Toward the Evaluation of Machine Translation Using Patent Information

Atsushi Fujii
Graduate School of Library,
Information and Media Studies
University of Tsukuba

Masao Utiyama
National Institute of Information
and Communications Technology

Mikio Yamamoto
Graduate School of Systems
and Information Engineering
University of Tsukuba

Takehito Utsuro
Graduate School of Systems
and Information Engineering
University of Tsukuba

Abstract

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. This paper describes our test collection, methods for evaluating machine translation, and preliminary experiments.

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 (Fujii et al., 2004; Fujii et al., 2006; Fujii et al., 2007b; Iwayama et al., 2006). In the Sixth NTCIR Workshop (Fujii et al., 2007b), 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 (Fujii et al., 2007a), 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. (2001) 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 de-

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

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

termined to organize a machine translation task for patents ("the Patent Translation Task") in the Seventh NTCIR Workshop (NTCIR-7). Because NTCIR-7 has not yet been completed, this paper describes the evaluation method in the Patent Translation Task and the result of preliminary experiments.

2 Overview of the Patent Translation Task

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

Second, the organizers provide the groups with a test data set of sentences in either Japanese or English. Each group is 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 evaluate the submission from each group. We use both intrinsic and extrinsic evaluation methods. In the intrinsic evaluation, we independently use both the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), which was proposed as an automatic evaluation measure for MT, and human judgment. In the extrinsic evaluation, we investigate 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 reuse these search topics for the evaluation of the MT. We also analyze 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 execute the above three steps in both a preliminary trial and the final evaluation, using the terms "dry run" and "formal run", respectively. If a problem is found in the dry run, we modify the task procedure for the formal run. We have finished analyzing the evaluation results for the dry run. The submission deadline for the formal run has now passed, but the evaluation of the submissions has not been finished. In this paper, we describe only the dry run.

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

3 Intrinsic Evaluation

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 (Fujii et al., 2007b), 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.

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 an unexamined Japanese 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

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

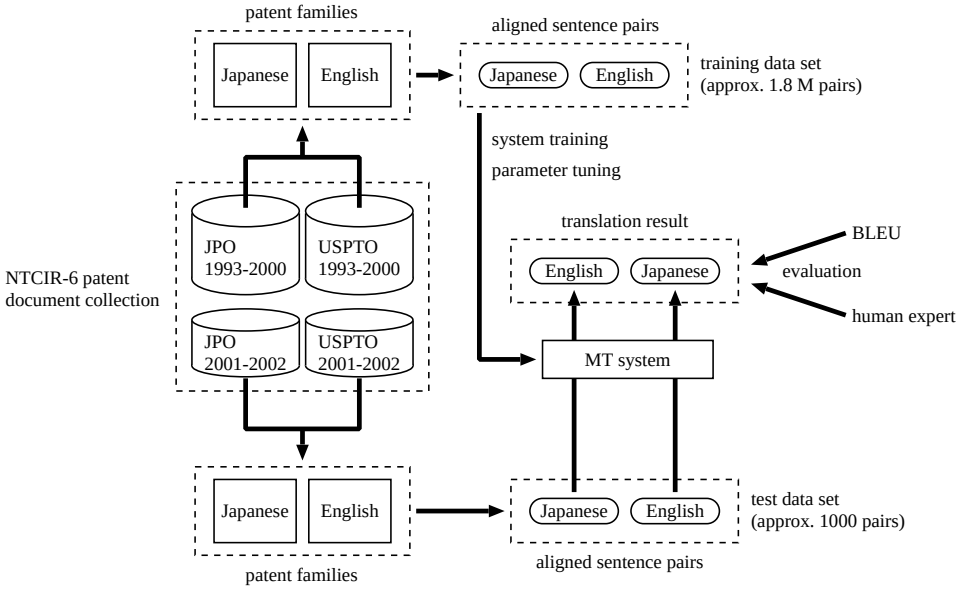

Figure 1: Overview of the intrinsic evaluation.

(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.

JPO patents. 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. Therefore, for these fields, we used a method (Utiyama and Isahara, 2007) 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 is used for both the dry run and the formal run.

The sentence pairs extracted from patent docu-

ments 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 are used to evaluate Japanese–English (or English–Japanese) MT. Unlike the training data set, we use different test sets for the dry run and the formal run.

To evaluate translation results submitted by participating groups, we independently use 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 use its counterpart sentence as the reference translation. We also ask several human experts to produce a reference translation for each test sentence independently, to enhance the objectivity of the evaluation by BLEU.

For human judgments, we ask 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 select 100 test sentences for human judgment purposes. We analyze the relationship between the evaluation by BLEU and the evaluation by human judgment.

The procedure for the dry run is 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 uses 822 test sentences, whereas the formal run uses 1381 test sentences.
- To compute the value of BLEU in the intrinsic evaluation, we used only a single reference. The reference sentence of a test sentence is 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.

4 Extrinsic Evaluation

In the extrinsic evaluation, we investigate the contribution of MT to CLPR. Each group is requested to machine translate search topics from English into Japanese. Each of the translated search topics is used to search a patent document collection in Japanese for the relevant documents. The evaluation results for CLPR are 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.

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 (Fujii et al., 2006), 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

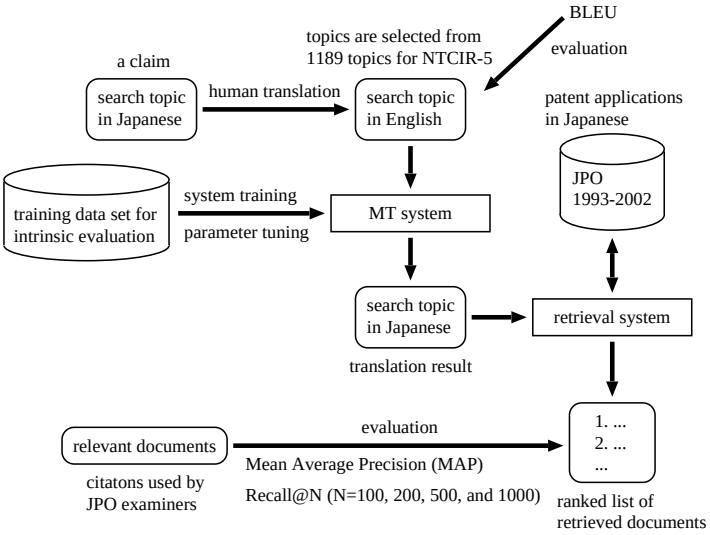

Figure 3: Overview of the extrinsic evaluation.

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 reuse 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.

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 non-patent 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 is requested to machine translate the search topics, the retrieval is performed by the organizers. As a result, we can standardize the retrieval system and the contribution of each group can be compared in terms of the translation accuracy alone. In addition, for most of the participating groups, who are research groups in natural

```
<TOPIC><NUM>1048</NUM>
<FDATE>19950629</FDATE>
<CLAIM>A milk-derived calcium-containing
composition comprising an inorganic salt
mainly composed of calcium obtained by bak-
ing 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.

language processing, the retrieval of 10 years' worth of patent documents is not a trivial task.

We use a system that was used in the NTCIR-5 Patent Retrieval Task (Fujii and Ishikawa, 2005) as the standard retrieval system. This system uses Okapi BM25 (Robertson et al., 1994) as the retrieval model and the International Patent Classification to restrict the number of retrieved documents.

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

As evaluation measures for CLPR, we use the Mean Average Precision (MAP), which has fre-

quently 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 use BLEU as an evaluation measure, for which we use the source search topics in Japanese as the reference translations.

In principle, for the extrinsic evaluation we can 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 select 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 Dry Run

5.1 Overview

The schedule of the dry run was as follows.

- 2008.01.10: Release of the intrinsic test data
- 2008.01.15: Release of the extrinsic test data
- 2008.02.14: Submission deadline
- 2008.02.29: Release of the evaluation results

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

As explained in Sections 2–4, the dry run involved three types of evaluation: Japanese–English intrinsic evaluation, English–Japanese intrinsic evaluation, and extrinsic evaluation. The numbers of groups participating in these evaluation types were nine, eight, and six, respectively. Because participation in the dry run was not mandatory, some groups intending to participate in the formal run did not submit their results for the dry run.

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	64.8	193
Intrinsic English	5	31.4	109
Extrinsic English	7	140.2	449

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 and 5.3.

In this paper, we anonymize the names of participating groups, with the name of each group being replaced by a capital letter, such as “A” or “B”. The names of participating groups will be disclosed at the NTCIR-7 final meeting in December 2008.

5.2 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 (S)”, “rule-based (R)”, and “example-based (E)” methods. The columns “BLEU” and “Human” denote the values for BLEU and human rating, respectively. Although the value for BLEU was calculated using all 822 test sentences, the value for human rating was averaged over the 100 sentences selected for human judgment purposes. The score with respect to adequacy and fluency, which are denoted as “Adequacy” and “Fluency”, respectively, ranges from 1 to 5. The value for human rat-

Table 2: Results of J–E intrinsic evaluation (Method: S = statistical, R = rule, E = example).

Group	Method	BLEU	Human	Adequacy	Fluency
A	S	23.29	3.36	1.76	1.6
B	S	23.14	3.4	1.78	1.62
C	S	19.54	3.18	1.68	1.5
D	R	18.66	5.49	2.95	2.54
E	S	18.27	3.46	1.79	1.67
F	R	17.20	5.58	3.01	2.57
G	E	15.55	3.46	1.86	1.6
H	S	8.32	2.37	1.21	1.16
I	S	1.05	2	1	1

ing, which is the sum of "Adequacy" and "Fluency", ranges from 1 to 10. The rows in Table 2, each of which corresponds to the result of a single group, are sorted according to the values for BLEU.

As shown in Table 2, groups that used a statistical method, such as "A", "B", and "C", tended to obtain large BLEU values, compared to groups that used rule-based and example-based methods. The difference in BLEU values between groups using a statistical method is due to the decoder and the size of the data used for training purposes. Groups that were not able to process the entire training data used a fragment of the training data.

Figure 5 shows each group's BLEU values with a 95% confidence interval; the values were computed by a bootstrap method (Koehn, 2004) using 1000-fold resampling. In Figure 5, three clusters are observable according to y-axis values: {A,B}, {C,D,E,F,G}, and {H,I}. Presumably, the groups in the first cluster (i.e., "A" and "B") used the entire training data and a sophisticated decoder. Looking at the second cluster, which ranges from "C" to "G", we see that the different methods (statistical, rule-based, and example-based) led to comparable BLEU results. The groups in the third cluster did not fully utilize resources. According to their system descriptions, group "H" used only a fragment of the training data and group "I" used the IBM Model-3.

Figure 6 shows the relationship between BLEU values and human rating. With regard to human rating, groups "D" and "F", which independently used a rule-based method, outperformed the other groups. While the difference between "D" and "F" in human rating is marginal, the difference between these

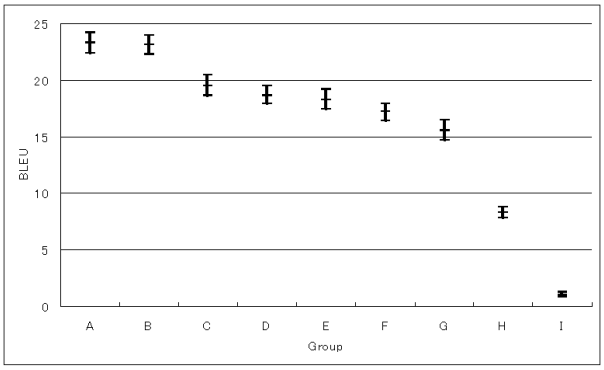

Figure 5: BLEU for Japanese–English intrinsic evaluation with a 95% confidence interval.

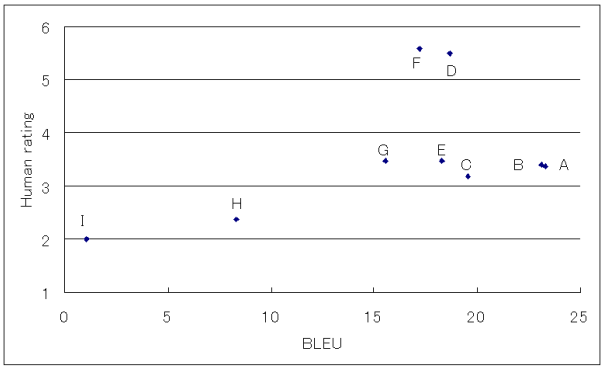

Figure 6: Relationship between BLEU and human rating for Japanese–English intrinsic evaluation.

two and the other groups is noticeable. According to their system descriptions, group "D" used a commercial MT system. We presume that group "F" also used a commercial MT system.

In Table 2, the values for "Adequacy" and "Fluency" for each group are almost the same. In other words, there was no method that is particularly effective for either adequacy or fluency.

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. In this section we focus on "Intrinsic"; we will elaborate on "Extrinsic" in Section 5.3. We use the same group names for both Tables 2 and 3.

In Table 3, group "J", which did not participate in the Japanese–English intrinsic evaluation, achieved

Table 3: Results of E–J intrinsic/extrinsic evaluation (Method: S = statistical, R = rule-based, E = example-based).

		Intrinsic				Extrinsic		
Group	Method	BLEU	Human	Adequacy	Fluency	BLEU	MAP rigid	MAP relax
J	S	28.69	3.87	2.03	1.84	—	—	—
A	S	26.65	3	1.61	1.39	21.29	0.5312	0.5922
C	S	25.06	2.93	1.56	1.37	18.28	0.4886	0.5413
B	S	24.11	3.16	1.67	1.49	17.84	0.4727	0.5207
E	S	21.63	2.95	1.53	1.42	13.55	0.4847	0.5316
G	E	19.58	3.08	1.61	1.47	10.63	0.4294	0.4624
F	R	15.58	5.4	2.99	2.41	11.96	0.3836	0.41
I	S	9.42	2.39	1.3	1.09	—	—	—
Mono	—	—	—	—	—	—	0.9000	0.9983

the best BLEU value. Groups "D" and "H" in Table 2 did not participate in the English–Japanese intrinsic evaluation.

Figure 7, which uses the same notation as Figure 5, shows the values of BLEU with a 95% confidence interval for each group. Comparing Figures 5 and 7, the relative superiority of "B" to "C" and that of "F" to "G" were reversed. Figure 8, which uses the same notation as Figure 6, shows the relationship between values for BLEU and human rating. As in Figure 6, group "F", which used a rule-based method, noticeably outperformed the other groups with respect to human rating.

In summary, statistical methods usually outperformed other methods with respect to BLEU, and rule-based methods outperformed other methods with respect to human rating, irrespective of the source and target languages.

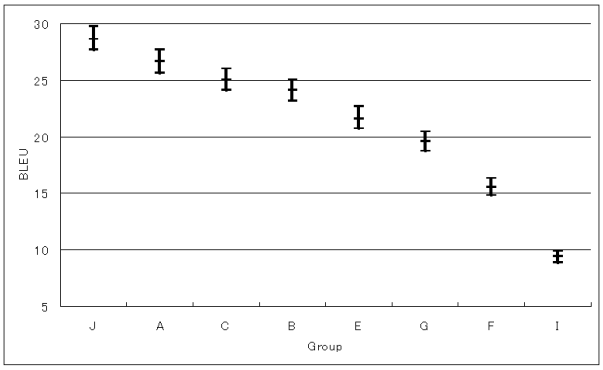

Figure 7: BLEU for English–Japanese intrinsic evaluation with a 95% confidence interval.

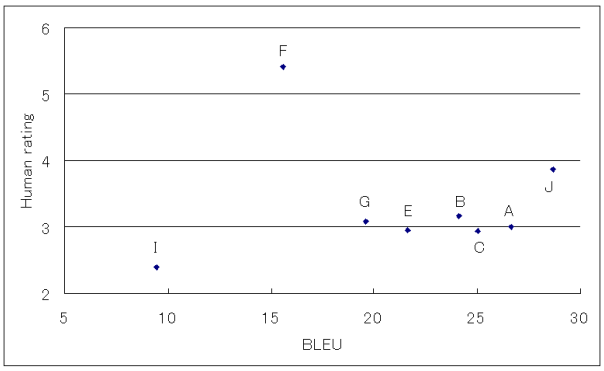

Figure 8: Relationship between BLEU and human rating for English–Japanese intrinsic evaluation.

5.3 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. From among the groups participating in the English–Japanese intrinsic evaluation, groups "J" and "I" did not participate in the extrinsic evaluation. According to their system descriptions, all groups participating in the extrinsic evaluation used the same method for the English–Japanese intrinsic evaluation.

In Table 3, "BLEU" in "Extrinsic", which denotes the BLEU values for the extrinsic evaluation, is different from "BLEU" in "Intrinsic". 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.

The difference between "MAP rigid" and "MAP

relax" is in the definition of correct answers for each search topic. For "MAP rigid", we used only relevant documents as the correct answers. However, for "MAP relax", we also used partially relevant documents as correct answers. Among the 100 search topics, 10 topics were associated with only partially relevant documents. Therefore, values in "MAP rigid" were calculated using only the other 90 search topics, whereas values in "MAP relax" were calculated using all 100 search topics.

In Table 3, the row "Mono" shows the results for monolingual retrieval, which is an upper bound to the retrieval effectiveness for CLPR. Because of lack of space, we focus on MAP and do not discuss Recall@N here. The relative superiority of groups in Recall@N was almost the same as in MAP.

Looking at Table 3, the relative superiority of the six groups with respect to BLEU was the same for the extrinsic evaluation as it was for the intrinsic evaluation, except for groups "F" and "G". 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.

Comparing "MAP rigid" to "MAP relax" in Table 3, the relative superiority between groups with respect to MAP is the same, irrespective of the correct-answer definition. The relative superiority in BLEU value was almost the same as that in MAP value. However, there was little correlation between the relative superiority in MAP value and that in human rating. Group "F", which received the best human rating, obtained the smallest MAP value.

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 (Keen, 1992). Table 4 shows the results for "MAP rigid", in which "$>$" 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 "Mono" with each of the CLPR results, all the differences in MAP values were significant at the 1% level. However, when comparing CLPR results, not all differences were significant. The difference was significant at the 1% level only for "A vs G", "A vs F", and "C vs F".

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

	A	C	E	B	G	F
Mono	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
A		—	—	$>$	$\gg$	$\gg$
C			—	—	—	$\gg$
E				—	—	$>$
B					—	$>$
G						—

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 9 shows the relationship between the difference in BLEU value and the level of statistical significance of the MAP value. In Figure 9, each bullet point corresponds to a comparison of two groups. The bullet points are classified into six clusters according to the evaluation type (intrinsic or extrinsic) and the level of statistical significance for MAP (1%, 5%, and not significant), with ovals showing the clusters. The y-axis denotes the difference between the two groups' BLEU values. The y-coordinate of each bullet point was calculated from the values in Table 3.

By considering the "intrinsic significant 1%" and "extrinsic significant 1%" clusters in Figure 9, we deduce the difference in BLEU value should be at least 6 to achieve a 1% level of significance for MAP values. However, because the y-coordinates of some bullet points in other clusters also range from 6 to 8,

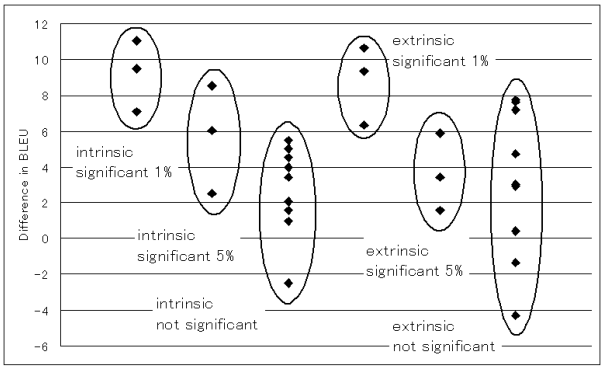

Figure 9: Relationship between difference in BLEU and statistical significance of MAP.

the difference in BLEU value should be more than 9 to safely achieve the 1% level of significance for MAP values.

At the same time, it is not clear to what extent this observation can be generalized. Because the number of groups participating in the extrinsic evaluation is small and the values for BLEU and MAP can depend on the data set used, further investigation is needed during the formal run to clarify the relationship between improvements in BLEU and 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 are performing the Patent Translation Task at the Seventh NTCIR Workshop. Our task comprises a dry run and a formal run, in which research groups submit their results for the same test data.

This paper has described the results and knowledge obtained from the evaluation of the dry 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.

Future work will include the evaluation of the formal run, for which we have already received submissions from a larger number of groups.

Acknowledgments

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

References

Atsushi Fujii and Tetsuya Ishikawa. 2005. 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.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2004. 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.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2006. 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.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007a. Introduction to the special issue on patent processing. *Information Processing & Management*, 43(5):1149–1153.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007b. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 359–365.

Shigeto Higuchi, Masatoshi Fukui, Atsushi Fujii, and Tetsuya Ishikawa. 2001. PRIME: A system for multilingual patent retrieval. In *Proceedings of MT Summit VIII*, pages 163–167.

Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. 2006. Evaluating patent retrieval in the third NTCIR workshop. *Information Processing & Management*, 42(1):207–221.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

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

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. In *Proceedings of MT Summit XI*, pages 475–482.