Exploiting Patent Information for the Evaluation of Machine Translation

Atsushi Fujii
University of Tsukuba

Masao Utiyama
National Institute of Information and Communications Technology

Mikio Yamamoto
University of Tsukuba

Takehito Utsuro
University of Tsukuba

Abstract

We have produced a test collection for machine translation (MT). Our test collection includes approximately 2 000 000 sentence pairs in Japanese and English, which were extracted from patent documents and can be used to train and evaluate MT systems. Our test collection also includes search topics for cross-lingual information retrieval, to evaluate the contribution of MT to retrieving patent documents across languages. We performed a task for MT at the NTCIR workshop and used our test collection to evaluate participating groups. This paper describes scientific knowledge obtained through our task.

Keywords: Machine translation, Patent information, Cross-lingual information retrieval

1 Introduction

Reflecting the rapid growth in the use of multilingual corpora, a number of data-driven Machine Translation (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 it is difficult to explore SMT with respect to Japanese.

However, patent documents can alleviate this data scarcity problem. Higuchi et al. (2001) used "patent families" as a parallel corpus for extracting 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. We organized a machine translation task for patents in the Seventh NTCIR Workshop (NTCIR-7). This paper describes our task, namely "the Patent Translation Task" and patent-specific and general scientific knowledge for MT obtained through this task.

We used both intrinsic and extrinsic evaluation methods. In the intrinsic evaluation, we used both the Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), which had been proposed as an automatic evaluation measure for MT, and human judgment. In the extrinsic evaluation, we evaluated the contribution of the MT to Cross-Lingual Information Retrieval (CLIR). In the Patent Retrieval Task at NTCIR-5 (Fujii et al., 2006), aimed at CLIR, search topics in Japanese were translated into English by human experts. We reused these search topics for the evaluation of the MT. We 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[1] and the IWSLT Workshop[2], is a distinctive feature of our research. We executed a preliminary trial and the final evaluation, using the terms "dry run" and "formal run", respectively. This paper describes only the formal run.

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

2 Intrinsic Evaluation

2.1 Evaluation Method

In the Patent Retrieval Task at NTCIR-6 (Fujii et al., 2007), 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.

From these document sets, we extracted patent families. In a patent family applied for under the Paris Convention, the member documents of a patent family are assigned the same priority number, and thus patent families can be identified automatically. 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 (Utiyama and Isahara, 2007) to 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. Thus, 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.

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.

To evaluate translation results submitted by participating groups, we 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 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 for producing multiple references in Section 2.2.

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. Thus, we did not produce additional references for the English–Japanese intrinsic evaluation.

For tokenization purposes, we used "ChaSen"[3] and the tokenizer in "ACL2007 the 2nd workshop on SMT"[4] for Japanese and English sentences, respectively. For human judgments, we asked experts to evaluate each translation result based on fluency and adequacy, using a five-point rating. Because manual evaluation for all submitted translations would be expensive, we randomly selected 100 test sentences for human judgment purposes.

2.2 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.

[3] http://chasen.naist.jp/hiki/ChaSen/

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

S600 According to our initial plan, we randomly selected 600 sentences from the 1381 Japanese test sentences, 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. Thus, 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, which we call "S300". However, we found that E6 had used an RBMT system for translation purposes.

In summary, all the reference translations for S600 and the reference translations by E6 for S300 are potentially 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 fully rely on RBMT systems, we can use the other reference translations with caution.

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. While 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. While 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. While this value is potentially influenced by RBMT systems, we can use as many reference translations and test sentences as possible.

In Section 4.2, we use terms "SRB", "MRB300", and "MRB600" for explaining the result of the Japanese–English intrinsic evaluation. 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 due to the budget constraint.

3 Extrinsic Evaluation

In the extrinsic evaluation, we evaluated the contribution of MT to CLIR. Each group was requested to machine translate search topics from English into Japanese. Each of the translated search topics was used to search patent documents in Japanese for the relevant documents.

In the Patent Retrieval Task at NTCIR-5, "invalidity search" was performed. The purpose was to search a Japanese patent collection, which is the collection described in Section 2, 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. In addition, with the aim of CLIR, 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. The following is an example of a topic claim translated into English.

> 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.

The claim used as the target of invalidation 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.

Although each group was requested to machine translate the search topics, the retrieval was performed by the organizers. Thus, we were able to standardize the retrieval system and the contribution of each group was evaluated in terms of the translation accuracy alone. We used a system that had also been used in NTCIR-5 (Fujii and Ishikawa, 2005) as the standard retrieval system. 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 accuracy.

As evaluation measures for CLIR, 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 the source search topics in Japanese were used as the reference translations.

In principle, we were able to use all of the 1189 search topics for NTCIR-5. However, because the length of a single claim is much longer than that of an ordinary sentence, we selected a subset of the search topics for the extrinsic evaluation. If we use search topics for which the average precision of the monolingual retrieval is small, the average precision of CLIR can be so small that it is difficult to distinguish the contributions of participating groups to CLIR. 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

Because we intended to use approximately 100 topics, we selected the first 100 topics for the dry run and the next 124 topics for the formal run.

4 Evaluation in the Formal Run

4.1 Overview

As explained in Sections 2–3, 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 in these evaluation types were 14, 12, and 12, respectively. Each group was allowed one month to translate the test data. To produce a baseline performance, the organizers submitted a result produced by Moses (Koehn and others, 2007), in which default parameters were used.

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. 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 description of their system, which will be used to analyze the evaluation results in Sections 4.2–4.4.

Table 1: Length of test inputs.

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

4.2 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", "MRB300", and "MRB600" in "BLEU" denote the values for each BLEU type. The numbers of test sentences used for these BLEU types are 1381, 300, and 600, respectively.

For human judgment, three experts independently

Table 2: Results of J–E intrinsic evaluation.

Group	Method	BLEU			Human
		SRB	MRB300	MRB600	
NTT	SMT	27.20	35.93	43.72	3.30
Moses *	SMT	27.14	36.02	43.40	3.18
(MIT)	SMT	27.14	37.31	44.69	3.40
NAIST-NTT	SMT	25.48	34.66	41.89	3.04
NiCT-ATR	SMT	24.79	32.29	39.40	2.78
KLE	SMT	24.49	33.59	40.20	2.94
(tsbmt)	RBMT	23.10	37.51	48.02	3.88
tori	SMT	22.29	27.92	35.02	3.01
Kyoto-U	EBMT	21.57	29.35	35.49	3.10
(MIBEL)	SMT	19.93	27.84	32.99	2.74
HIT2	SMT	19.48	29.33	33.60	2.86
JAPIO	RBMT	19.46	32.62	41.77	3.86
TH	SMT	15.90	24.20	28.72	2.13
FDU-MCandWI	SMT	9.55	19.94	20.27	2.08
(NTNU)	SMT	1.41	2.48	2.63	1.06

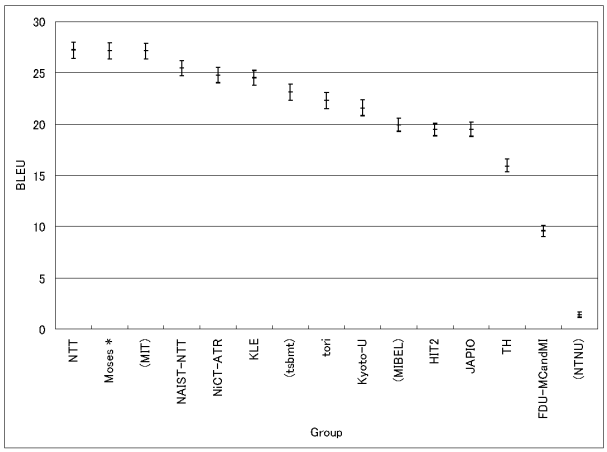

Figure 1: BLEU (SRB) with a 95% confidence interval for J–E intrinsic evaluation.

evaluated the same 100 sentences. The value for "Human", which is the average of adequacy and fluency, 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. "Moses *" denotes results for the submission produced by the organizers.

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 SMT 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. Groups that were not able to process the entire training data used a fragment of the training data.

Figure 1 shows each group's SRB values with a 95% confidence interval, calculated by a bootstrap method (Koehn, 2004) using 1000-fold resampling. In Figure 1, the SRB values for the top three groups are comparable and greater than those for the other groups, with a 95% confidence. The result for Moses, which had not been developed for Japanese, was in the top cluster, and Moses was effective for Japanese–English MT.

We discuss the values of BLEU obtained by multiple references. The value of BLEU generally increases, as the number of reference translations increases. This tendency was 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 also be observed in Figures 2 and 3, which use the same notation as Figure 1, and show the values of BLEU with a 95% confidence interval for MRB300 and MRB600, respectively. In Figure 2, the BLEU values for tsbmt and MIT are comparable and these groups outperformed the other groups. However, in Figure 3, tsbmt outperformed MIT and achieved the best BLEU value.

A reason for the above observations is that tsbmt and JAPIO used a RBMT method. Because as explained in Section 2.2, 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. How-

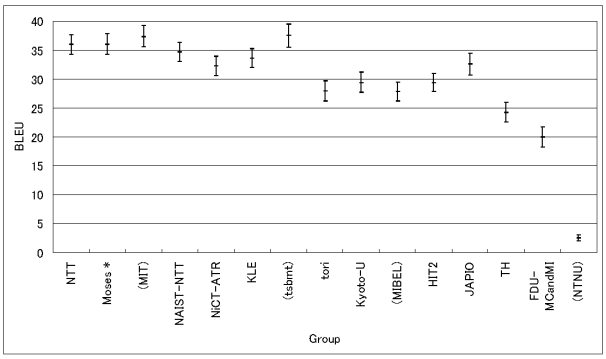

Figure 2: BLEU (MRB300) with a 95% confidence interval for J–E intrinsic evaluation.

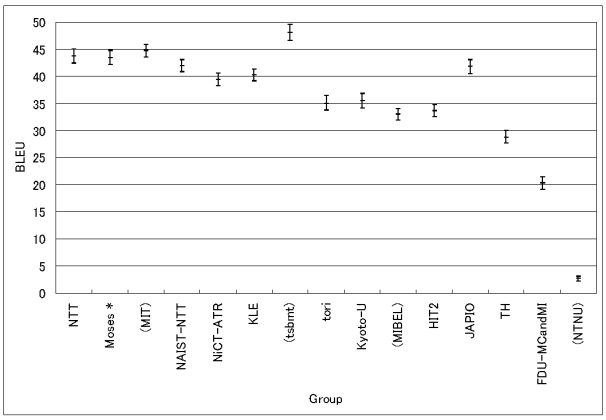

Figure 3: BLEU (MRB600) with a 95% confidence interval for J–E intrinsic evaluation.

ever, 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 4 graphs the value for "Human" in Table 2, in which the order of groups is the same as Figures 1–3. In Figure 4, tsbmt and JAPIO, which were not effective in SRB, 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 (Callison-Burch et al., 2006; Koehn and Monz, 2006). Figures 5 and 6 graph the value for adequacy and fluency, respectively. Although the relative superiority of the groups was almost the same in Figures 5 and 6, differences of the groups are more noticeable in Figure 5.

To further analyze this tendency, Figure 7 shows

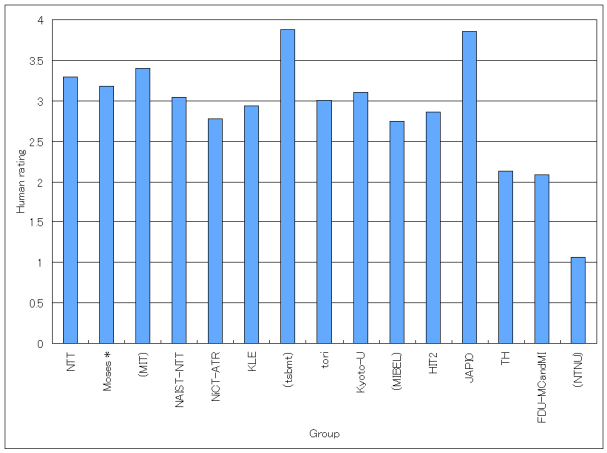

Figure 4: Human rating for J–E intrinsic evaluation.

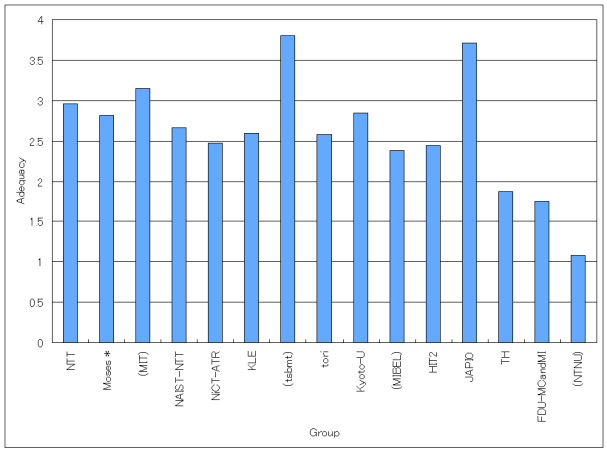

Figure 5: Adequacy for J–E intrinsic evaluation.

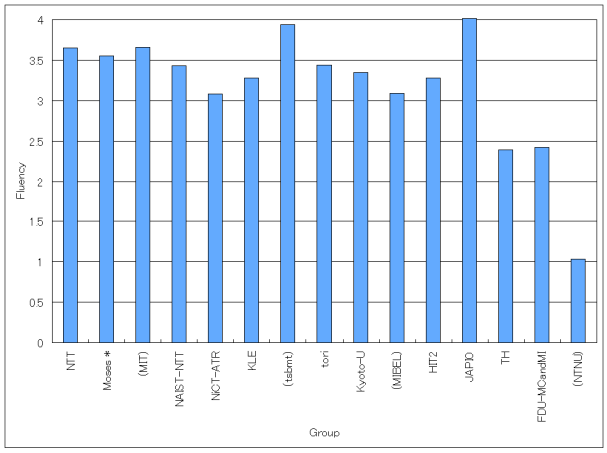

Figure 6: Fluency for J–E intrinsic evaluation.

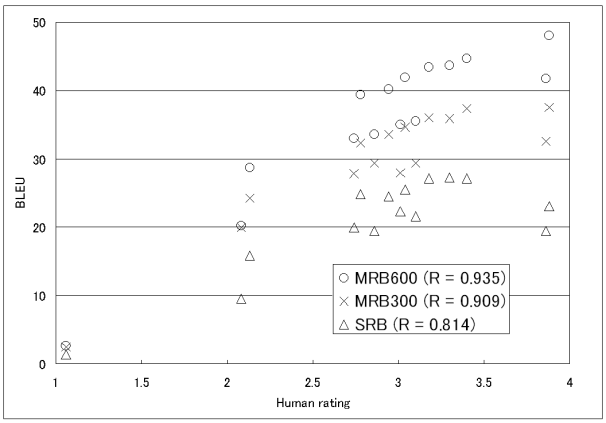

Figure 7: Relationship between BLEU and human rating for J–E intrinsic evaluation.

Table 3: Results of E–J int/ext evaluation.

		Intrinsic		Extrinsic	
Group	Method	BLEU	Human	BLEU	MAP
Moses *	SMT	30.58	3.30	20.70	.3140
HCRL	SMT	29.97	—	21.10	.3536
NiCT-ATR	SMT	29.15	2.89	19.40	.3494
NTT	SMT	28.07	3.14	18.69	.3456
NAIST-NTT	SMT	27.19	—	20.46	.3248
KLE	SMT	26.93	—	19.07	.2925
tori	SMT	25.33	—	17.54	.3187
(MIBEL)	SMT	23.72	—	18.67	.2873
HIT2	SMT	22.84	—	17.71	.2777
(Kyoto-U)	EBMT	22.65	2.48	13.75	.2817
(tsbmt)	RBMT	17.46	3.60	12.39	.2264
FDU-MCandWI	SMT	10.52	—	11.10	.2562
TH	SMT	2.23	—	1.39	.1000
Mono	—	—	—	—	.4797

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.

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 (Melamed et al., 2003). In such a case, while human judgments are not reusable, we need only reference translations, which are reusable, for evaluating MT methods. We also calculated the values of R for each BLEU type in terms of adequacy and fluency although these values are not shown in Figure 7. For adequacy, the values of R for SRB, MRB300, and MRB600 were 0.733, 0.846, and 0.887, respectively. For fluency, the values of R for SRB, MRB300, and MRB600 were 0.864, 0.940, and 0.951, respectively. This implies that BLEU is highly correlated with fluency more than adequacy.

4.3 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. We focus on "Intrinsic" and we will elaborate on "Extrinsic" in Section 4.4.

Because of time and budget constraints, we imposed two restrictions on the English–Japanese evaluation. First, human judgments were performed 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 RBMT, outperformed the other groups with respect to human rating. Figure 8, which uses the same notation as Figure 1, shows the values of BLEU with a 95% confidence interval for each group. In Figure 8, the relative superiority of top groups was different from that in Figure 1.

4.4 Extrinsic Evaluation

The "Extrinsic" column in Table 3 shows the results the values for BLEU and MAP for each group in the extrinsic evaluation. According to their system de-

scriptions, all the groups used the same method for both the intrinsic and extrinsic evaluations. As explained in Section 3, the English search topics for the extrinsic evaluation are human translations of search topics in Japanese. To calculate values for BLEU in the extrinsic evaluation, we used Japanese search topics 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. We found that the correlation coefficient between the values for BLEU in the intrinsic and extrinsic evaluation types was 0.964. BLEU for translating claims in patent applications is highly correlated with BLEU for translating other fields in patent applications, despite claims being described in a patent-specific language.

In Table 3, the row "Mono" shows the results for monolingual retrieval, which is an upper bound to the effectiveness for CLIR. The best MAP for CLIR by HCRL is 0.3536, which is 74% of that for Mono. We also used Recall@N as an evaluation measure for information retrieval (IR). We calculated the correlation coefficient between BLEU in the extrinsic evaluation and each IR evaluation measure. We found that the value of R for MAP was 0.936 whereas the values of R for Recall@N were below 0.9, irrespective of the value of N. Thus, we can use BLEU to predict the contribution of MT to CLIR with respect to MAP, without performing retrieval experiments. At the same time, human rating did not correlate with MAP because as in Table 3 tsbmt, whose MAP was the lowest, outperformed the other groups with respect to human rating.

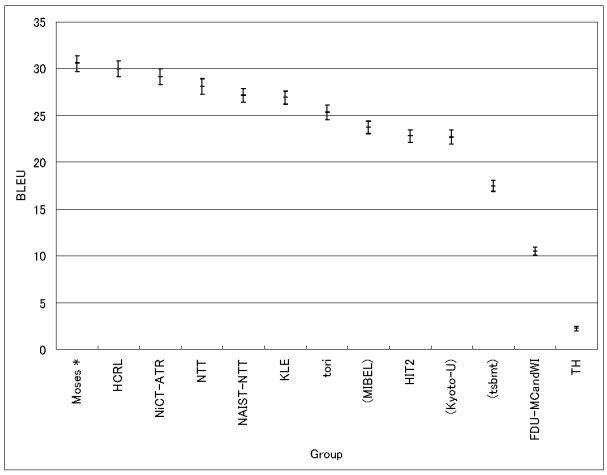

Figure 8: BLEU with a 95% confidence interval for E–J intrinsic evaluation.

5 Conclusion

We have produced a test collection for MT from patent documents, which is publicly available for research purposes. We have also obtained scientific knowledge through the evaluation for MT. Future work includes exploring appropriate evaluation measures for MT using different languages and different patent-related applications.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *EACL*, pages 249–256.

Atsushi Fujii and Tetsuya Ishikawa. 2005. Document structure analysis for the NTCIR-5 patent retrieval task. In *NTCIR*, pages 292–296.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2006. Test collections for patent retrieval and patent classification in the fifth NTCIR workshop. In *LREC*, pages 671–674.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2007. Overview of the patent retrieval task at the NTCIR-6 workshop. In *NTCIR*, pages 359–365.

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

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121.

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003*, pages 61–63.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

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