Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop

Isao Goto
National Institute of Information and Communications Technology
igoto@nict.go.jp

Bin Lu
City University of Hong Kong / Hong Kong Institute of Education
lubin2010@gmail.com

Ka Po Chow
Hong Kong Institute of Education
kpchow@ied.edu.hk

Eiichiro Sumita
National Institute of Information and Communications Technology
eiichiro.sumita@nict.go.jp

Benjamin K. Tsou
Hong Kong Institute of Education / City University of Hong Kong
btsou@ied.edu.hk

ABSTRACT

This paper gives an overview of the Patent Machine Translation Task (PatentMT) at NTCIR-9 by describing the test collection, evaluation methods, and evaluation results. We organized three patent machine translation subtasks: Chinese to English, Japanese to English, and English to Japanese. For these subtasks, we provided large-scale test collections, including training data, development data and test data. In total, 21 research groups participated and 130 runs were submitted. We conducted human evaluations for adequacy and acceptability, as well as automatic evaluation.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Machine translation

General Terms

Experimentation

Keywords

Patent machine translation, Human evaluation, NTCIR

1. INTRODUCTION

Patent information is important for communities all around the world, and there is a significant **practical need** for translations in order to understand patent information written in foreign languages and to apply for patents in foreign countries. Patents constitute one of the **challenging domains** for machine translation because patent sentences can be quite long and contain complex structures. The Patent Machine Translation Task (PatentMT), while cast in a framework of friendly competition, has the ultimate goal of fostering scientific cooperation. In this context, the organizers have proposed a research task and an open experiment infrastructure for the scientific community working on machine translation research. This task builds on the two previous patent translation tasks [7, 8].

There are two additions to this task that were not contained in the previous tasks.:

- Chinese to English

  We added a Chinese-to-English subtask. The need for translating patent information out of Chinese is increasing. In order to meet this need, we have built a large scale parallel corpus and distributed it to the participants.

- Acceptability evaluation

  We conducted a human evaluation for *acceptability* that aims at a more practical evaluation than adequacy. This human evaluation reveals how many test sentences can be understood from their translations.

The goals of PatentMT are:

- To develop challenging and significant practical research into patent machine translation.
- To investigate the performance of state-of-the-art machine translation in terms of patent translations involving Japanese, English, and Chinese.
- To compare the effects of different methods of patent translation by applying them to the same test data.
- To create publicly-available parallel corpora of patent documents and human evaluations of the MT results for patent information processing research.
- To drive machine translation research, which is an important technology for cross-lingual access of information written in unknown languages.

This paper is organized as follows: Section 2 explains the task design, Section 3 gives the participants and their submissions, Section 4 describes the human evaluation results, Section 5 shows the validation of human evaluations, Section 6 gives a meta-evaluation of the automatic evaluations, and Section 7 concludes the paper.

2. TASK DESIGN

We organized three patent machine translation subtasks: Chinese to English (CE), Japanese to English (JE), and English to Japanese (EJ). Participants chose the subtasks that

Table 1: Test collection

Use	Subtask	Contents
Training	CE	1 million patent parallel sentence pairs
		Monolingual patent corpus in English covering a span of 13 years (1993-2005)
	JE	Approximately 3.2 million patent parallel sentence pairs
		Monolingual patent corpus in English covering a span of 13 years (1993-2005)
	EJ	Approximately 3.2 million patent parallel sentence pairs
		Monolingual patent corpus in Japanese covering a span of 13 years (1993-2005)
Development	All	2,000 patent parallel sentence pairs
		Context documents
Test	All	2,000 patent test sentences
		Context documents
		2,000 reference sentences

they wished to participate in. The training data and test data were provided to participants. Participants translated the test data using their machine translation systems and submitted the translations to the PatentMT organizers. The PatentMT organizers evaluated the submitted translations and returned the evaluation results to the participants.

2.1 Test Collection

The test collections consist of training data, development data, test data, context documents, and reference data. The data was mostly from patent description sentences. (Patent documents consist of a title, abstract, claim, and description.)

2.1.1 Test collection for the CE subtask

The Chinese-English test collection was chosen from a large Chinese-English bilingual parallel corpus of sentence pairs [16]. The sets of training, development and test data are built in the following manner.

First, we divided our Chinese-English bilingual corpus into two sub-corpora with the following criterion: those sentence pairs from patents published on or prior to 2005 were used for the training data, while those on or after 2006 were used for the development and test data. Since the publication dates of English and Chinese corresponding patents may be different, the publication date of the English version was used.

We then sorted the list of patents randomly by assigning a random number to each patent pair and then sorted the patents according to this random number. Using this order, we examined each pair of patents and counted the number of sentences that aligned into pairs within it, then added these pairs to the data set until the required number of sentence pairs have been collected: 1 million sentence pairs for the training data set, and 2,000 sentence pairs each for both sets of development and test data.

For the test data set, we also manually removed similarly patterned sentence pairs in order to ensure that the test set would not contain a large quantity of sentences with patterns that closely resembled each other, since that could undermine sentence variety and limit the ability to evaluate the translation of sentences from a wider array of perspectives.

In all the data sets, we indicated the ID of the patent the sentences originate from, and included context data for the development and test data. The context data includes the titles and the international patent classification (IPC) code for the Chinese patents. For the development data set, the texts for the abstracts, claims and descriptions of both the Chinese and English patents are included, while in the test data set only the Chinese texts are included.

2.1.2 Test collection for the JE and EJ subtasks

We built new test data consisting of 2,000 sentences. We used the same training and development data as the NTCIR-8 Patent Translation Task [8]. The contents for the test collection are shown in Table 1.

The parallel data for training, development, and the test and reference candidates was automatically extracted from patent families in Japanese and English. Patent families are one of the methods for applying for patents in more than one country. They are sets of patent applications under the Paris Convention that use the same priority number. We used unexamined Japanese patent applications published by the Japan Patent Office (JPO) for patent sentences in Japanese and patent grant data published by the United States Patent and Trademark Office (USPTO) for patent sentences in English.

The training data was built from patent documents published between 1993 and 2005. We also provided monolingual patent documents in the target side language (patent grant data published by the USPTO or Japanese patent applications published by the JPO).

The development data consists of 2,000 sentence pairs built from patent documents published in 2006 and 2007. The patent documents containing the development data were provided as context documents for the development data.

The test data was built as follows: We randomly selected parallel sentences from a portion of the automatically-built 2006 and 2007 patent parallel sentence pairs. We manually judged whether the sentence pairs were correct translations, then selected 2,000 correct sentence pairs as the test data and their reference data. We provided the patent documents that the test sentences were extracted from as context documents for the test data. The context data includes the international patent classification (IPC) code.

2.2 Evaluation Methodology

We conducted human evaluations. Human evaluation was the primary evaluation, and we used human judgments to validate the automatic metrics because we contend, the same as Workshop on Statistical Machine Translation 2011 [1], that automatic evaluations are imperfect and are not reliable enough, especially when the system architectures are different.

Human evaluations were carried out by paid evaluation experts, using the criteria of adequacy and acceptability. For each criterion, three evaluators evaluated 100 sentences per system. The three evaluators evaluated different sentences. Thus, 300 sentences were evaluated per system.

In this evaluation, the evaluators looked at a source sentence and its translations to evaluate the results and did not use a reference. This is because we allowed slight differences between the source sentences and their reference sentences

when judging whether a translation was suitable for use as test data from an automatically-built parallel corpus, under the condition that the principle meaning of the two was the same (e.g. a difference in the presence or absence of a conjunction at the beginning of a sentence was permitted.)

2.2.1 Adequacy

We conducted a 5-scale (1 to 5) adequacy evaluation. The main purpose of the adequacy evaluation is to compare between the systems.

There are some criteria for adequacy. White [26] defined how much information from a fragment of a reference sentence is contained in the translation results. They insisted that fragmentation is intended to avoid biasing results in favor of linguistic compositional approaches (which may do relatively better on longer, clause level strings) or statistical approaches (which may do better on shorter strings not associated with syntactic constituency). This evaluation cannot evaluate whether the sentence meaning is correct or not because simply containing all of the fragments of the reference information does not guarantee the sentence meaning is correct. The NTCIR-7 Patent Translation Task [7] conducted adequacy evaluations using a criterion based on the degree of preservation of sentence-level meaning instead of the degree of fragments of the reference information contained.

We thought that the degree of sentence-level meaning preservation was better than that of fragments of reference information contained for the evaluation of translation quality. However, since the cost of checking sentence meanings is high, we evaluated quality based on relative comparison between systems with consideration of clause-level meanings for adequacy.

The instructions for the adequacy criterion are given in Appendix A. The adequacy grades were determined by a relative comparison of all the translated results.

The systems were ranked based on adequacy using the average system scores.

2.2.2 Acceptability

We conducted a 5-scale acceptability evaluation as shown in Fig 1. The main purpose of an acceptability evaluation is to clarify the percentage of translated sentences whose source sentence meanings can be understood from randomly selected test sentences. Acceptability is an evaluation of the sentence-level meaning. The acceptability criterion used in this evaluation aims more at practical evaluation than adequacy. For example, if a requirement of a translation system is that the source sentence meaning can be understood, translations of C or higher are useful, but if the requirement is that the source sentence meaning can be understood and the sentence is grammatically correct, then only translations of A or higher are useful. We can then know how many sentences are useful for each requirement. An adequacy criterion cannot answer these requirements.

Acceptability also contains an evaluation of fluency, since the differences in grading from C to AA are dependent on fluency. If the adequacy of a translation is very low, then the translation is nonsense even if its fluency is high. If the integrated evaluation score is calculated by averaging the adequacy score and the fluency score, then those translations could be overvalued. Acceptability avoids this problem, allowing us to consider fluency.

The instructions for the acceptability criterion are shown in Appendix B.

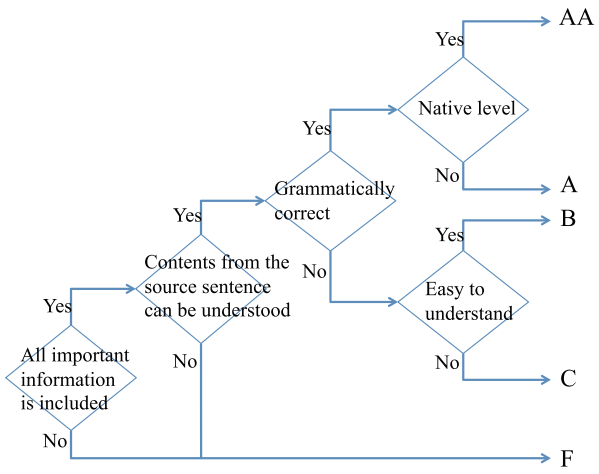

Figure 1: Acceptability.

We ranked the systems based on acceptability using pairwise comparison, which will now be explained. The *pairwise score* for a system A reflects how frequently it was judged to be better than or equal to other systems. Suppose there are five systems to be compared. For each input sentence, system A is included in four pairwise comparisons (against the other systems). System A is rewarded 1.0 for each of the comparisons in which system A is ranked the highest of the two. System A is rewarded 0.5 for each of the comparisons in which system A ties. System A's score is the total rewarded score in the pairwise comparisons, divided by the total number of pairwise comparisons involving system A.

There is a reason why the average score of acceptability for system ranking was not used. Here, we assume that the differences between the grades are based on general usability. It is important to be able to understand the contents from the source sentence. There is a large difference in usability between F and C. On the other hand, at the A-level, although the translations are at a non-native level, the contents from the source sentences can be understood and they are grammatically correct, having potential to be useful in many cases. Thus, it is thought that the difference of usability between A and AA is smaller than that between F and C. In addition, we think that useful grades depend on the specific usage. Therefore, it is difficult to give an appropriate score for each grade, so we avoided converting grades to scores and calculating averages.

2.2.3 Human Evaluation Procedure

We conducted human evaluation training before the main evaluation to normalize the evaluators' criteria. In the train-

Table 2: Schedule for PatentMT at NTCIR-9

Event	Date
Training corpus release	2011.1.5
Test data for JE/EJ release	2011.5.9
Test data for CE release	2011.5.19
Result submission for JE/EJ due	2011.5.22
Result submission for CE due	2011.6.1

Table 3: Participants and Subtasks Participated In

Group ID	Participant	Numbered Group ID		
		CE	JE	EJ
EIWA	Yamanashi Eiwa College[5]	G4	G01	
KLE	Pohang University of Science and Technology (POSTECH)[19]	G10	G05	G05
TORI	Tottori University[18]		G11	G08
RWTH	RWTH Aachen University[6]	G17	G10	
FRDC	Fujitsu R&D Center CO., LTD[30]	G5	G02	G02
NEU	Northeastern University[28]	G14	G08	
BUAA	Institute of Intelligent Information Processing, Beihang University[3]	G3		
UOTTS	The University of Tokyo[27]	G18	G12	G09
NCW	NTNU, NCCU, and WebGenie Information Ltd.[25]	G13		
ICT	Institute of Computing Technology, Chinese Academy of Sciences[29]	G7	G03	G03
BJTUX	Beijing Jiaotong University[12]	G2		G01
BBN	Raytheon BBN Technologies[17]	G1		
NAIST	Nara Institute of Science and Technology[14]		G07	
IBM	IBM Research[15]	G6		
KECIR	Shenyang Aerospace University	G9		
JAPIO	Japan Patent Information Organization[21]		G04	G04
KYOTO	Kyoto University[20]	G11	G06	G06
NTHU (IDEAS)	Institute for Information Industry, Chaoyang University of Technology and National Tsing Hua University[2]	G15		
NTT-UT	NTT Communication Science Labs. and the University of Tokyo[24]	G16	G09	G07
ISTIC	Institute of Scientific and Technical Information of China[9]	G8		
LIUM	University of Le Mans[23]	G12		

ing, all evaluators evaluated 100 translations and they held a meeting to determine common results for each subtask. The main evaluation was done after that. The common results produced at the training were used as the reference results for the main evaluation.

The instructions for the human evaluation procedure are shown in Appendix C.

2.2.4 *Automatic Evaluation*

We calculated automatic evaluation scores for three metrics: BLEU [22], NIST [4], and RIBES [11]. BLEU and NIST scores were calculated using NIST's `mteval-v13a.pl`[1]. RIBES scores were calculated using NTT's `RIBES.py` version 1.01[2]. Detailed procedures for the automatic evaluation are shown at the PatentMT web page[3].

2.3 Schedule

The task schedule is summarized in Table 2. We spent roughly four months for training and two weeks for translating.

3. PARTICIPANTS AND SUBMISSIONS

We received submissions from 21 groups. The number of groups for each subtask are: 18 for CE, 12 for JE, and 9 for EJ. Table 3 shows the participant organizations and the subtasks they participated in. The numbered Group ID shows the subtasks participated in, and is used for distinguishing groups in the participants' system description papers.

[1]http://www.itl.nist.gov/iad/mig/tools/
[2]http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html
[3]http://ntcir.nii.ac.jp/PatentMT/

In addition to the submissions from the participants, the organizers submitted results for baseline systems that consisted of 2 SMT systems, 5 commercial RBMT systems, and 1 online SMT system. The baseline systems are shown in Table 4. The SMT baseline systems consisted of publicly-available software, and the procedures for building the systems and translating using the systems were published on the PatentMT web page, so that the participants can build the SMT baseline systems and compare their results. The commercial RBMT systems and the Google online translation system were operated by the organizers. The translation results from the Google translation system were created by translating the test data via their web interface. We note that these RBMT companies and Google did not submit themselves. Since our objective is not to compare the commercial RBMT systems of companies who did not themselves participate, the SYSTEM-IDs of the commercial RBMT systems are anonymized in this paper.

Each participant is allowed to submit as many translated results ("runs") as they wish, but the submitted runs should be prioritized by the group. In this paper, we distinguish their runs using GROUP ID and a priority number connected by "-".

Some features from all of the submissions and their automatic evaluation scores are given in Appendix D. The resource information used by each run is indicated by:

- *Resource B*: The system used the bilingual training data provided by the organizers.
- *Resource M*: The system used the monolingual training data provided by the organizers.

Table 4: Baseline systems

SYSTEM-ID	SYSTEM	Type	CE	JE	EJ
BASELINE1	Moses' hierarchical phrase-based SMT system [10]	SMT	✓	✓	✓
BASELINE2	Moses' phrase-based SMT system [13]	SMT	✓	✓	✓
RBMTx	SYSTRAN 7 Premium Translator (Commercial RBMT)	RBMT	✓		
RBMTx	Huajian Multilingual EasyTrans version 3.0 (Commercial RBMT)	RBMT	✓		
RBMTx	The Honyaku 2009 premium patent edition (Commercial RBMT)	RBMT		✓	✓
RBMTx	ATLAS V14 (Commercial RBMT)	RBMT		✓	✓
RBMTx	PAT-Transer 2009 (Commercial RBMT)	RBMT		✓	✓
ONLINE1	Google online translation system	SMT	✓	✓	✓

- *Resource E*: The system used external knowledge other than data provided by the organizers or the system uses rule-based system.
- *Resource C*: The system uses context information.

4. HUMAN EVALUATION RESULTS

We evaluated adequacy at least for all of the first priority submissions. However, due to budget limitations, acceptability were evaluated for only selected systems.

4.1 Chinese to English

4.1.1 Adequacy Evaluation

Table 5 and Figure 2 show the results of the adequacy evaluation. Table 6 shows the results of a statistical significance test of the adequacy evaluation using a sign test.

From these results,we can observe the following:

- All the top systems are SMT systems. The top system, BBN-1, shows a significantly higher adequacy than the other systems.
- The adequacy score for Moses' hierarchical phrase-based SMT system (BASELINE1-1) is higher than that for Moses' phrase-based SMT system (BASELINE2-1) or the two RBMT baselines.
- Although both the two commercial RBMT systems (RBMT1-1 and RBMT2-1) and the Google online translation system (ONLINE1-1) did not have access to the training data, Google Translate shows better adequacy than the two commercial RBMT systems.

4.1.2 Acceptability Evaluation

Table 7 and Figure 3 show the results of the acceptability evaluation. Table 8 shows the results of the statistical significance test of the acceptability evaluation using a sign test.

From the results, we can see that the meaning in the source language could be understood (C-rank and above) for 79.7% of the translated sentences in the best-ranked system (BBN-1). This result significantly surpasses the others.

4.2 Japanese to English

4.2.1 Adequacy Evaluation

Table 9 and Figure 4 show the results of the adequacy evaluation. Table 10 shows the results of the statistical significance test of the adequacy evaluation using a sign test.

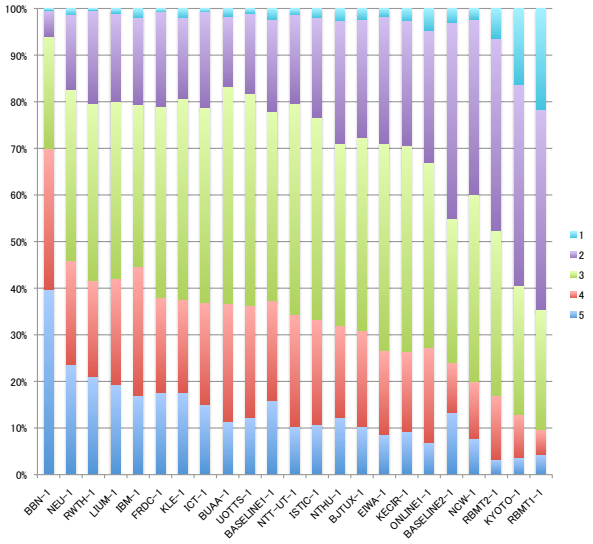

Figure 2: Results of CE adequacy.

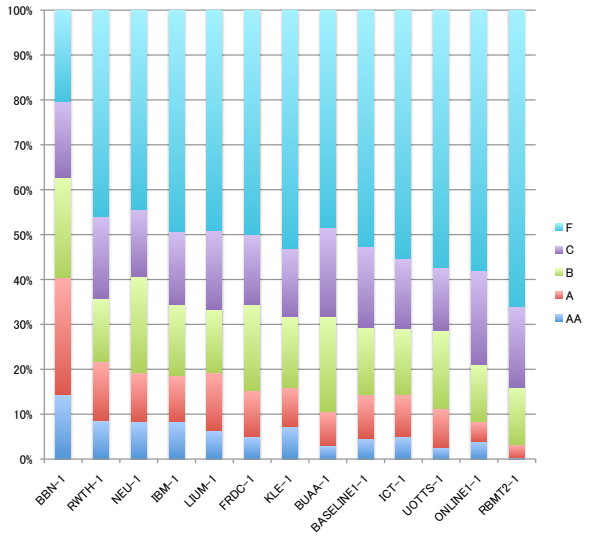

Figure 3: Results of CE acceptability.

The top five systems, JAPIO-1, RBMT-1, EIWA-1, RBMT3-1, RBMT2-1, are either commercial RBMT systems or sys-

Table 5: Results of CE adequacy

	Type	Resource			Average	Rate				
		B	M	E	score	5	4 or higher	3 or higher	2 or higher	1 or higher
BBN-1	SMT	✓	✓		4.033	0.397	0.700	0.940	0.997	1.000
NEU-1	SMT	✓	✓		3.510	0.237	0.460	0.827	0.987	1.000
RWTH-1	SMT	✓	✓		3.420	0.210	0.417	0.797	0.997	1.000
LIUM-1	SMT	✓	✓		3.403	0.193	0.420	0.800	0.990	1.000
IBM-1	SMT	✓	✓	✓	3.390	0.170	0.447	0.793	0.980	1.000
FRDC-1	SMT	✓	✓	✓	3.340	0.177	0.380	0.790	0.993	1.000
KLE-1	SMT	✓			3.340	0.177	0.377	0.807	0.980	1.000
ICT-1	SMT	✓	✓		3.300	0.150	0.370	0.787	0.993	1.000
BUAA-1	HYBRID	✓	✓		3.297	0.113	0.367	0.833	0.983	1.000
UOTTS-1	SMT	✓			3.293	0.123	0.363	0.817	0.990	1.000
BASELINE1-1	SMT	✓			3.290	0.160	0.373	0.780	0.977	1.000
NTT-UT-1	SMT	✓			3.230	0.103	0.343	0.797	0.987	1.000
ISTIC-1	HYBRID	✓	✓		3.187	0.107	0.333	0.767	0.980	1.000
NTHU-1	SMT	✓	?	✓	3.127	0.123	0.320	0.710	0.973	1.000
BJTUX-1	SMT	✓			3.113	0.103	0.310	0.723	0.977	1.000
EIWA-1	HYBRID	✓		✓	3.047	0.087	0.267	0.710	0.983	1.000
KECIR-1	SMT	✓	?	✓	3.037	0.093	0.263	0.707	0.973	1.000
ONLINE1-1	SMT			✓	2.967	0.070	0.273	0.670	0.953	1.000
BASELINE2-1	SMT	✓			2.893	0.133	0.240	0.550	0.970	1.000
NCW-1	SMT	✓			2.853	0.077	0.200	0.600	0.977	1.000
RBMT2-1	RBMT			✓	2.663	0.033	0.170	0.523	0.937	1.000
KYOTO-1	EBMT	✓			2.410	0.037	0.130	0.407	0.837	1.000
RBMT1-1	RBMT			✓	2.277	0.043	0.097	0.353	0.783	1.000

Table 6: Sign test of CE adequacy. "≫": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, "-": not significantly different.

	NEU-1	RWTH-1	LIUM-1	IBM-1	KLE-1	FRDC-1	ICT-1	BUAA-1	UOTTS-1	BASELINE1-1	NTT-UT-1	ISTIC-1	NTHU-1	BJTUX-1	EIWA-1	KECIR-1	ONLINE1-1	BASELINE2-2	NCW-1	RBMT2-2	KYOTO-1	RBMT1-1
BBN-1	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
NEU-1		-	>	-	>	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RWTH-1			-	-	-	-	-	-	>	-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
LIUM-1				-	-	-	-	-	-	-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
IBM-1					-	-	-	-	-	-	>	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
KLE-1						-	-	-	-	-	-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
FRDC-1							-	-	-	-	-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
ICT-1								-	-	-	-	>	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
BUAA-1									-	-	-	-	≫	>	≫	≫	≫	≫	≫	≫	≫	≫
UOTTS-1										-	-	≫	≫	>	≫	≫	≫	≫	≫	≫	≫	≫
BASELINE1-1											-	-	>	>	≫	≫	≫	≫	≫	≫	≫	≫
NTT-UT-1												-	-	-	≫	≫	≫	≫	≫	≫	≫	≫
ISTIC-1													-	-	>	≫	≫	≫	≫	≫	≫	≫
NTHU-1														-	-	-	≫	≫	≫	≫	≫	≫
BJTUX-1															-	-	≫	≫	≫	≫	≫	≫
EIWA-1																-	-	≫	≫	≫	≫	≫
KECIR-1																	-	≫	≫	≫	≫	≫
ONLINE1-1																		-	>	≫	≫	≫
BASELINE2-2																			-	>	≫	≫
NCW-1																				≫	≫	≫
RBMT2-2																					≫	≫
KYOTO-1																						-

tems that use commercial RBMT systems. From these results, the following are shown:

- The commercial RBMT systems had higher adequacies than the state-of-the-art SMT systems for patent machine translation from Japanese to English.
- The adequacy scores for normal hierarchical phrase-based SMT is slightly higher than that for normal phrase-based SMT in a comparison of BASELINE1-1 and BASELINE2-1.

4.2.2 Acceptability Evaluation

Table 11 and Figure 5 show the results of the acceptability evaluation. Table 12 shows the results of the statistical significance test of the acceptability evaluation using a sign test.

From the results, we can see that the source sentence meaning could be understood (C-rank and above) for 63% of the sentences in the best-ranked system using RBMT (JAPIO-1). For the best-ranked SMT system (NTT-UT-1), the source sentence meaning could be understood for 25%

Table 7: Results of CE acceptability

	Type	Resource			Pairwise	Rate				
		B	M	E	score	AA	A or higher	B or higher	C or higher	F or higher
BBN-1	SMT	✓	✓		0.744	0.143	0.403	0.627	0.797	1.000
RWTH-1	SMT	✓	✓		0.546	0.087	0.217	0.357	0.540	1.000
NEU-1	SMT	✓	✓		0.544	0.083	0.193	0.407	0.557	1.000
IBM-1	SMT	✓	✓	✓	0.513	0.083	0.187	0.343	0.507	1.000
LIUM-1	SMT	✓	✓		0.508	0.063	0.193	0.333	0.510	1.000
FRDC-1	SMT	✓	✓	✓	0.495	0.050	0.153	0.343	0.500	1.000
KLE-1	SMT	✓			0.491	0.073	0.160	0.317	0.470	1.000
BUAA-1	HYBRID	✓	✓		0.486	0.030	0.107	0.317	0.517	1.000
BASELINE1-1	SMT	✓			0.476	0.047	0.143	0.293	0.473	1.000
ICT-1	SMT	✓	✓		0.468	0.050	0.143	0.290	0.447	1.000
UOTTS-1	SMT	✓			0.441	0.027	0.113	0.287	0.427	1.000
ONLINE1-1	SMT			✓	0.422	0.040	0.083	0.210	0.420	1.000
RBMT2-1	RBMT			✓	0.365	0.003	0.033	0.160	0.340	1.000

Table 8: Sign test of CE acceptability. "≫": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, "-": not significantly different.

	RWTH-1	NEU-1	IBM-1	LIUM-1	FRDC-1	KLE-1	BUAA-1	BASELINE1-1	ICT-1	UOTTS-1	ONLINE1-1	RBMT2-1
BBN-1	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RWTH-1		-	-	>	≫	>	≫	>	≫	≫	≫	≫
NEU-1			-	-	>	>	>	≫	≫	≫	≫	≫
IBM-1				-	-	-	-	-	>	≫	≫	≫
LIUM-1					-	-	-	-	-	≫	≫	≫
FRDC-1						-	-	-	-	≫	≫	≫
KLE-1							-	-	-	≫	≫	≫
BUAA-1								-	-	-	≫	≫
BASELINE1-1									-	-	>	≫
ICT-1										>	>	≫
UOTTS-1											-	≫
ONLINE1-1												≫

Table 9: Results of JE adequacy

	Type	Resource			Average	Rate				
		B	M	E	score	5	4 or higher	3 or higher	2 or higher	1 or higher
JAPIO-1	RBMT			✓	3.667	0.297	0.590	0.807	0.973	1.000
RBMT1-1	RBMT			✓	3.530	0.273	0.530	0.763	0.963	1.000
EIWA-1	HYBRID	✓		✓	3.430	0.260	0.473	0.727	0.970	1.000
RBMT3-1	RBMT			✓	3.137	0.183	0.393	0.623	0.937	1.000
RBMT2-1	RBMT			✓	3.073	0.190	0.377	0.597	0.910	1.000
NTT-UT-1	SMT	✓			2.747	0.117	0.260	0.457	0.913	1.000
TORI-1	HYBRID	✓	✓	✓	2.730	0.100	0.217	0.490	0.923	1.000
RWTH-1	SMT	✓			2.663	0.107	0.237	0.407	0.913	1.000
BASELINE1-1	SMT	✓			2.617	0.083	0.200	0.403	0.930	1.000
NAIST-1	SMT	✓			2.610	0.097	0.213	0.417	0.883	1.000
FRDC-1	SMT	✓	✓		2.517	0.077	0.170	0.360	0.910	1.000
BASELINE2-1	SMT	✓			2.427	0.083	0.160	0.333	0.850	1.000
KYOTO-2	SMT	✓			2.413	0.057	0.117	0.350	0.890	1.000
KYOTO-1	EBMT	✓			2.380	0.060	0.160	0.307	0.853	1.000
UOTTS-1	SMT	✓			2.377	0.067	0.130	0.307	0.873	1.000
NEU-1	SMT	✓	✓		2.373	0.033	0.127	0.317	0.897	1.000
ONLINE1-1	SMT			✓	2.273	0.050	0.103	0.250	0.870	1.000
ICT-1	SMT	✓	✓		2.267	0.027	0.100	0.260	0.880	1.000
KLE-1	SMT	✓			2.040	0.037	0.117	0.193	0.693	1.000

of the translated sentences (C-rank and above).

There was a large difference in ability to retain the sentence-level meanings between the top-level commercial RBMT systems and the SMT systems.

Table 10: Sign test of JE adequacy. "≫": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, "-": not significantly different.

	RBMT1-1	EIWA-1	RBMT3-1	RBMT2-1	NTT-UT-1	TORI-1	RWTH-1	BASELINE1-1	NAIST-1	FRDC-1	BASELINE2-1	KYOTO-2	KYOTO-1	UOTTS-1	NEU-1	ONLINE1-1	ICT-1	KLE-1
JAPIO-1	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT1-1		>	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
EIWA-1			≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT3-1				-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT2-1					≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
NTT-UT-1						-	-	-	-	>	≫	≫	≫	≫	≫	≫	≫	≫
TORI-1							-	-	-	≫	≫	≫	≫	≫	≫	≫	≫	≫
RWTH-1								-	-	>	≫	≫	≫	≫	≫	≫	≫	≫
BASELINE1-1									-	-	≫	≫	≫	≫	>	≫	≫	≫
NAIST-1										-	>	-	>	≫	>	≫	≫	≫
FRDC-1											-	-	≫	>	-	≫	≫	≫
BASELINE2-1												-	-	-	-	-	-	≫
KYOTO-2													-	-	-	≫	≫	≫
KYOTO-1														-	-	-	-	≫
UOTTS-1															-	-	-	≫
NEU-1																-	-	≫
ONLINE1-1																	-	≫
ICT-1																		≫

Table 11: Results of JE acceptability

	Type	Resource			Pairwise	Rate				
		B	M	E	score	AA	A or higher	B or higher	C or higher	F or higher
JAPIO-1	RBMT			✓	0.712	0.083	0.223	0.510	0.633	1.000
RBMT1-1	RBMT			✓	0.674	0.077	0.193	0.470	0.570	1.000
EIWA-1	HYBRID	✓		✓	0.638	0.090	0.180	0.380	0.493	1.000
NTT-UT-1	SMT	✓			0.491	0.023	0.060	0.177	0.250	1.000
RWTH-1	SMT	✓			0.489	0.043	0.080	0.180	0.240	1.000
BASELINE1-1	SMT	✓			0.474	0.027	0.067	0.157	0.217	1.000
NAIST-1	SMT	✓			0.472	0.020	0.057	0.143	0.227	1.000
TORI-1	SMT	✓	✓	✓	0.460	0.043	0.063	0.123	0.183	1.000
FRDC-1	SMT	✓	✓		0.448	0.017	0.040	0.113	0.187	1.000
BASELINE2-1	SMT	✓			0.447	0.037	0.060	0.123	0.167	1.000
KYOTO-1	EBMT	✓			0.436	0.007	0.027	0.103	0.177	1.000
UOTTS-1	SMT	✓			0.425	0.027	0.040	0.083	0.130	1.000
ONLINE1-1	SMT			✓	0.418	0.013	0.027	0.057	0.117	1.000
NEU-1	SMT	✓	✓		0.416	0.017	0.023	0.060	0.120	1.000

Table 12: Sign test of JE acceptability. "≫": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, "-": not significantly different.

	RBMT1-1	EIWA-1	NTT-UT-1	RWTH-1	BASELINE1-1	NAIST-1	TORI-1	FRDC-1	BASELINE2-1	KYOTO-1	UOTTS-1	ONLINE1-1	NEU-1
JAPIO-1	-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT1-1		-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
EIWA-1			≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
NTT-UT-1				-	-	-	-	>	≫	≫	≫	≫	≫
RWTH-1					-	-	-	≫	≫	≫	≫	≫	≫
BASELINE1-1						-	-	-	-	>	≫	≫	≫
NAIST-1							-	-	-	-	≫	≫	≫
TORI-1								-	-	-	>	≫	>
FRDC-1									-	-	-	-	>
BASELINE2-1										-	-	-	>
KYOTO-1											-	-	>
UOTTS-1												-	-
ONLINE1-1													-

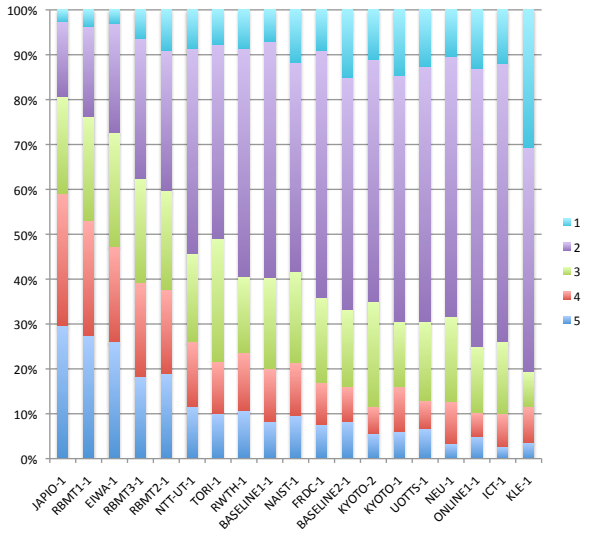

Figure 4: Results of JE adequacy.

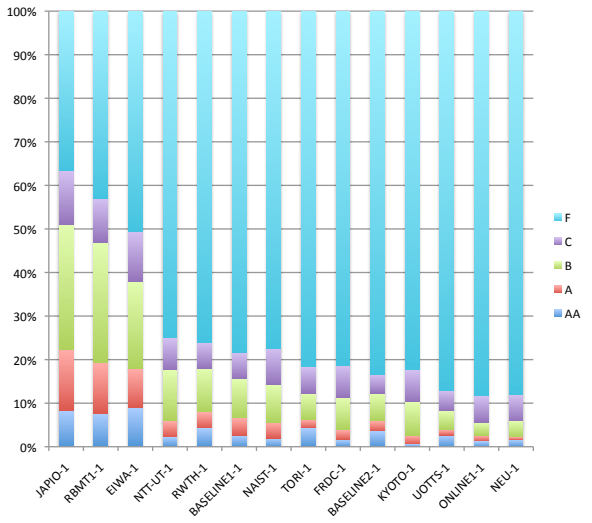

Figure 5: Results of JE acceptability.

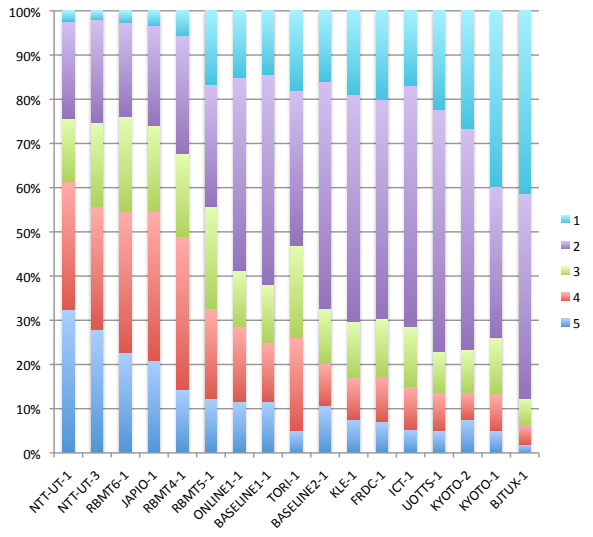

Figure 6: Results of EJ adequacy.

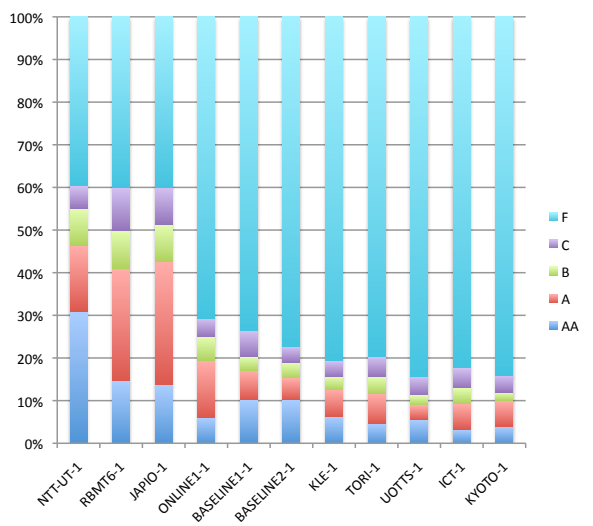

Figure 7: Results of EJ acceptability.

4.3 English to Japanese

4.3.1 Adequacy Evaluation

Table 13 and Figure 6 show the results of the adequacy evaluation. Table 14 shows the results of the statistical significance test of the adequacy evaluation using a sign test.

NTT-UT-1 and NTT-UT-3 are the top systems for the SMT systems and RBMT6-1, JAPIO-1, RBMT4-1, and RBMT5-1 are the top RBMT systems.

The following can be seen from the results:

- Some of the SMT systems achieved human evaluation scores (adequacy) equal or better than the top-level commercial RBMT systems. No SMT system did this at NTCIR-7, and it is thought that this is the first time for that this has been achieved.
- A feature of NTT-UT-1 and NTT-UT-3 is that the systems use a method that pre-orders English input sentences using parse results and head finalization rules, and translates in almost monotone word orders. The effectiveness of the method can be seen from the evaluation.
- The adequacy scores for the commercial RBMT systems were higher than those for the SMT systems other than NTT-UT-1 and NTT-UT-3.

4.3.2 Acceptability Evaluation

Table 15 and Figure 7 show the results of the acceptability evaluation. Table 16 shows the results of the statistical significance test of the acceptability evaluation using a sign test.

For the best SMT system (NTT-UT-1), the source sentence meaning could be understood (C and above) for 60%

Table 13: Results of EJ adequacy

	Type	Resource			Average	Rate				
		B	M	E	score	5	4 or higher	3 or higher	2 or higher	1 or higher
NTT-UT-1	SMT	✓	✓		3.670	0.323	0.613	0.757	0.977	1.000
NTT-UT-3	SMT	✓			3.563	0.280	0.557	0.747	0.980	1.000
RBMT6-1	RBMT			✓	3.507	0.227	0.547	0.760	0.973	1.000
JAPIO-1	RBMT			✓	3.463	0.210	0.547	0.740	0.967	1.000
RBMT4-1	RBMT			✓	3.253	0.143	0.490	0.677	0.943	1.000
RBMT5-1	RBMT			✓	2.840	0.123	0.327	0.557	0.833	1.000
ONLINE1-1	SMT			✓	2.667	0.117	0.287	0.413	0.850	1.000
BASELINE1-1	SMT	✓			2.603	0.117	0.250	0.380	0.857	1.000
TORI-1	HYBRID	✓	✓	✓	2.600	0.050	0.260	0.470	0.820	1.000
BASELINE2-1	SMT	✓			2.477	0.107	0.203	0.327	0.840	1.000
KLE-1	SMT	✓		✓	2.353	0.077	0.170	0.297	0.810	1.000
FRDC-1	SMT	✓	✓		2.347	0.070	0.173	0.303	0.800	1.000
ICT-1	SMT	✓	✓		2.320	0.053	0.150	0.287	0.830	1.000
UOTTS-1	SMT	✓			2.193	0.050	0.137	0.230	0.777	1.000
KYOTO-2	SMT	✓			2.180	0.077	0.137	0.233	0.733	1.000
KYOTO-1	EBMT	✓			2.047	0.050	0.133	0.260	0.603	1.000
BJTUX-1	SMT	✓			1.793	0.020	0.063	0.123	0.587	1.000

Table 14: Sign test of EJ adequacy. "≫": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, "-": not significantly different.

	NTT-UT-3	RBMT6-1	JAPIO-1	RBMT4-1	RBMT5-1	ONLINE1-1	BASELINE1-1	TORI-1	BASELINE2-1	KLE-1	FRDC-1	ICT-1	UOTTS-1	KYOTO-2	KYOTO-1	BJTUX-1
NTT-UT-1	≫	-	>	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
NTT-UT-3		-	-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT6-1			-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
JAPIO-1				≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT4-1					≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT5-1						-	>	≫	≫	≫	≫	≫	≫	≫	≫	≫
ONLINE1-1							-	-	>	≫	≫	≫	≫	≫	≫	≫
BASELINE1-1								-	-	≫	≫	>	≫	≫	≫	≫
TORI-1									-	≫	≫	≫	≫	≫	≫	≫
BASELINE2-1										-	-	-	≫	≫	≫	≫
KLE-1											-	-	-	≫	≫	≫
FRDC-1												-	-	≫	≫	≫
ICT-1													>	≫	≫	≫
UOTTS-1														-	≫	≫
KYOTO-2															>	≫
KYOTO-1																>

Table 15: Results of EJ acceptability

	Type	Resource			Pairwise	Rate				
		B	M	E	score	AA	A or higher	B or higher	C or higher	F or higher
NTT-UT-1	SMT	✓	✓		0.695	0.310	0.463	0.550	0.603	1.000
RBMT6-1	RBMT			✓	0.656	0.147	0.410	0.500	0.600	1.000
JAPIO-1	RBMT			✓	0.652	0.137	0.427	0.513	0.600	1.000
ONLINE1-1	SMT			✓	0.479	0.060	0.193	0.250	0.293	1.000
BASELINE1-1	SMT	✓			0.472	0.103	0.170	0.203	0.263	1.000
BASELINE2-1	SMT	✓			0.456	0.103	0.153	0.190	0.227	1.000
KLE-1	SMT	✓		✓	0.434	0.063	0.127	0.157	0.193	1.000
TORI-1	SMT	✓	✓	✓	0.432	0.047	0.117	0.157	0.203	1.000
UOTTS-1	SMT	✓			0.411	0.057	0.090	0.113	0.157	1.000
ICT-1	SMT	✓	✓		0.411	0.033	0.093	0.130	0.177	1.000
KYOTO-1	EBMT	✓			0.404	0.040	0.100	0.120	0.160	1.000

of the sentences. Of the systems using RBMT, the source sentence meaning could be understood (C or above) for 60% of the translated sentences in the best system (RBMT6-1).

The translation quality of the top SMT system (NTT-UT-1) was equal to or surpassing that of the top-level commercial RBMT systems for retaining the sentence-level meanings.

Table 16: Sign test of EJ acceptability. "≫": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, "-": not significantly different.

	RBMT6-1	JAPIO-1	ONLINE1-1	BASELINE1-1	BASELINE2-1	KLE-1	TORI-1	UOTTS-1	ICT-1	KYOTO-1
NTT-UT-1	>	>	≫	≫	≫	≫	≫	≫	≫	≫
RBMT6-1		-	≫	≫	≫	≫	≫	≫	≫	≫
JAPIO-1			≫	-	-	≫	≫	≫	≫	≫
ONLINE1-1						-	-	≫	≫	≫
BASELINE1-1						-	-	≫	≫	≫
BASELINE2-1						-	-	>	>	-
KLE-1							-	-	-	-
TORI-1								-	-	-
UOTTS-1									-	-
ICT-1										-

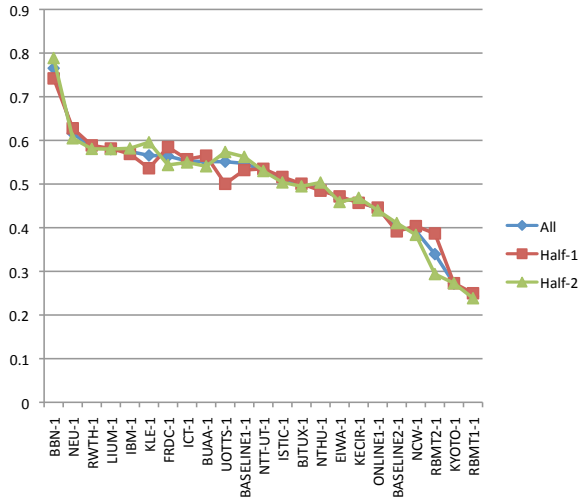

Figure 8: Comparison between data for CE adequacy.

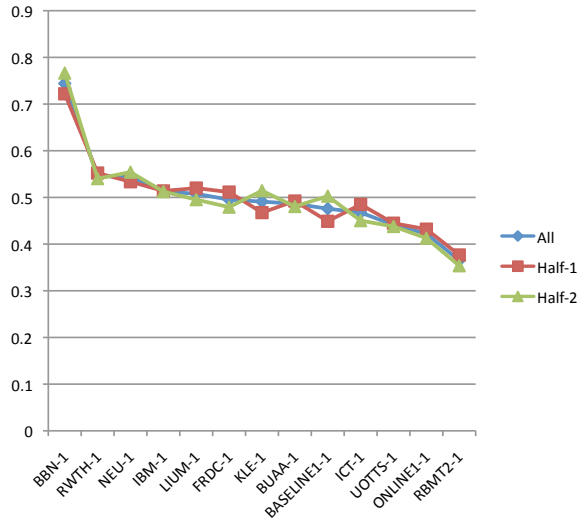

Figure 9: Comparison between data for CE acceptability.

Table 17: Pearson correlation coefficient between data

	Adequacy	Acceptability
CE	0.963	0.953
JE	0.940	0.972
EJ	0.985	0.982

5. VALIDATION OF HUMAN EVALUATION

In order to discuss the reliability of human evaluation, we will give the correlation of the meta-evaluation results between the divided data. In this section, pairwise scores were used for normalization purposes.

5.1 Difference of Data

Figures 8 to 13 show the evaluation results for the first half of the data (Half-1), for the second half of the data (Half-2), and for all of the data (All). This halved data contains half of the sentences evaluated by each evaluator. Table 17 shows the Pearson correlation coefficients of the system evaluation scores between the half data.

Although there are slight differences, there are no large differences that reverse high-ranked systems and low-ranked systems. The Pearson correlation coefficients are close to 1.0 for all of the data pairs.

Therefore, when the evaluators are the same and the data is different, there were no large differences in the evaluation, so the evaluations are thought to be consistent on this point.

The differences in the comparisons show the effects of *differences in the data* and *intra-evaluator consistency*. Consequently, the effects from either difference in the data or intra-evaluator consistency are thought to be less than the differences between results of Half-1 and Half-2.

5.2 Evaluator Differences

For each subtask and criterion, three evaluators evaluated the translations of different 100 source sentences. The system evaluation results from each evaluator and the system evaluation results from all of the evaluators are shown in Figures 14 to 19. Table 18 shows the Pearson correlation

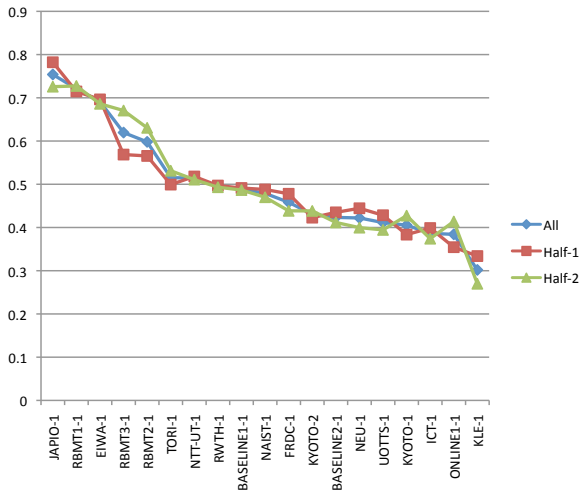

Figure 10: Comparison between data for JE adequacy.

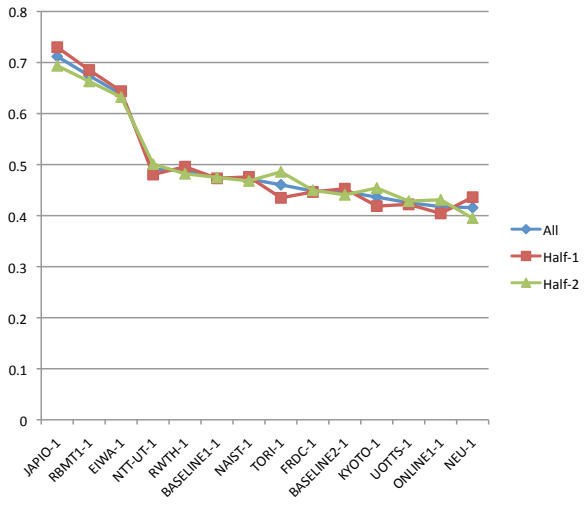

Figure 11: Comparison between data for JE acceptability.

coefficients for the system evaluation scores between evaluators.

Although there are slight differences, there are no large differences reversing high-ranked systems and low-ranked systems. Therefore, even when the evaluators and the data are different, there are no large differences in the evaluations, and the evaluations are thought to be consistent on this point. The differences in the comparisons show the effects of differences in the data and inter-evaluator consistency. Consequently, the effects from either difference in the data or inter-evaluator consistency are thought to be less than the differences between results of Evalutor-1, Evaluator-2, and Evaluator-3.

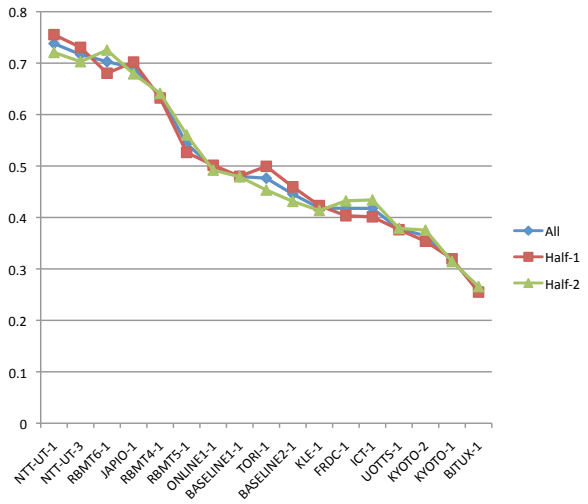

Figure 12: Comparison between data for EJ adequacy.

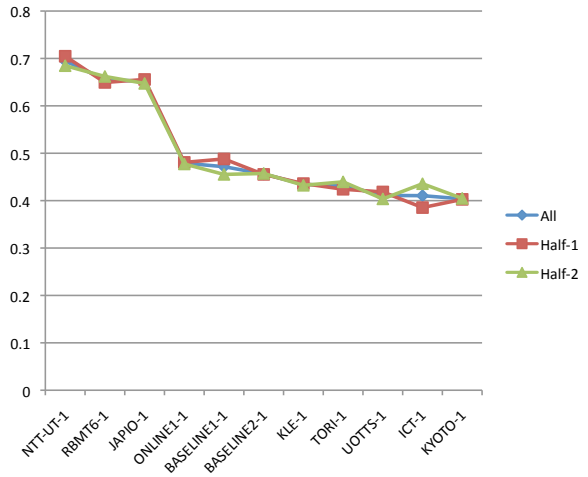

Figure 13: Comparison between data for EJ acceptability.

Table 18: Pearson correlation coefficient between evaluators by different data sets

	Evaluator	Adequacy	Acceptability
CE	1 & 2	0.929	0.918
	1 & 3	0.932	0.873
	2 & 3	0.966	0.898
JE	1 & 2	0.944	0.967
	1 & 3	0.882	0.935
	2 & 3	0.945	0.966
EJ	1 & 2	0.977	0.988
	1 & 3	0.963	0.979
	2 & 3	0.972	0.973

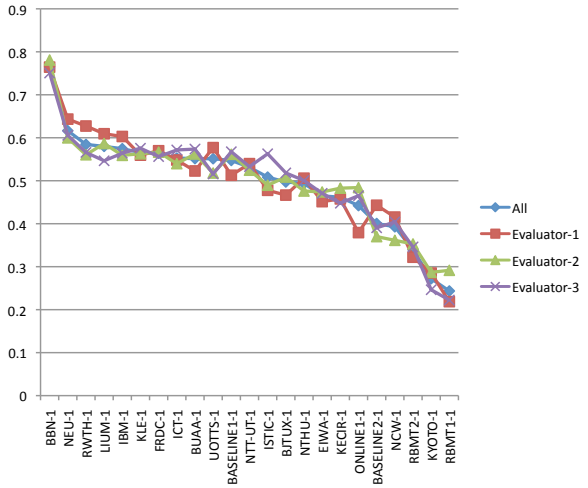

Figure 14: Comparison of the evaluators' evaluations of the different data sets for CE adequacy.

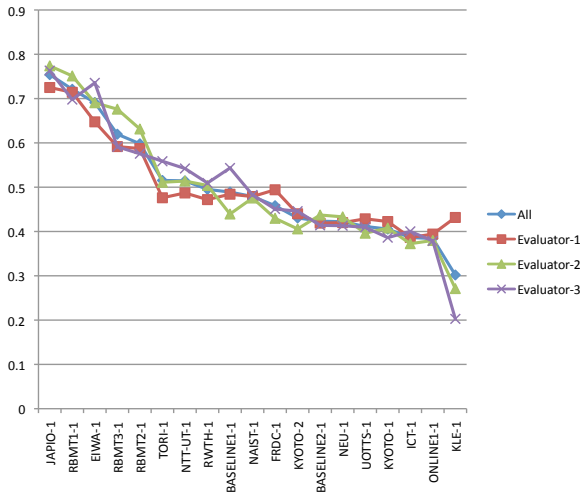

Figure 16: Comparison of the evaluators' evaluations of the different data sets for JE adequacy.

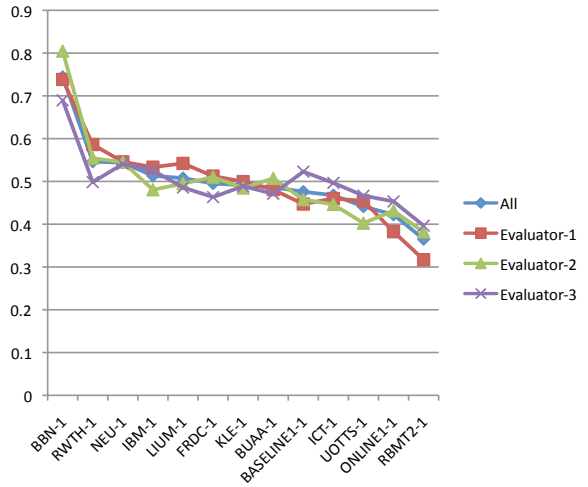

Figure 15: Comparison of the evaluators' evaluations of the different data sets for CE acceptability.

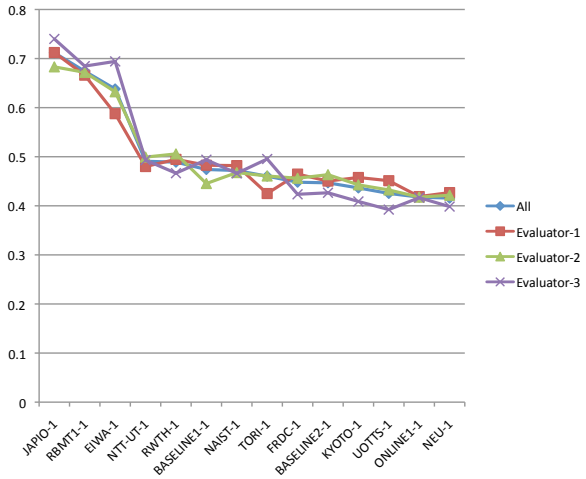

Figure 17: Comparison of the evaluators' evaluations of the different data sets for JE acceptability.

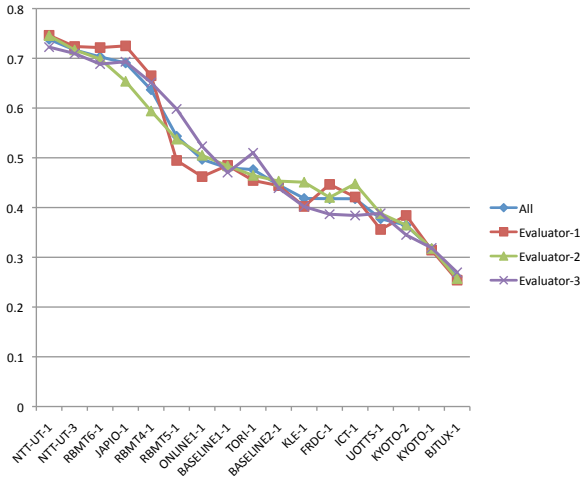

Figure 18: Comparison of the evaluators' evaluations of the different data sets for EJ adequacy.

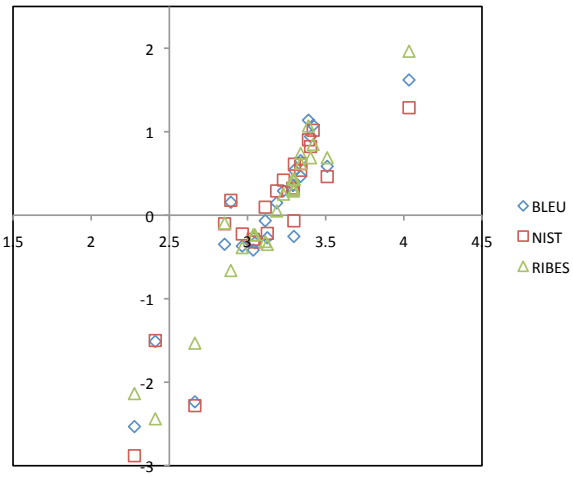

Figure 20: CE correlations between adequacy and automatic evaluation scores.

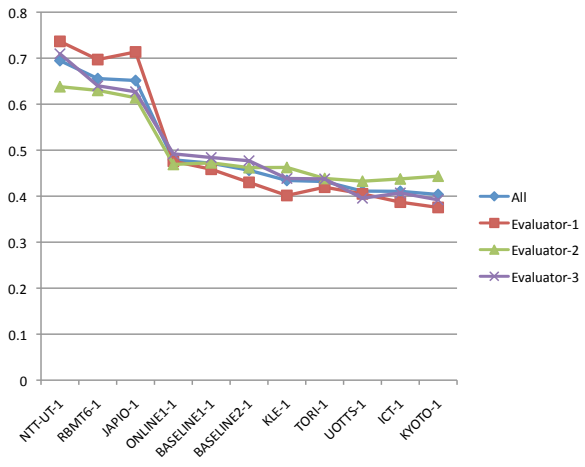

Figure 19: Comparison of the evaluators' evaluations of the different data sets for EJ acceptability.

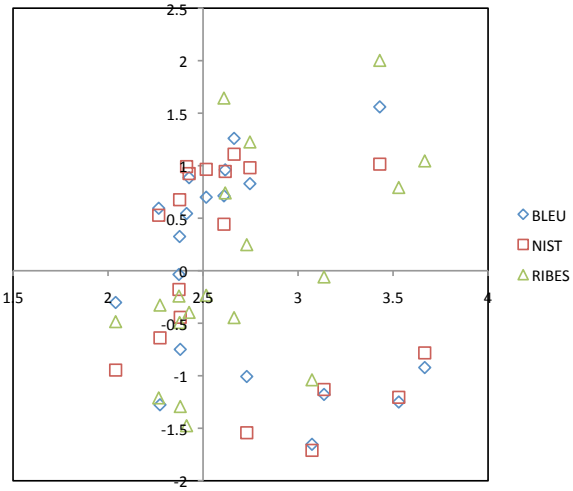

Figure 21: JE correlations between adequacy and automatic evaluation scores.

6. META-EVALUATION OF AUTOMATIC EVALUATION

We calculated the scoring from three automatic evaluation measures (BLEU, NIST, and RIBES) based on 2,000 test sentences for all the submissions. These automatic evaluation measures were partly calculated to investigate their reliability in the patent domain for the language pairs of CE, JE, and EJ.

The correlations between human evaluations and standardized automatic evaluation scores are shown in Figures 20 to 22. In these figures, the horizontal axis indicates the average adequacy score and the vertical axis indicates the pairwise score of automatic measures.

The Spearman rank-order correlation coefficients and the Pearson correlation coefficients between human evaluations (average adequacy scores) and automatic evaluation scores are shown in Table 19.

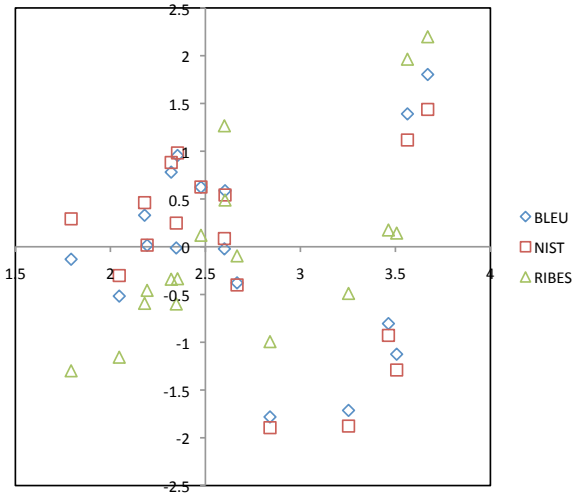

Figure 22: EJ correlations between adequacy and automatic evaluation scores.

Table 19: Correlation coefficients between adequacy and automatic evaluation scores

		Spearman	Pearson
CE	BLEU	0.931	0.915
	NIST	0.911	0.891
	RIBES	0.949	0.967
JE	BLEU	-0.042	-0.241
	NIST	-0.114	-0.286
	RIBES	0.632	0.579
EJ	BLEU	-0.029	-0.032
	NIST	-0.074	-0.209
	RIBES	0.716	0.683

In Figure 20 and Table 19, it can be seen that the three automatic evaluation measures have a high correlation with the human evaluation for the CE evaluation.

In Figures 21 and 22 and Table 19, it can be seen that the RIBES' correlation with human evaluation is higher than that of BLEU or NIST for JE and EJ evaluations including RBMT systems.

The correlations between the human evaluations and standardized automatic scores excluding the RMBT systems for JE and EJ are shown in Figures 23 and 24.

The Spearman rank-order correlation coefficients and the Pearson correlation coefficients between human evaluation and automatic scores excluding the RMBT systems for JE and EJ are shown in Table 20.

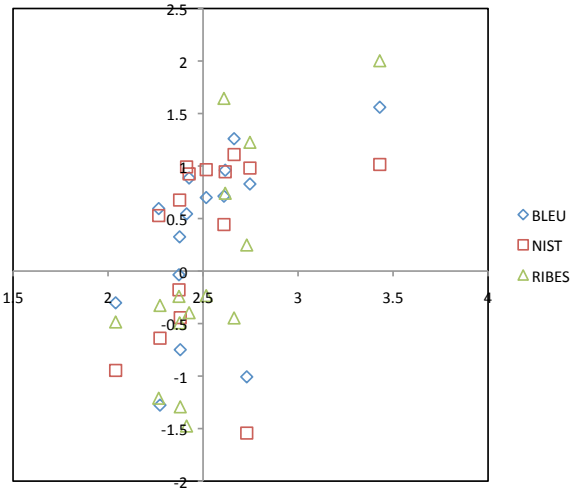

Figure 23: JE correlations between adequacy and automatic evaluation scores excluding RBMT systems.

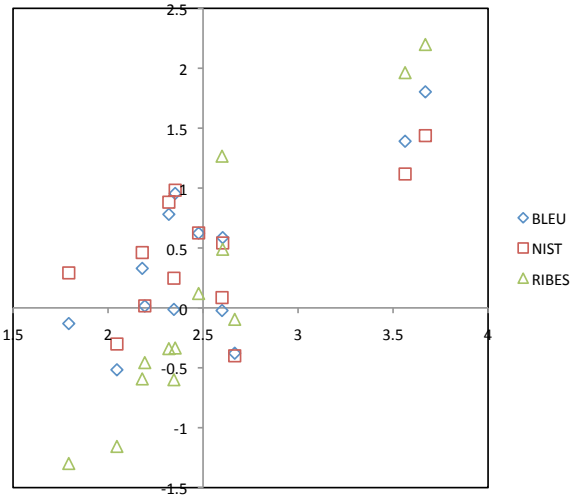

Figure 24: EJ correlations between adequacy and automatic evaluation scores excluding RBMT systems.

The correlations excluding RBMT systems for JE and EJ are higher than the correlations including the RBMT systems for the three automatic measures. Therefore, the reliability of the evaluations of the comparisons between systems without the RBMT systems is higher than the reliability of the evaluations of the comparisons between systems including the RBMT systems for the automatic evaluation of the quality of the JE and EJ patent translations.

Table 20: Correlation coefficients between adequacy and automatic evaluation scores excluding RBMT systems

		Spearman	Pearson
JE	BLEU	0.618	0.525
	NIST	0.543	0.362
	RIBES	0.679	0.741
EJ	BLEU	0.511	0.753
	NIST	0.412	0.603
	RIBES	0.929	0.943

7. CONCLUSION

In order to develop challenging and significant practical research into patent machine translation, we organized a Patent Machine Translation Task at NTCIR-9. For this task, we produced and provided test collections for Chinese/English and Japanese/English patent machine translations. This paper has described the results and knowledge obtained from the evaluations. We conducted human evaluations on the submitted and baseline results. Various innovative ideas were explored and their effectiveness in patent translation was shown in evaluations. For NTCIR-10, we would like to explore more practical evaluations and new topics in patent machine translation.

8. REFERENCES

[1] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, 2011.

[2] J. Chang, S.-T. Huang, H.-C. Yen, M.-J. Jiang, C.-C. Huang, J. S. Chang, and P.-C. Yang. [PatentMT] Summary Report of Team III_CYUT_NTHU. In *Proceedings of NTCIR-9*, 2011.

[3] W. Chao and Z. Li. ZZX_MT: the BeiHang MT System for NTCIR-9 PatentMT Task. In *Proceedings of NTCIR-9*, 2011.

[4] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

[5] T. Ehara. Machine translation system for patent documents combining rule-based translation and statistical post-editing applied to the PatentMT Task. In *Proceedings of NTCIR-9*, 2011.

[6] M. Feng, C. Schmidt, J. Wuebker, S. Peitz, M. Freitag, and H. Ney. The RWTH Aachen System for NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.

[7] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proceedings of NTCIR-7*, 2008.

[8] A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya, and S. Shimohata. Overview of the Patent Translation Task at the NTCIR-8 Workshop. In *Proceedings of NTCIR-8*, 2010.

[9] Y. He, C. Shi, and H. Wang. ISTIC Statistical Machine Translation System for Patent Machine Translation in NTCIR-9. In *Proceedings of NTCIR-9*, 2011.

[10] H. Hoang, P. Koehn, and A. Lopez. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 152–159, 2009.

[11] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, 2010.

[12] J. Jiang, J. Xu, Y. Lin, and Y. Zhang. System Description of BJTU-NLP SMT for NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.

[13] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, 2007.

[14] S. Kondo, M. Komachi, Y. Matsumoto, K. Sudoh, K. Duh, and H. Tsukada. Learning of Linear Ordering Problems and its Application to J-E Patent Translation in NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.

[15] Y.-S. Lee, B. Xiang, B. Zhao, M. Franz, S. Roukos, and Y. Al-Onaizan. IBM Chinese-to-English PatentMT System for NTCIR-9. In *Proceedings of NTCIR-9*, 2011.

[16] B. Lu, B. K. Tsou, T. Jiang, O. Y. Kwong, and J. Zhu. Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. In *Proceedings of the 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2010)*, Beijing, China, 2010.

[17] J. Ma and S. Matsoukas. BBN's Systems for the Chinese-English Sub-task of NTCIR-9 Patent MT Evaluation. In *Proceedings of NTCIR-9*, 2011.

[18] J. Murakami and M. Tokuhisa. Statistical Machine Translation with Rule based Machine Translation. In *Proceedings of NTCIR-9*, 2011.

[19] H. Na, J.-J. Li, S.-J. Kim, and J.-H. Lee. POSTECH's Statistical Machine Translation Systems for the NTCIR-9 PatentMT Task (English-to-Japanese). In *Proceedings of NTCIR-9*, 2011.

[20] T. Nakazawa and S. Kurohashi. EBMT System of KYOTO Team in PatentMT Task at NTCIR-9. In *Proceedings of NTCIR-9*, 2011.

[21] T. Ohio, T. Mitsuhashi, and T. Kakita. Use of the Japio Technical Field Dictionaries for NTCIR-PatentMT. In *Proceedings of NTCIR-9*, 2011.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[23] H. Schwenk and S. Abdul-Rauf. LIUM's Statistical Machine Translation System for the NTCIR

Chinese/English Patent Translation Task. In *Proceedings of NTCIR-9*, 2011.

[24] K. Sudoh, K. Duh, H. Tsukada, M. Nagata, X. Wu, T. Matsuzaki, and J. Tsujii. NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.

[25] Y.-H. Tseng, C.-L. Liu, C.-C. Tsai, J.-P. Wang, Y.-H. Chuang, and J. Jeng. Statistical Approaches to Patent Translation for PatentMT - Experiments with Various Settings of Training Data. In *Proceedings of NTCIR-9*, 2011.

[26] J. S. White, T. A. O'Connell, and F. E. O'Mara. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of AMTA*, 1994.

[27] X. Wu, T. Matsuzaki, and J. Tsujii. SMT Systems in the University of Tokyo for NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.

[28] T. Xiao, Q. Li, Q. Lu, H. Zhang, H. Ding, S. Yao, X. Xu, X. Fei, J. Zhu, F. Ren, and H. Wang. The NiuTrans Machine Translation System for NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.

[29] H. Xiong, L. Song, F. Meng, Y. Lü, and Q. Liu. The ICT's Patent MT System Description for NTCIR-9. In *Proceedings of NTCIR-9*, 2011.

[30] Z. Zheng, N. Ge, Y. Meng, and H. Yu. HPB SMT of FRDC Assisted by Paraphrasing for the NTCIR-9 PatentMT. In *Proceedings of NTCIR-9*, 2011.

APPENDIX

A. INSTRUCTIONS FOR THE ADEQUACY CRITERION

A.1 Evaluation Criterion

Adequacy is scored according to how well the meaning of a translation matches the meaning of the reference (source) translation for each sentence. Adequacy evaluations are done according to the following 5-level scale:

5	All meaning
4	Most meaning
3	Much meaning
2	Little meaning
1	None

A.2 Notes

1. Adequacy estimates the sentence meaning by evaluating fragments of a sentence.
2. The main reason for using fragments is to reduce evaluation costs. When sentences are long, fragment-level evaluation is easier than sentence-level.
3. Fragment size:
   (a) Clause-level (first priority) or
   (b) "subject and its predicate" level (second priority) or
   (c) phrase-level (third priority).
4. Supplementary definitions to reduce criterion ambiguity:
   (a) A score of 5 shows that the sentence-level meaning (subject, predicate and object) is correct.
   (b) Relative comparison:
   - A sentence whose sentence-level meaning is not correct would be evaluated as 1 to 4 by not only the absolute criterion (most, much, little, and none) but also a relative comparison among the multiple translation outputs.
   - The relative comparison must be consistent in all of the data.

B. INSTRUCTIONS FOR THE ACCEPTABILITY CRITERION

B.1 Evaluation Criterion

Acceptability evaluations are done using the 5-level scale in Figure 1.

B.2 Notes

1. Evaluations are done from the standpoint of whether the machine-translated English sentence conveys the important information and the content of the source sentence, not on the completeness of a literal translation.
2. What is "important information" ? "Important information" is the information that is necessary for a conversation between two people. This information is what the machine translation results need to convey in order for the conversation partner to understand the content of the source sentence.
3. What does "contents of the source sentence can be understood" mean? It refers to when two people can begin a conversation and the machine translated results allow the conversation partner to understand the contents of the conversation.
4. The first step and the second step of the chart can be merged, so that "F" is: either not all of the important information is included or the contents from the source sentence cannot be understood.
5. The level of correctness for the "Grammatically correct" step is whether the translation is grammatical enough to convey the meaning of the source sentence. Strict adequateness (e.g., Editor's emendation level) for each expression is not required here. Therefore, if there are sentences that include expressions which cannot be judged as fully expressing the patent or technological terms, but the meaning itself is expressed, then it can be evaluated as A.
6. On the "Native level" step, natural English sentences which do not need any correction are to be evaluated as AA. Therefore, all minimum required grammatical check points (including punctuation) for a natural English sentence are needed.
7. If there is a sentence in unnatural English lacking a subject (nominative), and if the sentence could be easily understood and grammatically correct if it were transformed from active sentence into the passive voice, it

can be evaluated as "B," as the sentence is grammatically incorrect.

8. The following type of differences are permissible: The character is the same but the character code is not the same. e.g. " １ ２ ３ " and "123" are regarded as the same.
9. Special characters such as Greek letters in the source sentences are replaced as letters enclosed by periods or enclosed by ampersands and semi-colons. These replacements are permissible. e.g. "5 μm" → "5 .mu.m" or "5 μm"
10. Some translations mistakenly include segments of characters from the source language. These segments are ignored if the translation works out appropriately without the segments.

C. INSTRUCTIONS FOR THE HUMAN EVALUATION PROCEDURE

C.1 Evaluation Method for Training and Main Evaluations

- The criteria for evaluation are based on the guidelines.
- One input sentence (or one reference sentence) and all of the system outputs are shown at the same time to compare systems.
- An evaluator evaluates all of the translations for the same input sentence.
- The MT output sentences for each input sentence are given random order to the evaluators.
- The evaluators could review the evaluations.

C.2 Training

Before the main evaluation, a trial evaluation is done. All of the evaluators evaluate translation results for the trial evaluations. The conditions for all evaluators were the same. After the trial evaluation, a consensus meeting is held in order to adjust the differences in the evaluations among all the evaluators and to decide on common evaluations for the translation results for the trial evaluation.

D. ALL SUBMISSIONS AND AUTOMATIC EVALUATION SCORES

Table 21: CE submissions and automatic evaluation scores

SYSTEM-ID (GROUP ID)	Priority	Type	Resource				BLEU	NIST	RIBES
			B	M	E	C			
BASELINE1	1	SMT	✓				0.3072	7.903	0.7719
BASELINE2	1	SMT	✓				0.2932	7.750	0.7284
BBN	1	SMT	✓	✓			0.3944	8.911	0.8327
BBN	2	SMT	✓				0.3664	8.595	0.8200
BJTUX	1	SMT	✓				0.2779	7.663	0.7422
BJTUX	2	SMT	✓				0.2808	7.701	0.7480
BUAA	1	HYBRID	✓	✓			0.2649	7.492	0.7673
BUAA	3	SMT	✓	✓			0.2631	7.477	0.7675
BUAA	2	SMT	✓	✓			0.2619	7.471	0.7671
EIWA	1	HYBRID	✓		✓		0.2597	7.228	0.7455
FRDC	1	SMT	✓	✓	✓		0.3146	8.126	0.7793
IBM	1	SMT	✓	✓	✓	✓	0.3611	8.509	0.7972
IBM	2	SMT	✓	✓	✓		0.3500	8.466	0.7936
IBM	3	SMT	✓	✓	✓	✓	0.3535	8.394	0.7954
IBM	4	SMT	✓	✓	✓	✓	0.3360	8.297	0.7759
IBM	5	SMT	✓	✓	✓	✓	0.3256	8.159	0.7766
IBM	6	SMT	✓	✓	✓	✓	0.3534	8.377	0.7938
IBM	7	SMT	✓	✓	✓	✓	0.3242	8.158	0.7759
IBM	9	SMT	✓	✓	✓	✓	0.3526	8.369	0.7949
IBM	10	SMT	✓	✓	✓		0.3442	8.255	0.7889
ICT	1	SMT	✓	✓			0.3197	8.203	0.7716
ICT	2	SMT	✓	✓			0.3152	8.125	0.7697
ICT	3	SMT	✓	✓			0.3157	8.136	0.7699
ICT	4	SMT	✓	✓			0.3078	8.124	0.7603
ICT	5	SMT	✓	✓			0.3076	8.033	0.7687
ICT	6	SMT	✓	✓			0.3064	8.005	0.7678
ICT	7	SMT	✓	✓			0.3092	8.075	0.7699
ICT	8	SMT	✓	✓			0.3071	8.032	0.7665
ISTIC	1	HYBRID	✓	✓			0.2927	7.867	0.7567
ISTIC	2	HYBRID	✓	✓			0.2851	7.766	0.7523
ISTIC	3	SMT	✓	✓			0.2833	7.794	0.7551
KECIR	1	SMT	✓	?	✓	✓	0.2536	7.260	0.7453
KECIR	2	SMT	✓	?	✓	✓	0.2588	7.411	0.7472
KECIR	3	SMT	✓	?	✓	✓	0.2184	7.104	0.7349
KLE	1	SMT	✓				0.3276	8.210	0.7841
KLE	2	SMT	✓				0.3135	8.001	0.7594
KYOTO	1	EBMT	✓				0.1780	5.991	0.6578
LIUM	1	SMT	✓	✓			0.3476	8.424	0.7820
NCW	1	SMT	✓				0.2584	7.455	0.7509
NCW	2	SMT	✓		✓		0.2424	7.351	0.7388
NCW	3	SMT	✓		✓		0.2433	7.341	0.7387
NCW	4	SMT	✓		✓		0.2307	7.113	0.7339
NCW	5	SMT	✓				0.2336	7.061	0.7411
NCW	6	SMT	✓				0.2092	6.547	0.7298
NCW	7	SMT	✓				0.2050	6.484	0.7308
NCW	8	SMT	✓				0.2147	6.792	0.7309
NCW	9	SMT	✓				0.1957	6.238	0.7226
NCW	10	SMT	✓				0.1816	6.131	0.6993
NCW	11	SMT	✓				0.1917	6.391	0.7027
NCW	12	SMT	✓		✓		0.1570	5.605	0.7035
NCW	13	SMT	✓		✓		0.1552	5.473	0.7024
NCW	14	SMT	✓		✓		0.0654	3.716	0.6379
NCW	15	SMT	✓				0.0583	3.139	0.6296
NCW	16	SMT	✓				0.0654	3.587	0.6375
NCW	17	SMT	✓				0.0632	3.487	0.6353
NCW	18	SMT	✓				0.0525	3.172	0.6082
NCW	19	SMT	✓				0.1200	4.638	0.6205
NCW	20	SMT	✓		✓		0.1032	4.063	0.5915
NCW	21	SMT	✓		✓		0.0170	2.007	0.4157
NCW	22	SMT	✓		✓		0.1251	4.849	0.6326
NEU	1	SMT	✓	✓			0.3229	8.047	0.7820
NEU	2	HYBRID	✓	✓			0.3273	8.085	0.7828
NTHU	1	SMT	✓	?	✓		0.2638	7.335	0.7408
NTHU	2	SMT	✓	?	✓		0.2634	7.322	0.7404
NTHU	3	SMT	✓	?	✓		0.2639	7.328	0.7408
NTHU	4	SMT	✓	?	✓		0.2637	7.316	0.7402
NTHU	5	SMT	✓	?	✓		0.2637	7.323	0.7405
NTHU	6	SMT	✓	?	✓		0.2637	7.323	0.7405
NTT-UT	1	SMT	✓				0.3026	8.003	0.7647
NTT-UT	2	SMT	✓		✓		0.3074	8.003	0.7628
ONLINE1	1	SMT			✓		0.2569	7.328	0.7393
RBMT1	1	RBMT			✓		0.1075	4.546	0.6698
RBMT2	1	RBMT			✓		0.1280	5.174	0.6938
RWTH	1	SMT	✓	✓			0.3569	8.629	0.7884
RWTH	2	SMT	✓	✓			0.3542	8.513	0.7961
RWTH	3	SMT	✓	✓			0.3440	8.427	0.7730
RWTH	4	SMT	✓				0.3399	8.404	0.7841
UOTTS	1	SMT	✓				0.3074	7.892	0.7662
UOTTS	2	SMT	✓				0.3067	7.874	0.7678

Table 22: JE submissions and automatic evaluation scores

SYSTEM-ID (GROUP ID)	Priority	Type	Resource				BLEU	NIST	RIBES
			B	M	E	C			
BASELINE1	1	SMT	✓				0.2895	7.770	0.7064
BASELINE2	1	SMT	✓				0.2861	7.756	0.6758
EIWA	1	HYBRID	✓		✓		0.3169	7.816	0.7404
FRDC	1	SMT	✓	✓			0.2776	7.783	0.6802
ICT	1	SMT	✓	✓			0.2728	7.492	0.6539
ICT	2	SMT	✓	✓			0.2690	7.603	0.6573
ICT	3	SMT	✓	✓			0.2655	7.504	0.6523
ICT	4	SMT	✓	✓			0.2671	7.488	0.6519
ICT	5	SMT	✓	✓			0.2606	7.467	0.6527
ICT	6	SMT	✓	✓			0.2684	7.603	0.6537
JAPIO	1	RBMT			✓	✓	0.2035	6.618	0.7146
KLE	1	SMT	✓			✓	0.2318	6.509	0.6735
KLE	2	SMT	✓			✓	0.2955	7.828	0.6564
KYOTO	1	EBMT	✓				0.2114	6.844	0.6517
KYOTO	2	SMT	✓				0.2705	7.801	0.6468
NAIST	1	SMT	✓				0.2782	7.435	0.7307
NEU	1	SMT	✓	✓			0.2440	7.021	0.6800
NEU	2	SMT	✓	✓			0.2488	7.274	0.6836
NEU	3	SMT	✓	✓			0.2238	6.806	0.6552
NTT-UT	1	SMT	✓				0.2835	7.793	0.7195
ONLINE1	1	SMT			✓		0.1873	6.714	0.6777
RBMT1	1	RBMT			✓		0.1885	6.336	0.7078
RBMT2	1	RBMT			✓		0.1701	5.999	0.6586
RBMT3	1	RBMT			✓		0.1918	6.386	0.6849
RWTH	1	SMT	✓				0.3032	7.879	0.6745
RWTH	2	SMT	✓	✓			0.2622	7.726	0.6581
RWTH	3	SMT	✓				0.3020	7.864	0.6701
RWTH	4	SMT	✓				0.2598	7.702	0.6564
TORI	1	HYBRID	✓	✓	✓		0.1996	6.111	0.6932
TORI	2	HYBRID	✓	✓	✓		0.2090	6.283	0.6972
TORI	3	HYBRID	✓	✓	✓		0.1684	5.329	0.6724
TORI	4	HYBRID	✓	✓	✓		0.1797	6.152	0.6041
TORI	5	HYBRID	✓	✓			0.1436	4.926	0.6607
UOTTS	1	SMT	✓				0.2605	7.590	0.6732
UOTTS	2	SMT	✓				0.2697	7.694	0.6976

Table 23: EJ submissions and automatic evaluation scores

SYSTEM-ID (GROUP ID)	Priority	Type	Resource				BLEU	NIST	RIBES
			B	M	E	C			
BASELINE1	1	SMT	✓				0.3166	7.795	0.7200
BASELINE2	1	SMT	✓				0.3190	7.881	0.7068
BJTUX	1	SMT	✓				0.2705	7.540	0.6559
BJTUX	1	SMT	✓				0.2584	7.497	0.6492
FRDC	1	SMT	✓	✓			0.2781	7.494	0.6810
ICT	1	SMT	✓	✓			0.3291	8.144	0.6903
ICT	2	SMT	✓	✓			0.3210	8.079	0.6861
ICT	3	SMT	✓	✓			0.3172	8.170	0.6888
ICT	4	SMT	✓	✓			0.3206	8.088	0.6888
ICT	5	SMT	✓	✓			0.3217	8.120	0.6893
ICT	6	SMT	✓	✓			0.3017	7.833	0.6615
JAPIO	1	RBMT			✓	✓	0.2272	6.289	0.7088
KLE	1	SMT	✓		✓		0.3403	8.247	0.6905
KLE	2	SMT	✓		✓		0.2982	7.844	0.6454
KLE	3	SMT	✓		✓		0.2851	7.613	0.6409
KLE	4	SMT	✓		✓		0.2839	7.676	0.6417
KLE	5	SMT	✓				0.3510	8.285	0.7429
KYOTO	1	EBMT	✓				0.2457	6.931	0.6610
KYOTO	2	SMT	✓				0.3001	7.714	0.6812
NTT-UT	1	SMT	✓	✓			0.3948	8.713	0.7813
NTT-UT	2	SMT	✓				0.3784	8.544	0.7777
NTT-UT	3	SMT	✓				0.3683	8.385	0.7729
ONLINE1	1	SMT			✓		0.2546	6.830	0.6991
RBMT4	1	RBMT			✓		0.1688	5.317	0.6850
RBMT5	1	RBMT			✓		0.1644	5.299	0.6669
RBMT6	1	RBMT			✓		0.2066	5.918	0.7076
TORI	1	HYBRID	✓	✓	✓		0.2775	7.328	0.7479
TORI	2	HYBRID	✓	✓	✓		0.2475	7.141	0.6782
TORI	3	HYBRID	✓	✓	✓		0.1610	5.169	0.6818
TORI	4	HYBRID	✓	✓	✓		0.2203	6.857	0.6509
TORI	5	HYBRID	✓	✓			0.0831	3.771	0.5902
UOTTS	1	SMT	✓				0.2799	7.258	0.6861
UOTTS	2	SMT	✓				0.2781	7.236	0.6899