Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop

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

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

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

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-10 by describing its evaluation methods, test collections, 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 144 runs were submitted. We performed four types of evaluations: Intrinsic Evaluation (IE), Patent Examination Evaluation (PEE), Chronological Evaluation (ChE), and Multilingual Evaluation (ME). We conducted human evaluations for IE and PEE.

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 three previous patent translation tasks [7, 8, 10].

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

- Patent Examination Evaluation (PEE)

  From the acceptability evaluation results at the NTCIR-9 PatentMT, we thought that there was a high possibility that the top-level machine translation systems would be useful for practical use in patent translation. Thus, we conducted an evaluation exploring practical MT performance in patent examination. This evaluation evaluated the usefulness of machine translation for patent examination.

- Chronological Evaluation (ChE)

  This evaluation compared results from NTCIR-10 and 9 to measure progress over time, using the NTCIR-9 test sets.

- Multilingual Evaluation (ME)

  This evaluation compared CE and JE translations using the same English references to see the source language dependency.

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 Chinese, Japanese, and English.
- To compare the effects of different methods of patent translation by applying them to the same test data.
- To explore practical MT performance in appropriate fields for patent machine translation.
- 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, Sections 4–7 describe the evaluation results, Section 8 shows the validation of human evaluations, Section 9 gives a meta-evaluation of the automatic evaluations, and Section 10 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 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. Finally, the participants presented their research results at the NTCIR-10 workshop.

Table 1: Evaluations

Evaluation Type	Description
Intrinsic Evaluation (IE)	Similar to the NTCIR-9 evaluation. The quality of translated sentences were evaluated using new test sets. Human and automatic evaluations were conducted.
Patent Examination Evaluation (PEE)	New: The usefulness of machine translation for patent examination was evaluated. This evaluation was conducted for the CE and JE subtasks.
Chronological Evaluation (ChE)	New: A comparison between NTCIR-10 and 9 to measure progress over time, using the NTCIR-9 test sets, for all the subtasks.
Multilingual Evaluation (ME)	New: A comparison of CE and JE translations using the same English references to see the source language dependency. This evaluation was conducted for the CE and JE subtasks.

2.1 Evaluation Methodology

At NTCIR-10, we conducted four types of evaluations: Intrinsic Evaluation (IE), Patent Examination Evaluation (PEE), Chronological Evaluation (ChE), and Multilingual Evaluation (ME), as shown in Table 1.

2.1.1 Intrinsic Evaluation (IE)

We conducted human evaluations for IE for all the subtasks. Human evaluation was the primary evaluation for IE, and we used human judgments to validate the automatic metrics because we contend, the same as for 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* in the same way as for NTCIR-9 PatentMT [10]. 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 reference sentence and corresponding translations to evaluate the results. The same evaluators evaluated for the CE and JE subtasks. The evaluators were target language native speakers.

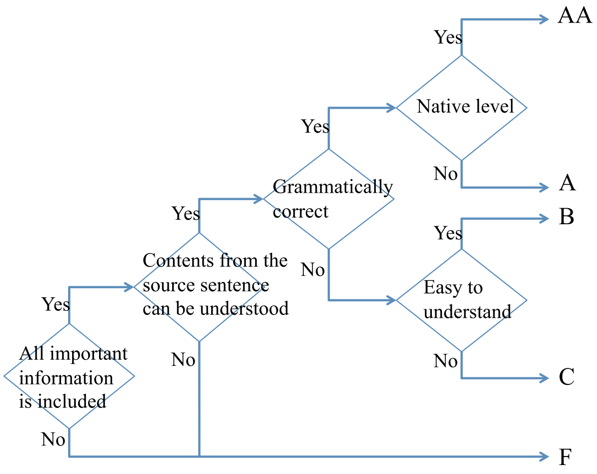
All important information is included
Yes
No
Contents from the source sentence can be understood
Yes
No
Grammatically correct
Yes
No
Native level
Yes
No
Easy to understand
Yes
No
AA
A
B
C
F

Figure 1: Acceptability.

Adequacy

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

We evaluated adequacy with consideration of clause-level meanings and relative comparison between systems. The instructions for the adequacy criterion are given in Appendix A.

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

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 for which the source sentence meanings can be understood from randomly selected test sentences. Acceptability is an evaluation of sentence-level meaning. The acceptability criterion used in this evaluation is aimed more at practical evaluation as opposed to adequacy. For example, if the requirement of a translation system is that the source sentence meaning can be understood, translations of C and higher are useful; however, if the requirement is that the source sentence meaning can be understood and the sentence is grammatically correct, then only translations of A and higher are useful. We can then know the number of sentences from a system would be useful for each requirement. An adequacy criterion cannot answer these requirements.

Acceptability also contains an evaluation of *fluency* that measures fluency in the target language, since it also affects the differences in grading from C to AA. If the adequacy of a translation is very low, then the translation is not correct even if the fluency is high. If the integrated evaluation score is calculated by averaging the adequacy and fluency scores, 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.

We ranked the systems based on acceptability using a pairwise comparison. 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 as 1.0 for each of the comparisons in which system A is ranked the highest of the two, and 0.5 for each of the comparisons in which system A is in a tie. System A's score is the total rewarded score in the pairwise comparisons divided by the total number of pairwise comparisons involving system A.[1]

Since, the main purpose of the acceptability evaluation is to measure translation quality for each system, the important results from the acceptability evaluation are not pairwise scores but the rates of each grade and above. For information access, the rates of C and above are thought to be important.

Human Evaluation Procedure

For the adequacy and acceptability evaluations, we conducted human evaluation training before the main evaluation to normalize the evaluators' criteria. In the training, all evaluators evaluated 200 translations for the CE and JE subtasks and 100 translations for the EJ subtask, and they held a meeting to determine the common results for each subtask. The main evaluation was done after that. The common results produced at the training were used as reference results for the main evaluation.

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

Automatic Evaluation

We calculated automatic evaluation scores for three metrics: RIBES [15], BLEU [21], and NIST [3]. RIBES scores were calculated using NTT's `RIBES.py` version 1.01[2]. BLEU and NIST scores were calculated using NIST's `mteval-v13a.pl`[3]. Detailed procedures for the automatic evaluation are shown at the PatentMT web page[4].

2.1.2 Patent Examination Evaluation (PEE)

We evaluated how useful machine translation would be for patent examinations for the CE and JE subtasks. Real reference patents that were used to reject patent applications were machine translated, and the translation results were evaluated to see if they would be useful for examining patent applications. The PEE concept is shown in Figure 2.

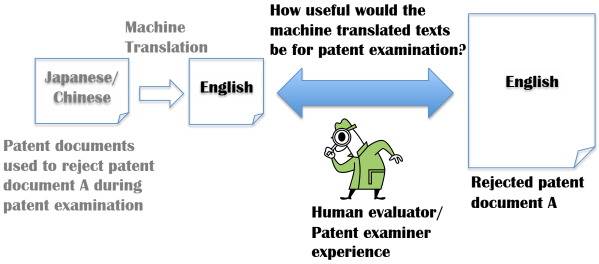
Machine Translation
How useful would the machine translated texts be for patent examination?
Japanese/ Chinese
English
English
Patent documents used to reject patent document A during patent examination
Human evaluator/ Patent examiner experience
Rejected patent document A

Figure 2: The PEE concept

PEE was carried out by two experienced patent examiners. These two evaluators worked as patent examiners at the Japan Patent Office and their English abilities are high. The Nippon Intellectual Property Translation Association (NIPTA) cooperated in conducting PEE.

The Real Framework of PEE

During patent examinations, patent examiners reject patents found to contain technology that is almost identical to that in existing patents or documents by referencing the existing patents or documents. Therefore, patent examiners need to understand the technology in existing patents or documents. When existing patents are written in a foreign language that a patent examiner cannot understand, a translation is needed to understand the existing patents. When the translation is done by machine translation, PEE evaluates the usefulness of the machine translation based on how many facts are understood in the translated results.

We conducted PEE using *shinketsu*, which are the final decisions from patent examinations at the Japan Patent Office. For the shinketsu whose final decision is a rejection, the following are included in many cases: a rejected patent application number, a reference patent number, description of the facts that the patent examiner recognized from the reference patent, and the reason for refusal. Therefore, by using shinketsu, we can obtain the facts that a patent examiner recognized from a reference patent. The real framework for PEE is shown in Figure 3. In Figure 3, the bottom shows preparation and the top shows the flow of translation and evaluation.

For the evaluation, evaluators first read a description of facts that a patent examiner recognized from a reference patent in the shinketsu, and read reference patent sentences that were machine translated. Then, evaluators evaluated the translation results. In practice, patent examiners first read patent application, then read the translations of existing patents, and finally they write down what facts the patent examiner recognized from the translations. Therefore, there were some differences in the procedures. However, we do not think that evaluations should be done without any previous knowledge of the reference patents, because reading the patent application before reading existing patents is a method of information gathering similar to information of reference patents since technology described in a rejected patent application and its reference patent would be similar. Therefore, although the procedures are not the same, we think that the patent examination evaluation would be effective in evaluating the usefulness of machine translation systems in patent examination.

[1] Note that the average score of acceptability was not used for system ranking. The reason is as follows. Here we assume that the differences between the grades are measured by 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. However, at the A-level, while the translations are at a non-native level, the contents from the source sentences can be understood and they are grammatically correct; thus, they have the potential to be useful in many cases. Thus, it is believed that the difference in usability between A and AA is smaller than that between F and C. In addition, we think that useful grades depend on specific usage. Therefore, it is difficult to give an appropriate score for each grade, and we avoided the conversion of grades to scores and calculation of averages.

[2] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html

[3] http://www.itl.nist.gov/iad/mig/tools/

[4] http://ntcir.nii.ac.jp/PatentMT-2/

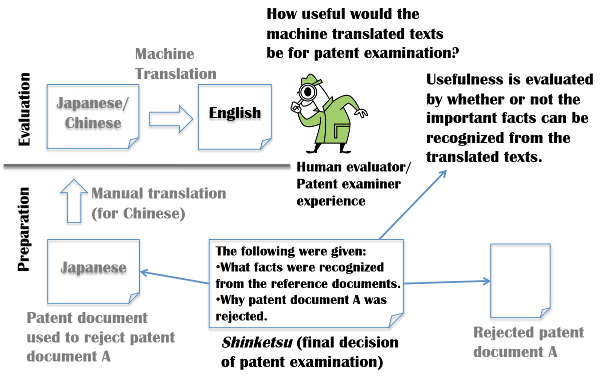
How useful would the machine translated texts be for patent examination?
Machine Translation
Evaluation
Japanese/ Chinese
English
Usefulness is evaluated by whether or not the important facts can be recognized from the translated texts.
Human evaluator/ Patent examiner experience
Manual translation (for Chinese)
Preparation
Japanese
The following were given:
•What facts were recognized from the reference documents.
•Why patent document A was rejected.
Patent document used to reject patent document A
Shinketsu (final decision of patent examination)
Rejected patent document A

Figure 3: The real framework for PEE

Evaluation Criteria for PEE

This evaluation is based on how many facts recognized by a patent examiner at a previous patent examination are understood from the translation results of a reference patent. The criteria consist of 6 grades. The evaluation criteria are shown in Table 2. The evaluation unit was a document. One evaluation value was given to one translated reference patent.

Table 2: The evaluation criteria for PEE

Grade	Description
VI	All facts useful for recognizing the cited invention were recognized and examination could be done using only the translation results.
V	At least half of the facts useful for recognizing the cited invention were recognized and the translation results were useful for examination.
IV	One or more facts useful for recognizing the cited invention were recognized and the translation results were useful for examination.
III	Falls short of reaching IV, but parts of the facts were recognized and it was proved that the cited invention could not be disregarded at the examination.
II	Parts of the facts were recognized but the translation results could not be seen as useful for examination.
I	None of the facts were recognized and the translation results were not useful for examination.

(Evaluation unit is document)

2.1.3 *Chronological Evaluation (ChE)*

We compared the translation results of NTCIR-10 and 9 to measure progress over time, using the NTCIR-9 test sets, for all the subtasks. This evaluation used the automatic evaluation measures of RIBES and BLEU described in Section 2.1.1.

2.1.4 *Multilingual Evaluation (ME)*

We compared CE and JE translations using the same English references to see the source language dependency. This evaluation was conducted for the CE and JE subtasks. For this purpose, we used a Chinese–Japanese–English trilingual test set. This evaluation used the automatic evaluation measures of RIBES and BLEU described in Section 2.1.1.

2.2 Test Collection

Table 3: 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 (IE)	All	2,300 patent test sentences
		Context documents
		2,300 reference sentences
Test (PEE)	CE & JE	29 patents
Test (ChE)	All	2,000 patent test sentences
		Context documents
		2,000 reference sentences
Test (ME)	CE & JE	2,000 patent test sentences
		Context documents (Japanese only)
		2,000 reference sentences

The test collections consisted of training data, development data, test data, context documents, and reference data. There was an exception: the test sets for PEE do not include reference data and context documents. The training, development, and test data was from patent description sentences. (Patent documents consist of a title, abstract, claim, and description.) The contents for the test collection are shown in Table 3.

2.2.1 *Training data for the CE subtask*

The Chinese–English test collection was chosen from a large Chinese–English bilingual parallel corpus of sentence pairs [17]. We used the same training and development data as NTCIR-9 [10]. The training and development data sets were built in the following manner:

The Chinese–English parallel sentence pairs were extracted from Patent Cooperation Treaty (PCT) patents in Chinese and English. First, we divided our Chinese–English bilingual corpus into two sub-corpora with the following criteria: 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 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 were collected: 1 million sentence pairs for the training data set, and 2,000 sentence pairs each for development data. The patent documents that the development data were extracted from were provided as context documents for the development data.

2.2.2 Training data for the JE and EJ subtasks

We used the same training and development data as the NTCIR-8 [8] and NTCIR-9 [10].

The parallel data for the training and development data was automatically extracted from patent families in Japanese and English [25]. 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 that the development data were extracted from were provided as context documents for the development data.

2.2.3 Test data for Intrinsic Evaluation

We built a new test set consisting of 2,300 bilingual sentence pairs for IE for each subtask. The test sentences were randomly selected from patent documents. If test data was simply selected from the automatically extracted parallel corpus, biases such as length or included expressions may result. To reduce bias, we selected test sentences using one method to select 2,000 test sentences and using another method to select 300 test sentences. The 300 test sentences were used for human evaluations. All of the 2,300 test sentences were used for automatic evaluation.

The 2,000 test sentences were selected by first, obtaining a sentence-length distribution of patents in the source language. We collected source language patents published in 2006 and 2007 where there were corresponding target language patents. We counted the number of sentences for each sentence length from the sentences in the description sections of the collected patents in the source language, and calculated the cumulative length distribution. We divided the cumulative length distribution into quarters (25% each). Next, we selected more than 400 patents in the source language published in 2006 and 2007 for each subtask where there were corresponding target language patents and which had not been used to produce either development data or previous NTCIR test data. We call the selected documents *the selected bilingual patents*. We randomly selected more than 500 sentences for each length division as test set candidates from the parallel sentences automatically extracted from the selected bilingual patents. We manually selected correct translation pairs from the candidates till the number of correct translation pairs was equal to 500 for each of the four length divisions, obtaining 2,000 bilingual test and reference sentence pairs.

The 300 test sentences were selected by randomly selecting sentences from all of the description sentences in the source language patents of the selected bilingual patents. We manually translated the selected 300 test sentences to produce reference sentences. When the sentences were translated, the original bilingual patent documents were provided to the translation company to check the translations of technical terms.

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.

2.2.4 Test data for Patent Examination Evaluation

We built test data consisting of 29 patents in Japanese and Chinese for PEE. We used 29 patents in Japanese that were used to reject other patent applications. We used *shinketsu*, which are the final decisions of patent examinations, to produce the test set. The shinketsu whose decisions were rejections include what facts were recognized from the reference patents at the examinations.

The test set was built as follows.

1. We collected shinketsu whose decisions were rejections.

2. We extracted descriptions of what facts were recognized from the reference patents.

3. From the reference patents, we extracted sentences that were evidence of the recognized facts in the shinketsu. This extraction was conducted by patent attorneys. The extracted sentences were used as test data for the JE subtask.

4. We manually translated the test data in Japanese into Chinese. The translated test data in Chinese was used as the test data for the CE subtask.

Example data for 1, 2, and 3 above is shown in Appendix D.

2.2.5 Test data for Chronological Evaluation

We used the NTCIR-9 test data as the NTCIR-10 PatentMT test data for ChE.

2.2.6 Test data for Multilingual Evaluation

We used the NTCIR-9 JE subtask test data as the NTCIR-10 PatentMT JE subtask test data for ME. We manually translated the NTCIR-9 JE subtask test data into Chinese. The translated test data in Chinese was used as the NTCIR-10 PatentMT CE subtask test data for ME.

2.3 Schedule

The task schedule is summarized in Table 4. We spent roughly 3.5 months for training and two weeks for translating.

Table 4: Schedule for PatentMT at NTCIR-10

Event	Date
Training corpus release	6/29/2012
Test data release	10/15/2012
Result submission deadline	10/28/2012

Table 5: Participants and Subtasks Participated In

Group ID	Participant	Subtask		
		CE	JE	EJ
ISTIC	Institute of Scientific and Technical Information of China [11]	✓	✓	✓
TORI	Tottori University [18]		✓	
RWTH	RWTH Aachen University [6]	✓	✓	
EIWA	Yamanashi Eiwa College [4]	✓	✓	✓
TSUKU	University of Tsukuba [31]			✓
JAPIO	Japan Patent Information Organization (Japio) [20]		✓	✓
UQAM	UQAM [22]		✓	✓
BJTUX	Beijing Jiaotong University [27]	✓	✓	✓
SJTU	Shanghai Jiao Tong University [30]	✓		
BUAA	BeiHang University, School of Computer Science & Engineering [2]	✓		
MIG	Department of Computer Science, National Chengchi University, Taiwan [26]	✓		
FUN-NRC	Future University Hakodate / National Research Council Canada [9]		✓	✓
KYOTO	Kyoto University [19]		✓	✓
NTITI	NTT Corporation / National Institute of Informatics [24]		✓	✓
OKAPU	Okayama Prefectural University [14]		✓	
HDU	Institute for Computational Linguistics, Heidelberg University [23]	✓	✓	
DCUMT	Dublin City University			✓
TRGTK	Torangetek Inc. [28]	✓	✓	✓
BBN	Raytheon BBN Technologies [13]	✓		
RWSYS	RWTH Aachen University / Systran [5]	✓		
SRI	SRI International [29]	✓		

Table 6: Baseline systems

SYSTEM-ID	SYSTEM	Type	CE	JE	EJ
BASELINE1	Moses' hierarchical phrase-based SMT system [12]	SMT	✓	✓	✓
BASELINE2	Moses' phrase-based SMT system [16]	SMT	✓	✓	✓
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 (October, 2012)	SMT	✓	✓	✓

3. PARTICIPANTS AND SUBMISSIONS

We received submissions from 21 groups. The number of groups for each subtask were: 12 for CE, 13 for JE, and 11 for EJ. Table 5 shows the participant organizations and the subtasks they participated in.

In addition to the submissions from the participants, the organizers submitted results for baseline systems that consisted of 2 SMT systems, 3 commercial RBMT systems, and 1 online SMT system[5]. The baseline systems are shown in Table 6. 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. 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 commercial RBMT systems from companies that did not themselves participate, the SYSTEM-IDs of the commercial RBMT systems are anonymized in this paper.

Each participant was allowed to submit as many translated results ("runs") as they wished for the intrinsic evaluation (IE), but at least one result had to have been produced using only the parallel corpus for training both the translation and language models whenever they used the corpus-based MT method. The submitted runs were to be prioritized by the participant for IE. For the other evaluations (PEE, ChE, and ME)[6], one translated result was required

[5] The Google translation system may have used training data that contained patents published in 2006 and 2007 that the test and reference sentences were extracted from. Therefore, the evaluation results of ONLINE1 did not fairly compare with the other results.

[6] Since the EJ subtask did not conduct PEE and ME, these results were not required for the EJ subtask.

to be submitted for each evaluation.[7,8]

In this paper, we distinguish runs by using a Run ID expressed by a Group ID (or a System ID for the baseline systems) and a priority number connected by a dash. Some features from the submissions and their automatic evaluation scores are given in Appendix E. 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.
- *Resource E*: The system used external information other than data provided by the organizers, or the system used a rule-based system.
- *Resource C*: The system used context information.

Type roughly indicates the type of system. The task definition defines the types as: "SMT"= statistical MT, "EBMT" = example-based MT, "RBMT" = rule-based MT, or "HYBRID" = hybrid MT. In this paper, HYBRID was used for systems using both RBMT and another type or types of MT.[9]

4. INTRINSIC EVALUATION RESULTS

In this section, we show the human evaluation results for the intrinsic evaluation. We evaluated the adequacy for at least all of the first priority submissions that were in time for the human evaluation schedule.[10] However, because of budget limitations, acceptability was evaluated for only selected systems.[11]

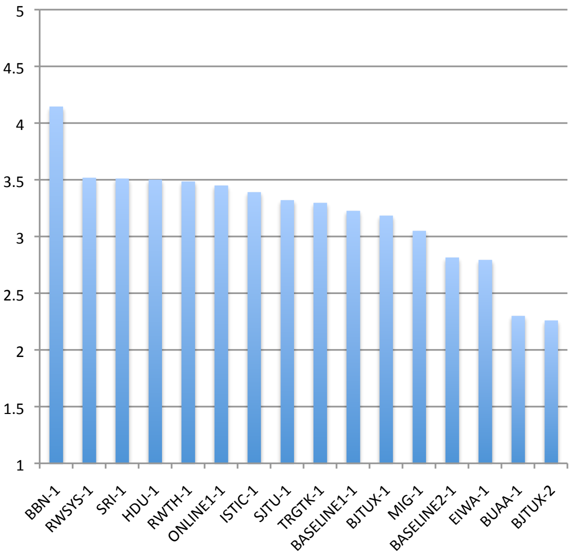

Figure 4: Results of CE adequacy.

Table 10: Sign test of CE acceptability. "$\gg$": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, and "-": not significantly different at $\alpha = 0.05$.

	RWTH-1	RWSYS-1	HDU-1	ONLINE1-1	SRI-1	TRGTK-1	ISTIC-1	SJTU-1
BBN-1	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
RWTH-1		-	-	>	$\gg$	$\gg$	$\gg$	$\gg$
RWSYS-1			-	-	-	$\gg$	$\gg$	$\gg$
HDU-1				-	-	>	$\gg$	$\gg$
ONLINE1-1					-	-	>	$\gg$
SRI-1						-	-	>
TRGTK-1							-	-
ISTIC-1								-

[7] There were submissions made after the result submission deadline. The groups whose submissions included submissions made after the result submission deadline were as follows. The CE subtask: BUAA. The JE subtask: UQAM and OKAPU. The EJ subtask: EIWA (ChE result), DCUMT and UQAM. Updates for corrections of format-level errors were not regarded as late submissions. Although submissions made after the result submission deadline did not meet the requirements, the organizers accepted these submissions as long as the submissions were in time for the official evaluations. We did not make a distinction between these submissions and submissions that met the deadline in our evaluations because our purpose is technology evaluation and this problem does not directly relate to technology.

[8] Some groups did not submit all of the required submissions. The groups whose submissions did not contain all of the required submissions were as follows. The CE subtask: BUAA, ISTIC, SRI, and TRGTK. The JE subtask: ISTIC, NTITI, TORI, and TRGTK. The EJ subtask: DCUMT, ISTIC, NTITI, and TRGTK. Note that when making this list, priority 1 systems that did not use an external resource were regarded as the corpus-based MT method. We did not distinguish between these groups' submissions and submissions that contained all the required files in our evaluations because our purpose is technology evaluation and this problem does not directly relate to technology.

[9] We re-assigned or added types in this paper for some submissions to meet our type category.

[10] Human evaluation could not be conducted for the submissions from UQAM and DCUMT because these submissions were not in time for the human evaluation schedule. We prepared human evaluation resources for adequacy for only the priority 1 results from all of the groups that registered to participate and from selected baseline systems. However, since some groups withdrew and two groups were unable to submit results in time for human evaluations, we used the surplus human evaluation resources caused by this reduction in the number of priority 1 results to evaluate results other than priority 1. We selected systems for the additional human evaluations using the system descriptions in the submission results according to the following criteria: (i) The method is different from the priority 1 method. (ii) The evaluation is expected to provide interesting information such as revealing the effectiveness of the methods by evaluating other types of systems, evaluating each component method in the priority 1 system combination, or comparing with baselines.

[11] We selected systems whose adequacy scores were high for the acceptability evaluation. We initially planed to evaluate only priority 1 results. However, in the EJ subtask, the best adequacy was achieved by a priority 2 result. Therefore, we made exceptions and also evaluated the best adequacy results.

Table 7: Results of CE adequacy

Run ID	Type	Resource			Average	Rate				
		B	M	E	score	5	4 or higher	3 or higher	2 or higher	1 or higher
BBN-1	SMT	✓	✓		4.15	0.520	0.740	0.887	1.000	1.000
RWSYS-1	HYBRID	✓	✓	✓	3.52	0.273	0.483	0.773	0.990	1.000
SRI-1	SMT	✓	✓		3.51	0.247	0.490	0.790	0.987	1.000
HDU-1	SMT	✓			3.50	0.270	0.510	0.743	0.980	1.000
RWTH-1	SMT	✓	✓		3.49	0.267	0.480	0.753	0.987	1.000
ONLINE1-1	SMT			✓	3.45	0.233	0.453	0.770	0.997	1.000
ISTIC-1	SMT	✓	✓		3.39	0.223	0.467	0.717	0.987	1.000
SJTU-1	SMT	✓			3.32	0.210	0.410	0.720	0.983	1.000
TRGTK-1	SMT	✓	✓		3.30	0.187	0.407	0.720	0.987	1.000
BASELINE1-1	SMT	✓			3.23	0.153	0.397	0.700	0.980	1.000
BJTUX-1	SMT	✓		✓	3.19	0.167	0.367	0.667	0.987	1.000
MIG-1	SMT	✓			3.05	0.133	0.280	0.647	0.993	1.000
BASELINE2-1	SMT	✓			2.82	0.127	0.240	0.490	0.960	1.000
EIWA-1	HYBRID	✓		✓	2.80	0.067	0.240	0.523	0.967	1.000
BUAA-1	SMT	✓	✓		2.30	0.020	0.070	0.343	0.870	1.000
BJTUX-2	EBMT	✓			2.26	0.017	0.063	0.353	0.830	1.000

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

	RWSYS-1	SRI-1	HDU-1	RWTH-1	ONLINE1-1	ISTIC-1	SJTU-1	TRGTK-1	BASELINE1-1	BJTUX-1	MIG-1	BASELINE2-1	EIWA-1	BUAA-1	BJTUX-2
BBN-1	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RWSYS-1		-	-	-	-	>	>	≫	≫	≫	≫	≫	≫	≫	≫
SRI-1			-	-	-	-	≫	≫	≫	≫	≫	≫	≫	≫	≫
HDU-1				-	-	-	>	>	≫	≫	≫	≫	≫	≫	≫
RWTH-1					-	>	>	≫	≫	≫	≫	≫	≫	≫	≫
ONLINE1-1						-	-	>	≫	≫	≫	≫	≫	≫	≫
ISTIC-1							-	-	>	≫	≫	≫	≫	≫	≫
SJTU-1								-	-	≫	≫	≫	≫	≫	≫
TRGTK-1									-	-	≫	≫	≫	≫	≫
BASELINE1-1										-	≫	≫	≫	≫	≫
BJTUX-1											>	≫	≫	≫	≫
MIG-1												≫	≫	≫	≫
BASELINE2-1													-	≫	≫
EIWA-1														≫	≫
BUAA-1															-

Table 9: Results of CE acceptability

Run ID	Type	Resource			Pairwise	Rate				
		B	M	E	score	AA	A or higher	B or higher	C or higher	F or higher
BBN-1	SMT	✓	✓		0.685	0.270	0.390	0.557	0.673	1.000
RWTH-1	SMT	✓	✓		0.533	0.147	0.207	0.350	0.467	1.000
RWSYS-1	HYBRID	✓	✓	✓	0.523	0.133	0.193	0.347	0.467	1.000
HDU-1	SMT	✓			0.494	0.090	0.190	0.307	0.430	1.000
ONLINE1-1	SMT			✓	0.482	0.083	0.117	0.270	0.477	1.000
SRI-1	SMT	✓	✓		0.478	0.077	0.153	0.280	0.433	1.000
TRGTK-1	SMT	✓	✓		0.444	0.070	0.130	0.250	0.397	1.000
ISTIC-1	SMT	✓	✓		0.436	0.050	0.117	0.257	0.367	1.000
SJTU-1	SMT	✓			0.425	0.070	0.143	0.240	0.353	1.000

4.1 Chinese to English

4.1.1 Adequacy Evaluation

Figure 4 and Table 7 show the results of the adequacy evaluation. Table 8 shows the results of the statistical significance test of the adequacy evaluation using a sign test. In the tables showing the results of a statistical significance test, the marks "≫", ">", and "-" indicate whether the Run ID to the left of a mark is significantly better than that above the mark.

From these results, we can observe the followings:

- All the top system are SMT or hybrid systems. The top system, BBN-1 [13], shows a significantly higher adequacy than the other systems. The second ranked RWSYS system combines various systems, including

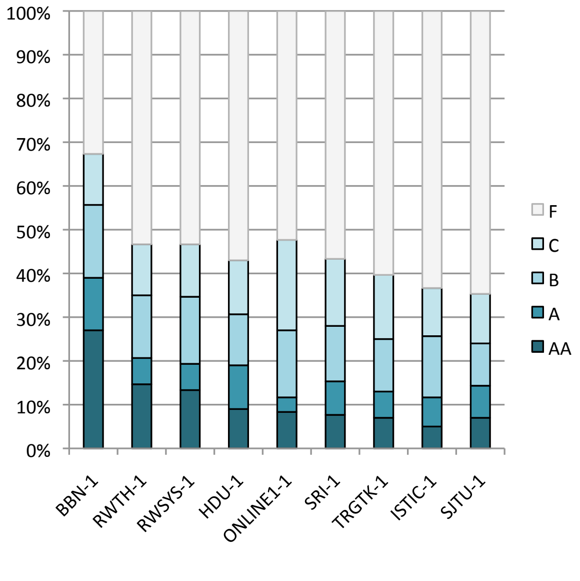

Figure 5: Results of CE acceptability.

phrase-based and hierarchical SMT, rule-based MT (RBMT) and MT with statistical post-editing [5].

- The adequacy score for Moses' hierarchical phrase-based SMT system (BASELINE1-1) is higher than that for Moses' phrase-based SMT system (BASELINE2-1).

4.1.2 Acceptability Evaluation

Figure 5 and Table 9 show the results of the acceptability evaluation. Table 10 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 67% of the translated sentences in the best-ranked system (BBN-1). This result significantly surpasses the others.

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

	TORI-1	EIWA-1	NTITI-1	RWTH-1	HDU-1	ONLINE1-1	FUN-NRC-1	KYOTO-1
JAPIO-1	>	≫	≫	≫	≫	≫	≫	≫
TORI-1		-	≫	≫	≫	≫	≫	≫
EIWA-1			>	≫	≫	≫	≫	≫
NTITI-1				-	>	≫	≫	≫
RWTH-1					-	>	≫	≫
HDU-1						-	-	>
ONLINE1-1							-	-
FUN-NRC-1								-

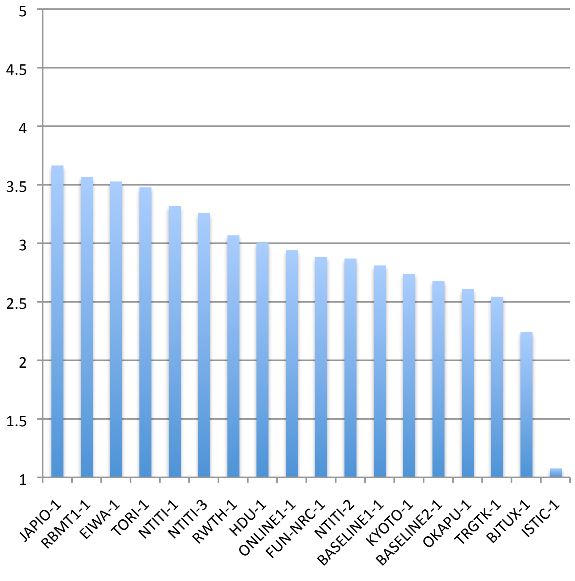

Figure 6: Results of JE adequacy.

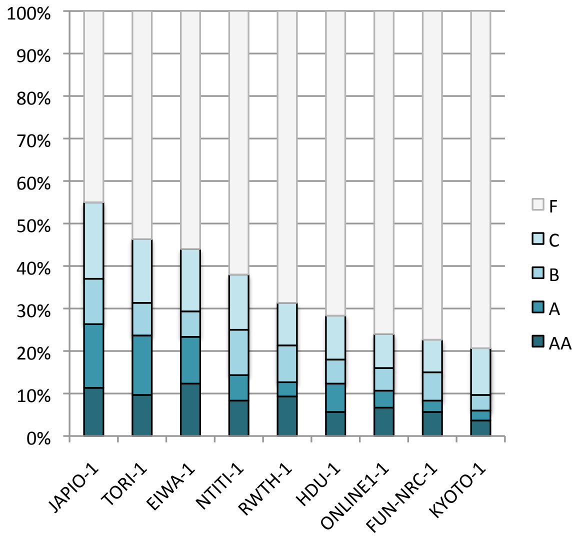

Figure 7: Results of JE acceptability.

4.2 Japanese to English

4.2.1 Adequacy Evaluation

Figure 6 and Table 11 show the results of the adequacy evaluation. Table 12 shows the results of the statistical significance test of the adequacy evaluation using a sign test.

The top four systems, JAPIO-1 [20], RBMT1-1, EIWA-1 [4], and TORI-1 [18] are either commercial RBMT systems or systems that use commercial RBMT systems. The best-ranked SMT system (NTITI-1) used a system combination that included a post-ordering method (NTITI-3) and a pre-ordering method (NTITI-2) [24]. The third-best ranked SMT system (RWTH-1) [6] used phrase-based SMT and a hierarchical phrase reordering model. From these results,

Table 11: Results of JE adequacy

Run ID	Type	Resource			Average	Rate				
		B	M	E	score	5	4 or higher	3 or higher	2 or higher	1 or higher
JAPIO-1	RBMT			✓	3.67	0.303	0.530	0.843	0.990	1.000
RBMT1-1	RBMT			✓	3.57	0.250	0.513	0.820	0.987	1.000
EIWA-1	HYBRID	✓		✓	3.53	0.227	0.497	0.817	0.990	1.000
TORI-1	HYBRID	✓		✓	3.48	0.223	0.490	0.780	0.987	1.000
NTITI-1	SMT	✓	✓		3.32	0.193	0.410	0.727	0.993	1.000
NTITI-3	SMT	✓	✓		3.26	0.183	0.380	0.707	0.990	1.000
RWTH-1	SMT	✓	✓		3.07	0.150	0.317	0.623	0.980	1.000
HDU-1	SMT	✓			3.01	0.157	0.293	0.587	0.973	1.000
ONLINE1-1	SMT			✓	2.94	0.107	0.257	0.603	0.977	1.000
FUN-NRC-1	SMT	✓	✓		2.89	0.103	0.227	0.583	0.973	1.000
NTITI-2	SMT	✓	✓		2.87	0.130	0.243	0.530	0.970	1.000
BASELINE1-1	SMT	✓			2.81	0.087	0.200	0.553	0.973	1.000
KYOTO-1	EBMT	✓			2.74	0.083	0.213	0.517	0.930	1.000
BASELINE2-1	SMT	✓			2.68	0.077	0.140	0.507	0.957	1.000
OKAPU-1	SMT	✓			2.61	0.043	0.153	0.463	0.950	1.000
TRGTK-1	SMT	✓	✓		2.55	0.050	0.100	0.443	0.953	1.000
BJTUX-1	SMT	✓		✓	2.25	0.027	0.040	0.293	0.887	1.000
ISTIC-1	SMT	✓	✓		1.08	0.000	0.000	0.007	0.073	1.000

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

	RBMT1-1	EIWA-1	TORI-1	NTITI-1	NTITI-3	RWTH-1	HDU-1	ONLINE1-1	FUN-NRC-1	NTITI-2	BASELINE1-1	KYOTO-1	BASELINE2-1	OKAPU-1	TRGTK-1	BJTUX-1	ISTIC-1
JAPIO-1	>	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RBMT1-1		-	-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
EIWA-1			-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
TORI-1				>	>	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
NTITI-1					-	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
NTITI-3						≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫	≫
RWTH-1							-	≫	>	≫	≫	≫	≫	≫	≫	≫	≫
HDU-1								-	-	>	-	≫	≫	≫	≫	≫	≫
ONLINE1-1									-	-	-	≫	≫	≫	≫	≫	≫
FUN-NRC-1										-	-	>	≫	≫	≫	≫	≫
NTITI-2											-	>	≫	≫	≫	≫	≫
BASELINE1-1												-	≫	≫	≫	≫	≫
KYOTO-1													-	-	≫	≫	≫
BASELINE2-1														-	-	≫	≫
OKAPU-1															-	≫	≫
TRGTK-1																≫	≫
BJTUX-1																	≫

Table 13: Results of JE acceptability

Run ID	Type	Resource			Pairwise	Rate				
		B	M	E	score	AA	A or higher	B or higher	C or higher	F or higher
JAPIO-1	RBMT			✓	0.630	0.113	0.263	0.370	0.550	1.000
TORI-1	HYBRID	✓		✓	0.580	0.097	0.237	0.313	0.463	1.000
EIWA-1	HYBRID	✓		✓	0.567	0.123	0.233	0.293	0.440	1.000
NTITI-1	SMT	✓	✓		0.515	0.083	0.143	0.250	0.380	1.000
RWTH-1	SMT	✓	✓		0.479	0.093	0.127	0.213	0.313	1.000
HDU-1	SMT	✓			0.457	0.057	0.123	0.180	0.283	1.000
ONLINE1-1	SMT			✓	0.439	0.067	0.107	0.160	0.240	1.000
FUN-NRC-1	SMT	✓	✓		0.429	0.057	0.083	0.150	0.227	1.000
KYOTO-1	EBMT	✓			0.404	0.037	0.060	0.097	0.207	1.000

the following are observed:

- The commercial RBMT systems had higher adequacies than the state-of-the-art SMT systems for patent machine translation from Japanese to English.
- A post-ordering method was effective because NTITI-3 used a post-ordering method and achieved an adequacy score close to that of the best-ranked SMT system (NTITI-1), and the best-ranked SMT system used the results of NTITI-3.

- A hierarchical phrase reordering model was effective for phrase-based SMT because RWTH-1 used a hierarchical phrase reordering model and phrase-based SMT, and it outperformed the phrase-based SMT baseline system (BASELINE2-1).

The reason that the SMT systems could not achieve adequacy scores as high as those from the top RBMT systems is thought to be because of word reordering. Since the word order in Japanese and English is significantly different (Japanese is a Subject-Object-Verb (SOV) language and English is a Subject-Verb-Object (SVO) language), word reordering is difficult for Japanese–English translation. The current SMT performs well for word selection, but not for the difficult word reordering of Japanese–English translation. On the other hand, the baseline commercial RBMT systems perform well for word reordering of Japanese–English translations. The results showing that RBMT systems were better than SMT systems were the same as the previous human evaluation results at NTCIR-7 [7] and NTCIR-9 [10].

4.2.2 *Acceptability Evaluation*

Figure 7 and Table 13 show the results of the acceptability evaluation. Table 14 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 55% of the sentences in the best-ranked system using RBMT (JAPIO-1). For the best-ranked SMT system (NTITI-1), the source sentence meaning could be understood for 38% of the translated sentences (C-rank and above).

At NTCIR-9, the rate for the best-ranked RBMT system was 63% and the rate for the best-ranked SMT system was 25%. Although there was still a large difference in the ability to retain the sentence-level meanings between the top-level commercial RBMT systems and the SMT systems for Japanese-to-English patent translation, the top SMT translation quality improved and the difference decreased.

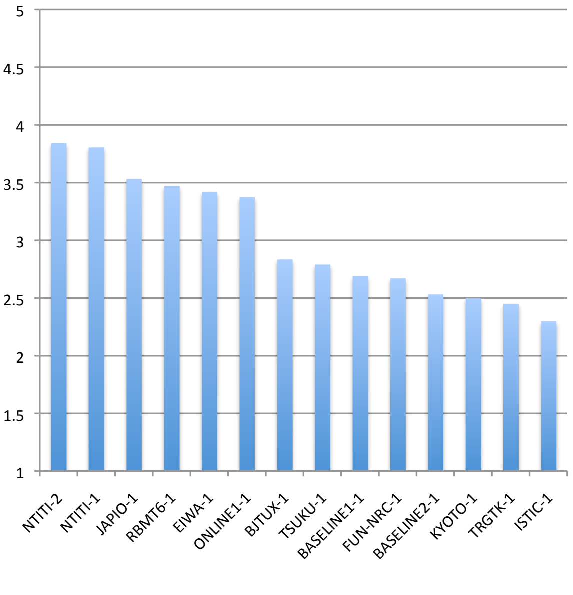

Figure 8: Results of EJ adequacy.

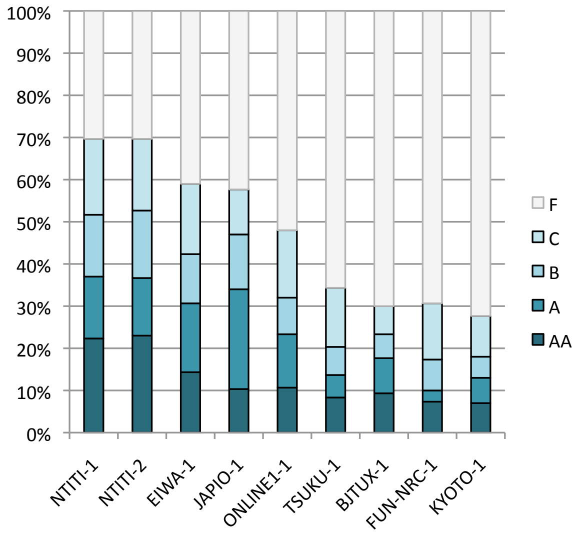

Figure 9: Results of EJ acceptability.

Table 18: Sign test of EJ acceptability. "≫": significantly different at $\alpha = 0.01$, ">": significantly different at $\alpha = 0.05$, and "-": not significantly different at $\alpha = 0.05$.

	NTITI-2	EIWA-1	JAPIO-1	ONLINE1-1	TSUKU-1	BJTUX-1	FUN-NRC-1	KYOTO-1
NTITI-1	-	≫	≫	≫	≫	≫	≫	≫
NTITI-2		≫	≫	≫	≫	≫	≫	≫
EIWA-1			-	>	≫	≫	≫	≫
JAPIO-1				≫	≫	≫	≫	≫
ONLINE1-1					≫	≫	≫	≫
TSUKU-1						-	-	-
BJTUX-1							-	-
FUN-NRC-1								-

4.3 English to Japanese

4.3.1 *Adequacy Evaluation*

Figure 8 and Table 15 show the results of the adequacy evaluation. Table 16 shows the results of the statistical significance test of the adequacy evaluation using a sign test.

In the top systems, NTITI-2 and NTITI-1 [24] are SMT systems, and JAPIO-1 [20], RBMT6-1, and EIWA-1 [4] are either commercial RBMT systems or systems that use commercial RBMT systems. From these results, the following are observed:

- The best-ranked SMT system (NTITI-2) was significantly better than the best-ranked RBMT system for patent machine translation from English to Japanese.
- A dependency parser and a pre-ordering method were effective for EJ translation because the best-ranked SMT system (NTITI-2) used a dependency parser and a pre-ordering method [24].
- The adequacy scores for the top-ranked commercial RBMT systems were significantly higher than those

Table 15: Results of EJ adequacy

Run ID	Type	Resource			Average	Rate				
		B	M	E	score	5	4 or higher	3 or higher	2 or higher	1 or higher
NTITI-2	SMT	✓	✓		3.84	0.303	0.707	0.857	0.977	1.000
NTITI-1	SMT	✓	✓		3.81	0.300	0.680	0.847	0.980	1.000
JAPIO-1	RBMT			✓	3.53	0.213	0.587	0.767	0.967	1.000
RBMT6-1	RBMT			✓	3.47	0.237	0.560	0.747	0.930	1.000
EIWA-1	HYBRID	✓		✓	3.42	0.240	0.517	0.723	0.940	1.000
ONLINE1-1	SMT			✓	3.38	0.223	0.470	0.740	0.943	1.000
BJTUX-1	SMT	✓			2.84	0.130	0.303	0.503	0.900	1.000
TSUKU-1	SMT	✓			2.79	0.110	0.313	0.490	0.880	1.000
BASELINE1-1	SMT	✓			2.69	0.117	0.250	0.450	0.873	1.000
FUN-NRC-1	SMT	✓	✓		2.67	0.103	0.207	0.443	0.920	1.000
BASELINE2-1	SMT	✓			2.53	0.103	0.200	0.387	0.843	1.000
KYOTO-1	EBMT	✓			2.50	0.093	0.240	0.427	0.740	1.000
TRGTK-1	SMT	✓	✓		2.45	0.080	0.163	0.333	0.873	1.000
ISTIC-1	SMT	✓	✓		2.30	0.060	0.130	0.290	0.820	1.000

Table 16: Sign test of EJ adequacy. “$\gg$”: significantly different at $\alpha = 0.01$, “$>$”: significantly different at $\alpha = 0.05$, and “-”: not significantly different at $\alpha = 0.05$.

	NTITI-1	JAPIO-1	RBMT6-1	EIWA-1	ONLINE1-1	BJTUX-1	TSUKU-1	BASELINE1-1	FUN-NRC-1	BASELINE2-1	KYOTO-1	TRGTK-1	ISTIC-1
NTITI-2	-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NTITI-1		$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
JAPIO-1			-	-	-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
RBMT6-1				-	-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
EIWA-1					-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
ONLINE1-1						$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
BJTUX-1							-	-	-	$\gg$	$\gg$	$\gg$	$\gg$
TSUKU-1								-	-	$\gg$	$\gg$	$\gg$	$\gg$
BASELINE1-1									-	$>$	$\gg$	$\gg$	$\gg$
FUN-NRC-1										$>$	$>$	$\gg$	$\gg$
BASELINE2-1											-	-	$\gg$
KYOTO-1												-	-
TRGTK-1													$>$

Table 17: Results of EJ acceptability

Run ID	Type	Resource			Pairwise	Rate				
		B	M	E	score	AA	A or higher	B or higher	C or higher	F or higher
NTITI-1	SMT	✓	✓		0.659	0.223	0.370	0.517	0.697	1.000
NTITI-2	SMT	✓	✓		0.658	0.230	0.367	0.527	0.697	1.000
EIWA-1	HYBRID	✓		✓	0.568	0.143	0.307	0.423	0.590	1.000
JAPIO-1	RBMT			✓	0.560	0.103	0.340	0.470	0.577	1.000
ONLINE1-1	SMT			✓	0.499	0.107	0.233	0.320	0.480	1.000
TSUKU-1	SMT	✓			0.409	0.083	0.137	0.203	0.343	1.000
BJTUX-1	SMT	✓			0.402	0.093	0.177	0.233	0.300	1.000
FUN-NRC-1	SMT	✓	✓		0.376	0.073	0.100	0.173	0.307	1.000
KYOTO-1	EBMT	✓			0.370	0.070	0.130	0.180	0.277	1.000

for SMT systems other than NTITI-2, NTITI-1, and ONLINE1-1.

Although English-to-Japanese translation is difficult for SMT because English and Japanese word order is significantly different, the pre-ordering method of NTITI-2 could handle this issue well. At the NTCIR-9, the top SMT system achieved an adequacy score comparable to that of the best-ranked RBMT system. In this evaluation, the top SMT system (NTITI-2) significantly outperformed the top-level RBMT systems.

4.3.2 Acceptability Evaluation

Figure 9 and Table 17 show the results of the acceptability evaluation. Table 18 shows the results of the statistical significance test of the acceptability evaluation using a sign test.

For the best SMT system (NTITI-1), the source sentence meaning could be understood (C and above) for 70% of the translated sentences. For the best RBMT system (JAPIO-1), the source sentence meaning could be understood (C and above) for 58% of the translated sentences.

The translation quality of the best-ranked SMT system (NTITI-1) was better than that of the top-level commercial

RBMT systems for retaining the sentence-level meanings.

5. PATENT EXAMINATION EVALUATION RESULTS

This section describes the Patent Examination Evaluation (PEE) results. Two evaluators who are experienced patent examiners evaluated. Each evaluator evaluated 20 patents for each system. 29 patents were used as test data. Translations of 11 patents were evaluated by both of the evaluators and translations of the remaining 18 patents were evaluated by one of the evaluators. For the patents evaluated by the two evaluators, we divided the number of documents in half to calculate the total evaluation results.

We selected three systems each for CE and JE translations to evaluate PEE, using the following criteria: (i) Top level adequacy and (ii) inclusion of many types of methods.[12]

Figure 10 and Table 19 show the results for CE translations and Figure 11 and Table 20 show the results for JE translations.

From the CE translation results, the best system (BBN-1) achieved 21% for VI and 88% for V and above. This indicated that the best system (BBN-1) is useful for patent examinations.

From the JE translations results, the best system (JAPIO-1) achieved 66% for VI and 100% for V and above. This results indicated that the best system (JAPIO-1) is useful for patent examination. Since JAPIO-1 was an RBMT system, it can be seen that the top RBMT system was consistent in its translation quality and was better than SMT for patent examination. The SMT system of NTITI-1 achieved 18% for VI and 64% for V and above. This indicated that the best-ranked SMT system is also useful for patent examinations to some extent.

The Japanese test sentences were existing real patent sentences, whereas the Chinese test data was produced by translating the Japanese test data into Chinese. Although the contents of the test data were the same, the translation results between Chinese-to-English translation and Japanese-to-English translation could not be fairly compared with generality due to bias such as domains and effects from the Japanese-to-Chinese manual translation. These biases would cause adverse effects on Chinese-to-English translation. The results of the Multilingual Evaluation (ME) described in Section 7 indicate these adverse effects. Therefore, the results would indicate that the actual performance for Chinese-to-English translation would be better than these evaluation results.

We received comprehensive comments for each system from the evaluators. The comprehensive comments are shown in Tables 21 and 22. Although evaluations for EIWA-1 were different between the evaluators, JAPIO-1 was evaluated highly by the both of the two evaluators. BBN-1 was also evaluated highly by the evaluators.

These evaluation results and translations can be used as standards of usefulness in patent examination. Concretely, by comparing new translation results of the PEE test data with these evaluated translations, their usefulness in patent examination for other systems can roughly be judged.

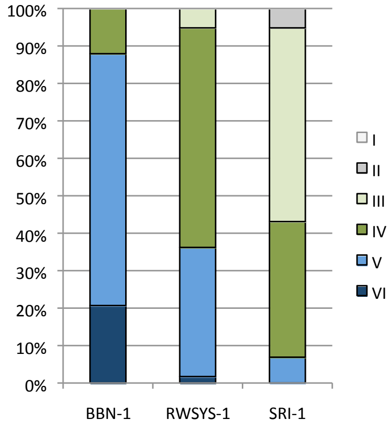

Figure 10: Results of PEE (CE).

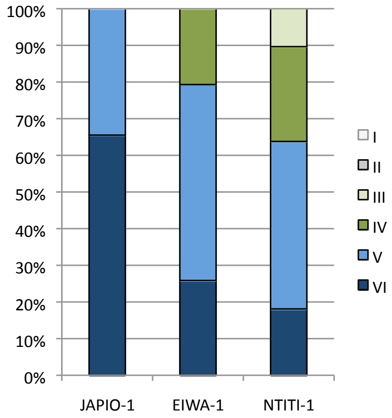

Figure 11: Results of PEE (JE).

6. CHRONOLOGICAL EVALUATION RESULTS

This section shows the chronological evaluation (ChE) results.[13] The NTCIR-9 test data was translated by the NTCIR-10 participants and these translations were compared to the submissions at NTCIR-9 to measure progress

[12] For CE translation, we selected BBN-1, RWSYS-1, and SRI-1 because these three systems achieved top-level adequacy and BBN-1 was the best-ranked system, RWSYS-1 was the best-ranked of the HYBRID systems that used a rule-based system, and SRI-1 was the second-ranked of the SMT systems on the adequacy results. For JE translation, we selected JAPIO-1, EIWA-1, and NTITI-1 because these three systems achieved top-level adequacy and JAPIO-1 was the best-ranked system, EIWA-1 was the best-ranked of the HYBRID systems, and NTITI was the best-ranked of the SMT systems on the adequacy results.

[13] The baseline systems were basically the same as those at NTCIR-9. The SMT baseline systems for CE translation were exactly the same. The differences between the SMT baseline systems at NTCIR-9 and NTCIR-10 for JE and EJ translation were from pre-processing. The differences for the RBMT baseline systems were from the pre-process and system configurations such as dictionary order.

Table 19: Results of PEE (CE)

Run ID	Type	Resource			Rate					
		B	M	E	VI	V or higher	IV or higher	III or higher	II or higher	I or higher
BBN-1	SMT	✓	✓		0.21	0.88	1.00	1.00	1.00	1.00
RWSYS-1	HYBRID	✓	✓	✓	0.02	0.36	0.95	1.00	1.00	1.00
SRI-1	SMT	✓	✓		0.00	0.07	0.43	0.95	1.00	1.00

Table 20: Results of PEE (JE)

Run ID	Type	Resource			Rate					
		B	M	E	VI	V or higher	IV or higher	III or higher	II or higher	I or higher
JAPIO-1	RBMT			✓	0.66	1.00	1.00	1.00	1.00	1.00
EIWA-1	HYBRID	✓		✓	0.26	0.79	1.00	1.00	1.00	1.00
NTITI-1	SMT	✓	✓		0.18	0.64	0.90	1.00	1.00	1.00

Table 21: PEE comprehensive comments (evaluator 1)

	Run ID	ID for evaluation	Comprehensive comments from evaluator 1
CE	BBN-1	MT1	Second-most consistent after MT4 in its translation quality. The system seemed to try to translate complicated input sentences depending on context and I would like to applaud this.
	RWSYS-1	MT2	There were fragmental translations. To understand the translations, sentences before or after, or common knowledge of technology were needed for many parts.
	SRI-1	MT3	Hard to read. It would not be practical for patent examination.
JE	JAPIO-1	MT4	Consistent in its translation quality. The system seemed to try to translate complicated input sentences depending on context and I would like to applaud this.
	EIWA-1	MT5	There were fragmental translations. To understand, sentences before or after, or common knowledge of technology were needed for many parts.
	NTITI-1	MT6	There were good results and not good results. Impression was inconsistent. If this problem were improved, it would be a good system.

Table 22: PEE comprehensive comments (evaluator 2)

	Run ID	ID for evaluation	Comprehensive comments from evaluator 2
CE	BBN-1	MT1	A little inconsistent. There were some English grammatical problems.
	RWSYS-1	MT2	There were good results and also not good results.
	SRI-1	MT3	The translations were hard to read.
JE	JAPIO-1	MT4	Even if the input Japanese sentences were abstruse, it sometimes could translate. Not only were the English translations good, but so was analyzing input Japanese sentences.
	EIWA-1	MT5	It was similar to MT4. It would be better than MT1.
	NTITI-1	MT6	There were good results and also not good results. It would be slightly better than MT2.

over time. We used the RIBES and BLEU automatic evaluation measures for this evaluation.

6.1 Chinese to English

Figures 12 and 13 are the RIBES and BLEU results for CE translation, respectively. From Figure 12, BBN, RWTH, ISTIC, ONLINE1, and BJTUX improved their RIBES scores from the NTCIR-9 RIBES scores, and from Figure 13, BBN, RWTH, ISTIC, and ONLINE1 improved their BLEU scores from the NTCIR-9 BLEU scores.

6.2 Japanese to English

Figures 14 and 15 are the RIBES and BLEU results for JE translation, respectively. From Figure 14, NTITI, RWTH, and KYOTO improved notably their RIBES scores from the NTCIR-9 RIBES scores, and from Figure 15, RWTH, NTITI, ONLINE1, and KYOTO improved notably their BLEU scores from the NTCIR-9 BLEU scores.

6.3 English to Japanese

Figures 16 and 17 are the RIBES and BLEU results for EJ translation respectively. From Figure 14, NTITI, ONLINE1, BJTUX, and KYOTO improved notably their RIBES scores from the NTCIR-9 RIBES scores, and from Figure 15, NTITI, BJTUX, and ONLINE1 improved notably their BLEU scores from the NTCIR-9 BLEU scores.

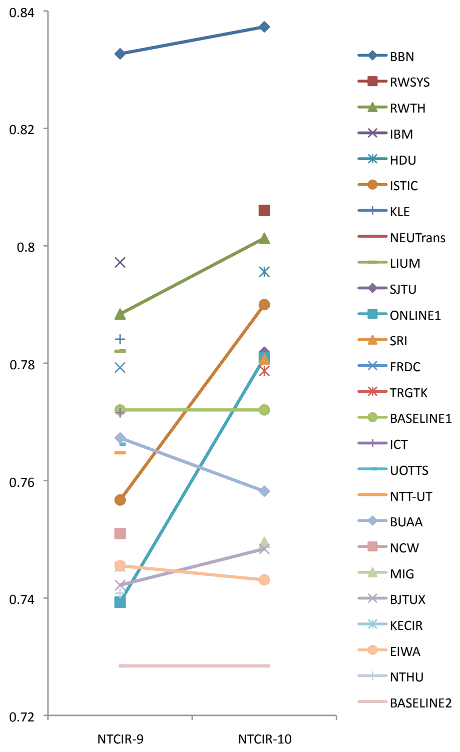

Figure 12: ChE results based on RIBES (CE).

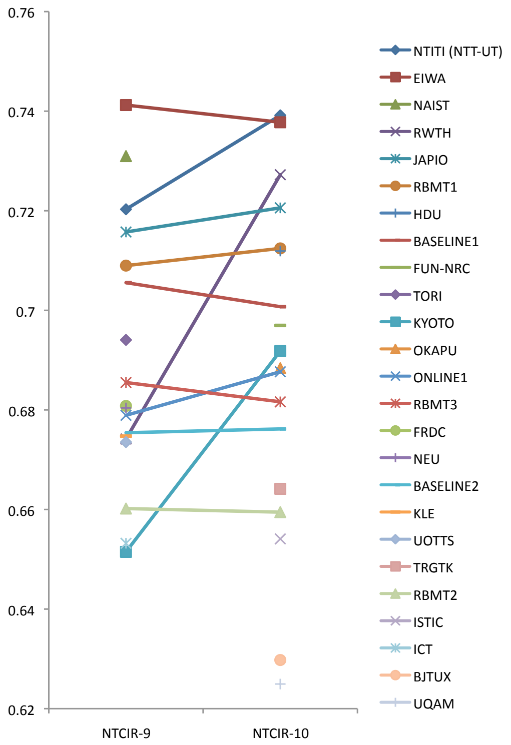

Figure 14: ChE results based on RIBES (JE).

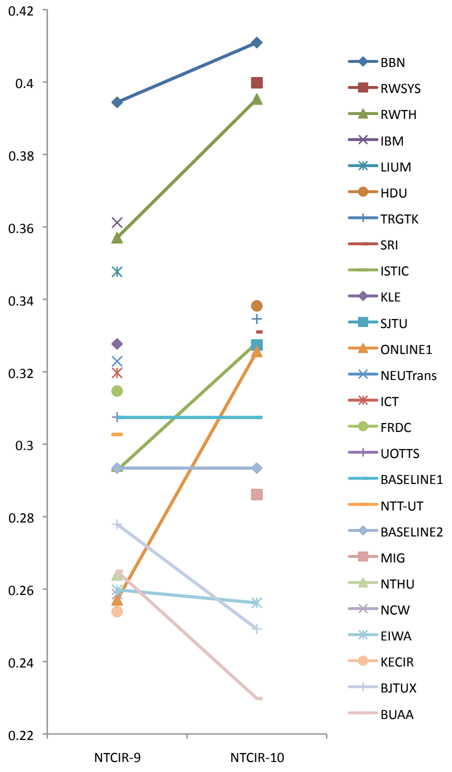

Figure 13: ChE results based on BLEU (CE).

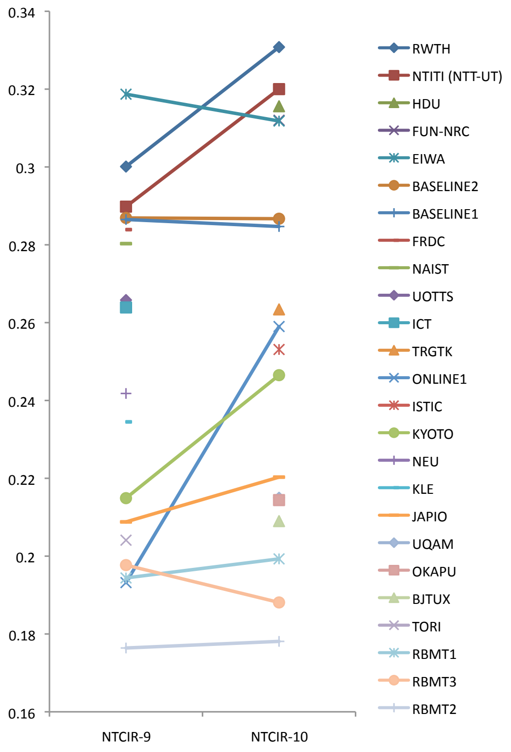

Figure 15: ChE results based on BLEU (JE).

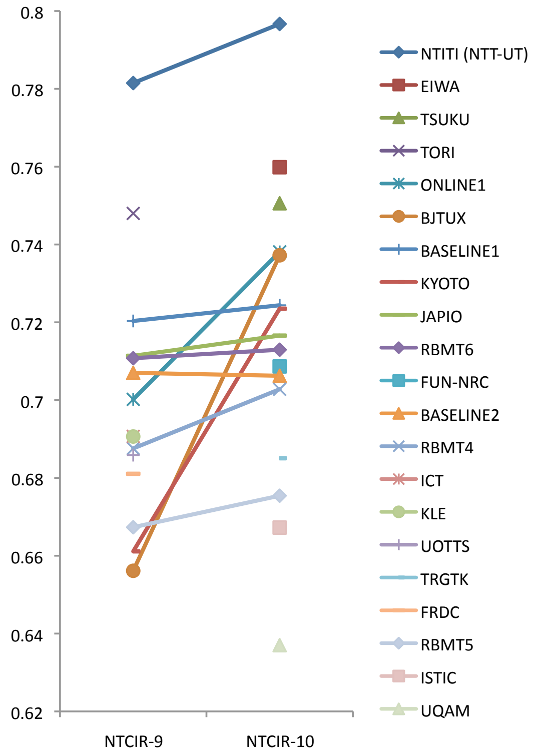

Figure 16: ChE results based on RIBES (EJ).

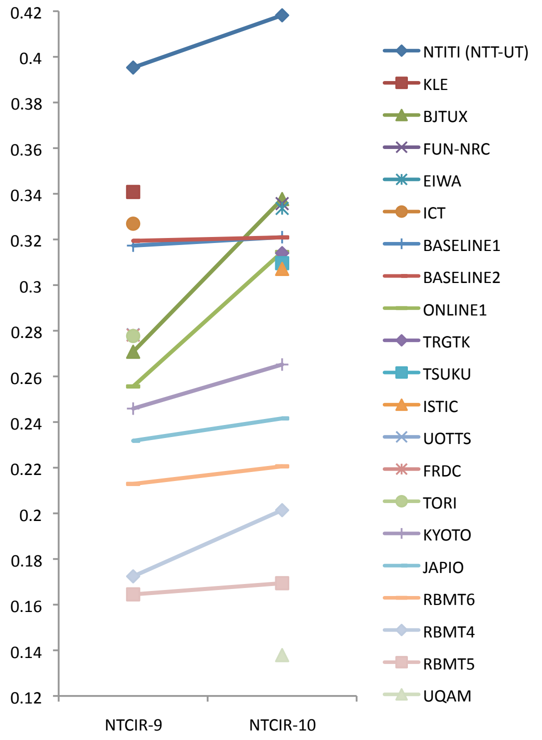

Figure 17: ChE results based on BLEU (EJ).

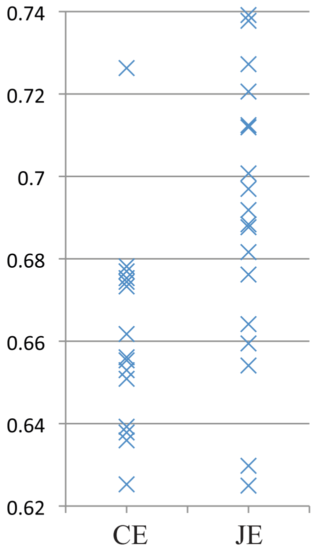

Figure 18: ME results based on RIBES.

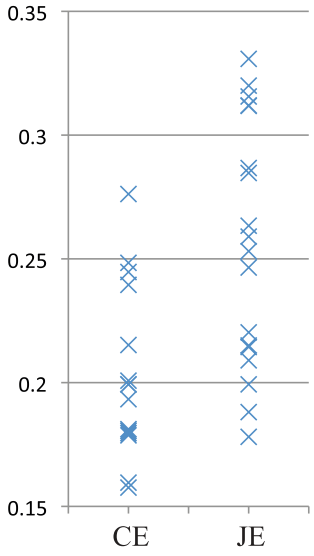

Figure 19: ME results based on BLEU.

7. MULTILINGUAL EVALUATION RESULTS

Multilingual evaluation compared CE and JE translation results using a trilingual test set that shares the same English reference sentences. This evaluation originally aimed to compare the source language dependency of CE and JE translations. However, evaluation results indicated the existence of effects from bias caused by producing the test data by translation.

The IE test data used for the human evaluations was produced by randomly selecting sentences from all of the description sections of the selected patents. Therefore, the test data can be regarded as randomly selected patent sentences with almost no bias. Adequacy and acceptability evaluations for CE and JE were conducted by the same evaluator under the same conditions. Therefore, although the test data was different, the corpus-level JE and CE adequacy and acceptability results can be roughly compared. The top adequacy for CE was 4.15 and the top acceptability rate where the source sentence meanings could be understood (C-rank and above) was 67%. On the other hand, the top adequacy for JE was 3.67 and the top acceptability rate where the source sentence meanings could be understood (C-rank and above) was 55%. From these results, the top CE translation quality was better than the top JE translation quality. Furthermore, the best automatic scores for the CE translation were better than the best automatic scores for the JE translation. Therefore, if conditions were the same and fair for CE and JE translation, the automatic scores for CE translations are expected to be higher than those for JE translations. However, the results of ME were the converse.

Figure 18 shows the RIBES scores and Figure 19 shows the BLEU scores for ME. These figures indicate that the scores for JE translation were higher than the scores for CE translation. These results were not consistent with the above expectation. This would be due to bias caused by how the CE test data was produced. Japanese test sentences were existing real patent sentences, but the Chinese test data was produced by manually translating the Japanese test data into Chinese. Therefore, while the domains of Japanese test sentences would match the Japanese-English training data, if there are differences between the domains of Japanese-English bilingual patents and the domains of Chinese-English bilingual patents, the Chinese test data produced by translating from the Japanese test data would not match the Chinese-English training data well. Another cause of bias would be translation. Although the translating was done by patent translation experts, the translation quality would not be exactly the same as real Chinese patents. The translation was conducted without context, so this translation process was also different from a real patent translation process.

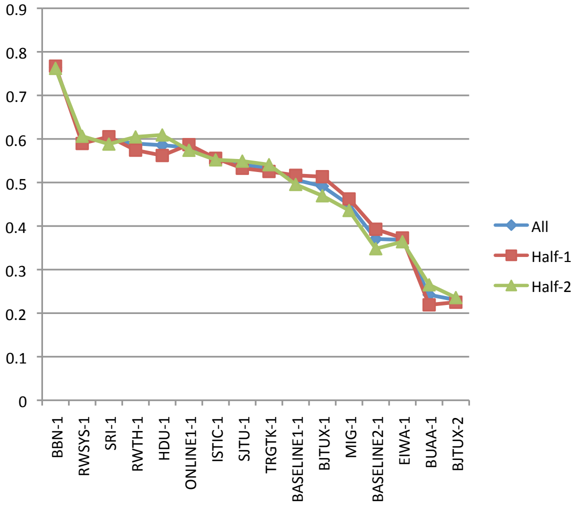

Figure 20: Comparison between data for CE adequacy.

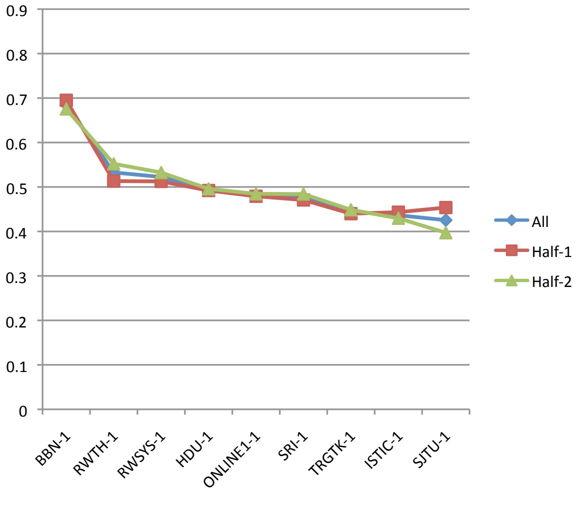

Figure 21: Comparison between data for CE acceptability.

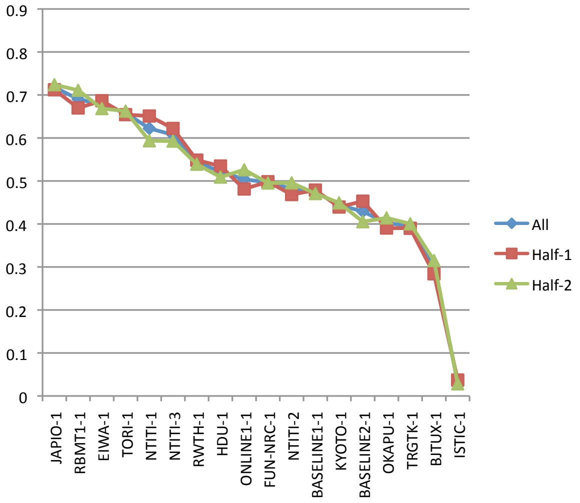

Figure 22: Comparison between data for JE adequacy.

8. VALIDATION OF HUMAN EVALUATION

To discuss reliability of the human evaluation for IE, we present the correlation between the evaluation results for divided data. We validated the reliability of human evaluation as follows:

1. The human evaluation data was divided into the first half data (Half-1) and the second half data (Half-2). Each contains half of all of the sentences evaluated by each evaluator.

2. Scores for the systems based on the halved data were calculated.

3. Correlation of system comparisons between the halved data was calculated.

Since the test data were built by random selection, it is

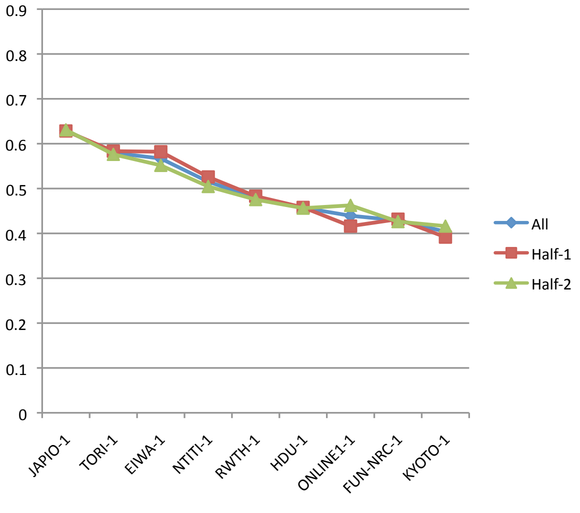

Figure 23: Comparison between data for JE acceptability.

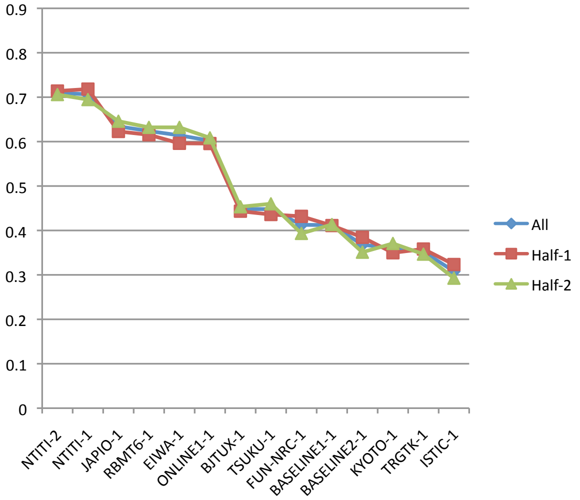

Figure 24: Comparison between data for EJ adequacy.

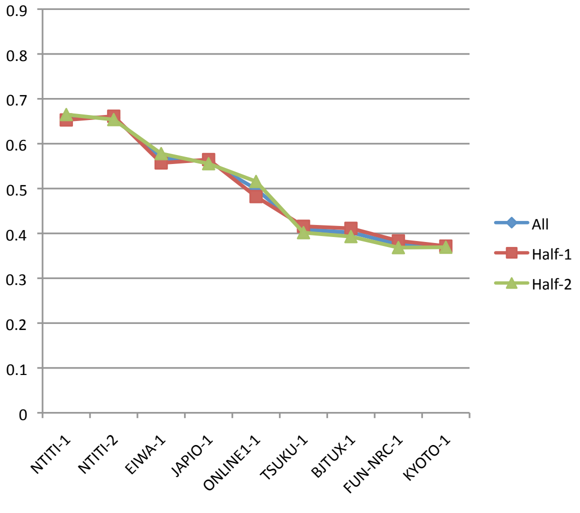

Figure 25: Comparison between data for EJ acceptability.

assumed that the evaluation is not affected by differences in the halved data. Under this assumption, the following is true: If the evaluation is reliable, the top systems based on the first half data will also be the top systems based on the second half data, and the lower-ranking systems based on the first half data will also be the lower-ranking systems based on the second half data, i.e., there is good correlation between system comparison results of the two halved data. On the other hand, if the evaluation is not reliable, the top systems based on the first half data would be the lower-ranking systems based on the second half data, or the lower-ranking systems based on the first half data would be the top systems based on the second half data, i.e., there is poor correlation between system comparison results of the two halved data. Therefore, we validated the reliability based on the correlation between the evaluation results for the divided data. In this section, pairwise scores for systems were used for normalization purposes. A pairwise score for a system reflects the frequency with which it was judged to be better than or equal to other systems. A detailed explanation of the pairwise score is given in Section 2.1.1 *Acceptability*.

Figures 20–25 show the evaluation results for the first half of the data (Half-1), the second half of the data (Half-2), and all of the data (All). In the figures, the vertical axis is the pairwise score, and the horizontal axis is the Run ID. Although there are slight differences between the half data, there are no large differences that reverse the high-ranked and low-ranked systems.

Table 23 shows the Pearson correlation coefficients of the system evaluation scores between the half data. These values indicate that there is a high correlation between the halved data.

These indicate that the evaluations of 150 sentences are thought to be consistent for system comparison, and this consistency shows the reliability of the evaluation results. The evaluation results of 300 sentences are thought to be more reliable than the evaluation results of 150 sentences because the number of sentences is larger.

Table 23: Pearson correlation coefficient between data

	Adequacy	Acceptability
CE	0.98	0.94
JE	0.98	0.97
EJ	0.99	0.99

In addition to the above main validation for reliability, we also checked the differences between evaluators. For each subtask and criterion, three evaluators evaluated the translations of 100 different source sentences. We checked the correlation between the evaluation results based on the 100 source sentences evaluated by the same evaluator. Table 24 shows the Pearson correlation coefficients for the system evaluation scores between evaluators. These values indicate that there is a high correlation between evaluators. Thus, even when the evaluators and the data are different, the evaluations are thought to be consistent for system comparison.

Table 24: Pearson correlation coefficient between evaluators by different data sets

	Evaluator	Adequacy	Acceptability
CE	1 & 2	0.96	0.96
	1 & 3	0.97	0.93
	2 & 3	0.97	0.91
JE	1 & 2	0.98	0.92
	1 & 3	0.99	0.95
	2 & 3	0.98	0.93
EJ	1 & 2	0.97	0.97
	1 & 3	0.96	0.92
	2 & 3	0.94	0.95

9. META-EVALUATION OF AUTOMATIC EVALUATION

We calculated the scoring from three automatic evaluation measures (RIBES, BLEU, and NIST) based on 2,300 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 26 to 28. In these figures, the horizontal axis indicates the average adequacy score and the vertical axis indicates the standardized automatic scores. 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 25.

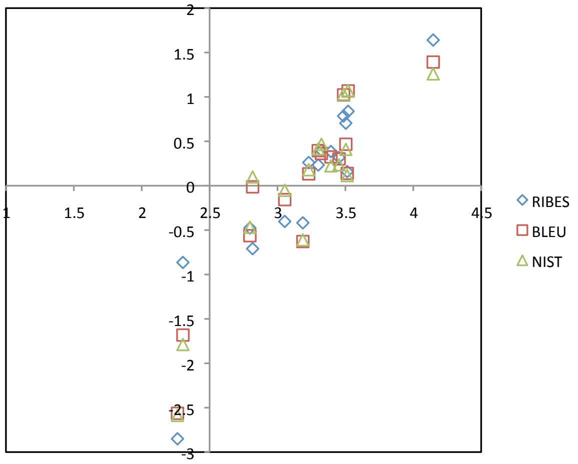

Figure 26: CE correlations between adequacy and automatic evaluation scores.

In Figure 26 and Table 25, it can be seen that the three automatic evaluation measures have a high correlation with the human evaluation for the CE evaluation.

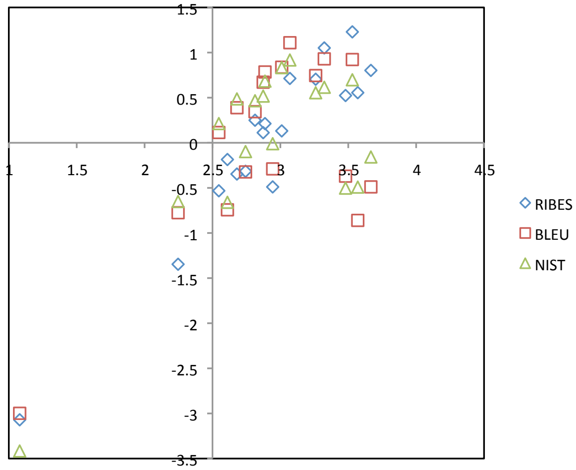

Figure 27: JE correlations between adequacy and automatic evaluation scores.

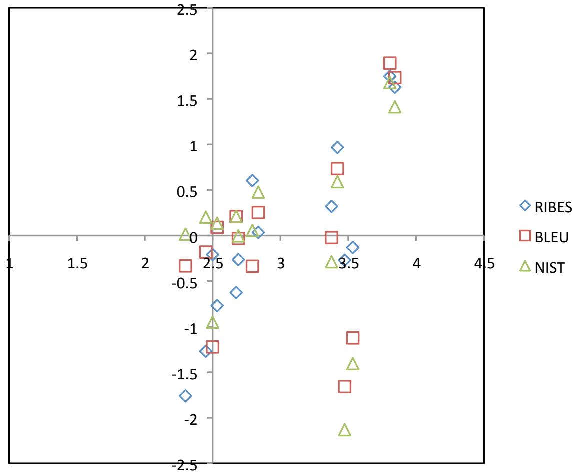

Figure 28: EJ correlations between adequacy and automatic evaluation scores.

Table 25: Correlation coefficients between adequacy and automatic evaluation scores

		Spearman	Pearson
CE	RIBES	0.89	0.91
	BLEU	0.89	0.91
	NIST	0.84	0.89
JE	RIBES	0.88	0.95
	BLEU	0.31	0.63
	NIST	0.36	0.69
EJ	RIBES	0.79	0.81
	BLEU	0.36	0.40
	NIST	0.22	0.15

In Figures 27 and 28 and Table 25, 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 29 and 30. 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 26.

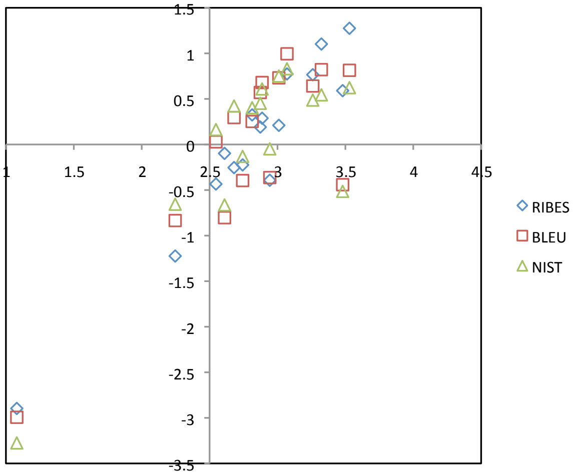

Figure 29: JE correlations between adequacy and automatic evaluation scores excluding RBMT systems.

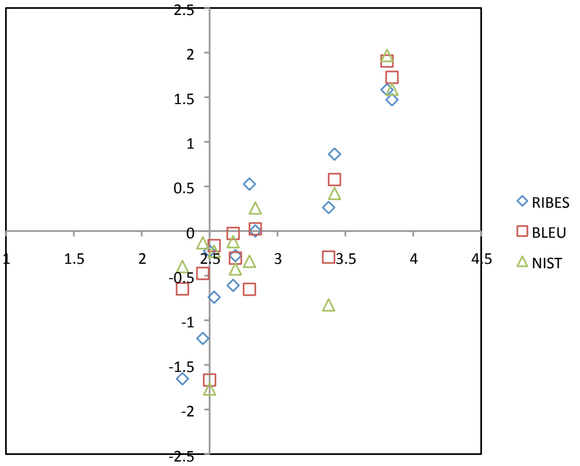

Figure 30: EJ correlations between adequacy and automatic evaluation scores excluding RBMT systems.

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.

These meta-evaluation results were consistent with the meta-evaluation results at NTCIR-9.

Table 26: Correlation coefficients between adequacy and automatic evaluation scores excluding RBMT systems

		Spearman	Pearson
JE	RIBES	0.88	0.96
	BLEU	0.69	0.83
	NIST	0.65	0.82
EJ	RIBES	0.93	0.92
	BLEU	0.76	0.84
	NIST	0.59	0.73

10. CONCLUSION

In order to develop challenging and significant practical research into patent machine translation, we organized a Patent Machine Translation Task at NTCIR-10. 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 to measure the translation quality of sentences. We also conducted patent examination evaluations that evaluated how useful machine translation would be for patent examinations. Various innovative ideas were explored and their effectiveness in patent translation was shown in evaluations.

11. ACKNOWLEDGMENTS

We would like to thank all of the evaluators for the human evaluations. We would like to thank the Nippon Intellectual Property Translation Association (NIPTA) and the contributors for their cooperation for producing the test set for PEE and the evaluation of PEE.

12. 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] W. Chao and Z. Li. ZZX_MT: the BeiHang MT System for NTCIR-10 PatentMT Task. In *Proceedings of NTCIR-10*, 2013.

[3] 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.

[4] T. Ehara. Machine translation system for patent documents combining rule-based translation and statistical postediting applied to the NTCIR-10 PatentMT Task. In *Proceedings of NTCIR-10*, 2013.

[5] M. Feng, M. Freitag, H. Ney, B. Buschbeck, J. Senellart, and J. Yang. The System Combination RWTH Aachen - SYSTRAN for the NTCIR-10 PatentMT Evaluation. In *Proceedings of NTCIR-10*, 2013.

[6] M. Feng, C. Schmidt, J. Wuebker, M. Freitag, and H. Ney. The RWTH Aachen System for NTCIR-10

PatentMT Evaluation. In *Proceedings of NTCIR-10*, 2013.

[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] A. Fujita and M. Carpuat. FUN-NRC: Paraphrase-augmented Phrase-based SMT Systems for NTCIR-10 PatentMT. In *Proceedings of NTCIR-10*, 2013.

[10] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of NTCIR-9*, 2011.

[11] Y. He, C. Shi, and H. Wang. ISTIC Statistical Machine Translation System for PatentMT in NTCIR-10. In *Proceedings of NTCIR-10*, 2013.

[12] 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.

[13] Z. Huang, J. Devlin, and S. Matsoukas. BBN's Systems for the Chinese-English Sub-task of the NTCIR-10 PatentMT Evaluation. In *Proceedings of NTCIR-10*, 2013.

[14] H. Isozaki. OkaPU's Japanese-to-English Translator for NTCIR-10 PatentMT. In *Proceedings of NTCIR-10*, 2013.

[15] 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.

[16] 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.

[17] 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.

[18] J. Murakami, I. Fujiwara, and M. Tokuhisa. Pattern-Based Statistical Machine Translation for NTCIR-10 PatentMT. In *Proceedings of NTCIR-10*, 2013.

[19] T. Nakazawa and S. Kurohashi. Description of KYOTO EBMT System in PatentMT at NTCIR-10. In *Proceedings of NTCIR-10*, 2013.

[20] T. Oshio, T. Mitsuhashi, and T. Kakita. Use of the Japio Technical Field Dictionaries and Commercial Rule-based engine for NTCIR-PatentMT. In *Proceedings of NTCIR-10*, 2013.

[21] 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.

[22] F. Sadat and Z. Fu. UQAM's System Description for the NTCIR-10 Japanese and English PatentMT Evaluation Tasks. In *Proceedings of NTCIR-10*, 2013.

[23] P. Simianer, G. Stupperich, L. Jehl, K. Waeschle, A. Sokolov, and S. Riezler. The HDU Discriminative SMT System for Constrained Data PatentMT at NTCIR10. In *Proceedings of NTCIR-10*, 2013.

[24] K. Sudoh, J. Suzuki, H. Tsukada, M. Nagata, S. Hoshino, and Y. Miyao. NTT-NII Statistical Machine Translation for NTCIR-10 PatentMT. In *Proceedings of NTCIR-10*, 2013.

[25] M. Utiyama and H. Isahara. A Japanese-English patent parallel corpus. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark, 2007.

[26] J.-P. Wang and C.-L. Liu. Using Parallel Corpora to Automatically Generate Training Data for Chinese Segmenters in NTCIR PatentMT Tasks. In *Proceedings of NTCIR-10*, 2013.

[27] P. Wu, J. Xu, Y. Yin, and Y. Zhang. System Description of BJTU-NLP MT for NTCIR-10 PatentMT. In *Proceedings of NTCIR-10*, 2013.

[28] H. Xiong and W. Luo. The TRGTK's System Description of the PatentMT Task at the NTCIR-10 Workshop. In *Proceedings of NTCIR-10*, 2013.

[29] B. Zhao, J. Zheng, W. Wang, and N. Scheffer. SRI Submissions to Chinese-English PatentMT NTCIR10 Evaluation. In *Proceedings of NTCIR-10*, 2013.

[30] H. Zhao, J. Zhang, M. Utiyama, and E. Sumita. An Improved Patent Machine Translation System Using Adaptive Enhancement for NTCIR-10 PatentMT Task. In *Proceedings of NTCIR-10*, 2013.

[31] Z. Zhu, J.-Y. Norimatsu, T. Tanaka, T. Inui, and M. Yamamoto. TSUKU Statistical Machine Translation System for the NTCIR-10 PatentMT Task. In *Proceedings of NTCIR-10*, 2013.

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 ones.
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 indicates 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 1–4 not only by 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 performed from the perspective of whether the machine-translated English sentence conveys the important information and the content of the source sentence and 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 needs to be conveyed by the machine translation results 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; therefore, "F" means that 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 indicates 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 considered to fully express 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 that 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 that lacks a subject (nominative), and if the sentence could be easily understood and is grammatically correct if it were transformed from the active sentence to the passive voice, it can be evaluated as "B," as the sentence is grammatically incorrect.
8. The following type of differences is permissible: The character is the same but the character code is not the same. e.g., "１２３" and "123" are considered to be 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 semicolons. These replacements are permissible. e.g., "５□ｍ" □ "5 .mu.m" or "5 &mu;m"
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 simultaneously 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 to the evaluators in a random order.
- The evaluators can 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 are the same. After the trial evaluation, a consensus meeting is held to make corrections to the differences in the evaluations obtained from all of the evaluators and to decide on common evaluations for the translation results for the trial evaluation.

D. EXAMPLE DATA FOR PEE

Table 27 shows example data for PEE.

Table 27: Example data for PEE

In shinketsu, the description of the facts that were recognized from the reference patent	The recognized facts divided into components.	Sentences in the reference patent yielding the evidence for the recognized facts. (Japanese test data)
これらの記載事項によると、引用例には、「内部において、先端側に良熱伝導金属部 43 が入り込んでいる中心電極 4 と、	内部において、先端側に良熱伝導金属部 43 が入り込んでいる中心電極 4	また、図 3 に示すごとく、中心電極 4 の内部においては、上記露出開始部 431 よりも先端側にも良熱伝導金属部 43 が入り込んでいる。
中心電極 4 の先端部に溶接されている貴金属チップ 45 と、	中心電極 4 の先端部に溶接されている貴金属チップ 45	また、中心電極 4 の先端部には、貴金属チップ 45 が溶接されている。
中心電極 4 を電極先端部 41 が碍子先端部 31 から突出するように挿嵌保持する絶縁碍子 3 と、	中心電極 4 を電極先端部 41 が碍子先端部 31 から突出するように挿嵌保持する絶縁碍子 3	上記中心電極 4 は、電極先端部 41 が碍子先端部 31 から突出するように絶縁碍子 3 に挿嵌保持されている。
絶縁碍子 3 を挿嵌保持する取付金具 2 と、	絶縁碍子 3 を挿嵌保持する取付金具 2	上記絶縁碍子 3 は、碍子先端部 31 が突出するように取付金具 2 に挿嵌保持される。
中心電極 4 の電極先端部 41 との間に火花放電ギャップ G を形成する接地電極 11 とを備えたスパークプラグにおいて、	中心電極 4 の電極先端部 41 との間に火花放電ギャップ G を形成する接地電極 11	上記接地電極 11 は、図 2 に示すごとく、電極先端部 41 との間に火花放電ギャップ G を形成する。
中心電極 4 の直径は、1.2∼2.2mm としたスパークプラグ。」の発明が記載されていると認められる。	中心電極 4 の直径は、1.2∼2.2mm	また、上記碍子固定部 22 の軸方向位置における中心電極 4 の直径は、例えば、1.2∼2.2mm とすることができる。

E. SUBMISSIONS AND AUTOMATIC EVALUATION SCORES

Tables 28–37 show all of the submissions except for PEE and automatic evaluation scores.

Table 28: CE submissions for IE and automatic evaluation scores

SYSTEM-ID (Group ID)	Priority	Type	Resource				RIBES	BLEU	NIST
			B	M	E	C			
BASELINE1	1	SMT	✓				0.7727	0.3252	8.303
BASELINE2	1	SMT	✓				0.7302	0.3134	8.208
BBN	1	SMT	✓	✓			0.8331	0.4268	9.561
BBN	2	SMT	✓				0.8284	0.3998	9.233
BJTUX	1	SMT	✓		✓		0.7429	0.2637	7.382
BJTUX	2	EBMT	✓				0.6363	0.1076	5.077
BUAA	1	SMT	✓	✓			0.7234	0.1787	6.008
BUAA	2	SMT	✓	✓			0.7140	0.1783	5.981
EIWA	1	HYBRID	✓		✓		0.7403	0.2690	7.548
HDU	1	SMT	✓				0.7921	0.3521	8.570
HDU	2	SMT	✓				0.7911	0.3539	8.721
ISTIC	1	SMT	✓	✓			0.7781	0.3406	8.354
MIG	1	SMT	✓				0.7436	0.3018	8.033
MIG	2	SMT	✓				0.7458	0.3017	8.020
MIG	3	SMT	✓				0.7457	0.3012	8.015
MIG	4	SMT	✓				0.7414	0.2866	7.789
ONLINE1	1	SMT			✓		0.7752	0.3388	8.367
RWSYS	1	HYBRID	✓	✓	✓		0.7980	0.4006	9.344
RWSYS	2	HYBRID	✓	✓	✓		0.7987	0.3994	9.339
RWTH	1	SMT	✓	✓			0.7956	0.3970	9.296
RWTH	2	SMT	✓	✓			0.7956	0.3975	9.299
RWTH	3	SMT	✓	✓			0.8000	0.3925	9.230
RWTH	4	SMT	✓				0.7832	0.3622	8.853
SJTU	1	SMT	✓			✓	0.7787	0.3437	8.637
SJTU	2	SMT	✓			✓	0.7661	0.3396	8.614
SRI	1	SMT	✓	✓			0.7682	0.3256	8.226
SRI	1	SMT	✓	✓			0.7651	0.3218	8.129
TRGTK	1	SMT	✓	✓			0.7714	0.3463	8.575
TRGTK	2	SMT	✓	✓			0.7739	0.3432	8.424

Table 29: JE submissions for IE and automatic evaluation scores

SYSTEM-ID (Group ID)	Priority	Type	Resource				RIBES	BLEU	NIST
			B	M	E	C			
BASELINE1	1	SMT	✓				0.6972	0.2856	7.973
BASELINE2	1	SMT	✓				0.6710	0.2886	7.999
BJTUX	1	SMT	✓		✓		0.6271	0.2093	6.548
EIWA	1	HYBRID	✓		✓		0.7402	0.3250	8.270
FUN-NRC	1	SMT	✓	✓			0.6955	0.3156	8.251
FUN-NRC	2	SMT	✓	✓			0.6929	0.3165	8.220
FUN-NRC	3	SMT	✓				0.6911	0.3058	8.111
FUN-NRC	4	SMT	✓				0.6906	0.3065	8.140
HDU	1	SMT	✓				0.6920	0.3192	8.442
HDU	2	SMT	✓				0.6965	0.3207	8.389
ISTIC	1	SMT	✓	✓			0.5514	0.0578	3.021
JAPIO	1	RBMT			✓	✓	0.7214	0.2288	7.178
KYOTO	1	EBMT	✓				0.6724	0.2401	7.254
KYOTO	2	EBMT	✓				0.6738	0.2381	7.228
NTITI	1	SMT	✓	✓			0.7324	0.3255	8.164
NTITI	2	SMT	✓	✓			0.6911	0.3079	8.039
NTITI	3	SMT	✓	✓			0.7171	0.3129	8.086
OKAPU	1	SMT	✓				0.6781	0.2115	6.537
ONLINE1	1	SMT			✓		0.6647	0.2424	7.366
RBMT1	1	RBMT			✓		0.7106	0.2035	6.752
RBMT2	1	RBMT			✓		0.6526	0.1863	6.423
RBMT3	1	RBMT			✓		0.6765	0.1924	6.563
RWTH	1	SMT	✓	✓			0.7175	0.3377	8.550
RWTH	2	SMT	✓				0.7144	0.3308	8.445
TORI	1	HYBRID	✓		✓		0.7092	0.2369	6.738
TORI	2	SMT	✓				0.5941	0.2454	7.050
TORI	3	SMT	✓				0.5955	0.2463	7.066
TORI	4	SMT	✓		✓		0.7099	0.2370	6.740
TRGTK	1	SMT	✓	✓			0.6628	0.2699	7.651
UQAM	1	SMT	✓			✓	0.6269	0.2180	7.072

Table 30: EJ submissions for IE and automatic evaluation scores

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.7231	0.3298	8.084
BASELINE2	1	SMT	✓				0.7042	0.3361	8.182
BJTUX	1	SMT	✓				0.7343	0.3445	8.421
DCUMT	1	SMT	✓	✓	✓	✓	0.6954	0.2786	7.588
EIWA	1	HYBRID	✓		✓		0.7692	0.3693	8.501
FUN-NRC	1	SMT	✓	✓			0.7096	0.3422	8.235
FUN-NRC	2	SMT	✓	✓			0.7089	0.3405	8.212
FUN-NRC	3	SMT	✓				0.7048	0.3289	8.098
FUN-NRC	4	SMT	✓				0.6651	0.2259	7.119
ISTIC	1	SMT	✓	✓			0.6672	0.3143	8.097
JAPIO	1	RBMT			✓	✓	0.7281	0.2736	7.100
KYOTO	1	EBMT	✓				0.7252	0.2685	7.419
KYOTO	2	EBMT	✓				0.7248	0.2662	7.378
NTITI	1	SMT	✓	✓			0.7984	0.4289	9.265
NTITI	2	SMT	✓	✓			0.7939	0.4207	9.078
ONLINE1	1	SMT			✓		0.7450	0.3303	7.885
RBMT4	1	RBMT			✓		0.7111	0.2244	6.295
RBMT5	1	RBMT			✓		0.6846	0.1858	5.763
RBMT6	1	RBMT			✓		0.7229	0.2461	6.591
TRGTK	1	SMT	✓	✓			0.6855	0.3221	8.228
TSUKU	1	SMT	✓				0.7556	0.3141	8.126
TSUKU	2	SMT	✓	✓			0.7566	0.3190	8.189
TSUKU	3	SMT	✓	✓			0.7566	0.3176	8.177
UQAM	1	SMT	✓			✓	0.6369	0.1497	5.668

Table 31: CE submissions for ChE and automatic evaluation scores

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.7720	0.3074	7.906
BASELINE2	1	SMT	✓				0.7284	0.2934	7.752
BBN	1	SMT	✓	✓			0.8373	0.4109	9.074
BJTUX	1	SMT	✓		✓		0.7484	0.2490	7.042
BUAA	1	SMT	✓	✓			0.7582	0.2298	7.193
EIWA	1	HYBRID	✓		✓		0.7431	0.2562	7.171
HDU	1	SMT	✓				0.7956	0.3382	8.225
ISTIC	1	SMT	✓	✓			0.7900	0.3280	8.242
MIG	1	SMT	✓				0.7495	0.2861	7.580
ONLINE1	1	SMT			✓		0.7809	0.3256	8.000
RWSYS	1	HYBRID	✓	✓	✓		0.8060	0.3998	9.044
RWTH	1	SMT	✓	✓			0.8013	0.3953	9.000
SJTU	1	SMT	✓			✓	0.7819	0.3274	8.196
SRI	1	SMT	✓	✓			0.7807	0.3310	8.133
TRGTK	1	SMT	✓	✓			0.7787	0.3346	8.213

Table 32: NTCIR-9 CE submissions and automatic evaluation scores calculated by the NTCIR-10 procedures

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.7720	0.3074	7.906
BASELINE2	1	SMT	✓				0.7284	0.2934	7.752
BBN	1	SMT	✓	✓			0.8327	0.3944	8.911
BJTUX	1	SMT	✓				0.7422	0.2779	7.664
BUAA	1	HYBRID	✓	✓			0.7673	0.2649	7.493
EIWA	1	HYBRID	✓		✓		0.7455	0.2598	7.229
FRDC	1	SMT	✓	✓	✓		0.7793	0.3147	8.126
IBM	1	SMT	✓	✓	✓	✓	0.7972	0.3612	8.509
ICT	1	SMT	✓	✓			0.7716	0.3197	8.203
ISTIC	1	HYBRID	✓	✓			0.7567	0.2928	7.868
KECIR	1	SMT	✓	?	✓	✓	0.7453	0.2538	7.263
KLE	1	SMT	✓				0.7841	0.3277	8.211
KYOTO	1	EBMT	✓				0.6607	0.1798	6.052
LIUM	1	SMT	✓	✓			0.7820	0.3476	8.424
NCW	1	SMT	✓				0.7510	0.2586	7.457
NEUTrans	1	SMT	✓	✓			0.7821	0.3229	8.047
NTHU	1	SMT	✓	?	✓		0.7408	0.2639	7.336
NTT-UT	1	SMT	✓				0.7647	0.3027	8.004
ONLINE1	1	SMT			✓		0.7393	0.2570	7.329
RBMT1	1	RBMT			✓		0.6699	0.1076	4.547
RBMT2	1	RBMT			✓		0.6945	0.1293	5.200
RWTH	1	SMT	✓	✓			0.7884	0.3570	8.629
UOTTS	1	SMT	✓				0.7662	0.3075	7.892

Table 33: JE submissions for ChE and automatic evaluation scores

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.7007	0.2847	7.724
BASELINE2	1	SMT	✓				0.6762	0.2867	7.745
BJTUX	1	SMT	✓		✓		0.6298	0.2090	6.434
EIWA	1	HYBRID	✓		✓		0.7378	0.3118	7.703
FUN-NRC	1	SMT	✓	✓			0.6970	0.3120	8.002
HDU	1	SMT	✓				0.7119	0.3156	8.153
ISTIC	1	SMT	✓	✓			0.6541	0.2531	7.194
JAPIO	1	RBMT			✓	✓	0.7206	0.2203	6.867
KYOTO	1	EBMT	✓				0.6918	0.2465	7.116
NTITI	1	SMT	✓	✓			0.7392	0.3200	7.899
OKAPU	1	SMT	✓				0.6884	0.2144	6.447
ONLINE1	1	SMT			✓		0.6877	0.2590	7.597
RBMT1	1	RBMT			✓		0.7124	0.1993	6.483
RBMT2	1	RBMT			✓		0.6595	0.1781	6.134
RBMT3	1	RBMT			✓		0.6816	0.1881	6.297
RWTH	1	SMT	✓	✓			0.7272	0.3308	8.187
TRGTK	1	SMT	✓	✓			0.6641	0.2634	7.420
UQAM	1	SMT	✓			✓	0.6249	0.2150	6.846

Table 34: NTCIR-9 JE submissions and automatic evaluation scores calculated by the NTCIR-10 procedures

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.7056	0.2865	7.660
BASELINE2	1	SMT	✓				0.6754	0.2869	7.690
EIWA	1	HYBRID	✓		✓		0.7412	0.3187	7.788
FRDC	1	SMT	✓	✓			0.6808	0.2839	7.794
ICT	1	SMT	✓	✓			0.6532	0.2639	7.395
JAPIO	1	RBMT			✓	✓	0.7157	0.2088	6.649
KLE	1	SMT	✓			✓	0.6747	0.2345	6.597
KYOTO	1	EBMT	✓				0.6515	0.2149	6.818
NAIST	1	SMT	✓				0.7310	0.2803	7.408
NEU	1	SMT	✓	✓			0.6804	0.2418	6.978
NTT-UT	1	SMT	✓				0.7203	0.2898	7.861
ONLINE1	1	SMT			✓		0.6789	0.1932	6.845
RBMT1	1	RBMT			✓		0.7090	0.1944	6.365
RBMT2	1	RBMT			✓		0.6602	0.1764	6.043
RBMT3	1	RBMT			✓		0.6855	0.1977	6.418
RWTH	1	SMT	✓				0.6742	0.3001	7.807
TORI	1	HYBRID	✓	✓	✓		0.6941	0.2041	6.172
UOTTS	1	SMT	✓				0.6735	0.2658	7.616

Table 35: EJ submissions for ChE and automatic evaluation scores

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.7244	0.3210	7.853
BASELINE2	1	SMT	✓				0.7063	0.3209	7.898
BJTUX	1	SMT	✓				0.7372	0.3377	8.183
EIWA	1	HYBRID	✓		✓		0.7599	0.3336	7.879
FUN-NRC	1	SMT	✓	✓			0.7087	0.3357	8.053
ISTIC	1	SMT	✓	✓			0.6672	0.3071	7.927
JAPIO	1	RBMT			✓	✓	0.7166	0.2416	6.529
KYOTO	1	EBMT	✓				0.7235	0.2652	7.182
NTITI	1	SMT	✓	✓			0.7967	0.4182	8.963
ONLINE1	1	SMT			✓		0.7381	0.3145	7.526
RBMT4	1	RBMT			✓		0.7028	0.2014	5.892
RBMT5	1	RBMT			✓		0.6754	0.1694	5.438
RBMT6	1	RBMT			✓		0.7130	0.2206	6.180
TRGTK	1	SMT	✓	✓			0.6851	0.3140	8.000
TSUKU	1	SMT	✓	✓			0.7506	0.3096	7.904
UQAM	1	SMT	✓			✓	0.6370	0.1379	5.543

Table 36: NTCIR-9 EJ submissions and automatic evaluation scores calculated by the NTCIR-10 procedures

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.7203	0.3173	7.805
BASELINE2	1	SMT	✓				0.7070	0.3194	7.887
BJTUX	1	SMT	✓				0.6561	0.2708	7.544
FRDC	1	SMT	✓	✓			0.6811	0.2781	7.494
ICT	1	SMT	✓	✓			0.6907	0.3269	8.121
JAPIO	1	RBMT			✓	✓	0.7114	0.2318	6.380
KLE	1	SMT	✓		✓		0.6906	0.3408	8.255
KYOTO	1	EBMT	✓				0.6611	0.2459	6.934
NTT-UT	1	SMT	✓	✓			0.7815	0.3953	8.719
ONLINE1	1	SMT			✓		0.7002	0.2556	6.844
RBMT4	1	RBMT			✓		0.6876	0.1724	5.394
RBMT5	1	RBMT			✓		0.6673	0.1645	5.302
RBMT6	1	RBMT			✓		0.7108	0.2129	6.058
TORI	1	HYBRID	✓	✓	✓		0.7480	0.2777	7.331
UOTTS	1	SMT	✓				0.6859	0.2783	7.243

Table 37: CE submissions for ME and automatic evaluation scores

SYSTEM-ID	Priority	Type	Resource				RIBES	BLEU	NIST
(Group ID)			B	M	E	C			
BASELINE1	1	SMT	✓				0.6509	0.1796	6.096
BASELINE2	1	SMT	✓				0.6253	0.1805	6.219
BBN	1	SMT	✓	✓			0.7263	0.2762	7.294
BJTUX	1	SMT	✓		✓		0.6379	0.1576	5.856
BUAA	1	SMT	✓	✓			0.6561	0.1787	6.147
EIWA	1	HYBRID	✓		✓		0.6359	0.1596	5.868
HDU	1	SMT	✓				0.6744	0.1993	6.282
ISTIC	1	SMT	✓	✓			0.6733	0.1993	6.425
MIG	1	SMT	✓				0.6393	0.1812	6.249
ONLINE1	1	SMT			✓		0.6753	0.2395	6.873
RWSYS	1	SMT	✓	✓	✓		0.6782	0.2484	7.052
RWTH	1	SMT	✓	✓			0.6770	0.2447	6.987
SJTU	1	SMT	✓			✓	0.6554	0.1933	6.387
SRI	1	SYSCOMB	✓	✓			0.6530	0.2008	6.288
TRGTK	1	SMT	✓	✓			0.6617	0.2152	6.616