Overview of the Patent Translation Task at the NTCIR-8 Workshop

Atsushi Fujii
Tokyo Institute of Technology

Masao Utiyama
National Institute of Information

Mikio Yamamoto
University of Tsukuba

Takehito Utsuro
University of Tsukuba

Terumasa Ehara
Yamanashi Eiwa College

Hiroshi Echizen-ya
Hokkai-Gakuen University

Sayori Shimohata
Oki Electric Industry CO., Ltd.

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 Eighth 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 3 200 000 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. In addition, our test collection includes machine translation results and their evaluation scores determined by human experts, which can be used to propose automatic evaluation methods for machine translation. This paper describes our test collection, methods for evaluating machine translation, and evaluation results for research groups participated in our task.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Patent information, Machine translation, Cross-lingual retrieval, Automatic evaluation, 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 had been performed repeatedly [2, 3, 5, 9]. 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. In NTCIR-7, we organized the Patent Translation Task [6, 7, 8] and used "patent families" as a parallel corpus for Japanese and English. 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.

In NTCIR-7, each participating group was requested to machine translate test sentences, for which Japanese and English were used as either source or target language. The organizers evaluated the translation result from each group

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

using 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[2] and the IWSLT Workshop[3], is a distinctive feature of our research.

In NTCIR-8, we also organized the Patent Translation Task. A major difference from NTCIR-7 is that we enhanced the size of a training data set. While the training and test data sets were produced from patent documents over 10 years in NTCIR-7, the data sets for NTCIR-8 were produced from 15 years of patent documents. In addition, we classified tasks associated with MT into the following three subtasks.

- Translation Subtask

  This subtask is almost the same as performed in NTCIR-7. Each participating group was requested to machine translate test sentences, for which Japanese and English were used as either source or target language. The translation results were evaluated by both the intrinsic and extrinsic evaluation methods. However, unlike NTCIR-7 we did not use human judgment for the intrinsic evaluation.

- Evaluation Subtask

  The purpose of this subtask is to explore automatic evaluation methods for MT. Although evaluation for MT using human judgments is expensive, automatic evaluation methods potentially alleviate this problem. We used machine translation results and their evaluation scores determined by human experts in NTCIR-7 as a test data set. Each group was requested to reproduce the human evaluation scores by an automatic method as much as possible.

- Retrieval Subtask

  This subtask is associated with the extrinsic evaluation in the Translation Subtask. While the purpose of the Translation Subtask is to machine translate test sentences, the purpose of the Retrieval Subtask is to search patent documents in Japanese related to a search topic in English. The test search topics were the same as in the extrinsic evaluation for the Translation Subtask. Each group was requested to perform either monolingual or cross-lingual retrieval. In the monolingual retrieval, the groups in the Retrieval Subtask were allowed to use search topics translated into Japanese by groups in the Translation Subtask.

Sections 2 and 3 explain the Translation and Evaluation Subtasks, respectively. However, we do not explain the Retrieval Subtask, for which no group submitted results.

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

2. TRANSLATION SUBTASK

2.1 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 [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.

Although we used the above document sets in NTCIR-7, we also used JPO and USPTO patent documents published during 2003–2007 for NTCIR-8. From these document sets, we automatically extracted patent families, as performed in NTCIR-7. 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.

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 1993–2005 to produce the training data set, we used patent documents published during 2006–2007 to produce the test data set.

The training data set has approximately 3 200 000 Japanese–English sentence pairs, which is one of the largest collections available for Japanese and English MT. 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. The numbers of test sentences for Japanese–English and English–Japanese were 1251 and 1119, respectively.

To evaluate translation results submitted by participating groups, we used BLEU. 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.

2.2 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 2 depicts the process flow of the extrinsic evaluation. We explain the entire process in terms of Figure 2.

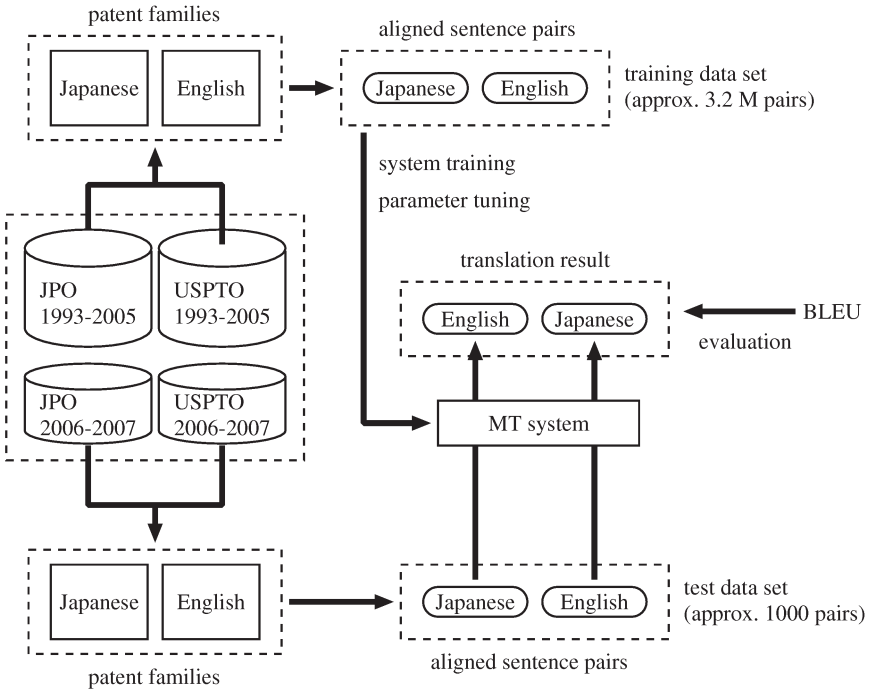

Figure 1: Overview of the intrinsic evaluation.

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-6 [5], invalidity search was performed. The purpose was to search a Japanese patent collection in 1993–2002, which is the collection described in Section 2.1, 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 1685 search topics.

For each search topic, one or more citations (i.e., prior arts) that were used for the rejection were used as relevant documents. In addition, with the aim of CLPR, these search topics were translated by human experts into English. In the extrinsic evaluation at NTCIR-8, we reused these search topics.

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-5/6 Patent Retrieval Task [1] as the standard retrieval system. This system 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 1685 search topics produced in NTCIR-6. 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. Thus, in practice we selected a subset of the search topics for NTCIR-8.

If we use search topics for which the average precision (AP) of the monolingual retrieval is small, the AP of CLPR methods can be so small that it is difficult to distinguish the contributions of participating groups to CLPR. At the same time, we discarded such topics for which the AP is 0.4 or more because these are "easy" topics. Therefore, we sorted

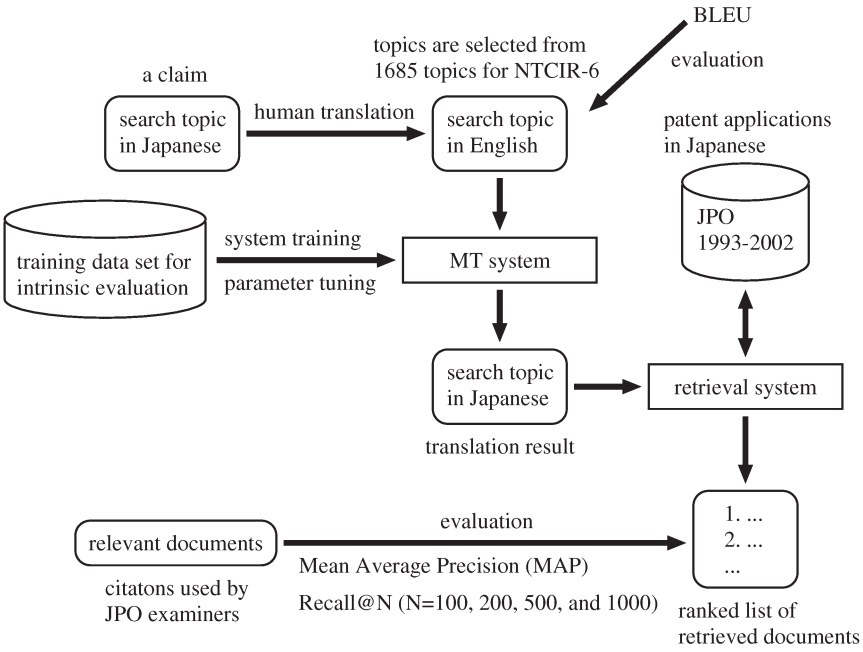

Figure 2: Overview of the extrinsic evaluation.

the 1685 search topics according to the AP of monolingual retrieval using the standard retrieval system, discarded topics for which the AP is 0.4 or more, and selected the first 91 topics for the remaining topics.

2.3 Evaluation Results

Table 1 shows the results of the Japanese–English intrinsic evaluation, in which the column "BLEU" denotes the BLEU values and the columns "BLEU-LOW" and "BLEU-HIGH" denote a 95% confidence interval calculated by a bootstrap method [10] using 1000-fold resampling.

In Table 1, we denote the names of runs submitted after the deadline 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.

Table 2, which uses the same notation as Table 1, shows the results for the English–Japanese intrinsic evaluation.

Table 3 shows the results of the extrinsic evaluation, and includes the values for BLEU, MAP, and Recall@N for each group. The BLEU values in Table 3 are different from those in Table 2. As explained in Section 2.2, the English search topics used for the extrinsic evaluation are human translations of search topics in Japanese. To calculate values for BLEU in Table 3, we used these search topics in Japanese as the reference translations. Additionally, in Table 3 the row "Monolingual" shows the results for monolingual retrieval, which is an upper bound to the effectiveness for CLPR.

Table 1: Results for J–E intrinsic evaluation.

Run	BLEU	BLEU-LOW	BLEU-HIGH
DCU-1	27.61	26.71	28.49
DCU-2	26.86	25.96	27.73
DCU-3	24.01	23.26	24.74
DCU-4	20.68	19.95	21.33
DCU-5	23.91	23.20	24.62
DCU-6	18.27	17.58	19.00
DCU-7	23.82	23.13	24.55
DCU-8	26.51	25.70	27.31
DCU-9	23.30	22.53	24.03
EIWA-1	34.30	33.35	35.25
KLE-1	27.75	26.85	28.69
KYOTO-1	21.23	20.47	22.02
NICT-1	30.32	29.40	31.12
NICT-2	30.14	29.27	31.01
NICT-3	24.96	24.16	25.77
NICT-4	25.79	24.96	26.61
(TORI-1)	25.65	24.71	26.54
TORI-2	21.56	20.60	22.57
TUTA-1	22.66	21.84	23.44
TUTA-2	26.27	25.46	27.01
Moses *	29.08	28.17	29.91

3. EVALUATION SUBTASK

Automatic evaluation technology for a machine translation result is one of the valuable and developing regions in the field. We already have the data including source sentences, reference sentences, test sentences (machine translated sentences) and human evaluated results for test sentences that are used in the NTCIR-7 patent translation task. Using these data, we organize the evaluation (AE) subtask in the NTCIR-8 patent translation task.

Participants evaluate the test sentences and submit the evaluation results to the organizers. It should be noticed that this evaluation will be sentence by sentence. It means that the AE subtask is not through a document level evaluation but through a sentence level evaluation. The organizers calculate correlation coefficients between each participant's evaluation results and human evaluation results and will send them back to the participants.

Table 2: Results for E–J intrinsic evaluation.

Run	BLEU	BLEU-LOW	BLEU-HIGH
DCU-1	33.03	32.00	34.04
DCU-2	32.50	31.47	33.49
DCU-3	29.53	28.60	30.46
DCU-4	26.38	25.55	27.19
DCU-5	26.38	25.52	27.21
DCU-6	27.93	27.05	28.68
DCU-7	30.08	29.25	31.00
DCU-8	26.22	25.40	27.10
DCU-9	27.23	26.36	28.08
DCU-10	26.21	25.39	27.03
DCU-11	26.45	25.61	27.36
DCU-12	30.53	29.62	31.49
DCU-13	26.83	26.02	27.64
DCU-14	1.27	1.03	1.50
KLE-1	29.18	28.41	29.94
KYOTO-1	24.13	23.39	24.97
NICT-1	35.37	34.46	36.29
NICT-2	35.87	34.99	36.80
TORI-1	26.02	25.14	26.89
TUTA-1	27.82	27.07	28.54
TUTA-2	28.50	27.66	29.29
Moses *	35.27	34.30	36.19

The data for the AE subtask are provided from the results of the NTCIR-7 patent translation task. In the AE subtask, the direction of machine translation is only from Japanese to English. Number of a reference sentence for one test datum is one. Then, the AE subtask is for the single reference evaluation.

Training data provided to the participants consist of five files. The first file includes 100 source Japanese sentences. The second file includes 100 reference English sentences. The third file includes 1100 test sentences which are outputs of 11 machine translation systems participating in the NTCIR-7 patent translation task, dry run. The fourth file includes 1100 adequacy values which correspond to the test sentences. The fifth file includes 1100 fluency values which correspond to the test sentences.

Test data provided to the participants consist of three files. The first file includes 100 source Japanese sentences. The second file includes 100 reference English sentences. The third file includes 1200 test sentences which are outputs of 12 machine translation systems participating in the NTCIR-7 patent translation task, formal run. The examples of the test data are as in Figures 3–5.

All submitted evaluation results of test sentences are meta-evaluated by the following method. We have made a human evaluation data for all of the test sentences of the test data. Measures used in the human evaluation are "adequacy" and "fluency". We use "median" of three human evaluation results which are provided by three human raters, i.e, one evaluation result from each human rater. Then, we compute correlation coefficients between submitted evaluation data and human evaluation data. We use Pearson's correlation and Spearman's rank correlation in this phase. So, we obtain the following four meta-evaluation results.

- Pearson's correlation coefficients to the adequacy data
- Spearman's rank correlation coefficients to the adequacy data
- Pearson's correlation coefficients to the fluency data
- Spearman's rank correlation coefficients to the fluency data

The participants can also obtain human evaluation data after they submitted their evaluation data.

We have only one participant for the AE subtask from HCU. The results are shown in Tables 4 and 5. In the tables, "Avg" means the average value for the correlation coefficients for the 12 test systems and "All" means the correlation coefficient for all data of the 12 test systems.

We only have one participant for the AE subtask. The correlation coefficients between automatic evaluation and human evaluations are distributed between 0.21 and 0.30 which are relatively low. We may need to set up the multi reference evaluation subtask.

We plan to provide all human evaluation data composed in the NTCIR-7 and NTCIR-8 patent translation task for the future researches of the automatic evaluation. They will be provided from NII's Informatics Research Data Repository.

4. 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 3 200 000 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.

In addition, our test collection includes machine translation results and their evaluation scores determined by human experts, which can be used to explore automatic evaluation methods for machine translation.

Using this test collection, we performed the Patent Translation Task at the Eighth NTCIR Workshop. This paper has described the results and knowledge obtained from the evaluation of the formal run submissions. Our test collection will be publicly available for research purposes after the final meeting of the Eighth NTCIR Workshop.

5. 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.

Table 3: Results for E–J extrinsic evaluation.

Run	BLEU	BLEU-LOW	BLEU-HIGH	MAP	R@100	R@200	R@500	R@1000
DCU-1	23.48	21.91	25.08	0.1759	0.5278	0.5919	0.6813	0.7373
DCU-2	23.71	22.11	25.27	0.1831	0.5254	0.5867	0.6775	0.7261
DCU-3	22.35	20.66	24.00	0.2082	0.549	0.6378	0.7037	0.7633
DCU-4	24.00	22.07	25.78	0.2072	0.5499	0.6261	0.723	0.7795
KLE-1	18.98	17.42	20.65	0.1707	0.5101	0.5865	0.6852	0.7411
KYOTO-1	17.25	15.89	18.57	0.1909	0.5258	0.5904	0.6955	0.7576
NICT-1	24.62	22.93	26.34	0.1914	0.5579	0.6466	0.7437	0.7829
TORI-1	13.35	11.97	14.69	0.1186	0.4094	0.4977	0.5873	0.6549
TUTA-1	21.21	19.61	22.85	0.1845	0.5639	0.6514	0.7255	0.8041
TUTA-2	21.88	20.20	23.67	0.1995	0.5652	0.6427	0.743	0.7877
Moses *	24.01	22.55	25.50	0.1943	0.5701	0.6626	0.7408	0.7917
Monolingual	—	—	—	0.3025	0.6784	0.7332	0.7956	0.8478

[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 *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, 2007.

[6] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the patent translation task at the NTCIR-7 workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, pages 389–400, 2008.

[7] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Producing a test collection for patent machine translation in the seventh NTCIR workshop. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1796–1799, 2008.

[8] 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.

[9] 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.

[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.

プリント機構は、感光体ドラム１１を備えている。
定着ローラ３９によって定着処理が施された転写紙は、図示しない排出機構を介して排出トレイ４３上に送られる。
レーザユニット２８から出力されるレーザ光は、感光体ドラム１１に照射される。

Figure 3: Top three source sentences of the test data.

the printing mechanism has a photoconductor drum 11 .
the transfer paper and fixed by a fixing roller 39 thus processed is discharged through a discharge tray 43 is sent to a mechanism (not shown) .
the laser beam outputted from the laser unit 8 irradiates the photosensitive drum 11 .

Figure 4: Top three reference sentences of the test data.

the printing mechanism comprises the photosensitive drum 11 .
the copy paper which has been subjected to fixing processing by the fixing roller 39 is fed onto a discharge tray 43 through a discharge mechanism (not shown) .
a photosensitive drum 11 is irradiated with the laser light outputted from the laser unit 28 .

Figure 5: Top three test sentences of the test data.

Table 4: Correlation coefficients to the adequacy data.

Run	Pearson		Spearman	
	Avg	All	Avg	All
HCU-1	0.2992	0.2463	0.2712	0.2234

Table 5: Correlation coefficients to the fluency data.

Run	Pearson		Spearman	
	Avg	All	Avg	All
HCU-1	0.2608	0.2285	0.2486	0.2126