Overview of the 1st Workshop on Asian Translation

Toshiaki Nakazawa
Japan Science and Technology Agency
nakazawa@pa.jst.jp

Hideya Mino
National Institute of Information and Communications Technology
hideya.mino@nict.go.jp

Isao Goto
NHK
goto.i-es@nhk.or.jp

Sadao Kurohashi
Kyoto University
kuro@i.kyoto-u.ac.jp

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

Abstract

This paper presents the results of the 1st workshop on Asian translation (WMT2014) shared tasks, which included J↔E translation subtasks and J↔C translation subtasks. As the first year of WAT, 12 institutions participated to the shared tasks. More than 300 translation results have been submitted to the automatic evaluation server, and selected submissions were manually evaluated.

1 Introduction

The Workshop on Asian Translation (WAT) is a new open evaluation campaign focusing on Asian languages. We would like to invite a broad range of participants and conduct various forms of machine translation experiments and evaluation. Collecting and sharing our knowledge will allow us to understand the essence of machine translation and the problems to be solved. We are working toward the practical use of machine translation among all Asian countries.

For the 1st WAT, we chose scientific papers as the targeted domain, and selected the languages Japanese, Chinese and English.

What makes WAT unique:

- Open innovation platform
  The test data is fixed and open, so you can repeat evaluations on the same data and confirm changes in translation accuracy over time. WAT has no deadline for the automatic translation quality evaluation (continuous evaluation), so you can submit translation results at any time.

- Domain and language pairs
  WAT is the world's first workshop that uses scientific papers as a domain and Japanese-Chinese as a language pair. In the future, we will add more Asian languages, such as Korean, Vietnamese, Indonesian, Thai, Myanmar and so on.

- Evaluation method
  Evaluation will be done by both automatic and human evaluation. For human evaluation, WAT will use crowdsourcing, which is low cost and allows multiple evaluations.

LangPair	Train	Dev	DevTest	Test
ASPEC-JE	3,008,500	1,790	1,784	1,812
ASPEC-JC	672,315	2,090	2,148	2,107

Table 1: Statistics of ASPEC.

2 Dataset

WAT uses Asian Scientific Paper Excerpt Corpus (ASPEC)[1] as the dataset. ASPEC is constructed by the Japan Science and Technology Agency (JST) in collaboration with the National Institute of Information and Communications Technology (NICT). It consists of a Japanese-English scientific paper abstract corpus (ASPEC-JE), which is used for J↔E subtasks, and a Japanese-Chinese scientific paper excerpt corpus (ASPEC-JC), which is used for J↔C subtasks. The statistics of each corpus are described in Table1.

2.1 ASPEC-JE

The training data of ASPEC-JE was constructed by the NICT from approximately 2 million Japanese-English scientific paper abstracts owned by the JST. Because the paper abstracts are kind

[1] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 1–19,
Tokyo, Japan, 4th October 2014.

of comparable corpora, the sentence correspondences are automatically found using the method of (Utiyama and Isahara, 2007). Each sentence pair is accompanied with the similarity score and the field symbol. The similarity scores are calculated by the method of (Utiyama and Isahara, 2007). The field symbols are single letters A-Z and show the scientific field of each document[2]. The correspondance between the symbols and field names, along with the frequency and occurance ratios for the training data, are given in the README of ASPEC-JE.

The development, development-test and test data were extracted from parallel sentences from Japanese-English paper abstracts owned by JST that are not contained in the training data. Each data set contains 400 documents. Furthermore, the data has been selected to contain the same relative field coverage across each data set. The document alignment was conducted automatically and only documents with a 1-to-1 alignment are included. It is therefore possible to restore the original documents. The format is the same as for the training data except that there is no similarity score.

2.2 ASPEC-JC

ASPEC-JC is a parallel corpus consisting of Japanese scientific papers from the literature database and electronic journal site J-STAGE of JST that have been translated to Chinese after receiving permission from the necessary academic associations. The parts selected were abstracts and paragraph units from the body text, as these contain the highest overall vocabulary coverage.

The development, development-test and test data are extracted at random from documents containing single paragraphs across the entire corpus. Each set contains 400 paragraphs (documents). Therefore there are no documents sharing the same data across the training, development, development-test and test sets.

3 Baseline Systems

Human evaluations were conducted as pairwise comparisons between the translation results for a specific baseline system and translation results for each participant's system. That is, the specific baseline system was the standard of human evaluation. A phrase-based statistical machine translation (SMT) system was adopted as the specific baseline system at WAT 2014.

In addition to the results for the baseline phrase-based SMT system, we produced results for the baseline systems that consisted of a hierarchical phrase-based SMT system, a string-to-tree syntax-based SMT system, a tree-to-string syntax-based SMT system, five commercial rule-based machine translation (RBMT) systems, and two online translation systems. 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 WAT 2014 web page[3]. We used Moses (Koehn et al., 2007; Hoang et al., 2009) as the implementation of the baseline SMT systems. The Berkeley parser (Petrov et al., 2006) was used to obtain syntactic annotations. The baseline systems are shown in Table 2.

The commercial RBMT systems and the online translation systems were operated by the organizers. We note that these RBMT companies and online translation companies did not submit themselves. Since our objective is not to compare commercial RBMT systems or online translation systems from companies that did not themselves participate, the system description of these systems are anonymized in this paper.

We describe the detail of the baseline SMT systems.

3.1 Data for Training

We used the following data for the training of the SMT baseline systems.

- Training data for the language model: All of the target language sentences in the parallel corpus.
- Training data for the translation model: Sentences that were 40 words or less in length. (For Japanese–English training data, we only used train-1.txt, which consisted of 1 million parallel sentence pairs with high similarity scores.)
- Development data for tuning: All of the development data.

3.2 Common Settings for Baseline SMT

We used the following tools for tokenization.

- Juman version 7.0[4] for Japanese segmentation.

[2] http://opac.jst.go.jp/bunrui/index.html

[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/

[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

System ID	System	Type	JE	EJ	JC	CJ
SMT Phrase	Moses' Phrase-based SMT	SMT	✓	✓	✓	✓
SMT Hiero	Moses' Hierarchical Phrase-based SMT	SMT	✓	✓	✓	✓
SMT S2T	Moses' String-to-Tree Syntax-based SMT and Berkeley parser	SMT	✓		✓	
SMT T2S	Moses' Tree-to-String Syntax-based SMT and Berkeley parser	SMT		✓		✓
RBMT X	The Honyaku V15 (Commercial system)	RBMT	✓	✓		
RBMT X	ATLAS V14 (Commercial system)	RBMT	✓	✓		
RBMT X	PAT-Transer 2009 (Commercial system)	RBMT	✓	✓		
RBMT X	J-Beijing 7 (Commercial system)	RBMT			✓	✓
RBMT X	Hohrai 2011 (Commercial system)	RBMT			✓	✓
Online X	Google translate (July, 2014)	(SMT)	✓	✓	✓	✓
Online X	Bing translator (July, 2014)	(SMT)	✓	✓	✓	✓

Table 2: Baseline Systems

- Stanford Word Segmenter version 2014-01-04[5] (Chinese Penn Treebank (CTB) model) for Chinese segmentation.
- The Moses toolkit for English tokenization.

To obtain word alignments, GIZA++ and grow-diag-final-and heuristics were used. We used 5-gram language models with modified Kneser-Ney smoothing, which were built using a tool in the Moses toolkit (Heafield et al., 2013).

3.3 Phrase-based SMT

We used the following Moses' configuration for the phrase-based SMT system.

- distortion-limit = 20
- msd-bidirectional-fe lexicalized reordering
- Phrase score option: GoodTuring

The default values were used for the other system parameters.

3.4 Hierarchical Phrase-based SMT

We used the following Moses' configuration for the hierarchical phrase-based SMT system.

- max-chart-span = 1000
- Phrase score option: GoodTuring

The default values were used for the other system parameters.

3.5 String-to-Tree Syntax-based SMT

We used Berkeley parser to obtain target language syntax. We used the following Moses' configuration for the string-to-tree syntax-based SMT system.

- max-chart-span = 1000
- Phrase score option: GoodTuring
- Phrase extraction options: MaxSpan = 1000, MinHoleSource = 1, and NonTermConsecSource.

The default values were used for the other system parameters.

3.6 Tree-to-String Syntax-based SMT

We used Berkeley parser to obtain source language syntax. We used the following Moses' configuration for the baseline tree-to-string syntax-based SMT system.

- max-chart-span = 1000
- Phrase score option: GoodTuring
- Phrase extraction options: MaxSpan = 1000, MinHoleSource = 1, MinWords = 0, NonTermConsecSource, and AllowOnlyUnalignedWords.

The default values were used for the other system parameters.

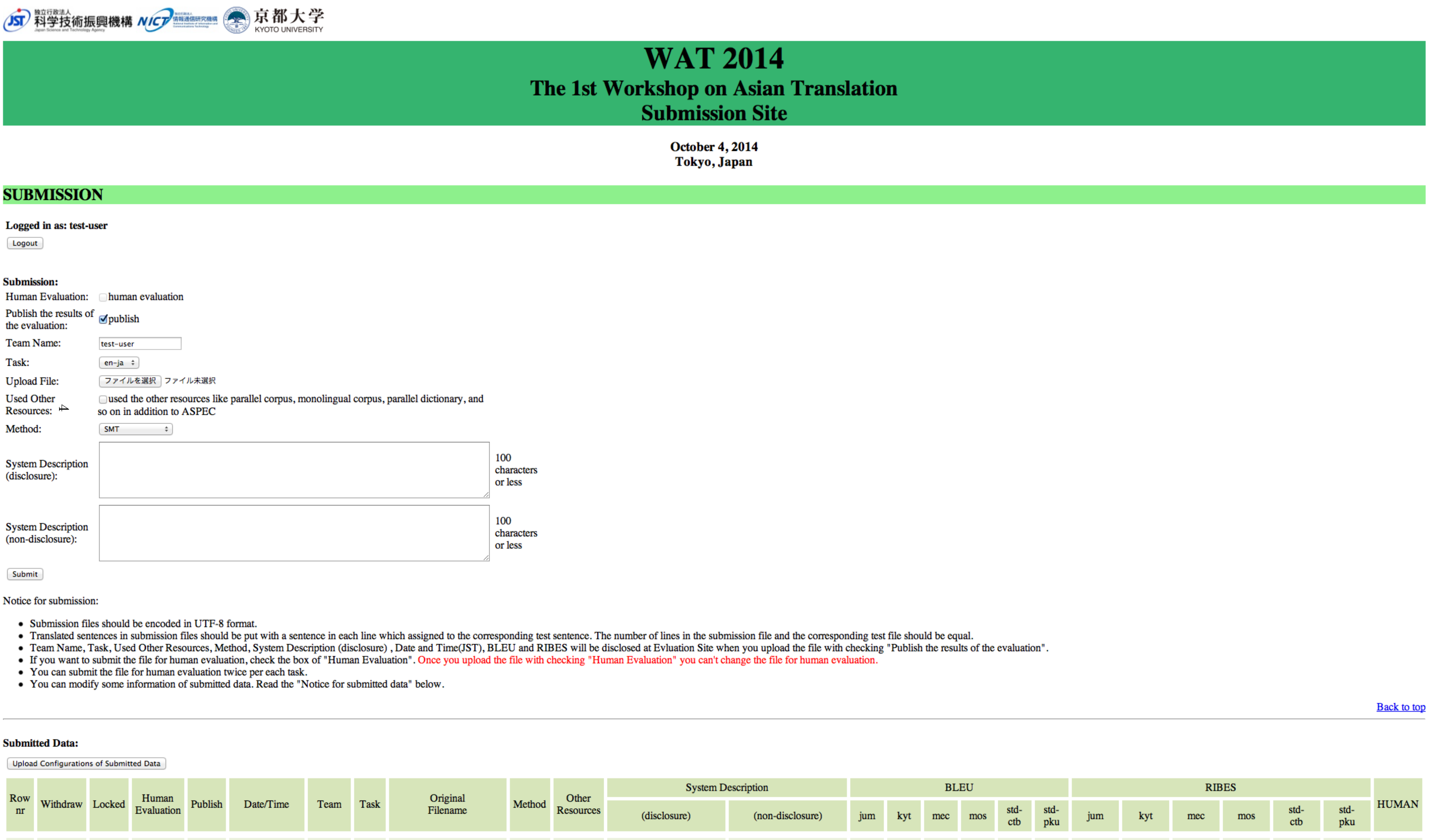
4 Automatic Evaluation

4.1 Procedure of Calculating Automatic Evaluation Score

We calculated automatic evaluation scores of the translation results applying two popular metrics: BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010). BLEU scores were calculated with *multi-bleu.perl* distributed with the Moses toolkit (Koehn et al., 2007). RIBES scores were calculated with *RIBES.py* version 1.02.4 [6]. All scores of each task were calculated using one reference. Before the calculation of the automatic evaluation scores, the translation results have been tokenized with word segmentation tools on each language.

[5]http://nlp.stanford.edu/software/segmenter.shtml

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



WAT 2014
The 1st Workshop on Asian Translation
Submission Site

October 4, 2014
Tokyo, Japan

SUBMISSION

Logged in as: test-user

Logout

Submission:

Human Evaluation: human evaluation

Publish the results of the evaluation: publish

Team Name: test-user

Task: en-ja

Upload File: ファイルを選択 ファイル未選択

Used Other Resources: used the other resources like parallel corpus, monolingual corpus, parallel dictionary, and so on in addition to ASPEC

Method: SMT

System Description (disclosure): 100 characters or less

System Description (non-disclosure): 100 characters or less

Submit

Notice for submission:

- Submission files should be encoded in UTF-8 format.
- Translated sentences in submission files should be put with a sentence in each line which assigned to the corresponding test sentence. The number of lines in the submission file and the corresponding test file should be equal.
- Team Name, Task, Used Other Resources, Method, System Description (disclosure) , Date and Time(JST), BLEU and RIBES will be disclosed at Evluation Site when you upload the file with checking "Publish the results of the evaluation".
- If you want to submit the file for human evaluation, check the box of "Human Evaluation". Once you upload the file with checking "Human Evaluation" you can't change the file for human evaluation.
- You can submit the file for human evaluation twice per each task.
- You can modify some information of submitted data. Read the "Notice for submitted data" below.

Back to top

Submitted Data:

Upload Configurations of Submitted Data

Row nr	Withdraw	Locked	Human Evaluation	Publish	Date/Time	Team	Task	Original Filename	Method	Other Resources	System Description		BLEU						RIBES						HUMAN
											(disclosure)	(non-disclosure)	jum	kyt	mec	mos	std-ctb	std-pku	jum	kyt	mec	mos	std-ctb	std-pku	

Figure 1: The submission web page for participants

For Japanese segmentation we use three different tools, which are Juman version 7.0 (Kurohashi et al., 1994), KyTea 0.4.6 (Neubig et al., 2011) with Full SVM model [7] and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0 [8]. For Chinese segmentation we use two different tools, which are KyTea 0.4.6 with Full SVM Model in MSR model and Stanford Word Segmenter version 2014-06-16 with Chinese Penn Treebank (CTB) and Peking University (PKU) model [9] (Tseng, 2005). For English segmentation we use tokenizer.perl [10] in the Moses toolkit.

The detailed procedures for the automatic evaluation are shown at WAT2014 evaluation web page [11].

4.2 Automatic Evaluation System

The participants submit the translation results via an automatic evaluation system deployed at WAT2014 web page, which give them automatic evaluation scores of the results they upload. Figure 1 shows the submitting interface for participants. The system requires the participants to provide the following information when they upload the translation results:

- Subtask ($J \leftrightarrow E$, $J \leftrightarrow C$)
- Method (SMT, RBMT, SMT and RBMT, EBMT, Other)
- Existence of the use of other resources in addition to ASPEC
- Permission of publishing the automatic evaluation scores on WAT2014 web page

The server of the system keeps all submitted information including translation results or scores and participants can confirm the only information they uploaded. The information of translation results which the participant permits to publish is disclosed on the web page. In addition to submitting the translation results for automatic evaluation, participants submit the results for human evaluation with the same web interface. This automatic evaluation system will be available even after WAT2014. Everybody can use the system by registering on the registration web page [12].

5 Human Evaluation

5.1 Using Crowdsourcing

As all the MT researchers know, the human evaluation costs a lot of time and money. One of the solutions to reduce them is using crowdsourcing. Other machine translation evaluation campaigns such as IWSLT (2011, 2012) and WMT (2012, 2013) also used crowdsourcing for the human evaluation. Recently, there are so many crowdsourcing services in the world: Amazon Mechanical Turk[13], CrowdFlower[14], Yahoo Crowdsourcing[15], Lancers[16] and so on. Among these services, we used Lancers for the human evaluation of WAT2014.

There are two reasons of choosing Lancers. One is that we can set the category of the crowdsourcing task ('Translation' for this case). We can reach the appropriate workers by setting the appropriate categories. The other reason is that we can ask the task to the identity-verified workers. This function guarantee the quality of the workers. These two advantages can keep the evaluation quality at higher level.

5.2 Human Evaluation Method

For the human evaluation, we randomly chose *documents* from the Test set of ASPEC data, in total 400 sentence pairs for JE and JC. We excluded the documents which contains a sentence with longer than 100 Japanese characters. Each submission is compared with the baseline translation (Phrase-based SMT, described in Section 3) and given *HUMAN* score.

5.2.1 Pairwise Evaluation of Sentences

We conducted pairwise evaluation of each test sentence of the 400 sentences. The input sentence and two translations (the baseline and a submission) are shown to the workers, and the workers are asked to judge which translation is better than the other, or they are of the same quality. The order of the two translations are at random. Figure 2 shows the illustration of the evaluation.

[7] http://www.phontron.com/kytea/model.html
[8] http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz
[9] http://nlp.stanford.edu/software/segmenter.shtml
[10] https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl
[11] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html
[12] http://lotus.kuee.kyoto-u.ac.jp/WAT/registration/index.html
[13] https://www.mturk.com
[14] http://www.crowdflower.com
[15] http://crowdsourcing.yahoo.co.jp (Japanese service)
[16] http://www.lancers.jp (Japanese service)

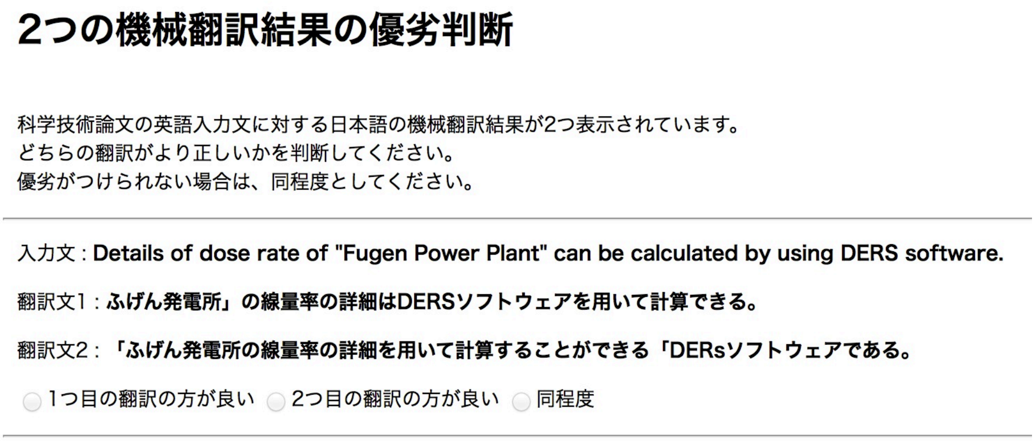


Figure 2: The illustration of the crowdsourcing evaluation. The workers are asked to judge which translation is better, or the same.

Worker 1	A	A	A	A	A	A	Tie	Tie	Tie	B
Worker 2	A	A	A	Tie	Tie	B	Tie	Tie	B	B
Worker 3	A	Tie	B	Tie	B	B	Tie	B	B	B
Decision	A	A	A	A	Tie	B	Tie	B	B	B

Table 3: The combinations of human judgements and the final decision of each sentence pairs from system A and B.

5.2.2 Voting

The crowdsourcing workers are not specialists, thus the quality of the judgements are not necessarily precise. To guarantee the quality of the evaluation, each sentence is evaluated by 3 different workers and the final decision is made by the voting of the judgements. Table 3 shows all the combinations of the worker judgements and the final decision.

5.2.3 HUMAN Score Calculation

Suppose W to be the number of *wins* compared to the baseline, L to be the number of *losses* and T to be the number of *ties*, the HUMAN score, which is the official human evaluation score of WAT2014, can be calculated by the following formula:

$$HUMAN = 100 \times \frac{W - L}{W + L + T}$$

From the definition, the HUMAN score ranges between -100 and 100.

5.2.4 Confidence Interval Estimation

As there are several ways to estimate the confidence interval, we chose the bootstrap resampling (Koehn, 2004) to estimated 95% confidence interval. The procedure is as follows:

1. randomly select 300 sentences from the 400 human evaluation sentences, and calculate the HUMAN score on the selected sentences
2. iterate the previous step 1000 times and get 1000 HUMAN scores
3. sort the 1000 scores and estimate the 95% confidence interval by discarding top and bottom 25 scores

5.3 Cost of Evaluation

One big benefit of using crowdsourcing is that we can reduce the cost of evaluations. In WAT2014, one judgement costs 5 JPY. The evaluation of a submission requires 3 (judgements) × 400 (sentence pairs) = 1,200 judgements and it costs 5 × 1,200 = 6,000 JPY. The time for the evaluation differs depending on the translation direction. On the average, one evaluation finished in a couple of days.

6 Participants List

Table 4 shows the list of participants to WAT2014. There are not only the Japanese organizations, but some organizations came from outside Japan. 12 teams submitted one or more translation results to the automatic evaluation server, and 11 teams submitted one or more translation results to the human evaluation.

7 Evaluation Results

In this section, the evaluation results of WAT2014 are reported from several perspectives. Parts of the results of both automatic and human evaluations are also accessible at WAT2014 website[17].

7.1 Official Automatic Evaluation Results

Figure 3 shows the official automatic evaluation results of the representative submissions and baseline systems. The automatic evaluation results of all the submissions are shown in Section Appendix A.

7.2 Official Human Evaluation Results

HUMAN scores

Figure 4 shows the official human evaluation results. The error bars in the figures show the 95% confidence interval (see Section 5.2.4). Note that overlapping the error bars between two submissions does not necessarily mean that there is no significant difference. If an error bar crosses the x-axis (HUMAN score = 0), it means that there is no significant difference between the submission and the baseline (SMT Phrase).

From the results, the followings can be observed:

- The best SMT system achieved better quality than RBMT system.
- The translation quality of the widely used systems was Phrase-based SMT < Hierarchical PBSMT < Syntax-based SMT (S2T and T2S).
- Forest-to-String Syntax-based SMT system (Neubig, 2014) achieved the best quality for all the translation directions.

Statistical Significance Testing between Submissions

Tables 5, 6, 7 and 8 show the results of statistical significance testings of JE, EJ, JC and CJ translations respectively where all the pairs of submissions are tested. $\ggg$, $\gg$ and $>$ mean that the system in the row is *better* than the system in the column by $p < 0.01$, 0.05, 0.1 respectively. The testings are also done by the bootstrap resampling as follows:

1. randomly select 300 sentences from the 400 human evaluation sentences, and calculate the HUMAN scores on the selected sentences for both systems
2. iterate the previous step 1000 times and count the number of wins (W), losses (L) and ties (T)
3. calculate $p = \frac{L}{W+L}$

Inter-annotator Agreement

To assess the reliability of agreement between the crowdsourcing workers, we calculated the Fleiss' kappa (Fleiss and others, 1971) values. The results are shown in Table 9. We can see that the Kappa values are larger for X $\rightarrow$ J translations than J $\rightarrow$ X translations. This may be because we used Japanese crowdsourcing service for the evaluation and the majority of the crowdsourcing workers are Japanese. The MT evaluation of their mother tongue is much easier than the others in general.

Case Study: Direct Comparison and Relative Comparison

Looking at evaluation results of WEBLIO-EJ1 1 and 2 submissions (see Table 12), the automatic and human evaluations are inconsistent: the WEBLIO-EJ1 2 is consistently better than WEBLIO-EJ1 1 in the automatic evaluation, however it is much worse in the human evaluation. According to the descriptions of the two submissions, the difference of the two is whether it uses the forest input or not. It is natural that using the forest input improves the translation quality, thus we conducted the human evaluation of WEBLIO-EJ1 2 compared to WEBLIO-EJ1 1, which means we used WEBLIO-EJ1 1 as the baseline for the human evaluation.

The HUMAN score was 2.50 $\pm$ 4.17 which means there is no significant difference between the two, and this result is far from the results of the official results. Actually, taking the confidence intervals into consideration, this conclusion can be derived under some probability.

The Fleiss' kappa value was 0.528 and it is much higher than the other E$\rightarrow$J human evaluations. This may be because the outputs of the two systems are quite similar and it is very easy for the

[17] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

	non-removal	removal
JE BLEU	0.46489	0.95098
JE RIBES	0.78255	0.83691
EJ BLEU	0.41524	0.84418
EJ RIBES	0.75105	0.85730
JC BLEU	0.49240	0.07937
JC RIBES	0.38695	0.10198
CJ BLEU	0.78713	0.82592
CJ RIBES	0.70081	0.83209

Table 10: The changes of correlations (R^2) before and after removing RBMT and online systems.

workers to judge which translation is better. If two translations have both better and worse parts than the other, the workers would evaluate differently from person to person.

7.3 Correlation between Automatic and Human Evaluations

Figure 5 shows the correlations between automatic evaluation measures (BLEU/RIBES) and the HUMAN score. It is well known that the automatic and human evaluations do not have good correlations for RBMT and online systems. Removing these systems from the graph changes the correlation values (R^2) like in Table 10. The correlation becomes much better after removing the RBMT and online systems for all the translation directions other than J→C.

8 Submitted Data

The number of published automatic evaluaition results of 12 teams exceeded 100 by the day of WAT2014 workshop and 37 translation results for human evaluation was submitted by 11 teams. We will organize the all submitted data for human evaluation and make it public.

9 Conclusion and Future Perspective

This paper summarized the WAT2014 machine translation evaluation campaing. We had 12 participants worldwide, and collected a large number of submissions which are useful to improve the current machine translation systems by analyzing the submissions and finding the issues.

For the next WAT workshop, we are planning to conduct context-aware MT evaluations. The test data of WAT is prepared using the paragraph as a unit, while almost all other evaluation campaigns use the sentence as a unit. Therefore, it is suitable to investigate the importance of the context for the translation.

Also, we are very happy to include other languages if there are available resources.

Appendix A Submissions

Tables 11, 12, 13 and 14 summarize all the submissions listed in the automatic evaluation server at the point of WAT2014 workshop (4th, October, 2014). The OTHER RESOURCES column shows the use of other resources such as parallel corpora, monolingual corpora and parallel dictionaries in addition to ASPEC.

Team ID	Organization	JE	EJ	JC	CJ
NAIST (Neubig, 2014)	Nara Institute of Science and Technology	✓	✓	✓	✓
EIWA (Ehara, 2014)	Yamanashi Eiwa College	✓			✓
Kyoto-U (Richardson et al., 2014)	Kyoto University	✓	✓	✓	✓
WEBLIO-EJ1 (Zhu, 2014)	Weblio, Inc.		✓		
TMU (Ohwada et al., 2014)	Tokyo Metropolitan University	✓			
BJTUNLP (Cai et al., 2014)	Beijing Jiaotong University			✓	
NII (Hoshino et al., 2014)	National Institute of Informatics	✓			
SAS_MT (Wang et al., 2014)	SAS Research and Development Co., Ltd		✓		✓
Sense (Tan and Bond, 2014)	Saarland University & Nanyang Technological University	✓	✓	✓	✓
NICT (Ding et al., 2014)	National Institute of Information and Communication Technology			✓	
TOSHIBA (Sonoh et al., 2014)	Toshiba Corporation	✓		✓	
WASUIPS (Yang and Lepage, 2014)	Waseda University			✓*	✓*

Table 4: The list of participants which submitted translation results to WAT2014 and their participations to each subtasks. (*Only submitted to automatic evaluations.)

	NAIST 2	Kyoto-U 1	SMT S2T	TOSHIBA 1	RBMT D	EIWA	Kyoto-U 2	TOSHIBA 2	Online D	SMT Hiero	Sense	NII 1	NII 2	TMU 1	TMU 2
NAIST 1	-	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
NAIST 2		⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
Kyoto-U 1			-	-	-	-	-	>	⋙	⋙	⋙	⋙	⋙	⋙	⋙
SMT S2T				-	-	-	>	>	⋙	⋙	⋙	⋙	⋙	⋙	⋙
TOSHIBA 1					-	-	-	-	⋙	⋙	⋙	⋙	⋙	⋙	⋙
RBMT D						-	-	-	⋙	⋙	⋙	⋙	⋙	⋙	⋙
EIWA							-	-	⋙	⋙	⋙	⋙	⋙	⋙	⋙
Kyoto-U 2								-	⋙	⋙	⋙	⋙	⋙	⋙	⋙
TOSHIBA 2									≫	⋙	⋙	⋙	⋙	⋙	⋙
Online D										≫	⋙	⋙	⋙	⋙	⋙
SMT Hiero											≫	⋙	⋙	⋙	⋙
Sense												≫	⋙	⋙	⋙
NII 1													⋙	⋙	⋙
NII 2														-	-
TMU 1															-

Table 5: Statistical significance testing of JE results.

	NAIST 2	WEBLIO-EJ1 1	Online A	Kyoto-U 1	WEBLIO-EJ1 2	SMT T2S	Kyoto-U 2	SMT Hiero	SAS_MT	Sense	RBMT B
NAIST 1	≫	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
NAIST 2		⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
WEBLIO-EJ1 1			-	≫	⋙	⋙	⋙	⋙	⋙	⋙	⋙
Online A				>	≫	⋙	⋙	⋙	⋙	⋙	⋙
Kyoto-U 1					-	>	≫	≫	⋙	⋙	⋙
WEBLIO-EJ1 2						-	-	>	⋙	⋙	⋙
SMT T2S							-	-	≫	⋙	⋙
Kyoto-U 2								-	≫	⋙	⋙
SMT Hiero									>	⋙	⋙
SAS_MT										⋙	⋙
Sense											-

Table 6: Statistical significance testing of EJ results.

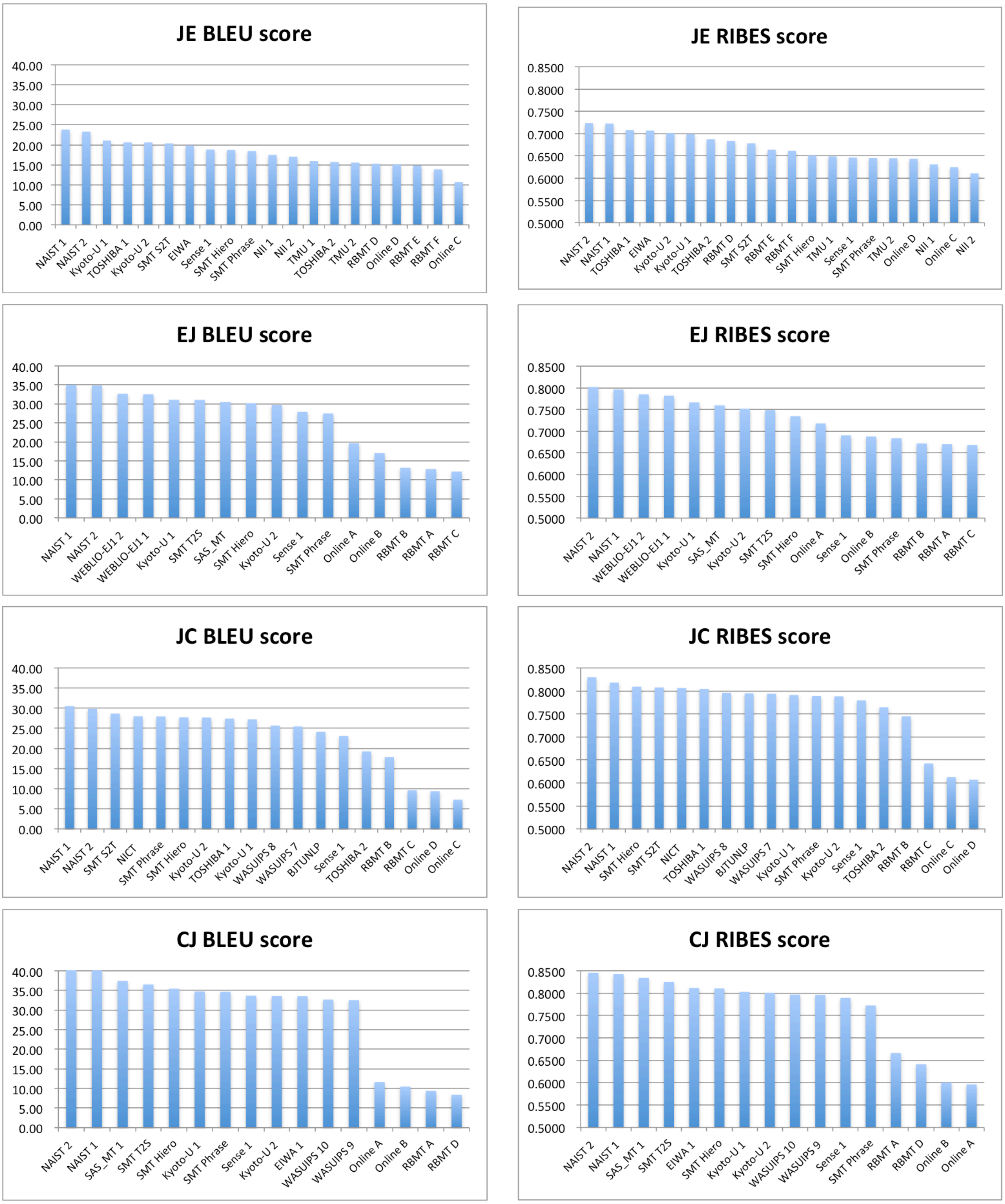


Figure 3: The official automatic evaluation results.

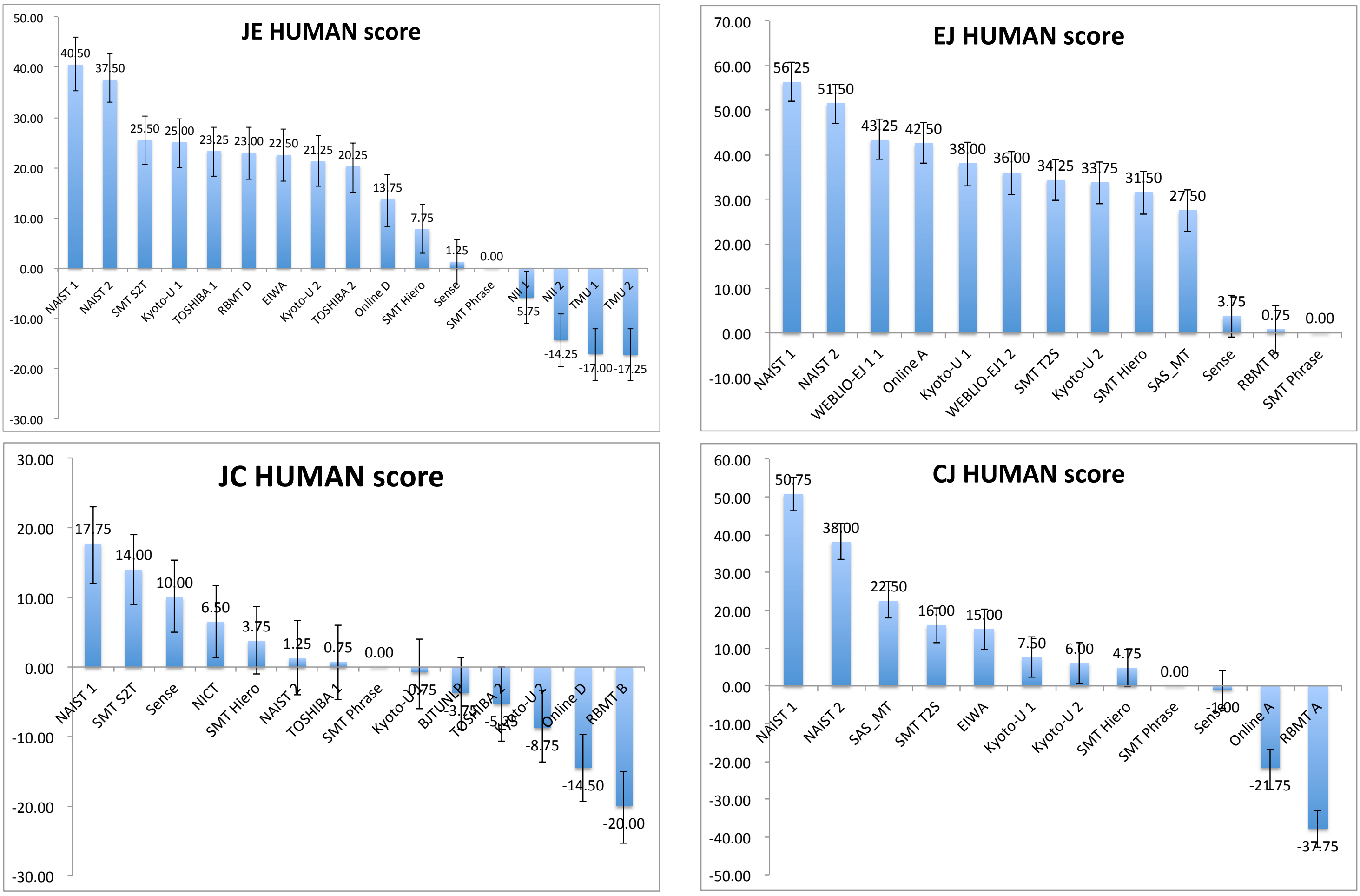


Figure 4: The official human evaluation results.

	SMT S2T	Sense	NICT	SMT Hiero	NAIST 2	TOSHIBA 1	Kyoto-U 1	BJTUNLP	TOSHIBA 2	Kyoto-U 2	Online D	RBMT B
NAIST 1	-	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
SMT S2T		-	≫	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
Sense			-	≫	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
NICT				-	>	≫	≫	⋙	⋙	⋙	⋙	⋙
SMT Hiero					-	-	-	≫	⋙	⋙	⋙	⋙
NAIST 2						-	-	>	≫	⋙	⋙	⋙
TOSHIBA 1							-	>	≫	⋙	⋙	⋙
Kyoto-U 1								-	>	⋙	⋙	⋙
BJTUNLP									-	>	⋙	⋙
TOSHIBA 2										-	⋙	⋙
Kyoto-U 2											≫	⋙
Online D												>

Table 7: Statistical significance testing of JC results.

	NAIST 2	SAS_MT	SMT T2S	EIWA	Kyoto-U 1	Kyoto-U 2	SMT Hiero	Sense	Online A	RBMT A
NAIST 1	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
NAIST 2		⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
SAS_MT			≫	≫	⋙	⋙	⋙	⋙	⋙	⋙
SMT T2S				-	⋙	⋙	⋙	⋙	⋙	⋙
EIWA					⋙	⋙	⋙	⋙	⋙	⋙
Kyoto-U 1						-	-	⋙	⋙	⋙
Kyoto-U 2							-	≫	⋙	⋙
SMT Hiero								≫	⋙	⋙
Sense									⋙	⋙
Online A										⋙

Table 8: Statistical significance testing of CJ results.

JE	
System ID	Kappa
NAIST 1	0.162
NAIST 2	0.047
SMT S2T	0.099
Kyoto-U 1	0.070
TOSHIBA 1	0.098
RBMT D	0.075
EIWA	0.083
Kyoto-U 2	0.139
TOSHIBA 2	0.078
Online D	0.055
SMT Hiero	0.119
Sense	0.245
NII 1	0.119
NII 2	0.086
TMU 1	0.091
TMU 2	0.136
ave.	0.106

EJ	
System ID	Kappa
NAIST 1	0.280
NAIST 2	0.250
WEBLIO-EJ1 1	0.238
Online A	0.219
Kyoto-U 1	0.216
WEBLIO-EJ1 2	0.240
SMT T2S	0.240
Kyoto-U 2	0.229
SMT Hiero	0.277
SAS_MT	0.248
Sense	0.395
RBMT B	0.217
ave.	0.254

JC	
System ID	Kappa
NAIST 1	0.077
SMT S2T	0.069
Sense	0.087
NICT	0.066
SMT Hiero	0.202
NAIST 2	0.093
TOSHIBA 1	0.089
Kyoto-U 1	0.091
BJTUNLP	0.198
TOSHIBA 2	0.066
Kyoto-U 2	0.163
Online D	0.035
RBMT B	0.083
ave.	0.101

CJ	
System ID	Kappa
NAIST 1	0.168
NAIST 2	0.203
SAS_MT	0.167
SMT T2S	0.236
EIWA	0.175
Kyoto-U 1	0.199
Kyoto-U 2	0.180
SMT Hiero	0.274
Sense	0.228
Online A	0.239
RBMT A	0.130
ave.	0.200

Table 9: The Fleiss' kappa values of human evaluation results.

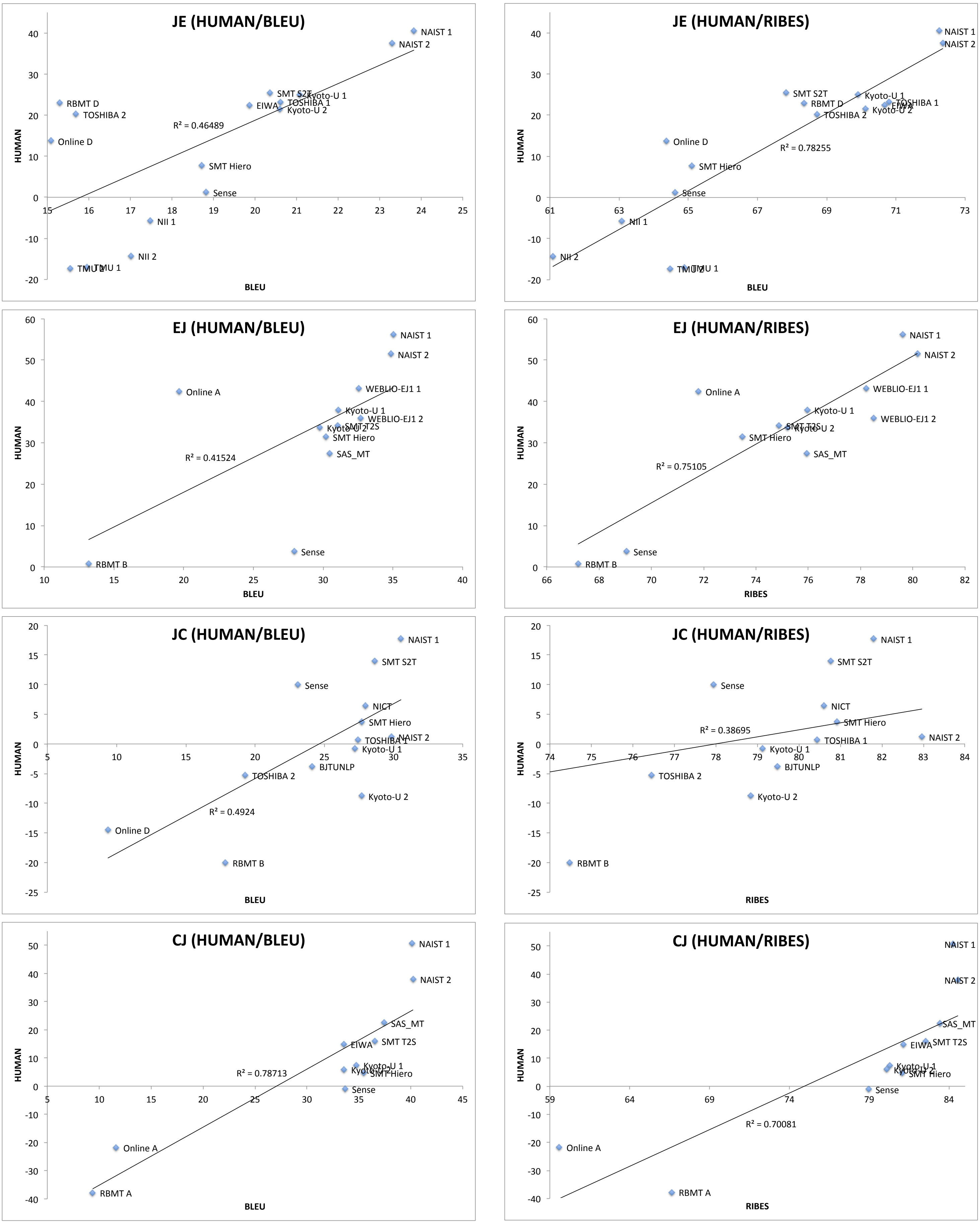


Figure 5: The correlations between BLEU/RIBES and HUMAN scores.

SYSTEM ID	ID	METHOD	OTHER RESOURCES	BLEU	RIBES	HUMAN	SYSTEM DESCRIPTION
SMT Hiero	2	SMT	NO	18.72	0.651066	+7.75	Hierarchical Phrase-based SMT
SMT Phrase	6	SMT	NO	18.45	0.645137	——	Phrase-based SMT
SMT S2T	9	SMT	NO	20.36	0.678253	+25.50	String-to-Tree SMT
Online D	35	Other	YES	15.08	0.643588	+13.75	Online D
RBMT E	76	Other	YES	14.82	0.663851	——	RBMT E
RBMT F	79	Other	YES	13.86	0.661387	——	RBMT F
Online C	87	Other	YES	10.64	0.624827	——	Online C
RBMT D	96	Other	YES	15.29	0.683378	+23.00	RBMT D
NAIST 1	46	SMT	YES	23.82	0.722599	+40.50	Travatar-based Forest-to-String SMT System with Extra Dictionaries
NAIST 2	119	SMT	NO	23.29	0.723541	+37.50	Travatar-based Forest-to-String SMT System
NAIST 3	125	SMT	NO	23.47	0.723670	——	Travatar-based Forest-to-String SMT System (Tuned BLEU+RIBES)
EIWA	116	SMT and RBMT	YES	19.86	0.706686	+22.50	Combination of RBMT and SPE(statistical post editing)
Kyoto-U 3	136	EBMT	NO	20.02	0.689829	——	Our baseline system using 3M parallel sentences
Kyoto-U 2	256	EBMT	NO	20.60	0.701154	+21.25	Our new baseline system after several modifications
Kyoto-U 1	262	EBMT	NO	21.07	0.698953	+25.00	Our new baseline system after several modifications + 20-best parses, KN7, RNNLM reranking
TMU 2	300	SMT	NO	15.55	0.644698	-17.25	Our baseline system with preordering method
TMU 1	301	SMT	NO	15.95	0.648879	-17.20	Our baseline system with another preordering method
TMU 3	307	SMT	NO	15.40	0.613119	——	Our baseline system
NII 1	271	SMT	NO	17.47	0.630825	-5.75	Our Baseline
NII 2	272	SMT	NO	17.01	0.610833	-14.25	Our Baseline with Preordering
Sense 1	164	SMT	NO	18.82	0.646204	+1.25	Paraphrase max10
Sense 2	185	SMT	NO	18.57	0.640393	——	Baseline SMT
Sense 3	191	SMT	NO	18.00	0.641377	——	Context sensitive SMT
Sense 4	205	SMT	NO	18.87	0.646133	——	SMT with lexicon
Sense 5	206	SMT	NO	18.91	0.637375	——	SMT with lexicon X5
TOSHIBA 2	240	RBMT	YES	15.69	0.687122	+20.25	RBMT system
TOSHIBA 1	241	SMT and RBMT	YES	20.61	0.707936	+23.25	RBMT with SPE(Statistical Post Editing) system

Table 11: JE submissions

SYSTEM ID	ID	METHOD	OTHER RESOURCES	BLEU			RIBES			HUMAN	SYSTEM DESCRIPTION
				juman	kytea	mecab	juman	kytea	mecab		
SMT Phrase	5	SMT	NO	27.48	29.80	28.27	0.683735	0.691926	0.695390	——	Phrase-based SMT
SMT T2S	12	SMT	NO	31.05	33.44	32.10	0.748883	0.758031	0.760516	+34.25	Tree-to-String SMT
Online A	34	Other	YES	19.66	21.63	20.17	0.718019	0.723486	0.725848	+42.50	Online A
RBMT B	66	Other	YES	13.18	14.85	13.48	0.671958	0.680748	0.682683	+0.75	RBMT B
RBMT A	68	Other	YES	12.86	14.43	13.16	0.670167	0.676464	0.678934	——	RBMT A
Online B	91	Other	YES	17.04	18.67	17.36	0.687797	0.693390	0.698126	——	Online B
RBMT C	95	Other	YES	12.19	13.32	12.14	0.668372	0.672645	0.676018	——	RBMT C
SMT Hiero	367	SMT	NO	30.19	32.56	30.94	0.734705	0.746978	0.747722	+31.50	Hierarchical Phrase-based SMT
NAIST 1	118	SMT	NO	35.03	37.16	35.81	0.796079	0.801520	0.806581	+56.25	Travatar-based Forest-to-String SMT System
NAIST 2	126	SMT	NO	34.84	37.15	35.67	0.801742	0.807010	0.811081	+51.50	Travatar-based Forest-to-String SMT System (Tuned BLEU+RIBES)
Kyoto-U 3	134	EBMT	NO	28.93	31.61	29.59	0.743969	0.755744	0.756545	——	Our baseline system using 3M parallel sentences
Kyoto-U 4	186	EBMT	NO	30.25	32.78	30.84	0.755629	0.765251	0.766495	——	Using n-best parses and RNNLM
Kyoto-U 2	253	EBMT	NO	29.76	32.46	30.46	0.752058	0.764049	0.766435	+33.75	Our new baseline system after several modifications
Kyoto-U 1	267	EBMT	NO	31.09	33.55	31.73	0.766435	0.770908	0.771545	+38.00	Our new baseline system after several modifications + 20-best parses, KN7, RNNLM reranking
WEBLIO-EJ1 1	132	SMT	NO	32.53	34.87	33.26	0.782066	0.786902	0.792616	+43.25	Weblio Pre-reordering SMT System Baseline
WEBLIO-EJ1 2	202	SMT	NO	32.69	35.04	33.40	0.785015	0.790066	0.795027	+36.00	Weblio Pre-reordering SMT System (with forest inputs)
SAS_MT	264	SMT	NO	30.47	33.00	31.47	0.759415	0.770948	0.771605	+27.50	Syntactic reordering Hierarchical SMT (using part of data)
Sense 2	163	SMT	NO	27.88	30.27	28.72	0.690718	0.699334	0.703139	——	Paraphrase max10
Sense 1	184	SMT	NO	27.92	30.18	28.66	0.690464	0.700583	0.703049	+3.75	Baseline SMT
Sense 3	190	SMT	NO	26.59	28.46	27.15	0.684467	0.694678	0.697257	——	Context sensitive SMT
Sense 4	265	SMT	NO	27.00	29.15	27.81	0.681194	0.689623	0.693560	——	SMT with 20x lexicon
Sense 5	274	SMT	NO	27.33	29.54	28.16	0.679666	0.688801	0.691011	——	SMT with lexicon X5

Table 12: EJ submissions

SYSTEM ID	ID	METHOD	OTHER RESOURCES	BLEU			RIBES			HUMAN	SYSTEM DESCRIPTION
				kytea	stanford (ctb)	stanford (pku)	kytea	stanford (ctb)	stanford (pku)		
SMT Hiero	3	SMT	NO	27.71	27.70	27.35	0.809128	0.809561	0.811394	+3.75	Hierarchical Phrase-based SMT
SMT Phrase	7	SMT	NO	27.96	28.01	27.68	0.788961	0.790263	0.790937	——	Phrase-based SMT
SMT S2T	10	SMT	NO	28.65	28.65	28.35	0.807606	0.809457	0.808417	+14.00	String-to-Tree SMT
Online D	37	Other	YES	9.37	8.93	8.84	0.606905	0.606328	0.604149	-14.50	Online D
Online C	216	Other	YES	7.26	7.01	6.72	0.612808	0.613075	0.611563	——	Online C
RBMT B	243	RBMT	NO	17.86	17.75	17.49	0.744818	0.745885	0.743794	-20.00	RBMT B
RBMT C	244	RBMT	NO	9.62	9.96	9.59	0.642278	0.648758	0.645385	——	RBMT C
NAIST 1	122	SMT	NO	30.53	30.46	30.25	0.818040	0.819406	0.819492	+17.75	Travatar-based Forest-to-String SMT System
NAIST 2	123	SMT	NO	29.83	29.77	29.54	0.829627	0.830839	0.830529	+1.25	Travatar-based Forest-to-String SMT System (Tuned BLEU+RIBES)
Kyoto-U 3	18	EBMT	NO	26.69	26.48	26.30	0.796402	0.798084	0.798383	——	Our baseline system
Kyoto-U 1	257	EBMT	NO	27.21	27.02	26.83	0.791270	0.792166	0.790743	-0.75	Our new baseline system after several modifications
Kyoto-U 2	259	EBMT	NO	27.67	27.44	27.34	0.788321	0.789069	0.788206	-8.75	Our new baseline system after several modifications + 20-best parses, KN7, RNNLM reranking
BJTUNLP	224	SMT	NO	24.12	23.76	23.55	0.794834	0.796186	0.793054	-3.75	SMT
Sense 2	175	SMT	NO	27.92	28.03	27.67	0.793876	0.796589	0.797332	——	SMT
Sense 1	201	SMT	NO	23.09	22.94	23.04	0.779495	0.779502	0.780262	+10.00	Character based SMT
NICT	260	SMT	NO	27.98	28.18	27.84	0.806070	0.808684	0.807809	+6.50	Pre-reordering for phrase-based SMT (dependency parsing + manual rules)
TOSHIBA 2	236	RBMT	YES	19.28	18.93	18.82	0.764491	0.765346	0.763931	-5.25	RBMT system
TOSHIBA 1	238	SMT and RBMT	YES	27.42	26.82	26.79	0.804444	0.803302	0.803980	+0.75	RBMT with SPE(Statistical Post Editing) system
WASUIPS 1	371	SMT	NO	22.71	22.49	22.39	0.776323	0.777615	0.777327	——	Our baseline system (segmentation tools: urheen and mecab, moses: 1.0).
WASUIPS 2	373	SMT	YES	24.70	24.25	24.28	0.790030	0.790460	0.790898	——	Our baseline system + additional quasi-parallel corpus (segmentation tools: urheen and mecab, moses: 1.0).
WASUIPS 3	376	SMT	NO	25.44	25.04	24.98	0.794244	0.793945	0.794823	——	Our baseline system (segmentation tools: urheen and mecab, moses: 2.1.1).
WASUIPS 4	377	SMT	YES	25.60	25.10	25.07	0.794716	0.795786	0.795594	——	Our baseline system + additional quasi-parallel corpus segmentation tools: urheen and mecab, moses: 2.1.1).
WASUIPS 5	381	SMT	NO	22.01	21.81	21.61	0.767418	0.767414	0.766092	——	Our baseline system (segmentation tools: kytea, moses: 1.0).
WASUIPS 6	382	SMT	YES	22.20	22.02	21.91	0.771952	0.773341	0.772107	——	Our baseline system + additional quasi-parallel corpus (segmentation tools: kytea, moses: 1.0).
WASUIPS 7	385	SMT	NO	25.45	25.10	25.01	0.793819	0.793308	0.793029	——	Our baseline system (segmentation tools: kytea, moses: 2.1.1).
WASUIPS 8	386	SMT	YES	25.68	25.01	25.11	0.795721	0.795504	0.795129	——	Our baseline system + additional quasi-parallel corpus (segmentation tools: kytea, moses: 2.1.1).
WASUIPS 9	389	SMT	NO	25.08	24.81	24.64	0.790498	0.791430	0.790142	——	Our baseline system (segmentation tools: stanford-ctb and juman, moses: 2.1.1).
WASUIPS 10	390	SMT	YES	25.63	25.30	25.18	0.794646	0.795307	0.794024	——	Our baseline system + additional quasi-parallel corpus (segmentation tools: stanford-ctb and juman, moses: 2.1.1).

Table 13: JC submissions

SYSTEM ID	ID	METHOD	OTHER RESOURCES	BLEU			RIBES			HUMAN	SYSTEM DESCRIPTION
				juman	kytea	mecab	juman	kytea	mecab		
SMT Hiero	4	SMT	NO	35.43	35.91	35.64	0.810406	0.798726	0.807665	+4.75	Hierarchical Phrase-based SMT
SMT Phrase	8	SMT	NO	34.65	35.16	34.77	0.772498	0.766384	0.771005	—	Phrase-based SMT
SMT T2S	13	SMT	NO	36.52	37.07	36.64	0.825292	0.820490	0.825025	+16.00	Tree-to-String SMT
Online A	36	Other	YES	11.63	13.21	11.87	0.595925	0.598172	0.598573	-21.75	Online A
Online B	215	Other	YES	10.48	11.26	10.47	0.600733	0.596006	0.600706	—	Online B
RBMT A	239	RBMT	NO	9.37	9.87	9.35	0.666277	0.652402	0.661730	-37.75	RBMT A
RBMT D	242	RBMT	NO	8.39	8.70	8.30	0.641189	0.626400	0.633319	—	RBMT D
NAIST 1	120	SMT	NO	40.11	41.29	40.30	0.842477	0.834824	0.842235	+50.75	Travatar-based Forest-to-String SMT System
NAIST 2	124	SMT	NO	40.21	40.82	40.15	0.845486	0.838092	0.845625	+38.00	Travatar-based Forest-to-String SMT System (Tuned BLEU+RIBES)
EIWA 2	137	RBMT	YES	18.69	18.33	18.32	0.740183	0.720281	0.732466	—	RBMT plus user dictionary
EIWA 1	138	SMT and RBMT	YES	33.53	33.74	33.87	0.811350	0.800506	0.808504	+15.00	RBMT with user dictionary plus SPE(statistical post editing)
Kyoto-U 3	133	EBMT	NO	33.26	35.09	33.62	0.791680	0.787105	0.791269	—	Using n-best parses and RNNLM
Kyoto-U 4	135	EBMT	NO	32.68	33.30	32.45	0.786229	0.783016	0.786352	—	Our baseline system
Kyoto-U 2	258	EBMT	NO	33.57	34.43	33.45	0.800949	0.795390	0.800986	+6.00	Our new baseline system after several modifications
Kyoto-U 1	268	EBMT	NO	34.75	35.89	34.83	0.802629	0.798631	0.802930	+7.50	Our new baseline system after several modifications + 20-best parses, KN7, RNNLM reranking
SAS_MT 2	232	SMT	NO	36.58	36.22	36.10	0.822180	0.807535	0.817368	—	Syntactic reordering phrase-based SMT (SAS token tool)
SAS_MT 1	263	SMT	NO	37.42	37.65	37.07	0.834170	0.825551	0.833048	+22.50	Syntactic reordering Hierarchical SMT (using SAS token tool)
Sense 2	174	SMT	NO	34.56	35.08	34.64	0.771975	0.766470	0.771081	—	SMT
Sense 1	200	SMT	NO	33.66	33.86	33.46	0.789495	0.774338	0.784012	-1.00	Character based SMT
WASUIPS 1	369	SMT	NO	27.66	28.09	28.20	0.779183	0.762949	0.770846	—	Our baseline system (segmentation tools: urheen and mecab, moses: 1.0).
WASUIPS 2	370	SMT	YES	30.44	30.92	30.86	0.789824	0.773142	0.781475	—	Our baseline system + additional quasi-parallel corpus (segmentation tools: urheen and mecab, moses: 1.0).
WASUIPS 3	374	SMT	NO	31.87	32.26	32.26	0.794303	0.777876	0.786422	—	Our baseline system (segmentation tools: urheen and mecab, moses: 2.1.1).
WASUIPS 4	375	SMT	YES	32.19	32.55	32.54	0.795838	0.780027	0.787591	—	Our baseline system + additional quasi-parallel corpus (segmentation tools: urheen and mecab, moses: 2.1.1).
WASUIPS 5	379	SMT	NO	27.37	28.28	27.43	0.774423	0.753749	0.767073	—	Our baseline system (segmentation tools: kytea, moses: 1.0).
WASUIPS 6	380	SMT	YES	27.86	28.89	28.00	0.776550	0.756721	0.769409	—	Our baseline system + additional quasi-parallel corpus (segmentation tools: kytea, moses: 1.0).
WASUIPS 7	383	SMT	NO	32.08	33.09	32.18	0.793230	0.775168	0.787665	—	Our baseline system (segmentation tools: kytea, moses: 2.1.1).
WASUIPS 8	384	SMT	YES	32.43	33.36	32.48	0.796220	0.778075	0.789657	—	Our baseline system + additional quasi-parallel corpus (segmentation tools: kytea, moses: 2.1.1).
WASUIPS 9	387	SMT	NO	32.52	32.69	32.47	0.796059	0.780402	0.790107	—	Our baseline system (segmentation tools: stanford-ctb and juman, moses: 2.1.1).
WASUIPS 10	388	SMT	YES	32.65	32.81	32.59	0.796777	0.781733	0.791219	—	Our baseline system + additional quasi-parallel corpus (segmentation tools: stanford-ctb and juman, moses: 2.1.1).

Table 14: CJ submissions

References

Jingsheng Cai, Yujie Zhang, Hua Shan, and Jinan Xu. 2014. System Description: Dependency-based Preordering for Japanese-Chinese Machine Translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Chenchen Ding, Masao Utiyama, Eiichiro Sumita, and Mikio Yamamoto. 2014. Word Order Does NOT Differ Significantly Between Chinese and Japanese. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Terumasa Ehara. 2014. A machine translation system combining rule-based machine translation and statistical post-editing. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. 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.

Sho Hoshino, Hubert Soyer, Yusuke Miyao, and Akiko Aizawa. 2014. Japanese to English Machine Translation using Preordering and Compositional Distributed Semantics. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 944–952, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.*

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

T. Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 529–533, Stroudsburg, PA, USA. Association for Computational Linguistics.

Graham Neubig. 2014. Forest-to-String SMT for Asian Language Translation: NAIST at WAT 2014. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Kenichi Ohwada, Ryosuke Miyazaki, and Mamoru Komachi. 2014. Predicate-Argument Structure-based Preordering for Japanese-English Statistical Machine Translation of Scientific Papers. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

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

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July. Association for Computational Linguistics.

John Richardson, Fabien Cromières, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. KyotoEBMT System Description for the 1st Workshop on Asian Translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Satoshi Sonoh, Satoshi Kinoshita, Hiroyuki Tanaka, and Satoshi Kamatani. 2014. Toshiba MT System Description for the WAT2014 Workshop. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Liling Tan and Francis Bond. 2014. Manipulating Input Data in Machine Translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Huihsin Tseng. 2005. A conditional random field word segmenter. In *In Fourth SIGHAN Workshop on Chinese Language Processing*.

Masao Utiyama and Hitoshi Isahara. 2007. A japanese-english patent parallel corpus. In *MT summit XI*, pages 475–482.

Rui Wang, Xu Yang, and Yan Gao. 2014. The SAS Statistical Machine Translation System for WAT 2014. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Wei Yang and Yves Lepage. 2014. Consistent Improvement in Translation Quality of Chinese–Japanese Technical Texts by Adding Additional Quasi-parallel Training Data. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.

Zhongyuan Zhu. 2014. Weblio Pre-reordering Statistical Machine Translation System. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*.