). This paper describes our task, namely "the Patent Translation Task".

2 Overview of the Patent Translation Task

The Patent Translation Task comprised the following three steps. First, the organizers, who are the authors of this paper, provided groups participating in the Patent Translation Task with a training data set of aligned sentence pairs in Japanese and English. Each participating group was allowed to use this data set to train their MT system, whether it is a data-driven SMT or a conventional knowledge-intensive rule-based MT.

[1] http://research.nii.ac.jp/ntcir/index-en.html

[2] http://www.statmt.org/wmt07/baseline.html

Second, the organizers provided the groups with a test data set of sentences in either Japanese or English. Each group was requested to machine translate each sentence from its original language into the other language and submit their translation results to the organizers.

Third, the organizers evaluated the submission from each group. We used both intrinsic and extrinsic evaluation methods. In the intrinsic evaluation, we independently used both the Bilingual Evaluation Understudy (BLEU) [11], which had been proposed as an automatic evaluation measure for MT, and human judgment. In the extrinsic evaluation, we investigated the contribution of the MT to CLPR. In the Patent Retrieval Task at NTCIR-5, aimed at CLPR, search topics in Japanese were translated into English by human experts. We reused these search topics for the evaluation of the MT. We also analyzed the relationship between different evaluation measures.

The use of extrinsic evaluation, which is not performed in existing MT-related evaluation activities, such as the NIST MetricsMATR Challenge[3] and the IWSLT Workshop[4], is a distinctive feature of our research.

We executed the above three steps in both a preliminary trial and the final evaluation, using the terms "dry run" and "formal run", respectively. While Fujii et al. [6] described the dry run, this paper describes the formal run.

Sections 3 and 4 explain the intrinsic and extrinsic evaluation methods, respectively. Section 5 describes the evaluation results for the formal run.

3 Intrinsic Evaluation

3.1 Evaluation Method

Figure 1 depicts the process flow of the intrinsic evaluation. We explain the entire process in terms of Figure 1.

In the Patent Retrieval Task at NTCIR-6 [5], the following two document sets were used.

- Unexamined Japanese patent applications published by the JPO during the 10-year period 1993–2002. There are approximately 3 500 000 of these documents.

- Patent grant data published by the USPTO during the 10-year period 1993–2002. There are approximately 1 300 000 of these documents. Because the USPTO documents include only patents that have been granted, there are fewer of these documents than of the above JPO documents.

[3] http://www.nist.gov/speech/tests/metricsmatr/
[4] http://www.slc.atr.jp/IWSLT2008/

From these document sets, we automatically extracted patent families. From among the various ways to apply for patents in more than one country, we focused only on patent applications claiming priority under the Paris Convention. In a patent family applied for under the Paris Convention, the member documents of a patent family are assigned the same priority number, and patent families can therefore be identified automatically.

Figure 2 shows an example of a patent family, in which the upper and lower parts are fragments (bibliographic information and abstracts) of a JPO patent application and a USPTO patent, respectively. In Figure 2, item "(31)" in the Japanese document and item "[21]" in the English document each denote the priority number, which is "295127" in both cases.

Using priority numbers, we extracted approximately 85 000 USPTO patents that originated from JPO patent applications. While patents are structured in terms of several fields, in the "Background of the Invention" and the "Detailed Description of the Preferred Embodiments" fields, text is often translated on a sentence-by-sentence basis. For these fields, we used a method [13] to automatically align sentences in Japanese with their counterpart sentences in English.

In the real world, a reasonable scenario is that an MT system is trained using existing patent documents and is then used to translate new patent documents. Therefore, we produced training and test data sets based on the publication year. While we used patent documents published during 1993–2000 to produce the training data set, we used patent documents published during 2001–2002 to produce the test data set.

The training data set has approximately 1 800 000 Japanese–English sentence pairs, which is one of the largest collections available for Japanese and English MT. To evaluate the accuracy of the alignment, we randomly selected 3000 sentence pairs from the training data and asked a human expert to judge whether each sentence pair represents a translation or not. Approximately 90% of the 3000 pairs were correct translations. This training data set was used for both the dry run and the formal run.

The sentence pairs extracted from patent documents published during 2000–2001 numbered approximately 630 000. For the test data set, we selected approximately 1000 sentence pairs that had been judged as correct translations by human experts. In the selected pairs, the Japanese (or English) sentences were used to evaluate Japanese–English (or English–Japanese) MT. Unlike the training data set, we used different test sets for the dry run and the formal run.

To evaluate translation results submitted by participating groups, we independently used BLEU and human judgment. To calculate the value of BLEU for the test sentences, we need one or more reference translations. For each test sentence, we used its counterpart

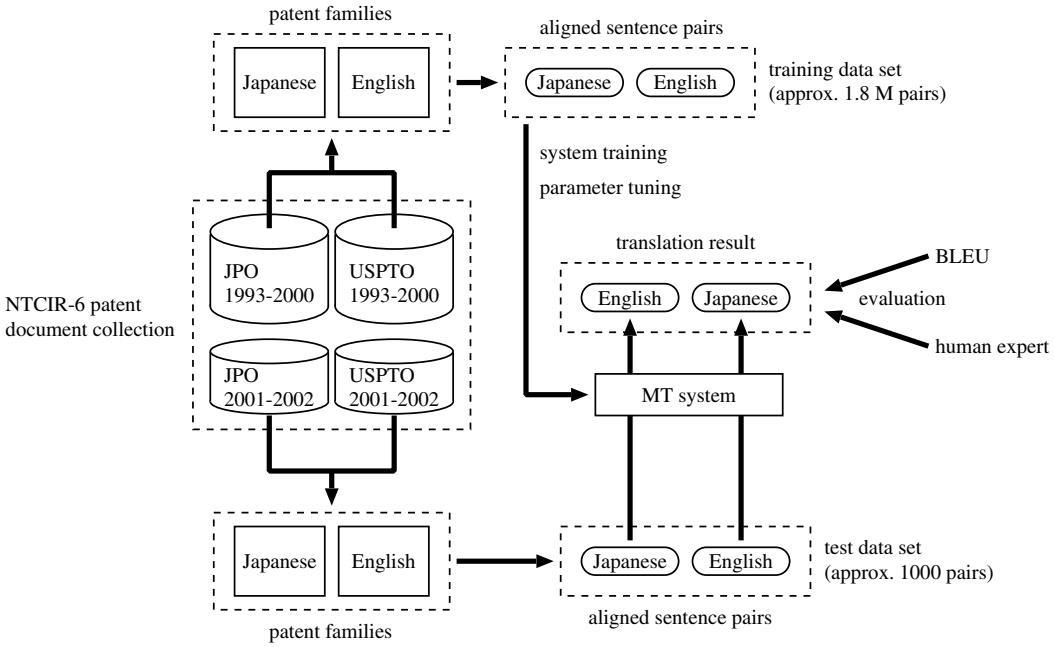

Figure 1. Overview of the intrinsic evaluation.

(11)【公開番号】特開平８－１１４２７８
(43)【公開日】平成８年（１９９６）５月７日
(54)【発明の名称】マイクロアクチュエータ
(21)【出願番号】特願平７－２３９２３０
(22)【出願日】平成７年（１９９５）８月２４日
(31)【優先権主張番号】２９５，１２７
(32)【優先日】１９９４年８月２４日
(33)【優先権主張国】米国（ＵＳ）
(57)【要約】
【課題】断熱構造を備えるマイクロアクチュエータ。
【解決手段】フローチャネルを介して運搬される流体流を制御する超小型バルブの形態をなすマイクロアクチュエータであり、サーマルアクチュエータによって選択的に駆動される熱駆動部材を有し、これが駆動されることによって熱エネルギを生成する第１基板と、対向する第１、第２主要面を有する第２基板よりなる。第２基板が第１主要面で第１基板に取付けられる。第２の主要面は第２基板が支持体に取付けられると絶縁セルを画定し、これによってマイクロアクチュエータの熱容量を減少させ、第１基板を支持体から熱遮断する。

[11] Patent Number 5,529,279
[45] Date of Patent June 25, 1996
[54] Thermal isolation structures for microactuators
[57] Abstract
A microactuator preferably in the form of a microminiature valve for controlling the flow of a fluid carried by a flow channel includes a first substrate having a thermally-actuated member selectively operated by a thermal actuator such that the first substrate thereby develops thermal energy, and a second substrate having opposed first and second major surfaces. The second substrate is attached to the first substrate at the first major surface. The second major surface defines an isolation cell for enclosing a volume when the second substrate is attached to the support to thereby reduce the thermal mass of the microactuator and to thermally isolate the first substrate from the support.
[21] Appl. No.: 295127
[22] Filed: August 24, 1994

Figure 2. Example of JP-US patent family.

sentence as the reference translation. We also asked several human experts to produce a reference translation for each test sentence in Japanese independently, to enhance the objectivity of the evaluation by BLEU. We elaborate on the method to produce multiple references in Section 3.2.

We produced additional references only for the Japanese–English intrinsic evaluation. In the patent families we extracted, Japanese applications were first produced and then translated into English. The writing quality of these texts is not always satisfactory because texts are not always produced by English-speaking translators and are sometimes produced by editing outputs of MT systems. If human experts back-translate these low-quality texts into Japanese, the quality of references would not be satisfactory. Therefore, we did not produce additional references for the English–Japanese intrinsic evaluation.

We used "Bleu Kit"[5] to calculate BLEU values. For tokenization purposes, we used "ChaSen"[6] and the tokenizer in "ACL2007 the 2nd workshop on SMT"[7] for Japanese and English sentences, respectively.

For human judgments, we asked human experts to evaluate each translation result based on fluency and adequacy, using a five-point rating. However, because manual evaluation for all submitted translations would be expensive, we randomly selected 100 test sentences for human judgment purposes. We analyzed the relationship between the evaluation by BLEU and the

[5] http://www.mibel.cs.tsukuba.ac.jp/~norimatsu/bleu_kit/
[6] http://chasen-legacy.sourceforge.jp/
[7] http://www.statmt.org/wmt07/baseline.html

evaluation by human judgment.

The procedure for the dry run was fundamentally the same as that for the formal run. However, mainly because of time constraints, we imposed the following restrictions on the dry run.

- The dry run used 822 test sentences, whereas the formal run used 1381 test sentences.
- To calculate the value of BLEU in the intrinsic evaluation, we used only a single reference. The reference sentence of a test sentence was the counterpart translation in our test collection. The correctness of each counterpart translation had been verified by a human expert.
- For the human judgment, a single expert evaluated 100 translated sentences for each group. In the formal run, three human experts independently evaluated the same sentences.

3.2 A Story of Producing Multiple References

To increase the number of reference translations for each test sentence, we initially intended to target 600 test sentences. However, due to a number of problems, we produced reference translations for the following two sets of test sentences.

- S600

  According to our initial plan, we randomly selected 600 sentences from the 1381 Japanese test sentences for the formal run, and three experts (E1, E2, and E3) independently translated all the 600 sentences into English. We call these 600 Japanese sentences "S600". However, a post-work interview found that the three experts had used a rule-based MT (RBMT) system for translation purposes, although they had not fully relied on that system and had consulted translations on a word-by-word basis, if necessary.

- S300

  As explained above, the reference translations for S600 are somewhat influenced by the RBMT system used. We concerned that values of BLEU calculated by these reference translations potentially favor RBMT systems. To avoid this problem, we asked different three experts (E4, E5, and E6) to translate a subset of S600. Mainly because of time and budget constraints, we targeted only 300 sentences. We call these 300 Japanese sentences "S300". However, we found that E6 had used an RBMT system for translation purposes.

In summary, all the reference translations produced for S600 and the reference translations produced by E6 for S300 are influenced by RBMT systems.

In addition, it is often the case that a human expert edits a machine translated text, to produce a patent application. Thus, the counterpart English sentences for Japanese test sentences are also potentially influenced by RBMT systems.

To minimize the influence of RBMT systems, we can use only the reference translations produced by E4 and E5 for S300 in the evaluation. At the same time, because experts did not strongly rely on RBMT systems, we can also use the other reference translations with caution.

In principle, we can calculate BLEU values with different combinations of sentence sets and reference translations. In practice, we used the following three types of BLEU values for the Japanese–English intrinsic evaluation. Each BLEU type is associated with an advantage and a disadvantage.

- Single-Reference BLEU (SRB)

  This value is calculated by the counterpart sentences for the 1381 test sentences. Although only a single reference translation is used for each test sentence, we can use all test sentences available.

- Multi-Reference BLEU for S300 (MRB300)

  This value is calculated by the reference translations produced by E4 and E5 for S300. Although we can target only 300 test sentences, we can use as many reference translations as possible, while avoiding the influence of RBMT systems.

- Multi-Reference BLEU for S600 (MRB600)

  This value is calculated by the reference translations produced by E1, E2, and E3, and the counterpart sentences for S600. Although this value is potentially influenced by RBMT systems, we can use as many reference translations and test sentences as possible.

We use terms "SRB", "MRB300", and "MRB600" for explaining the result of the Japanese–English intrinsic evaluation in Section 5.2. However, we do not use these terms to explain the result of the English–Japanese intrinsic evaluation, for which additional reference translations were not produced.

4 Extrinsic Evaluation

In the extrinsic evaluation, we investigated the contribution of MT to CLPR. Each group was requested to machine translate search topics from English into Japanese. Each of the translated search topics was used to search a patent document collection in Japanese for the relevant documents. The evaluation results for CLPR were compared with those for a monolingual retrieval in Japanese. Figure 3 depicts the process flow of the extrinsic evaluation. We explain the entire process in terms of Figure 3.

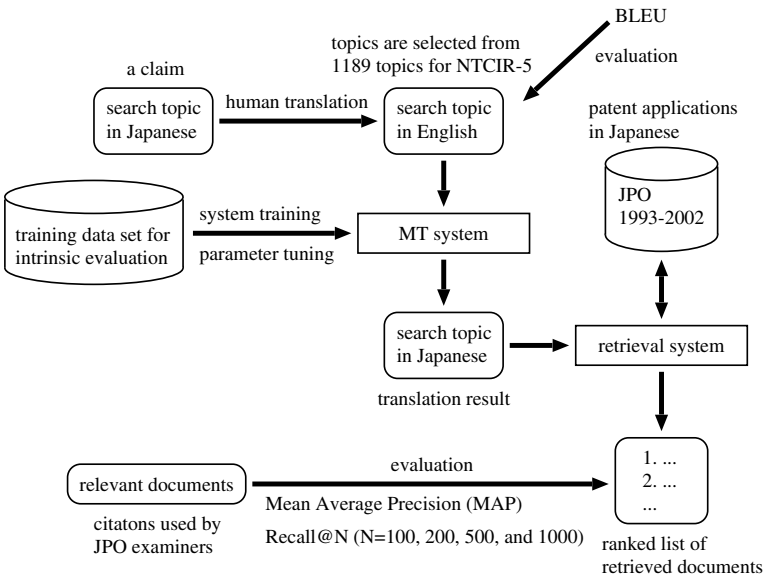

Figure 3. Overview of the extrinsic evaluation.

Processes for patent retrieval differ significantly, depending on the purpose of the retrieval. One process is the "technology survey", in which patents related to a specific technology, such as "blue light-emitting diode", are searched for. This process is similar to ad hoc retrieval tasks targeting nonpatent documents.

Another process is the "invalidity search", in which prior arts related to a patent application are searched for. Apart from academic research, invalidity searches are performed by examiners in government patent offices and searchers in the intellectual property divisions of private companies.

In the Patent Retrieval Task at NTCIR-5 [3], invalidity search was performed. The purpose was to search a Japanese patent collection, which is the collection described in Section 3, for those patents that can invalidate the demand in an existing claim. Therefore, each search topic is a claim in a patent application. Search topics were selected from patent applications that had been rejected by the JPO. There are 1189 search topics.

For each search topic, one or more citations (i.e., prior arts) that were used for the rejection were used as relevant or partially relevant documents. The degree of relevance of the citation with respect to a topic was determined based on the following two ranks.

- The citation used to reject an application was regarded as a "relevant document" because the decision for the rejection was made confidently.

- A citation used to reject an application with another citation was regarded as a "partially relevant document" because each citation is partially related to the claim in the application.

By definition, each search topic is associated with either a single relevant document or multiple partially relevant documents. Within the 1189 search topics, 619 topics are associated with relevant documents and the remaining 570 topics are associated with partially relevant documents.

In addition, with the aim of CLPR, these search topics were translated by human experts into English during NTCIR-5. In the extrinsic evaluation at NTCIR-7, we reused these search topics. Each search topic file includes a number of additional SGML-style tags. Figure 4 shows an example of a topic claim translated into English, in which <NUM> denotes the topic identifier.

```
<TOPIC><NUM>1048</NUM>
<FDATE>19950629</FDATE>
<CLAIM>A milk-derived calcium-containing
composition comprising an inorganic salt mainly
composed of calcium obtained by baking a
milk-derived prepared matter containing milk
casein-bonding calcium and/or colloidal calcium.
</CLAIM></TOPIC>
```

Figure 4. Example search topic produced at NTCIR-5.

In Figure 4, the claim used as the target of invalidation is specified by <CLAIM>, which is also the target of translation. In retrieval tasks for nonpatent documents, such as Web pages, a query is usually a small number of keywords. However, because each search topic in our case is usually a long and complex noun phrase including clauses, the objective is almost

translating sentences. The date of filing is specified by <FDATE>. Because relevant documents are prior arts, only the patents published before this date can potentially be relevant.

Although each group was requested to machine translate the search topics, the retrieval was performed by the organizers. As a result, we were able to standardize the retrieval system and the contribution of each group was evaluated in terms of the translation accuracy alone. In addition, for most of the participating groups, who are research groups in natural language processing, the retrieval of 10 years' worth of patent documents was not a trivial task.

We used a system that had also been used in the NTCIR-5 Patent Retrieval Task [1] as the standard retrieval system. This system, which uses Okapi BM25 [12] as the retrieval model, sorts documents according to the score and retrieves up to the top 1000 documents for each topic. This system also uses the International Patent Classification to restrict the retrieved documents.

Because the standard retrieval system performed word indexing and did not use the order of words in queries and documents, the order of words in a translation did not affect the retrieval effectiveness. In CLPR, a word-based dictionary lookup method can potentially be as effective as the translation of sentences.

As evaluation measures for CLPR, we used the Mean Average Precision (MAP), which has frequently been used for the evaluation of information retrieval, and Recall for the top N documents (Recall@N). In the real world, an expert in patent retrieval usually investigates hundreds of documents. Therefore, we set $N = 100$, 200, 500, and 1000. We also used BLEU as an evaluation measure, for which we used the source search topics in Japanese as the reference translations.

In principle, for the extrinsic evaluation we were able to use all of the 1189 search topics produced in NTCIR-5. However, because the length of a single claim is usually much longer than that of an ordinary sentence, the computation time for the translation can be prohibitive. Therefore, in practice we independently selected a subset of the search topics for the dry run and the formal run.

If we use search topics for which the average precision of the monolingual retrieval is small, the average precision of CLPR methods can be so small that it is difficult to distinguish the contributions of participating groups to CLPR. Therefore, we sorted the 1189 search topics according to the Average Precision (AP) of monolingual retrieval using the standard retrieval system and found the following distribution.

- AP $\geq$ 0.9: 100 topics
- 0.9 $>$ AP $\geq$ 0.3: 124 topics
- AP $<$ 0.3: 965 topics

We selected the first 100 topics for the dry run and the next 124 topics for the formal run.

5 Evaluation in the Formal Run

5.1 Overview

The schedule of the formal run was as follows.

- 2008.04.25: Release of the test data
- 2008.05.30: Submission deadline

The participating groups were allowed one month to translate the test data.

As explained in Sections 2–4, the formal run involved three types of evaluation: Japanese–English intrinsic evaluation, English–Japanese intrinsic evaluation, and English–Japanese extrinsic evaluation. The numbers of groups participated in these evaluation types were 14, 12, and 12, respectively. In addition, to produce a baseline performance, the organizers submitted a result produced by Moses, in which default parameters were used, to each evaluation type.

Table 1 gives statistics with respect to the length of test sentences and search topics. While we counted the number of characters for sentences in Japanese, we counted the number of words for sentences and search topics in English.

Table 1. Length of test sentences and search topics.

	Min.	Avg.	Max.
Intrinsic Japanese	11	60.1	302
Intrinsic English	5	29.0	117
Extrinsic English	13	115.4	412

For each evaluation type, each group was allowed to submit more than one result and was requested to assign a priority to each result. For the sake of conciseness, we show only the highest priority results for each group with each evaluation type. Each group was also requested to submit a brief description of their MT system, which will be used to analyze the evaluation results in Sections 5.2–5.4.

5.2 J–E Intrinsic Evaluation

Table 2 shows the results of the Japanese–English intrinsic evaluation, in which the column "Method" denotes the method used by each group, namely "Statistical MT (SMT)", "Rule-Based MT (RBMT)", and "Example-Based MT (EBMT)". The columns "BLEU" and "Human" denote the values for BLEU and human rating, respectively. The columns "SRB",

Table 2. Results of J–E intrinsic evaluation.

Group	Method	BLEU			Human	Adequacy	Fluency
		SRB	MRB300	MRB600			
NTT	SMT	27.20	35.93	43.72	3.30	2.96	3.65
Moses *	SMT	27.14	36.02	43.40	3.18	2.81	3.55
(MIT)	SMT	27.14	37.31	44.69	3.40	3.15	3.66
NAIST-NTT	SMT	25.48	34.66	41.89	3.04	2.66	3.43
NiCT-ATR	SMT	24.79	32.29	39.40	2.78	2.47	3.08
KLE	SMT	24.49	33.59	40.20	2.94	2.59	3.28
(tsbmt)	RBMT	23.10	37.51	48.02	3.88	3.81	3.94
tori	SMT	22.29	27.92	35.02	3.01	2.58	3.44
Kyoto-U	EBMT	21.57	29.35	35.49	3.10	2.85	3.35
(MIBEL)	SMT	19.93	27.84	32.99	2.74	2.38	3.09
HIT2	SMT	19.48	29.33	33.60	2.86	2.44	3.28
JAPIO	RBMT	19.46	32.62	41.77	3.86	3.71	4.02
TH	SMT	15.90	24.20	28.72	2.13	1.87	2.39
FDU-MCandWI	SMT	9.55	19.94	20.27	2.08	1.75	2.42
(NTNU)	SMT	1.41	2.48	2.63	1.06	1.08	1.04

"MRB300", and "MRB600" in "BLEU" denote the values for different types of BLEU. The numbers of test sentences used for these BLEU types are 1381, 300, and 600, respectively. See Section 3.2 for details of different types of BLEU values.

For human judgment, three experts independently evaluated the same 100 sentences. The score with respect to adequacy and fluency, which are denoted as "Adequacy" and "Fluency", respectively, ranges from 1 to 5. Each score is an average over the 100 sentences and also the three experts. The value for human rating, which is the average of "Adequacy" and "Fluency", also ranges from 1 to 5. The rows in Table 2, each of which corresponds to the result of a single group, are sorted according to the values for SRB.

A number of groups submitted their results with the highest priority after the deadline. We denote the names of these groups in parentheses. In addition, "Moses *" denotes results for the submission produced by the organizers. These results are not official results and should be discarded for strict comparisons.

As shown in Table 2, groups that used an SMT method, such as "NTT", "Moses", and "MIT", tended to obtain large values for SRB, compared to groups that used RBMT and EBMT methods. The difference in SRB values between groups using an SMT method is due to the decoder and the size of the data used for training purposes. Top groups generally used a regular or hierarchical phrase-base SMT method. However, "FDU-MCandWI" used the IBM Model 4, which is a word-based SMT method. In addition, groups that were not able to process the entire training data used a fragment of the training data.

Figure 5 shows each group's SRB values with a 95% confidence interval; the values were calculated by a bootstrap method [10] using 1000-fold resampling. In Figure 5, the SRB values for the top three groups are comparable and greater than those for the other groups, with a 95% confidence. The result obtained by Moses, which had not been developed for Japanese, was in the top cluster, and thus Moses performed effectively for Japanese–English MT.

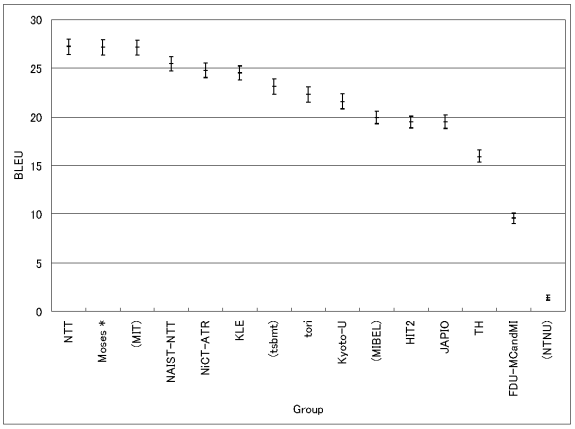

Figure 5. BLEU (SRB) with a 95% confidence interval for Japanese–English intrinsic evaluation.

We discuss the values of BLEU obtained by multiple references. The value of BLEU, which is the extent to which word n-grams in each test sentence match to those in one or more reference translations, generally increases, as the number of reference translations increases. This tendency is also observed in our evaluation. In Table 2, the values for MRB600 are generally larger than those for MRB300 and SRB.

In Table 2, increases of "tsbmt" and "JAPIO" in MRB300 and MRB600 are noticeable. This tendency can easily be observed in Figures 6 and 7, which use

the same notation as Figure 5, and show the values of BLEU with a 95% confidence interval for MRB300 and MRB600, respectively. In Figure 6, the BLEU values for tsbmt and MIT are comparable and these groups outperformed the other groups. However, in Figure 7, tsbmt outperformed MIT and achieved the best BLEU value.

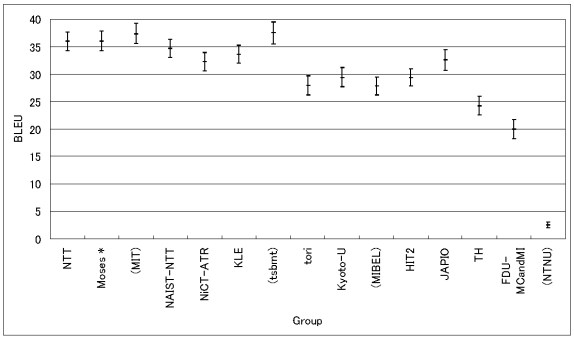

Figure 6. BLEU (MRB300) with a 95% confidence interval for Japanese–English intrinsic evaluation.

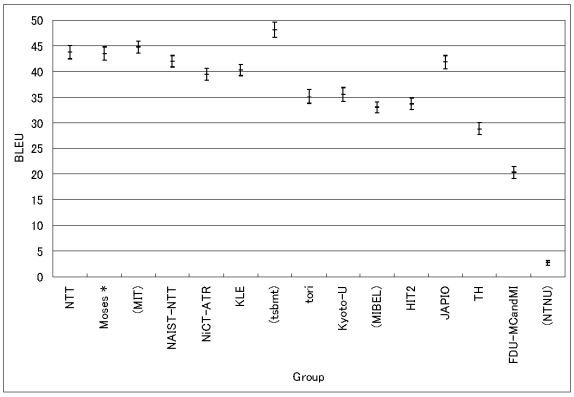

Figure 7. BLEU (MRB600) with a 95% confidence interval for Japanese–English intrinsic evaluation.

A reason for the above observations is that tsbmt and JAPIO used an RBMT method. Because as explained in Section 3.2, the values for MRB600 are potentially influenced by RBMT systems, it can be predicted that MRB600 favors RBMT methods. However, the values for MRB300 are not influenced by RBMT systems. This is possibly due to the characteristics of the reference translations for MRB300 and the training data set used. The participating SMT systems had been trained on our training data set, consisting of Japanese sentences and their counterpart English sentences. Because the characteristics of the counterpart sentences for the test and training data sets are similar, these SMT systems outperformed the RBMT systems in SBR. However, because the reference translations for MRB300 are independent of the counterpart sentences in the training data set, unlike RBMT systems, these SMT systems did not perform effectively.

Figure 8 graphs the value for "Human" in Table 2, in which the order of groups is the same as Figures 5–7. In Figure 8, tsbmt and JAPIO, which were not effective in BLEU, outperformed the other groups with respect to human rating. BLEU is generally suitable for comparing the effectiveness of SMT methods, but not suitable for evaluating other types of methods.

To further investigate this tendency, Figure 9 shows the relationship between the values for human rating and each BLEU type. In Figure 9, we also show the correlation coefficient ("R") between human rating and each BLEU type. The value of R for SRB is 0.814, which is smaller than those for MRB300 and MRB600. This is mainly due to the two outliers on the right side that correspond to the results for tsbmt and JAPIO.

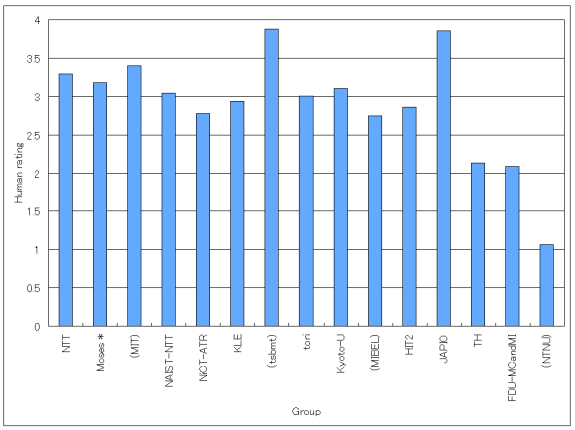

Figure 8. Human rating for Japanese–English intrinsic evaluation.

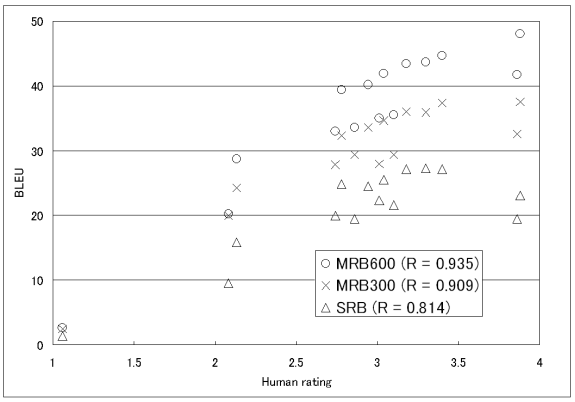

Figure 9. Relationship between BLEU and human rating for Japanese–English intrinsic evaluation.

However, the values of R for MRB300 and MRB600 are more than 0.9, showing a high correlation between human rating and BLEU. By using multiple references, the evaluation result by BLEU became similar to that by human rating. In such a case, while human judgments are not reusable, we need only reference translations, which are reusable, for evaluating MT methods.

Finally, in Table 2, the relative superiority of the groups is almost the same in "Adequacy" and "Fluency". In other words, there was no method that is particularly effective for either adequacy or fluency.

5.3 E–J Intrinsic Evaluation

Table 3 shows the results for the English–Japanese intrinsic evaluation and the extrinsic evaluation, which are denoted as "Intrinsic" and "Extrinsic", respectively. Because the source language was English for both evaluation types, we compare the results for "Intrinsic" and "Extrinsic" in a single table. The rows in Table 3, each of which corresponds to the result of a single group, are sorted according to the values for BLEU in "Intrinsic".

Unlike the Japanese–English evaluation in Table 2, "MIT", "JAPIO", and "NTNU" did not participate in the English–Japanese evaluation, and "HCRL" participated only in the English–Japanese evaluation. In this section we focus on "Intrinsic" and we will elaborate on "Extrinsic" in Section 5.4.

Mainly because of time and budget constraints, we imposed two restrictions on the English–Japanese evaluation. First, human judgments were performed only for a small number of groups, for which we selected one or more top groups in terms of BLEU from each method type (i.e., SMT, RBMT, and EBMT). Second, we did not produce additional reference translations and used only the counterpart sentences for the 1381 test sentences as the reference.

In Table 3, SMT methods are generally effective in terms of BLEU and Moses achieved the best BLEU value. However, tsbmt, which used an RBMT method, outperformed the other groups with respect to human rating.

Figure 10, which uses the same notation as Figure 5, shows the values of BLEU with a 95% confidence interval for each group. In Figure 10, the relative superiority of top groups was different from that in Figure 5.

5.4 Extrinsic Evaluation

The "Extrinsic" column in Table 3 shows the results of the extrinsic evaluation, and includes the values for BLEU and MAP for each group. According to their system descriptions, all groups participating in the extrinsic evaluation used the same method for the English–Japanese intrinsic evaluation.

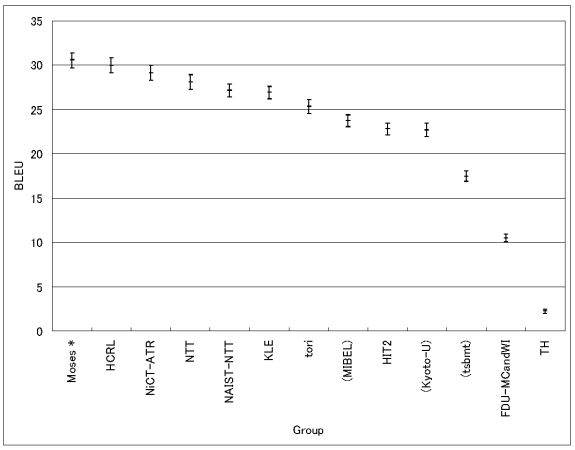

Figure 10. BLEU with a 95% confidence interval for English–Japanese intrinsic evaluation.

In Table 3, "Extrinsic BLEU", which denotes the BLEU values for the extrinsic evaluation, is different from "Intrinsic BLEU". As explained in Section 4, the English search topics used for the extrinsic evaluation are human translations of search topics in Japanese. To calculate values for BLEU in the extrinsic evaluation, we used these search topics in Japanese as the reference translations.

In Table 3, the relative superiority of the groups with respect to BLEU was almost the same for the extrinsic evaluation as it was for the intrinsic evaluation. Figure 11 shows the relationship between the values for BLEU in the intrinsic and extrinsic evaluation types, in which the correlation coefficient is 0.964. Therefore, the accuracy of translating claims in patent applications is correlated with the accuracy of translating other fields in patent applications, despite claims being described in a patent-specific language.

In Table 3, to calculate the values for MAP, we used both relevant and partially relevant documents as the correct answers. See Section 4 for details concerning relevant and partially relevant documents. In Table 3, the row "Mono" shows the results for monolingual retrieval, which is an upper bound to the effectiveness for CLPR. The best MAP for CLPR obtained by HCRL is 0.3536, which is 74% of that for Mono.

In addition to MAP, we also used Recall@N as an evaluation measure for information retrieval (IR). Figure 12 shows the relationship between the values for BLEU in the extrinsic evaluation and each IR evaluation measure. In Figure 12, among the IR evaluation measures, the value of R for MAP is the largest. In other words, we can use BLEU to predict the contribution of MT systems to CLPR with respect to MAP,

Table 3. Results of E–J intrinsic/extrinsic evaluation.

		Intrinsic				Extrinsic	
Group	Method	BLEU	Human	Adequacy	Fluency	BLEU	MAP
Moses *	SMT	30.58	3.30	2.90	3.69	20.70	0.3140
HCRL	SMT	29.97	—	—	—	21.10	0.3536
NiCT-ATR	SMT	29.15	2.89	2.59	3.20	19.40	0.3494
NTT	SMT	28.07	3.14	2.74	3.54	18.69	0.3456
NAIST-NTT	SMT	27.19	—	—	—	20.46	0.3248
KLE	SMT	26.93	—	—	—	19.07	0.2925
tori	SMT	25.33	—	—	—	17.54	0.3187
(MIBEL)	SMT	23.72	—	—	—	18.67	0.2873
HIT2	SMT	22.84	—	—	—	17.71	0.2777
(Kyoto-U)	EBMT	22.65	2.48	2.42	2.54	13.75	0.2817
(tsbmt)	RBMT	17.46	3.60	3.53	3.67	12.39	0.2264
FDU-MCandWI	SMT	10.52	—	—	—	11.10	0.2562
TH	SMT	2.23	—	—	—	1.39	0.1000
Mono	—	—	—	—	—	—	0.4797

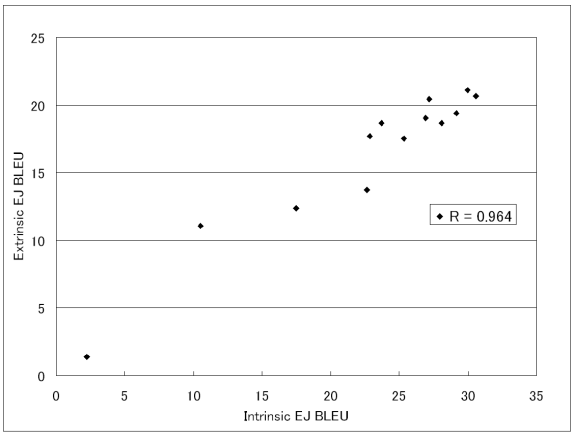

Figure 11. Relationship between BLEU in intrinsic and extrinsic evaluation types.

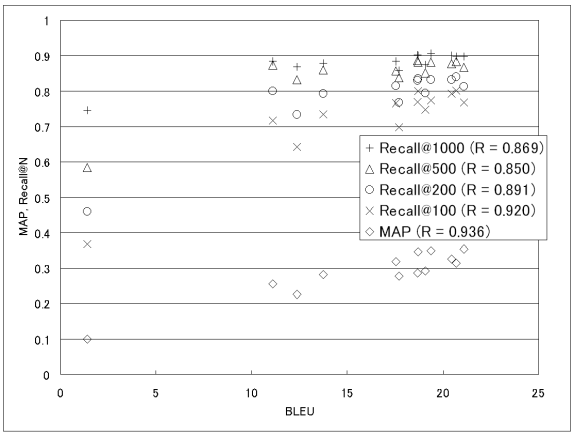

Figure 12. Relationship between BLEU and IR evaluation measures for English–Japanese extrinsic evaluation.

without performing retrieval experiments.

In line with the literature for information retrieval, we used the two-sided paired t-test for statistical testing, which investigates whether differences in MAP values are meaningful or simply because of chance [9]. Table 4 shows the results, in which the rows are sorted according to the values of "Extrinsic BLEU" in Table 3. In Table 4, "$>$" and "$\gg$" indicate that the difference between two groups in MAP value was significant at the 5% and 1% levels, respectively, and "—" indicates that the difference between two groups in MAP value was not significant. In Table 4, comparing CLPR results, not all differences in MAP values were significant.

The extent to which the BLEU value should be improved to achieve a statistically significant improvement in MAP value is a scientific question. To answer this question, Figure 13 shows the relationship between the difference in BLEU value and the level of statistical significance of the MAP value. In Figure 13, each bullet point corresponds to a comparison of two groups. The bullet points are classified into three clusters according to the level of statistical significance for MAP, namely "Not significant", "Significant at 5%", and "Significant at 1%". The y-axis denotes the difference between the two groups' BLEU values. The y-coordinate of each bullet point was calculated from the values for "Extrinsic BLEU" in Table 3.

By comparing the three clusters in Figure 13, we deduce the difference in BLEU value should be more than 10 to safely achieve the 1% level of significance for MAP values. In the dry run [6], this threshold was 9 and thus the result was almost the same as the formal run. However, it is not clear to what extent these observations can be generalized. Because the values for BLEU and MAP can depend on the data set used, further investigation is needed to clarify the relationship between improvements in BLEU and MAP.

Table 4. Results of t-test for MAP: "$\gg$": 1%, "$>$": 5%, "—": not significantly different

	Moses	NiCT-ATR	NTT	NAIST-NTT	tori	KLE	MIBEL	Kyoto-U	HIT2	FDU-MCandWI	tsbmt	TH
HCRL	—	—	—	—	—	$>$	$>$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
Moses		—	—	—	—	—	—	—	—	—	$\gg$	$\gg$
NiCT-ATR			—	—	—	—	$>$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NTT				—	—	$>$	$>$	$>$	$\gg$	$\gg$	$\gg$	$\gg$
NAIST-NTT					—	—	—	$>$	$>$	$>$	$\gg$	$\gg$
tori						—	—	—	—	$>$	$\gg$	$\gg$
KLE							—	—	—	—	$>$	$\gg$
MIBEL								—	—	—	$>$	$\gg$
Kyoto-U									—	—	$>$	$\gg$
HIT2										—	—	$\gg$
FDU-MCandWI											—	$\gg$
tsbmt												$\gg$

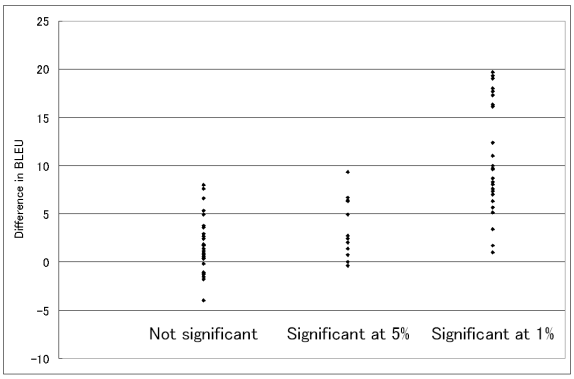

Figure 13. Relationship between difference in BLEU and statistical significance of MAP.

6 Conclusion

To aid research and development in machine translation, we have produced a test collection for Japanese/English machine translation. To obtain a parallel corpus, we extracted patent documents for the same or related inventions published in Japan and the United States.

Our test collection includes approximately 2 000 000 sentence pairs in Japanese and English, which were extracted automatically from our parallel corpus. These sentence pairs can be used to train and evaluate machine translation systems. Our test collection also includes search topics for cross-lingual patent retrieval, which can be used to evaluate the contribution of machine translation to retrieving patent documents across languages.

Using this test collection, we performed the Patent Translation Task at the Seventh NTCIR Workshop. Our task comprised a dry run and a formal run, in which research groups submitted their results for the same test data.

This paper has described the results and knowledge obtained from the evaluation of the formal run submissions. Our research is the first significant exploration into utilizing patent information for the evaluation of machine translation. Our test collection will be publicly available for research purposes after the final meeting of the Seventh NTCIR Workshop.

In the Eighth NTCIR Workshop, we plan to perform the Patent Translation Task using a larger document set and explore outstanding issues in the current research.

Acknowledgments

This research was supported in part by the Collaborative Research of the National Institute of Informatics.

References

[1] A. Fujii and T. Ishikawa. Document structure analysis for the NTCIR-5 patent retrieval task. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 292–296, 2005.

[2] A. Fujii, M. Iwayama, and N. Kando. The patent retrieval task in the fourth NTCIR workshop. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 560–561, 2004.

[3] A. Fujii, M. Iwayama, and N. Kando. Test collections for patent retrieval and patent classification in the fifth NTCIR workshop. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 671–674, 2006.

[4] A. Fujii, M. Iwayama, and N. Kando. Introduction to the special issue on patent processing. *Information Processing & Management*, 43(5):1149–1153, 2007.

[5] A. Fujii, M. Iwayama, and N. Kando. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Pro-*

ceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pages 359–365, 2007.

[6] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Toward the evaluation of machine translation using patent information. In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas*, pages 97–106, 2008.

[7] S. Higuchi, M. Fukui, A. Fujii, and T. Ishikawa. PRIME: A system for multi-lingual patent retrieval. In *Proceedings of MT Summit VIII*, pages 163–167, 2001.

[8] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. Evaluating patent retrieval in the third NTCIR workshop. *Information Processing & Management*, 42(1):207–221, 2006.

[9] E. M. Keen. Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28(4):491–502, 1992.

[10] P. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, 2004.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[12] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*, 1994.

[13] M. Utiyama and H. Isahara. A Japanese-English patent parallel corpus. In *Proceedings of MT Summit XI*, pages 475–482, 2007.