Overview of the 4th Workshop on Asian Translation

Toshiaki Nakazawa
Japan Science and Technology Agency
nakazawa@nlp.ist.i.kyoto-u.ac.jp

Shohei Higashiyama and **Chenchen Ding**
National Institute of Information and Communications Technology
{shohei.higashiyama, chenchen.ding}@nict.go.jp

Hideya Mino and **Isao Goto**
NHK
{mino.h-gq, goto.i-es}@nhk.or.jp

Hideto Kazawa
Google
kazawa@google.com

Yusuke Oda
Nara Institute of Science and Technology
oda.yusuke.on9@is.naist.jp

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

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

Abstract

This paper presents the results of the shared tasks from the 4th workshop on Asian translation (WAT2017) including J↔E, J↔C scientific paper translation subtasks, C↔J, K↔J, E↔J patent translation subtasks, H↔E mixed domain subtasks, J↔E newswire subtasks and J↔E recipe subtasks. For the WAT2017, 12 institutions participated in the shared tasks. About 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. Following the success of the previous workshops WAT2014 (Nakazawa et al., 2014), WAT2015 (Nakazawa et al., 2015) and WAT2016 (Nakazawa et al., 2016), WAT2017 brings together machine translation researchers and users to try, evaluate, share and discuss brand-new ideas of machine translation. We have been working toward practical use of machine translation among all Asian countries.

For the 4th WAT, we adopted new translation subtasks with English-Japanese news corpus and English-Japanese recipe corpus in addition to the subtasks at WAT2016 [1]. Furthermore, we invited research papers on topics related to machine translation, especially for Asian languages. The submitted research papers were peer reviewed by three program committee members and the committee accepted 4 papers, which focus on on neural machine translation, and construction and evaluation of language resources. We also launched the small NMT task, which aims to build a small NMT system that keeps a reasonable translation quality. There are, however, no submissions to the task this year.

WAT is the uniq workshop on Asian language transration with the following characteristics:

- Open innovation platform
  Due to the fixed and open test data, we can repeatedly evaluate translation systems on the same dataset over years. There is no deadline of translation result submission with respect to automatic evaluation of translation quality and WAT receives submissions at any time.

- Domain and language pairs
  WAT is the world's first workshop that targets scientific paper domain, and Chinese↔Japanese and Korean↔Japanese language pairs. In the future, we will add more Asian languages such as Vietnamese, Thai, Burmese and so on.

- Evaluation method
  Evaluation is done both automatically

[1] This year we did not conduct Indonesian-English newswire subtask, which is conducted in WAT2016, due to corpus license reasons.

Proceedings of the 4th Workshop on Asian Translation, pages 1–54,
Taipei, Taiwan, November 27, 2017.

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

Table 1: Statistics for ASPEC.

and manually. For automatic evaluation, we use three metrics: BLEU, RIBES and AMFM. As human evaluation, we evaluate translation results by pairwise evaluation and JPO adequacy evaluation. JPO adequacy evaluation is conducted for the selected submissions according to the pairwise evaluation results.

2 Dataset

WAT2017 uses the Asian Scientific Paper Excerpt Corpus (ASPEC) [2], JPO Patent Corpus (JPC) [3], JIJI Corpus [4], IIT Bombay English-Hindi Corpus (IITB Corpus) [5] and Recipe Corpus [6] as the dataset.

2.1 ASPEC

ASPEC was constructed by the Japan Science and Technology Agency (JST) in collaboration with the National Institute of Information and Communications Technology (NICT). The corpus 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 for each corpus are shown in Table 1.

2.1.1 ASPEC-JE

The training data for ASPEC-JE was constructed by NICT from approximately two million Japanese-English scientific paper abstracts owned by JST. The data is a comparable corpus and sentence correspondences are found automatically using the method from (Utiyama and Isahara, 2007). Each sentence pair is accompanied by a similarity score that are calculated by the method and a field ID that indicates a scientific field. The correspondence between field IDs and field names, along with the frequency and occurrence ratios for the training data, are descripted in the README file of ASPEC-JE.

The development, development-test and test data were extracted from parallel sentences from the Japanese-English paper abstracts that exclude the sentences in the training data. Each dataset consists of 400 documents and contains sentences in each field at the same rate. 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 the training data except that there is no similarity score.

2.1.2 ASPEC-JC

ASPEC-JC is a parallel corpus consisting of Japanese scientific papers, which come from the literature database and electronic journal site J-STAGE by JST, and their translation to Chinese with permission from the necessary academic associations. Abstracts and paragraph units are selected from the body text so as to 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). There are no documents sharing the same data across the training, development, development-test and test sets.

2.2 JPC

JPC was constructed by the Japan Patent Office (JPO). The corpus consists of Chinese-Japanese patent description corpus (JPC-CJ), Korean-Japanese patent description corpus (JPC-KJ) and English-Japanese patent description corpus (JPC-EJ) with the sections of Chemistry, Electricity, Mechanical engineering, and Physics on the basis of International Patent Classification (IPC). Each corpus is partitioned into training, development, development-test and test data. This corpus is used for patent subtasks C↔J, K↔J and E↔J. The statistics for each corpus are shown

[2] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/index.html
[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html
[4] http://lotus.kuee.kyoto-u.ac.jp/WAT/jiji-corpus/index.html
[5] http://www.cfilt.iitb.ac.in/iitb_parallel/index.html
[6] http://lotus.kuee.kyoto-u.ac.jp/WAT/recipe-corpus/index.html

Lang	Train	Dev	DevTest	Test
CJ	1,000,000	2,000	2,000	2,000
KJ	1,000,000	2,000	2,000	2,000
EJ	1,000,000	2,000	2,000	2,000

Table 2: Statistics for JPC.

in Table 2.

The Sentence pairs in each data were randomly extracted from a description part of comparable patent documents under the condition that a similarity score between two sentences is greater than or equal to the threshold value 0.05. The similarity score was calculated by the method from (Utiyama and Isahara, 2007) as with ASPEC. Document pairs which were used to extract sentence pairs for each data were not used for the other data. Furthermore, the sentence pairs were extracted so as to be the same number among the four sections. The maximize number of sentence pairs which are extracted from one document pair was limited to 60 for training data and 20 for the development, development-test and test data.

The training data for JPC-CJ was made with sentence pairs of Chinese-Japanese patent documents published in 2012. For JPC-KJ and JPC-EJ, the training data was extracted from sentence pairs of Korean-Japanese and English-Japanese patent documents published in 2011 and 2012. The development, development-test and test data for JPC-CJ, JPC-KJ and JPC-EJ were respectively made with 100 patent documents published in 2013.

2.3 JIJI Corpus

JIJI Corpus was constructed by Jiji Press, Ltd. in collaboration with NICT. The corpus consists of news text that comes from Jiji Press news of various categories including politics, economy, nation, business, markets, sports and so on. The corpus is partitioned into training, development, development-test and test data, which consists of Japanese-English sentence pairs. The statistics for each corpus are shown in Table 3.

The sentence pairs in each data are identified in the same manner as that for ASPEC using the method from (Utiyama and Isahara, 2007).

Lang	Train	Dev	DevTest	Test
EJ	200,000	2,000	2,000	2,000

Table 3: Statistics for JIJI Corpus.

Lang	Train	Dev	Test	Mono
H	–	–	–	45,075,279
EH	1,492,827	520	2,507	–
JH	152,692	1,566	2,000	–

Table 4: Statistics for IITB Corpus. "Mono" indicates monolingual Hindi corpus.

2.4 IITB Corpus

IIT Bombay English-Hindi corpus contains English-Hindi parallel corpus (IITB-EH) as well as monolingual Hindi corpus collected from a variety of sources and corpora developed at the Center for Indian Language Technology, IIT Bombay over the years. This corpus is used for mixed domain subtasks H↔E. Furthermore, mixed domain subtasks H↔J were added as a pivot language task with a parallel corpus created using publicly available corpora (IITB-JH) [7]. Most sentence pairs in IITB-JH come from the Bible corpus. The statistics for each corpus are shown in Table 4.

2.5 Recipe Corpus

Recipe Corpus was constructed by Cookpad Inc. Each recipe consists of a title, ingredients, steps, a description and a history. Every text in titles, ingredients and steps consists of a parallel sentence while one in descriptions and histories is not always a parallel sentence. Although all of the texts in the training set can be used for training, only titles, ingredients and steps in the test set is used for evaluation. The statistics for each corpus are described in Table 5.

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.

[7] http://lotus.kuee.kyoto-u.ac.jp/WAT/Hindi-corpus/WAT2017-Ja-Hi.zip

Lang	TextType	Train	Dev	DevTest	Test
EJ	Title	14,779	500	500	500
	Ingredient	127,244	4,274	4,188	3,935
	Step	108,993	3,303	3,086	2,804

Table 5: Statistics for Recipe Corpus.

That is, the specific baseline system was the standard for human evaluation. A phrase-based statistical machine translation (SMT) system was adopted as the specific baseline system at WAT 2017, which is the same system as that at WAT 2014 to WAT 2016.

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, seven commercial rule-based machine translation (RBMT) systems, and two online translation systems. We also experimentally produced results for the baseline systems that consisted of an neural machine translation system using the implementation of (Vaswani et al., 2017). The SMT baseline systems consisted of publicly available software, and the procedures for building the systems and for translating using the systems were published on the WAT web page[8]. 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 6.

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. Because our objective is not to compare commercial RBMT systems or online translation systems from companies that did not themselves participate, the system IDs of these systems are anonymous in this paper.

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

System ID	System	Type	ASPEC				JPC						IITB		JIJI		RECIPE	
			JE	EJ	JC	CJ	JE	EJ	JC	CJ	JK	KJ	HE	EH	JE	EJ	JE	EJ
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	PC-Transer V13 (Commercial system)	RBMT													✓	✓	✓	✓
RBMT X	J-Beijing 7 (Commercial system)	RBMT			✓	✓			✓	✓								
RBMT X	Hohrai 2011 (Commercial system)	RBMT			✓	✓				✓								
RBMT X	J Soul 9 (Commercial system)	RBMT									✓	✓						
RBMT X	Korai 2011 (Commercial system)	RBMT									✓	✓						
Online X	Google translate	Other	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Online X	Bing translator	Other	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AIAYN	Google's implementation of "Attention Is All You Need"	NMT	✓	✓														

Table 6: Baseline Systems

3.1 Training Data

We used the following data for training 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 ASPEC Japanese–English training data, we only used train-1.txt, which consists of one 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[9] for Japanese segmentation.
- Stanford Word Segmenter version 2014-01-04[10] (Chinese Penn Treebank (CTB) model) for Chinese segmentation.
- The Moses toolkit for English and Indonesian tokenization.
- Mecab-ko[11] for Korean segmentation.
- Indic NLP Library[12] for Hindi segmentation.

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 for JE, EJ, JC, and CJ
  - 0 for JK, KJ, HE, and EH
  - 6 for IE and EI
- 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 the 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 the 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 for Calculating Automatic Evaluation Score

We evaluated translation results by three metrics: BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and AMFM (Banchs et al., 2015). BLEU scores were calculated using `multi-bleu.perl` which was distributed

[9] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[10] http://nlp.stanford.edu/software/segmenter.shtml
[11] https://bitbucket.org/eunjeon/mecab-ko/
[12] https://bitbucket.org/anoopk/indic_nlp_library

with the Moses toolkit (Koehn et al., 2007). RIBES scores were calculated using `RIBES.py` version 1.02.4 [13]. AMFM scores were calculated using scripts created by the technical collaborators of WAT2017. All scores for each task were calculated using the corresponding reference.

Before the calculation of the automatic evaluation scores, the translation results were tokenized with word segmentation tools for each language. For Japanese segmentation, we used three different tools: Juman version 7.0 (Kurohashi et al., 1994), KyTea 0.4.6 (Neubig et al., 2011) with Full SVM model [14] and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0 [15]. For Chinese segmentation, we used two different tools: KyTea 0.4.6 with Full SVM Model in MSR model and Stanford Word Segmenter (Tseng, 2005) version 2014-06-16 with Chinese Penn Treebank (CTB) and Peking University (PKU) model [16]. For Korean segmentation we used mecab-ko [17]. For English segmentation, we used `tokenizer.perl` [18] in the Moses toolkit. For Hindi segmentation, we used Indic NLP Library [19]. The detailed procedures for the automatic evaluation are shown on the WAT2017 evaluation web page [20].

4.2 Automatic Evaluation System

The participants submit translation results via an automatic evaluation system deployed on the WAT2017 web page, which automatically gives evaluation scores for the uploaded results. Figure 1 shows the submission interface for participants. The system requires participants to provide the following information when they upload translation results:

- Subtask:
  Scientific papers subtask (J↔E, J↔C),
  Patents subtask (C↔J, K↔J, E↔J),
  Newswire subtask (J↔E),
  Mixed domain subtask (H↔E, H↔J) or
  Recipe subtask (J↔E);
- Method:
  SMT, RBMT, SMT and RBMT, EBMT, NMT or Other;
- Use of other resources in addition to the provided data ASPEC / JPC / IITB Corpus / JIJI Corpus / Recipe Corpus;
- Permission to publish automatic evaluation scores on the WAT2017 web page.

Although participants can confirm only the information that they filled or uploaded, the server for the system stores all submitted information including translation results and scores. Information about translation results that participants permit to be published is disclosed via the WAT2017 evaluation web page. Participants can also submit the results for human evaluation using the same web interface. This automatic evaluation system will remain available even after WAT2017. Anybody can register an account for the system by following the procesures in the registration web page [21].

[13] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html
[14] http://www.phontron.com/kytea/model.html
[15] http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz
[16] http://nlp.stanford.edu/software/segmenter.shtml
[17] https://bitbucket.org/eunjeon/mecab-ko/
[18] https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl
[19] https://bitbucket.org/anoopk/indic_nlp_library
[20] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html
[21] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2017/registration/index.html

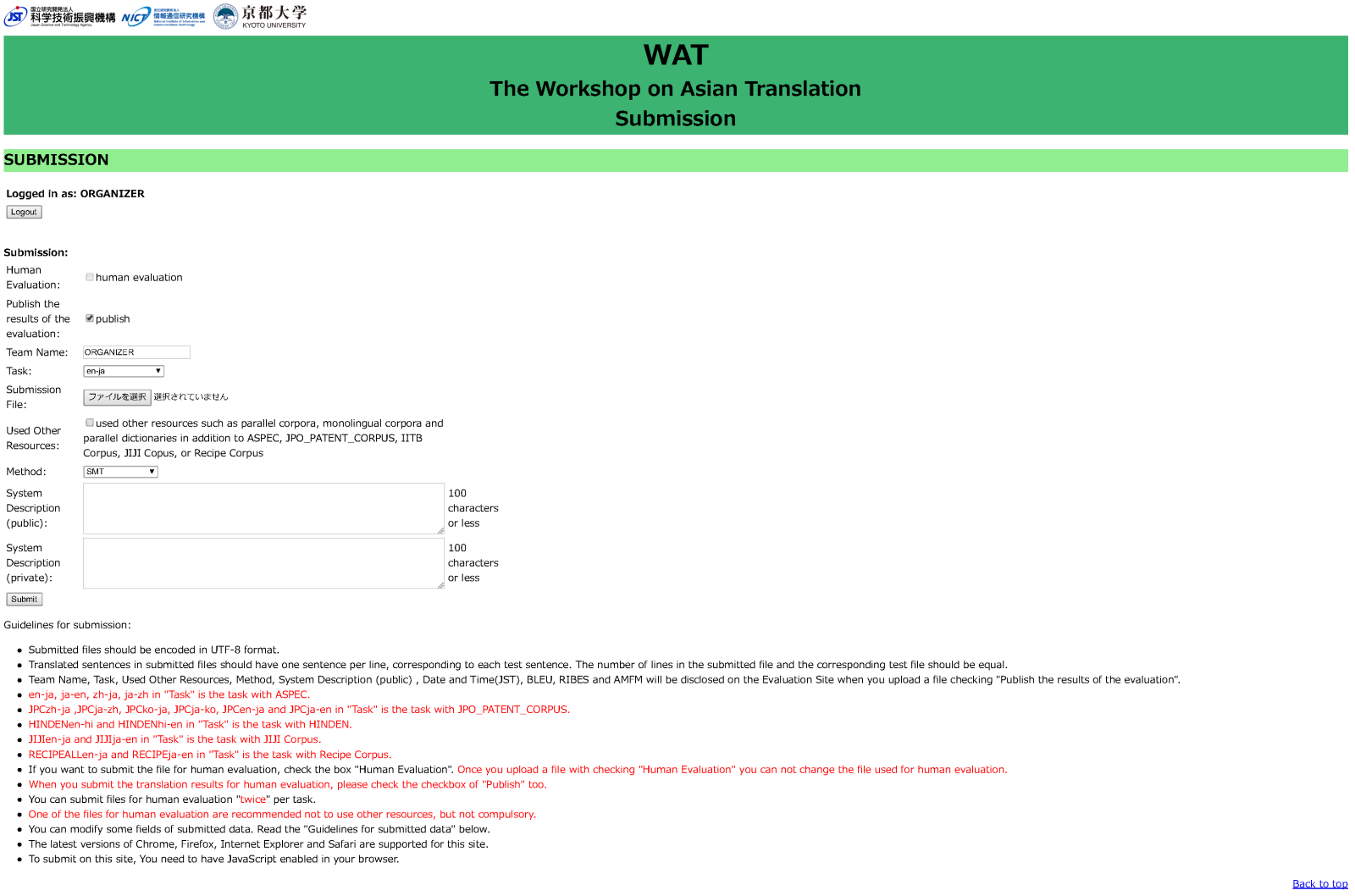

Figure 1: The submission web page for participants

5 Human Evaluation

In WAT2017, we conducted 2 kinds of human evaluations: *pairwise evaluation* and *JPO adequacy evaluation*.

5.1 Pairwise Evaluation

The pairwise evaluation is the same as the last year, but not using the crowdsourcing this year. We asked professional translation company to do pairwise evaluation. The cost of pairwise evaluation per sentence is almost the same to that of last year.

We randomly chose 400 sentences from the Test set for the pairwise evaluation. We used the same sentences as the last year for the continuous subtasks. Each submission is compared with the baseline translation (Phrase-based SMT, described in Section 3) and given a *Pairwise* score.

5.1.1 Pairwise Evaluation of Sentences

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

5.1.2 Voting

To guarantee the quality of the evaluations, each sentence is evaluated by 5 different annotators and the final decision is made depending on the 5 judgements. We define each judgement $j_i (i = 1, \cdots, 5)$ as:

$$j_i = \begin{cases} 1 & \text{if better than the baseline} \\ -1 & \text{if worse than the baseline} \\ 0 & \text{if the quality is the same} \end{cases}$$

The final decision D is defined as follows using $S = \sum j_i$:

$$D = \begin{cases} win & (S \geq 2) \\ loss & (S \leq -2) \\ tie & (otherwise) \end{cases}$$

5.1.3 Pairwise Score Calculation

Suppose that W is the number of *wins* compared to the baseline, L is the number of *losses* and T is the number of *ties*. The Pairwise score can be calculated by the following formula:

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

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

5.1.4 Confidence Interval Estimation

There are several ways to estimate a confidence interval. We chose to use bootstrap resampling (Koehn, 2004) to estimate the 95% confidence interval. The procedure is as follows:

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

5.2 JPO Adequacy Evaluation

The participants' systems, which achieved the top 3 highest scores among the pairwise evaluation results of each subtask[22], were also evaluated with the JPO adequacy evaluation. The JPO adequacy evaluation was carried out by translation experts with a quality evaluation criterion for translated patent documents which the Japanese Patent Office (JPO) decided. For each system, two annotators evaluate the test sentences to guarantee the quality.

5.2.1 Evaluation of Sentences

The number of test sentences for the JPO adequacy evaluation is 200. The 200 test sentences were randomly selected from the 400 test sentences of the pairwise evaluation. The test sentence include the input sentence, the submitted system's translation and the reference translation.

[22] The number of systems varies depending on the subtasks.

5	All important information is transmitted correctly. (100%)
4	Almost all important information is transmitted correctly. (80%–)
3	More than half of important information is transmitted correctly. (50%–)
2	Some of important information is transmitted correctly. (20%–)
1	Almost all important information is NOT transmitted correctly. (–20%)

Table 7: The JPO adequacy criterion

5.2.2 Evaluation Criterion

Table 7 shows the JPO adequacy criterion from 5 to 1. The evaluation is performed subjectively. "Important information" represents the technical factors and their relationships. The degree of importance of each element is also considered to evaluate. The percentages in each grade are rough indications for the transmission degree of the source sentence meanings. The detailed criterion can be found on the JPO document (in Japanese) [23].

6 Participants List

Table 8 shows the list of participants for WAT2017. This includes not only Japanese organizations, but also some organizations from outside Japan. 12 teams submitted one or more translation results to the automatic evaluation server or human evaluation.

[23] http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

Team ID	Organization	ASPEC					JPC					IITBC			JIJI		RECIPE						
		JE	EJ	JC	CJ		JE	EJ	JC	CJ	KJ	HE	EH		JE	EJ	TTL JE	TTL EJ	ING JE	ING EJ	STE JE	STE EJ	
Kyoto-U (Cromieres et al., 2017)	Kyoto University	✓	✓	✓	✓																		
TMU (Matsumura and Komachi, 2017)	Tokyo Metropolitan University	✓	✓	✓																			
EHR (Ehara, 2017)	Ehara NLP Research Laboratory							✓		✓	✓												
NTT (Morishita et al., 2017)	NTT Communication Science Laboratories	✓	✓					✓		✓	✓				✓	✓							
JAPIO (Kinoshita et al., 2017)	Japan Patent Information Organization						✓	✓		✓	✓												
NICT-2 (Imamura and Sumita, 2017)	National Institute of Information and Communications Technology	✓	✓	✓	✓		✓	✓		✓	✓				✓								
XMUNLP (Wang et al., 2017)	Xiamen University											✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	
UT-IIS (Neishi et al., 2017)	The University of Tokyo		✓																				
CUNI (Kocmi et al., 2017)	Charles University, Institute of Formal and Applied Linguistics	✓					✓								✓								
IITB-MTG (Singh et al., 2017)	Indian Institute of Technology Bombay											✓	✓										
u-tkb (Long et al., 2017)	University of Tsukuba						✓	✓	✓	✓													
NAIST-NICT (Oda et al., 2017)	NAIST/NICT		✓																				

Table 8: List of participants who submitted translation results to WAT2017 and their participation in each subtasks.

7 Evaluation Results

In this section, the evaluation results for WAT2017 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2017 website[24].

7.1 Official Evaluation Results

Figures 2, 3, 4 and 5 show the official evaluation results of ASPEC subtasks, Figures 6, 7, 8, 9 and 10 show those of JPC subtasks, Figures 11 and 12 show those of IITBC subtasks, Figures 13 and 14 show those of JIJI subtasks and Figures 15, 16, 17, 18, 19 and 20 show those of RECIPE subtasks. Each figure contains automatic evaluation results (BLEU, RIBES, AM-FM), the pairwise evaluation results with confidence intervals, correlation between automatic evaluations and the pairwise evaluation, the JPO adequacy evaluation result and evaluation summary of top systems.

The detailed automatic evaluation results for all the submissions are shown in Appendix A. The detailed JPO adequacy evaluation results for the selected submissions are shown in Table 9. The weights for the weighted κ (Cohen, 1968) is defined as $|Evaluation1 - Evaluation2|/4$.

From the evaluation results, the following can be observed:

- The translation quality of this year is better than that of last year for all the subtasks.
- There is no big difference between the neural network based translation models according to the JPO adequacy evaluation results for ASPEC subtasks.

7.2 Statistical Significance Testing of Pairwise Evaluation between Submissions

Tables 10, 11 and 12 show the results of statistical significance testing of ASPEC subtasks, Tables 13, 14 and 15 show those of JPC subtasks, Table 16 shows those of IITBC subtasks, Table 17 shows those of JIJI subtasks and Tables 18, 19 and 20 show those of RECIPE subtasks. $\ggg$, $\gg$ and $>$ mean that the system in the row is *better* than the system in the column at a significance level of $p < 0.01$, 0.05 and 0.1 respectively. Testing is also done by the bootstrap resampling as follows:

1. randomly select 300 sentences from the 400 pairwise evaluation sentences, and calculate the Pairwise 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 workers, we calculated the Fleiss' κ (Fleiss et al., 1971) values. The results are shown in Table 21. We can see that the κ values are larger for X → J translations than for J → X translations. This may be because the majority of the workers are Japanese, and the evaluation of one's mother tongue is much easier than for other languages in general.

8 Submitted Data

The number of published automatic evaluation results for the 14 teams exceeded 300 before the start of WAT2017, and 67 translation results for pairwise evaluation were submitted by 12 teams. Furthermore, we selected several translation results from each subtask according to the pairwise evaluation scores and evaluated them for JPO adequacy evaluation. We will organize the all of the submitted data for human evaluation and make this public.

9 Conclusion and Future Perspective

This paper summarizes the shared tasks of WAT2017. We had 12 participants worldwide, and collected a large number of useful submissions for improving the current machine translation systems by analyzing the submissions and identifying the issues.

For the next WAT workshop, we plan to change the baseline system from the PBSMT to NMT because the pairwise scores are saturated for some of the subtasks. Also, we

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

are planning to do extrinsic evaluation of the translations.

Unfortunately, there was no participants for the small NMT task this year. We will brush-up the task definition and invite participants for the next WAT.

Appendix A Submissions

Tables 22 to 41 summarize all the submissions listed in the automatic evaluation server at the time of the WAT2017 workshop (27th, November, 2017). The OTHER column shows the use of resources such as parallel corpora, monolingual corpora and parallel dictionaries in addition to ASPEC, JPC, IITB Corpus, JIJI Corpus, RECIPE Corpus.

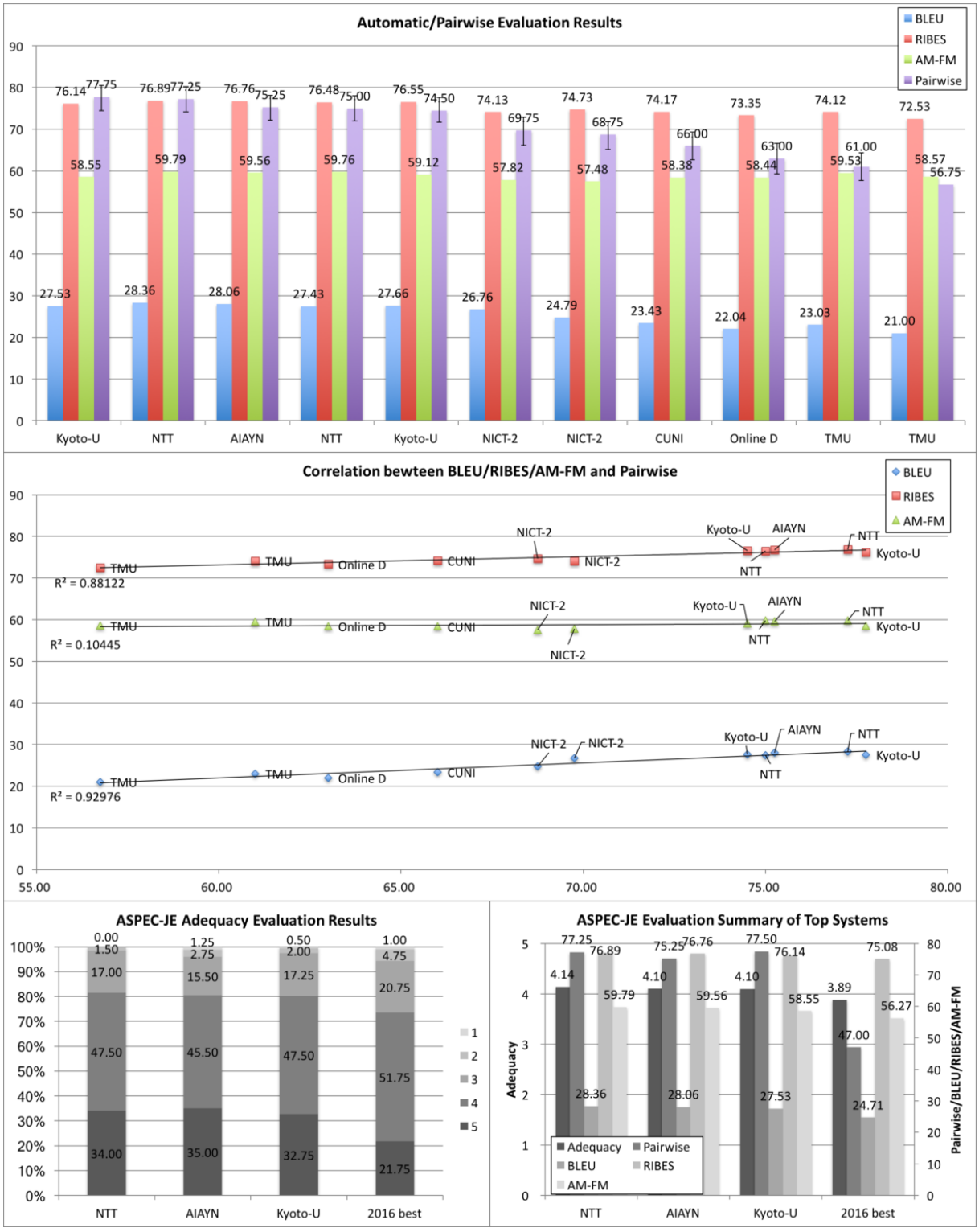

Figure 2: Official evaluation results of ASPEC-JE.

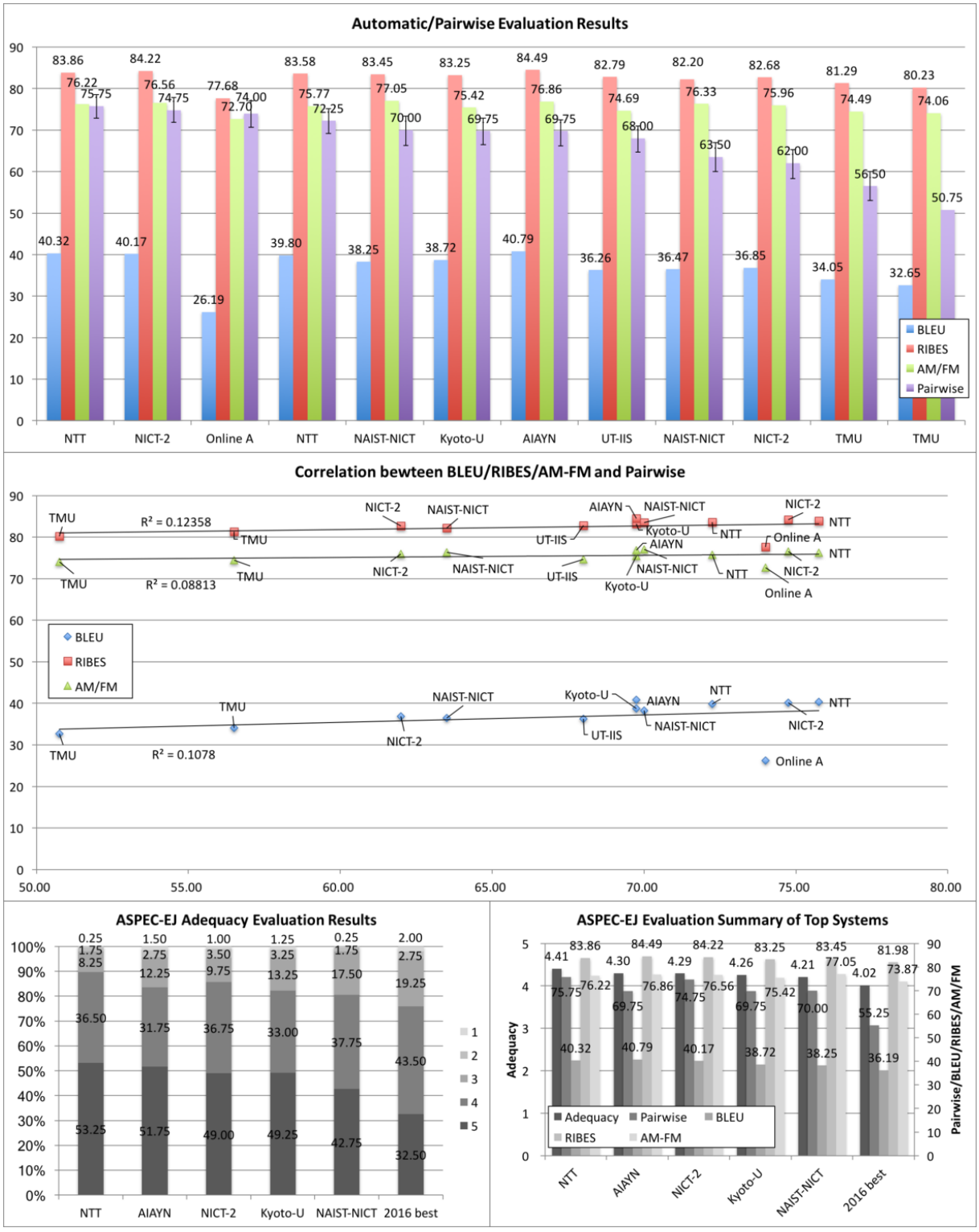

Figure 3: Official evaluation results of ASPEC-EJ.

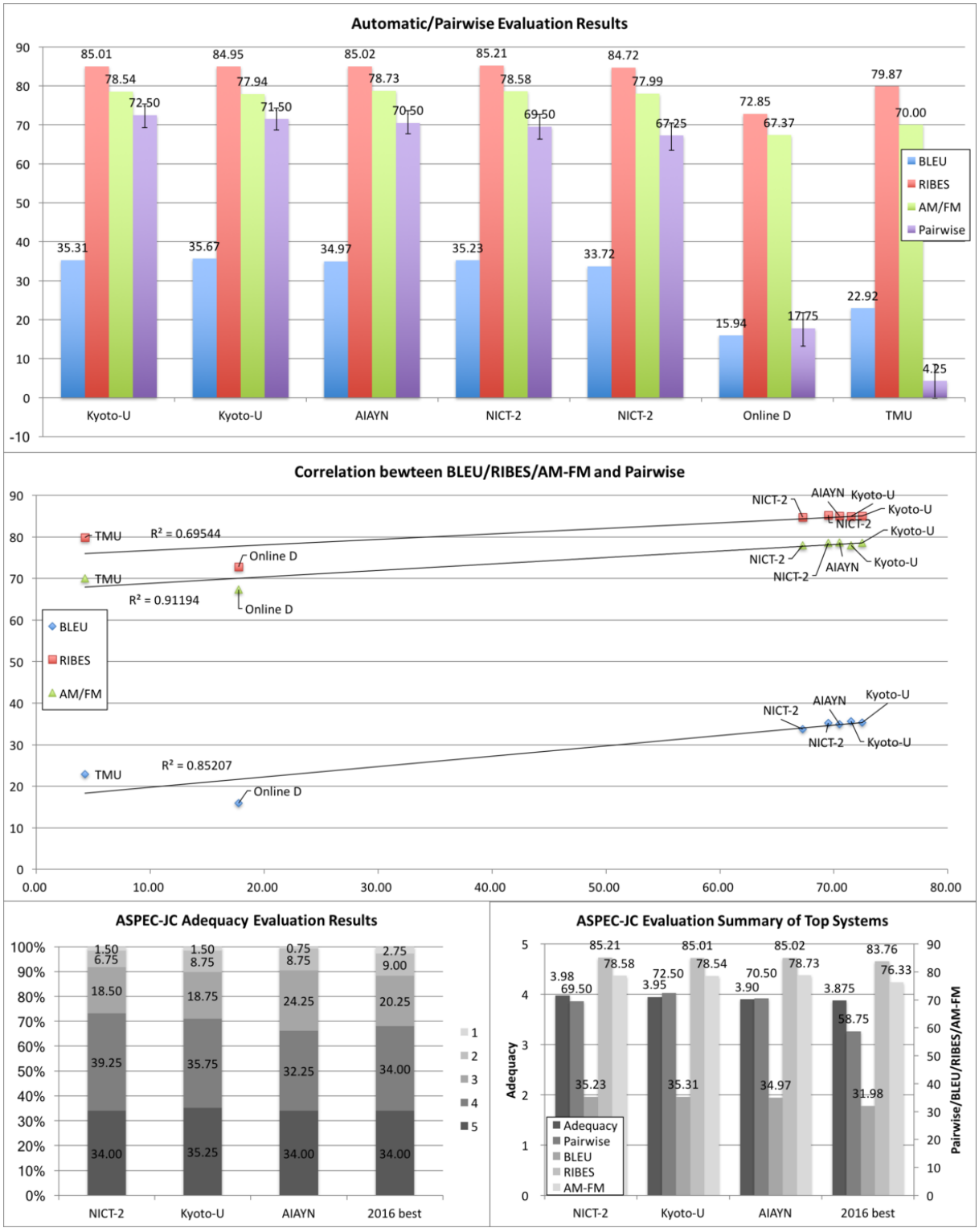

Figure 4: Official evaluation results of ASPEC-JC.

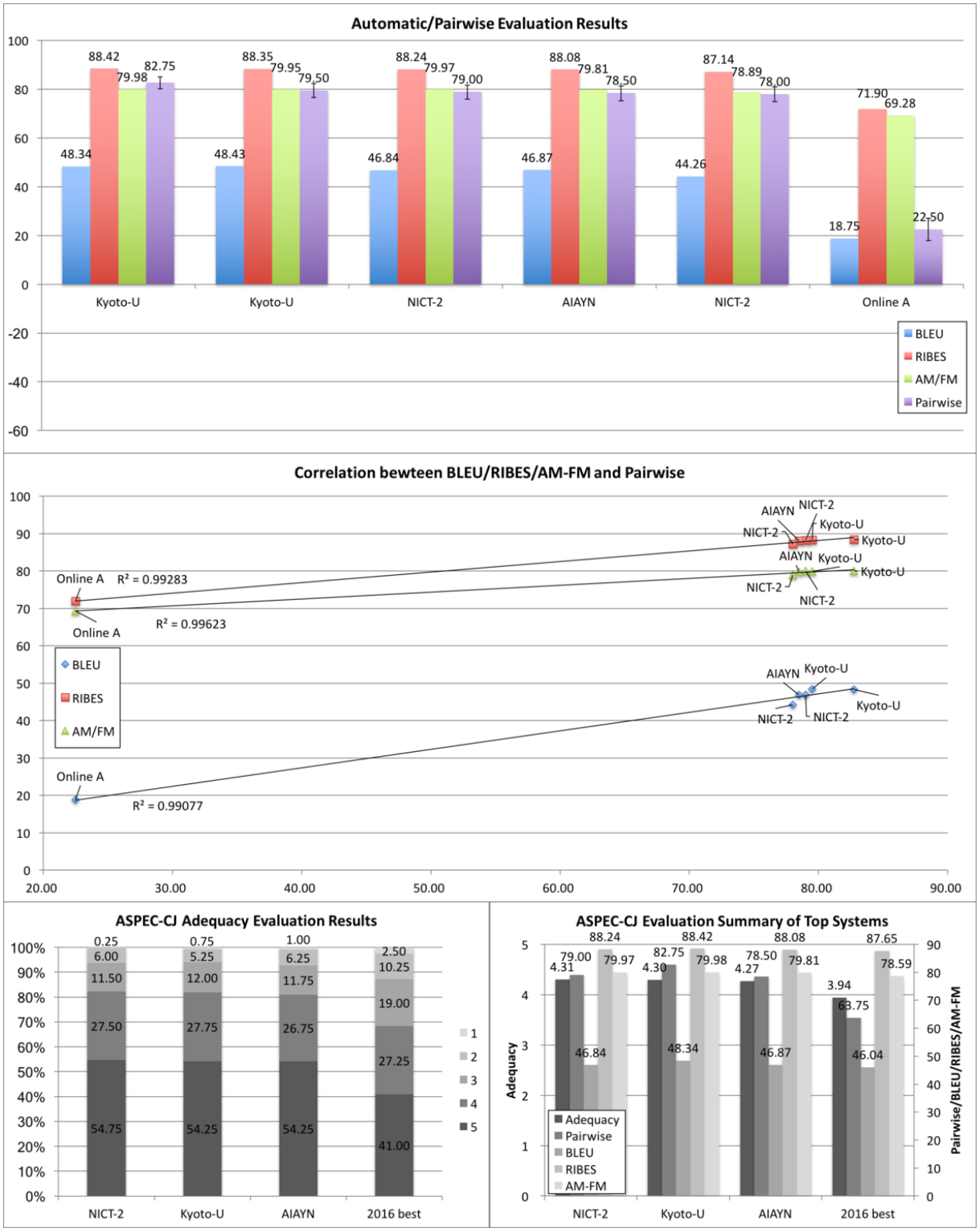

Figure 5: Official evaluation results of ASPEC-CJ.

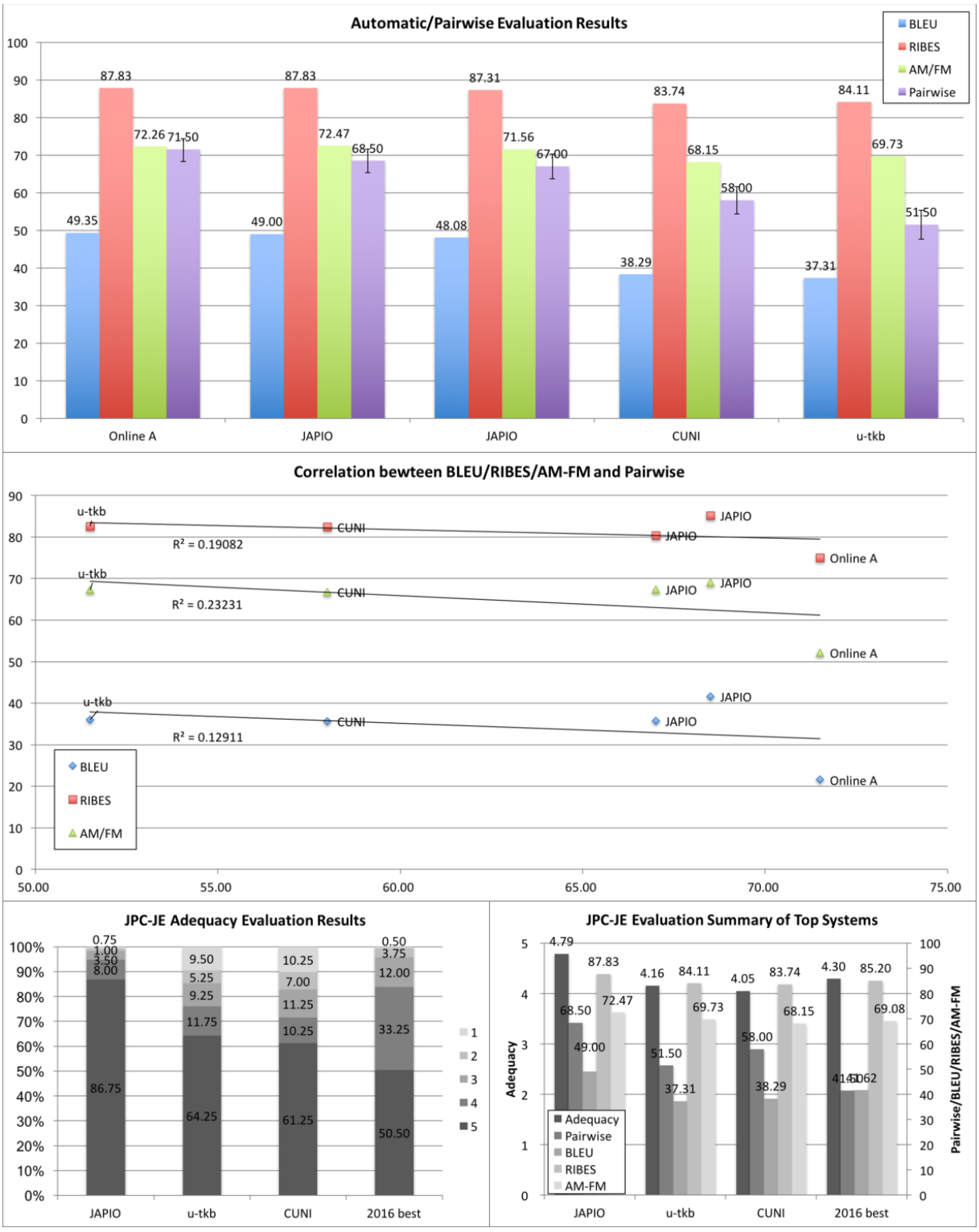

Figure 6: Official evaluation results of JPC-JE.

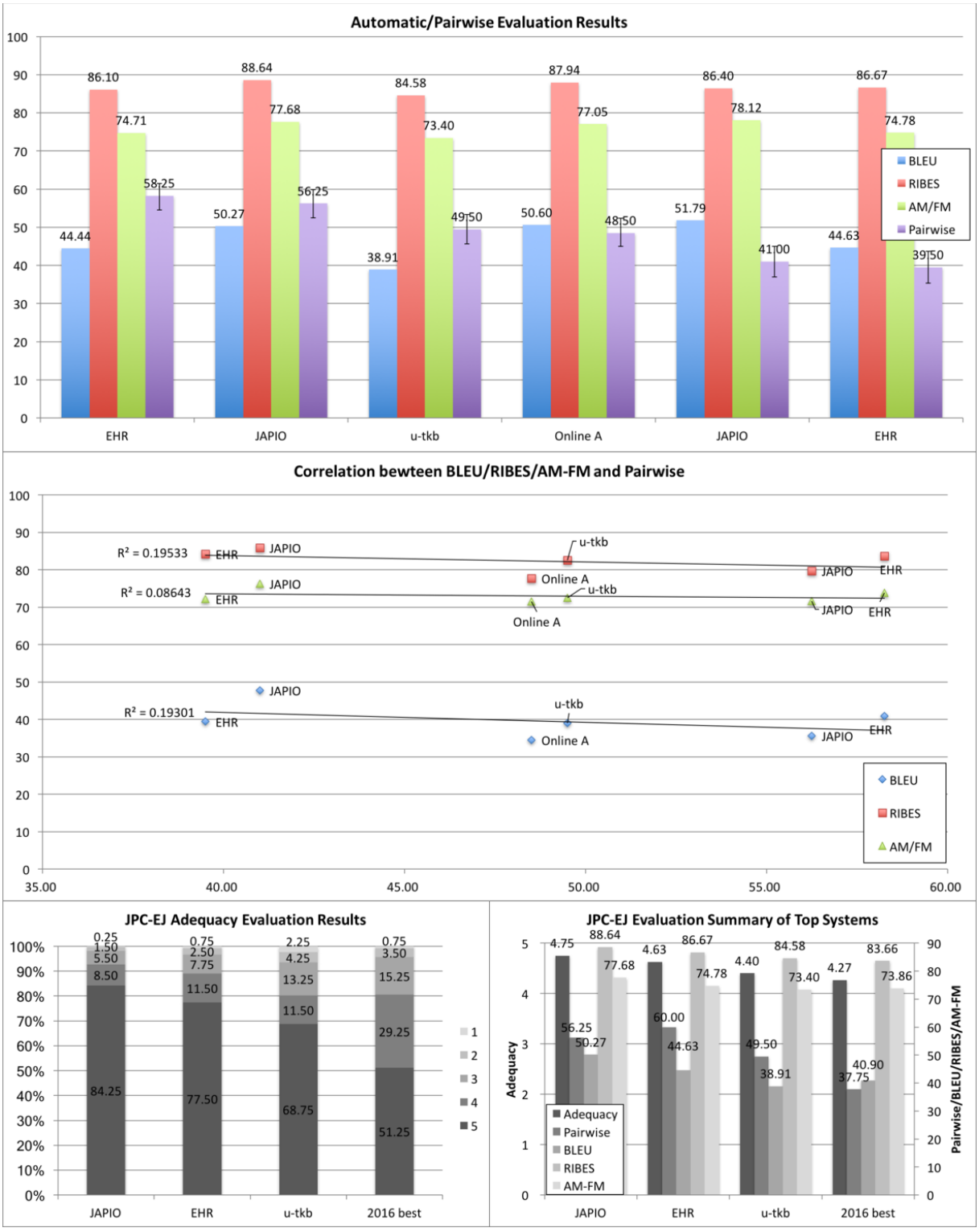

Figure 7: Official evaluation results of JPC-EJ.

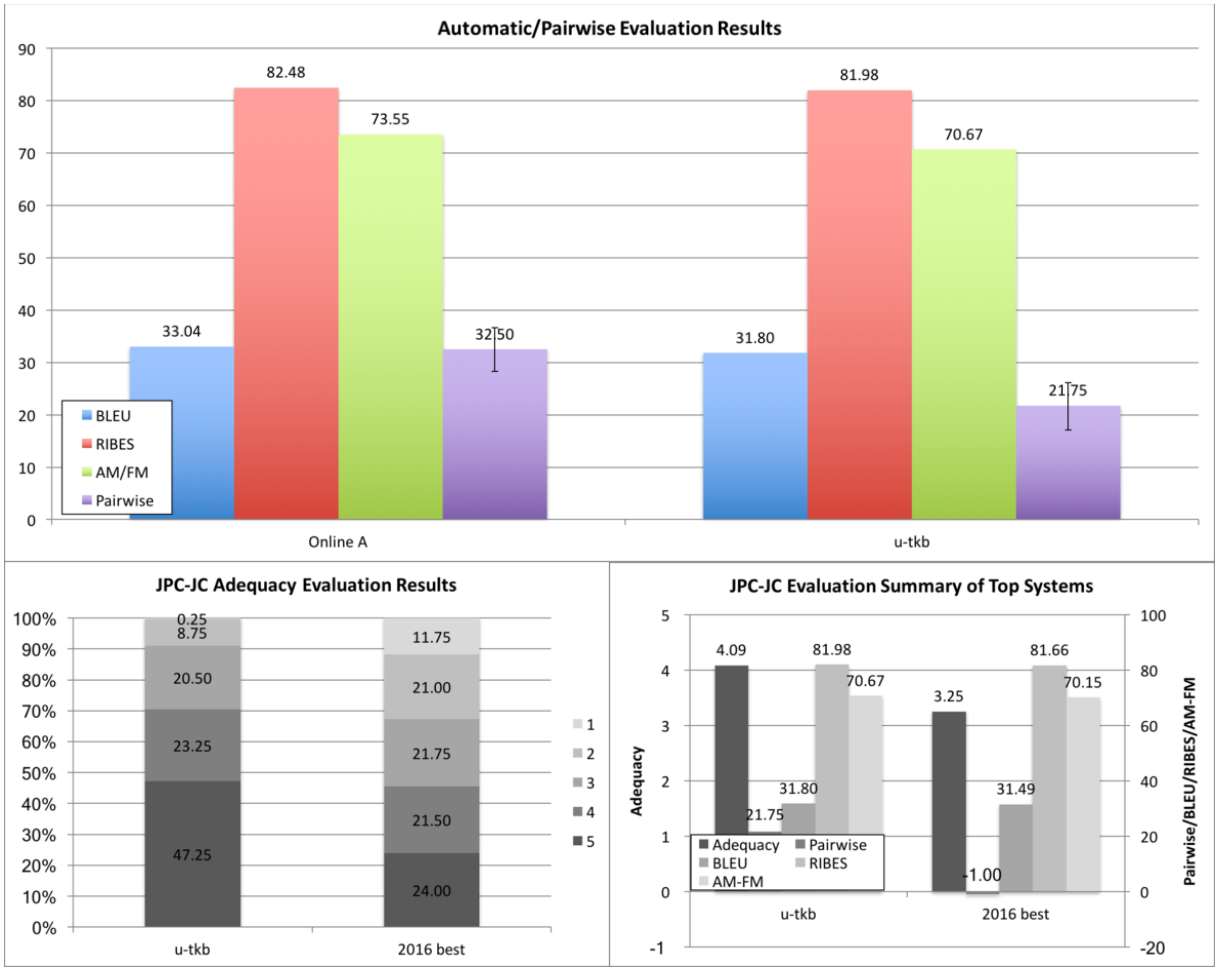

Figure 8: Official evaluation results of JPC-JC.

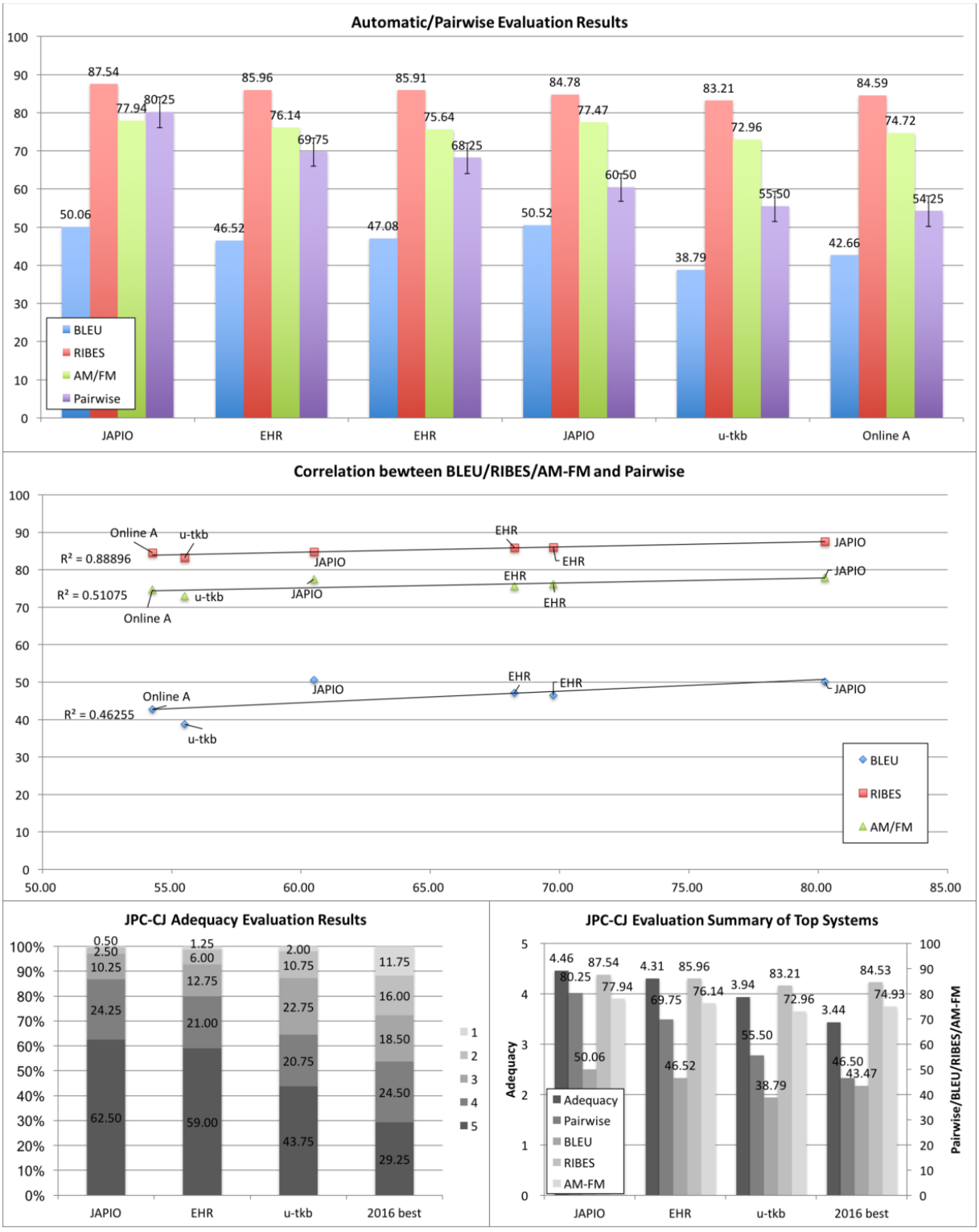

Figure 9: Official evaluation results of JPC-CJ.

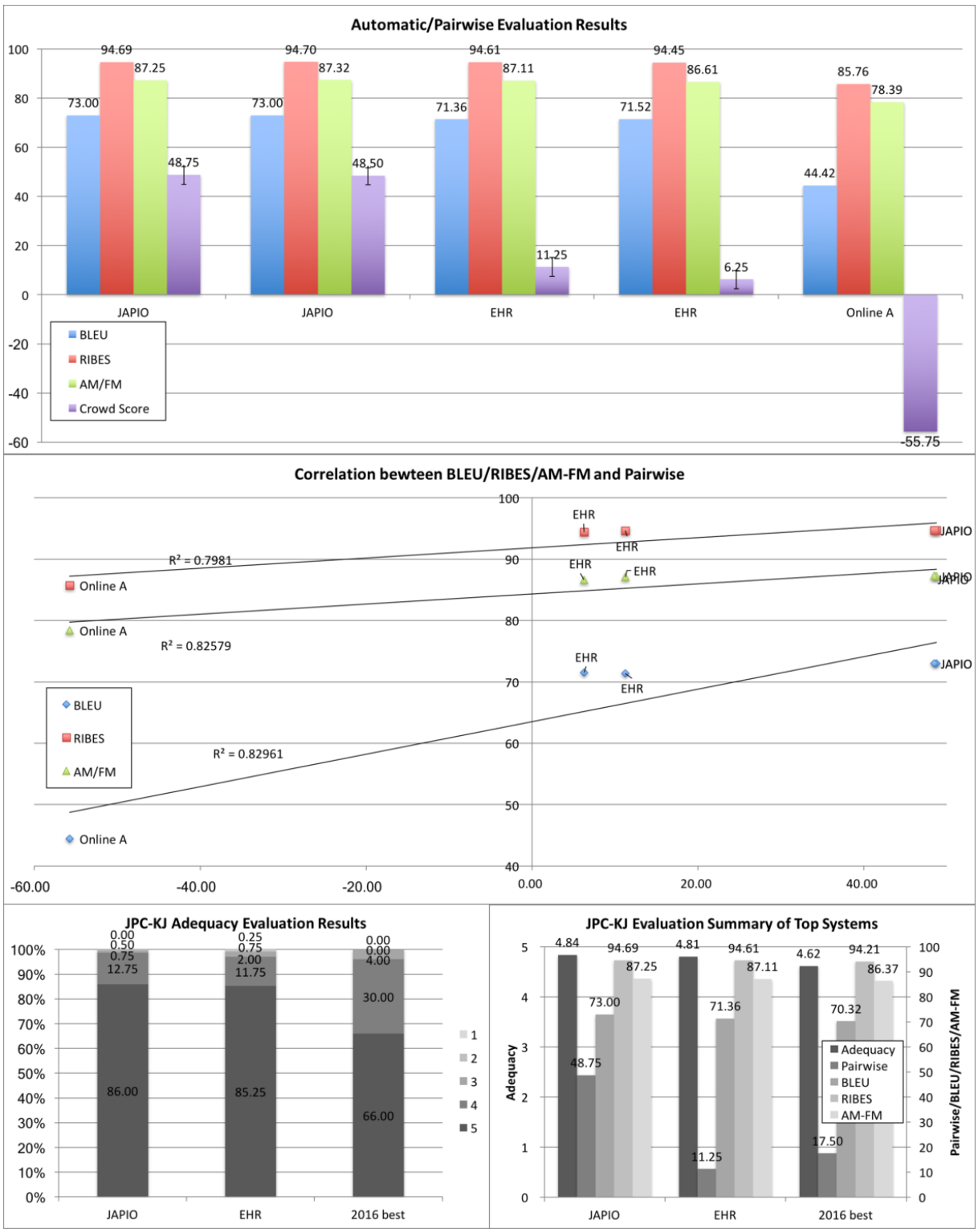

Figure 10: Official evaluation results of JPC-KJ.

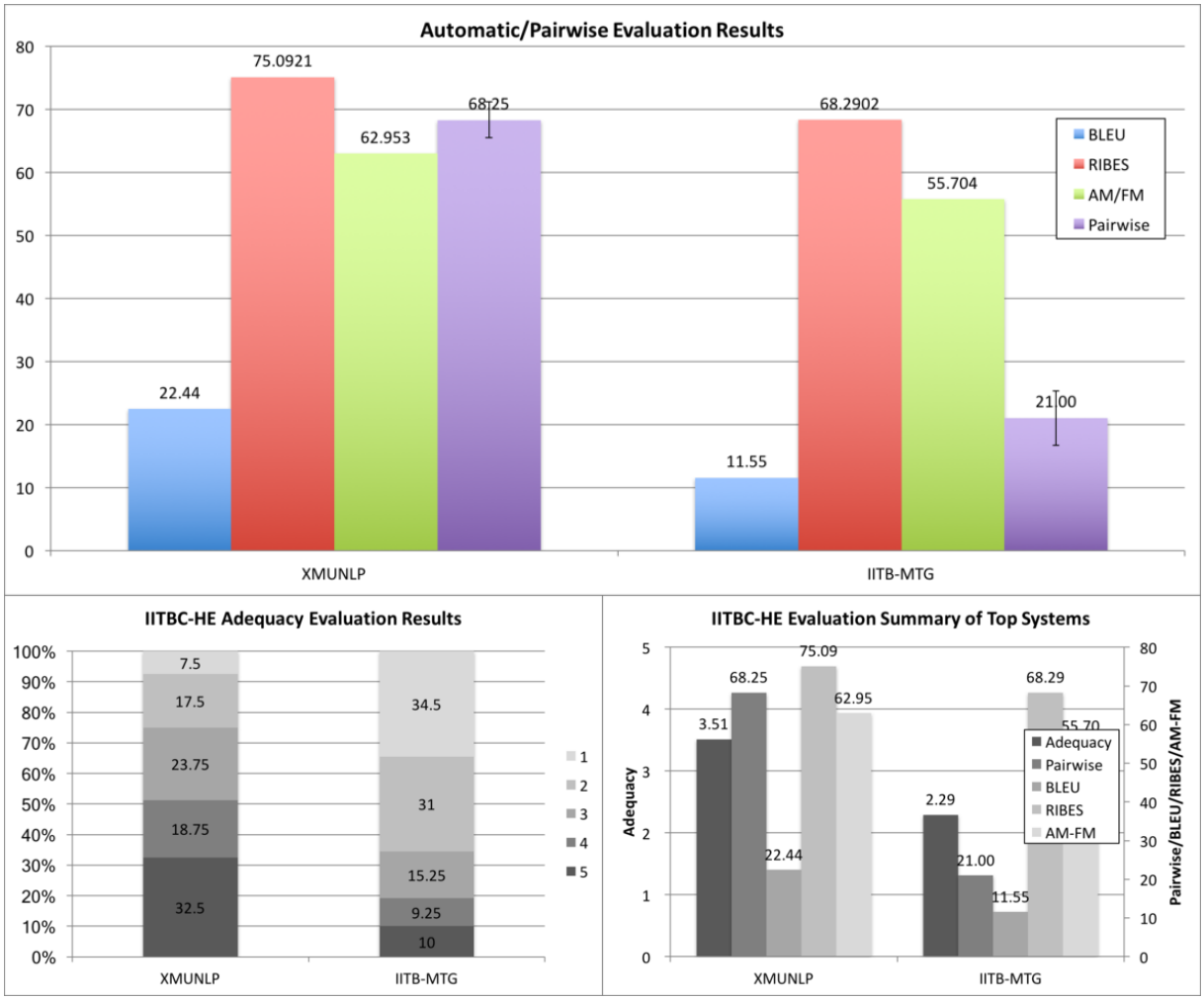

Figure 11: Official evaluation results of IITBC-HE.

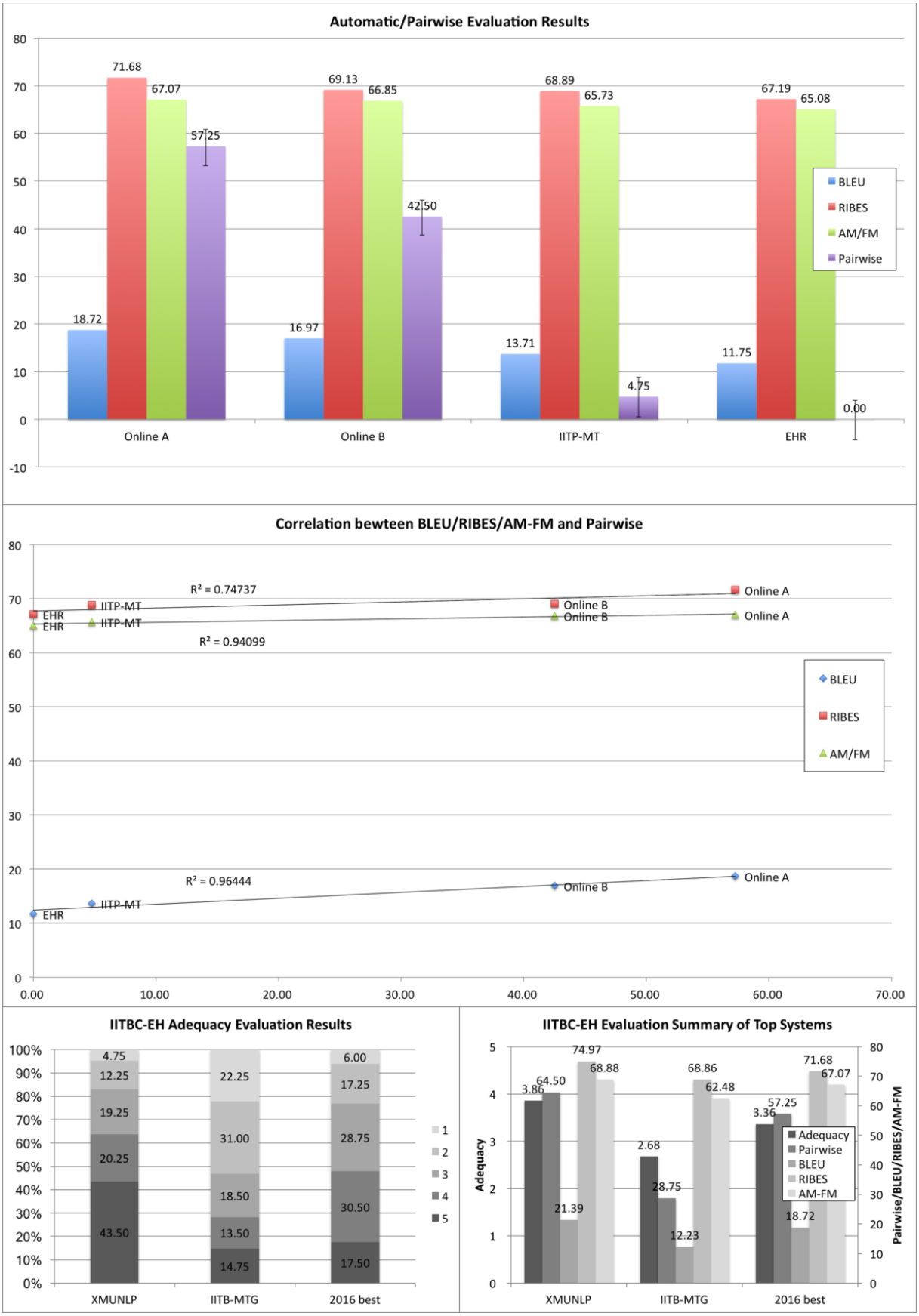

Figure 12: Official evaluation results of IITBC-EH.

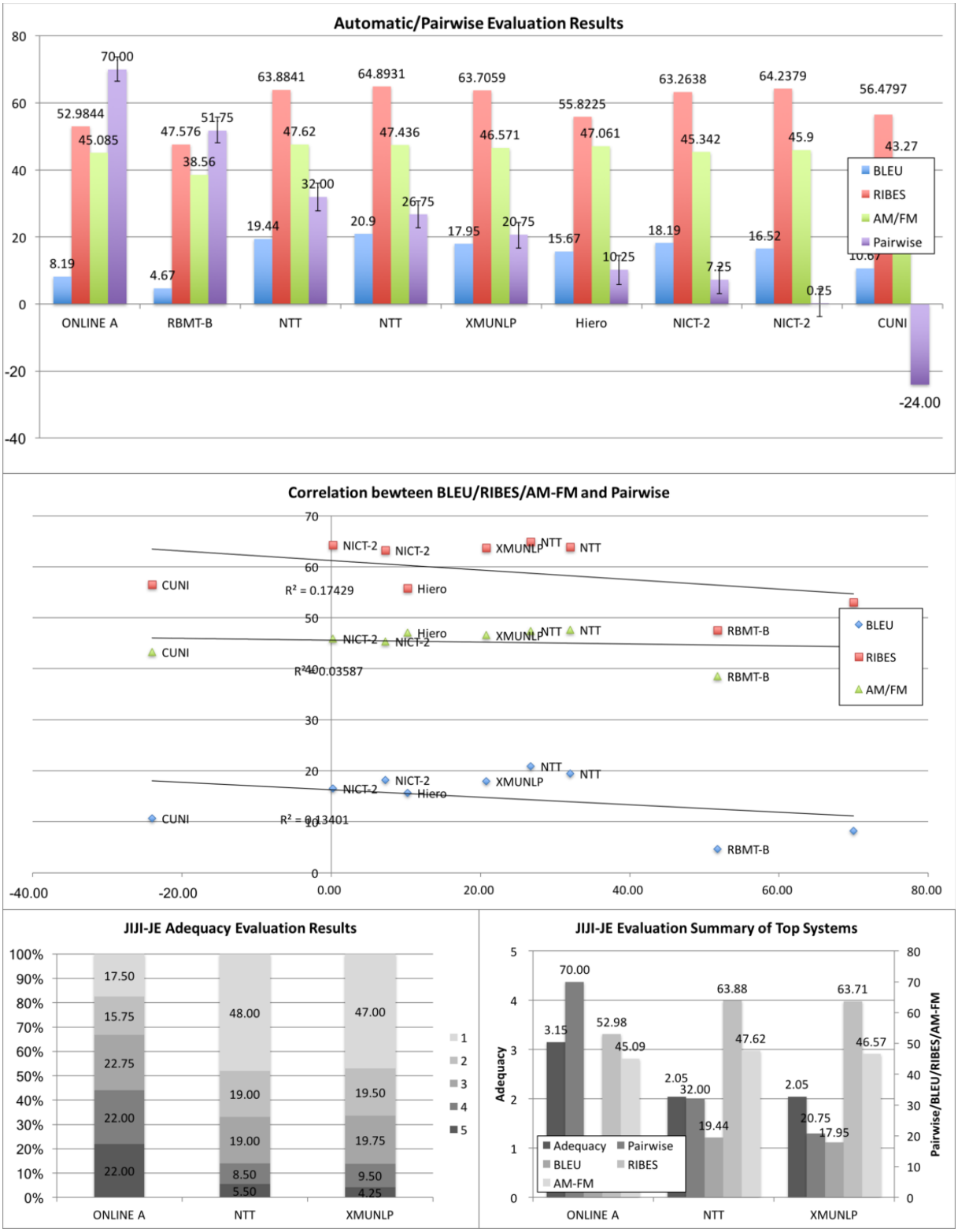

Figure 13: Official evaluation results of JIJI-JE.

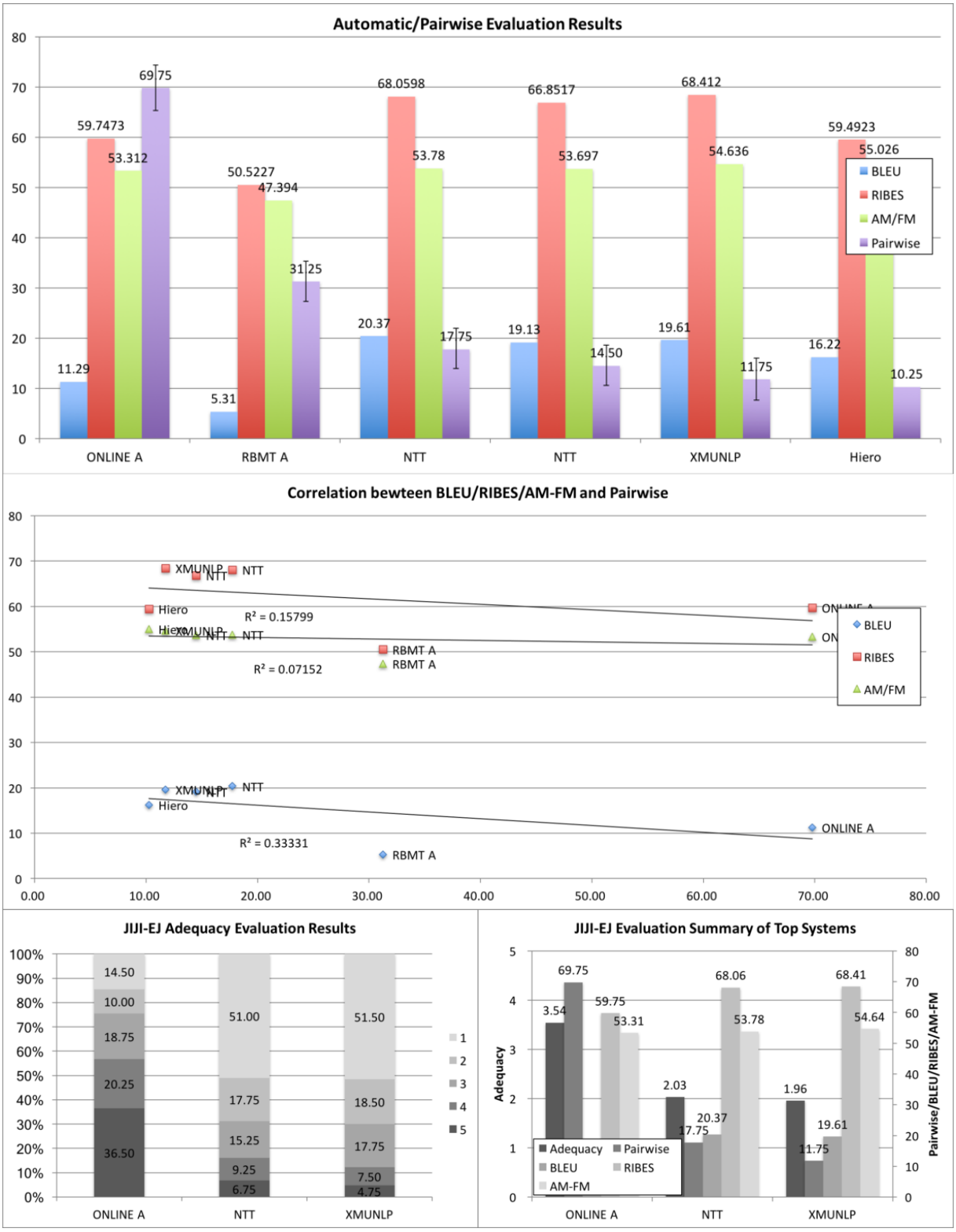

Figure 14: Official evaluation results of JIJI-EJ.

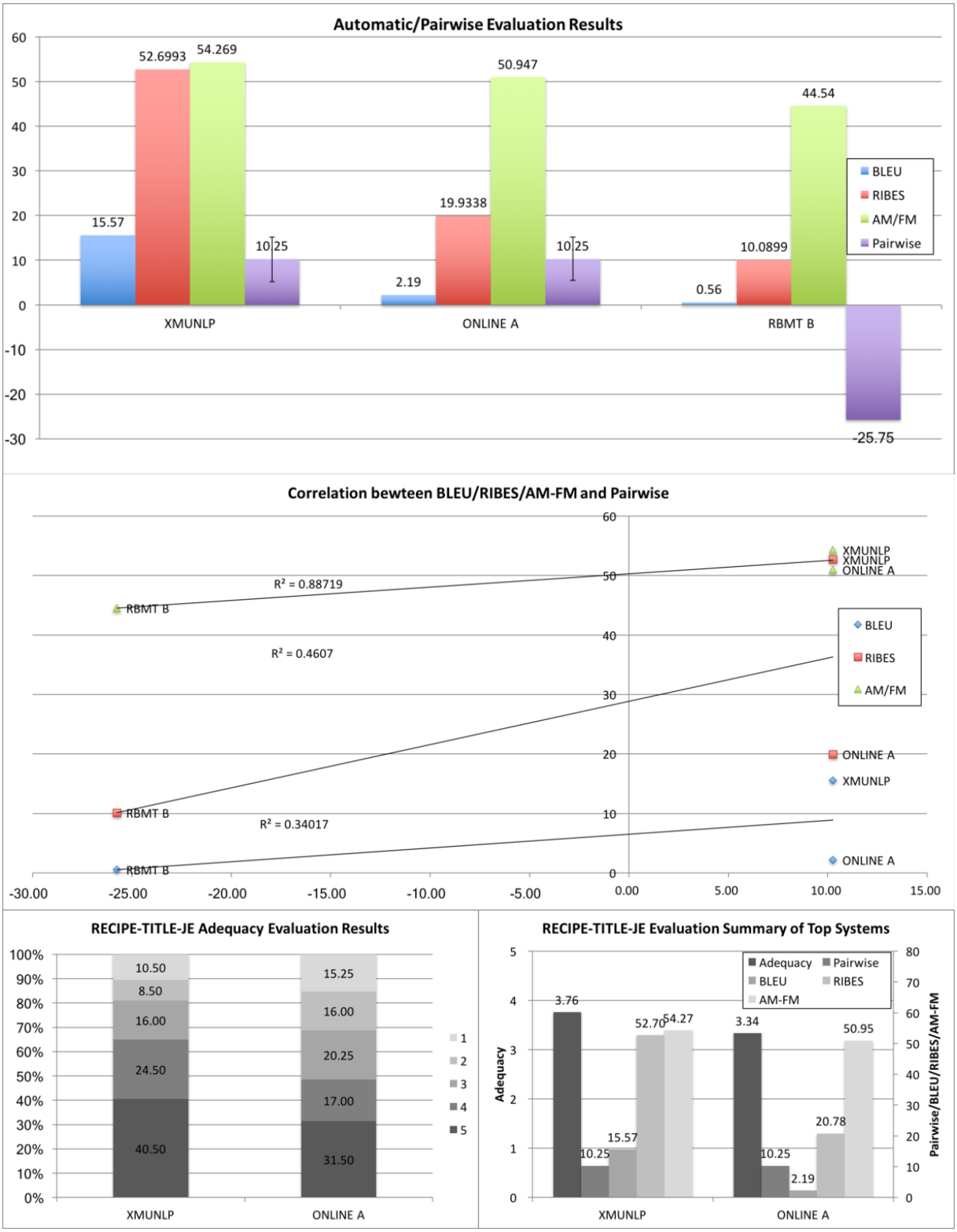

Figure 15: Official evaluation results of RECIPE-TTL-JE.

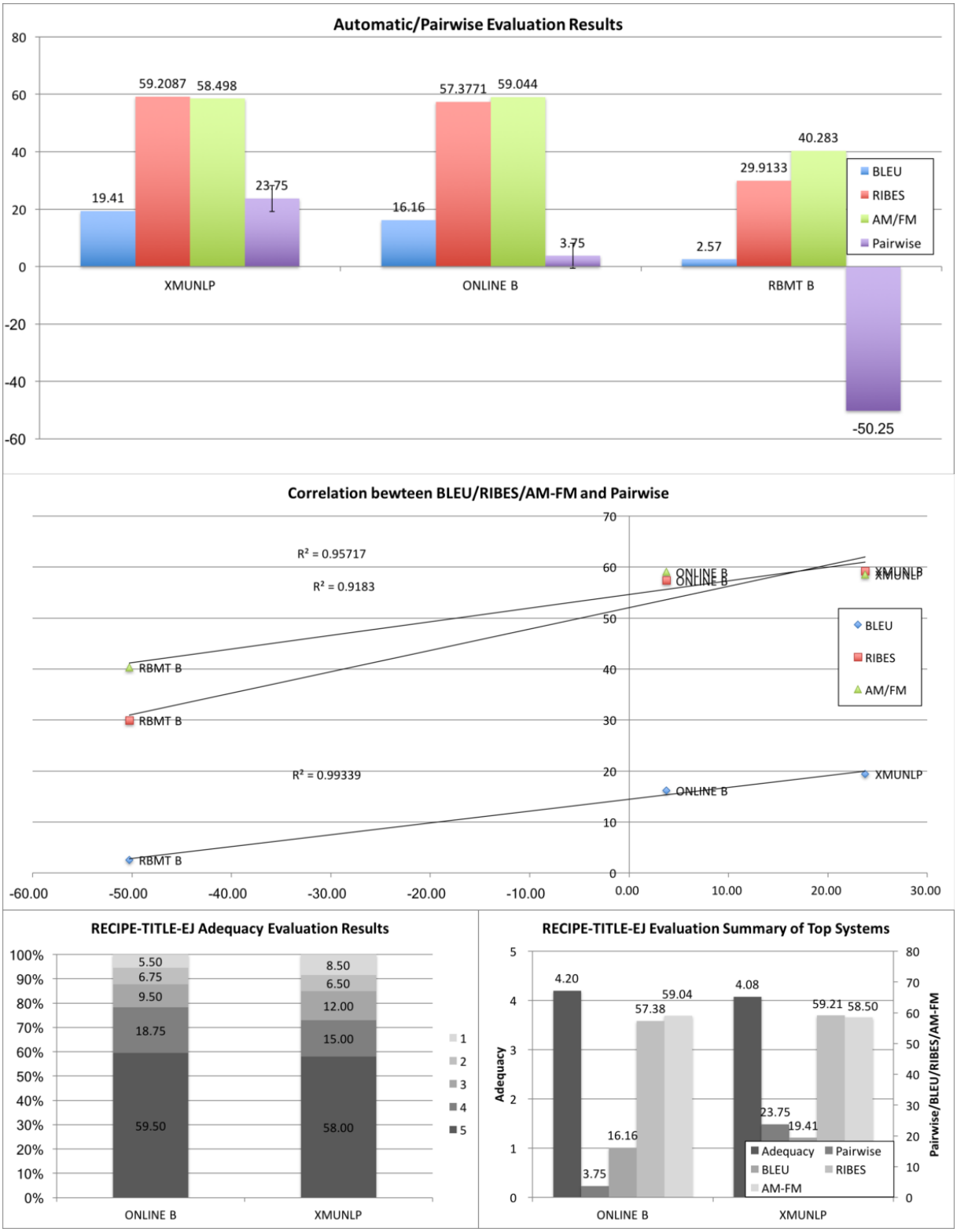

Figure 16: Official evaluation results of RECIPE-TTL-EJ.

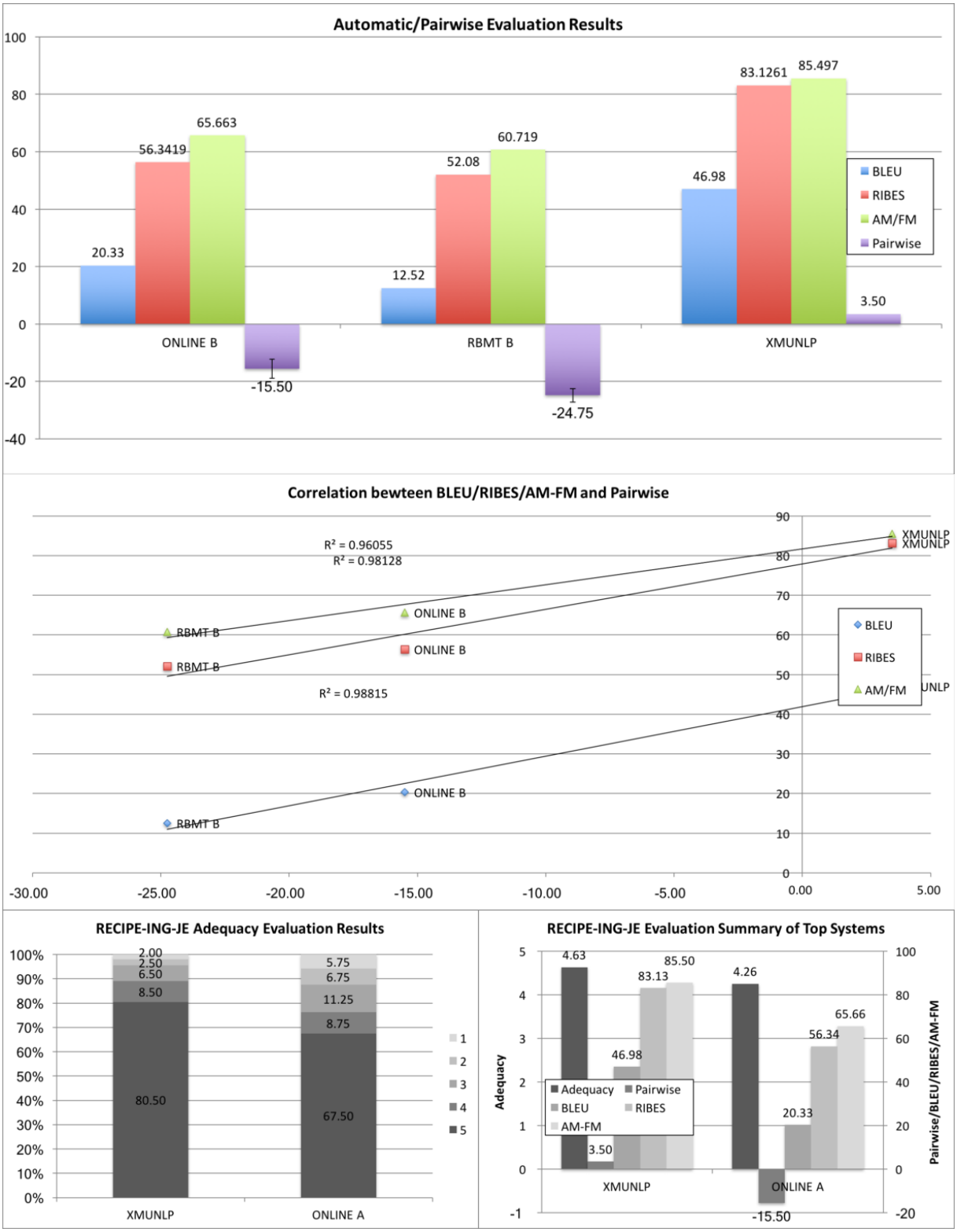

Figure 17: Official evaluation results of RECIPE-ING-JE.

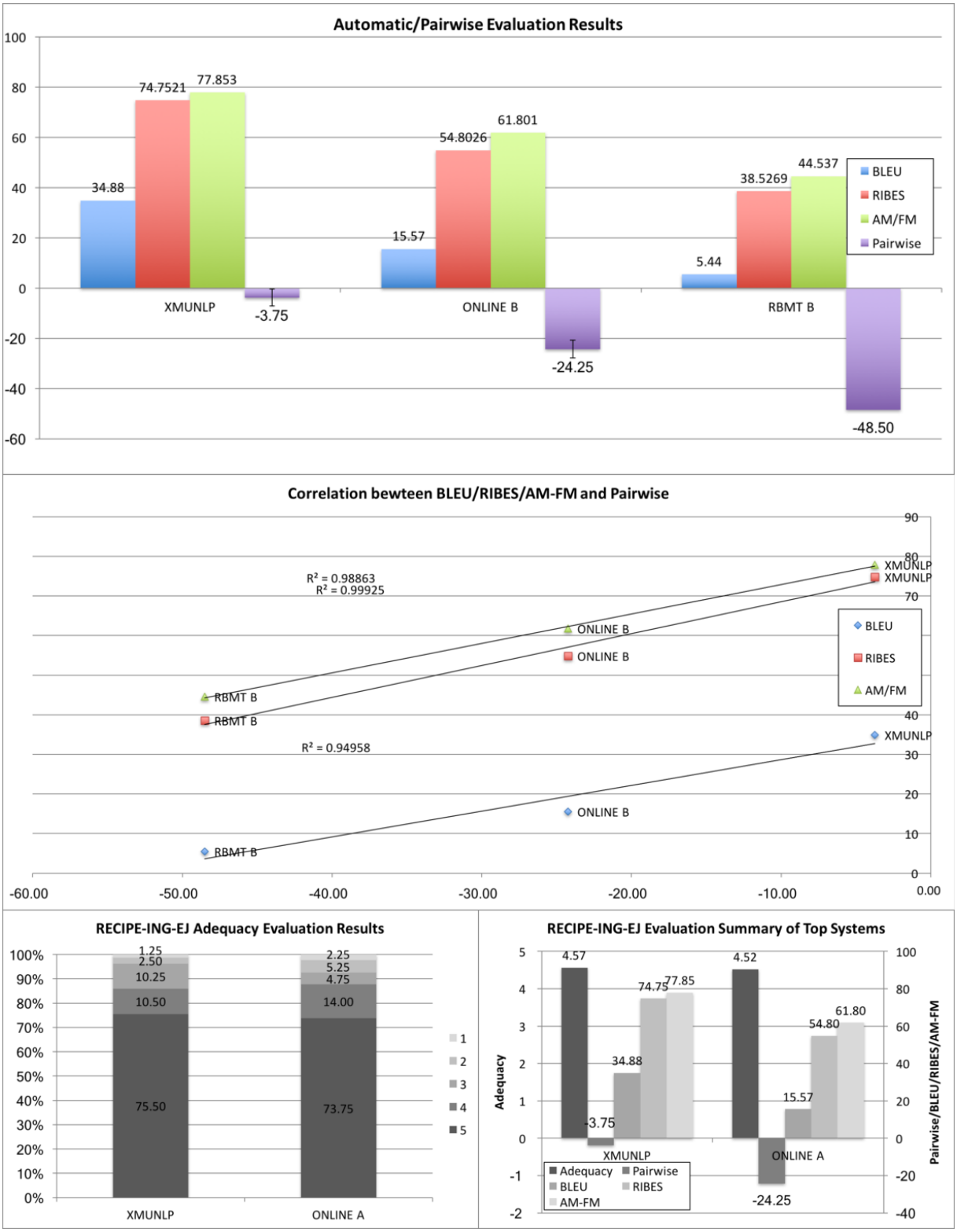

Figure 18: Official evaluation results of RECIPE-ING-EJ.

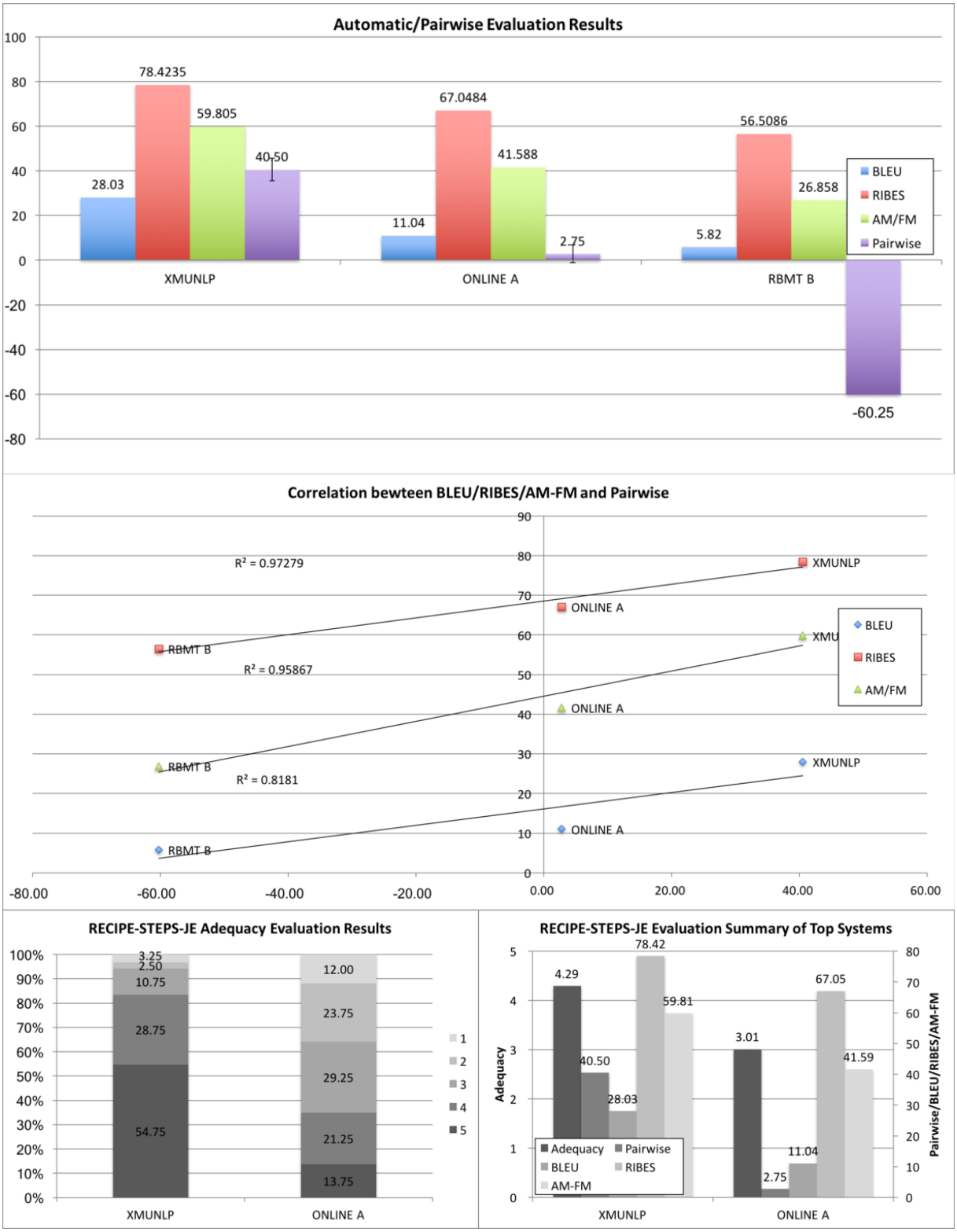

Figure 19: Official evaluation results of RECIPE-STE-JE.

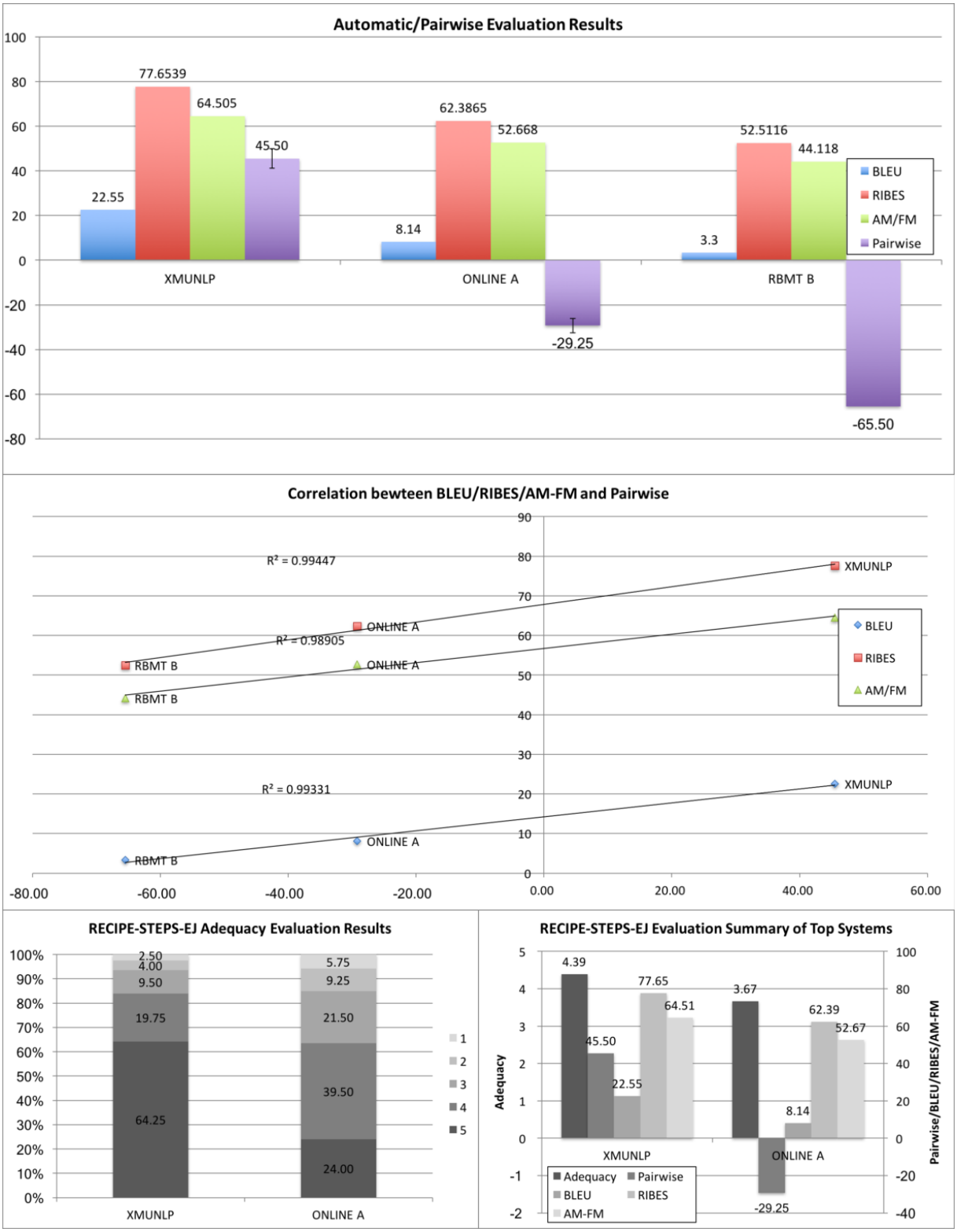

Figure 20: Official evaluation results of RECIPE-STE-EJ.

Subtask	SYSTEM ID	DATA ID	Annotator A average	Annotator A variance	Annotator B average	Annotator B variance	all average	weighted κ	weighted κ
ASPEC-JE	NTT	1681	4.15	0.58	4.13	0.52	4.14	0.29	0.41
	AIAYN	1736	4.16	0.67	4.05	0.75	4.10	0.26	0.42
	Kyoto-U	1717	4.11	0.69	4.09	0.54	4.10	0.26	0.40
	2016 best	1246	3.76	0.68	4.01	0.67	3.89	0.21	0.31
ASPEC-EJ	NTT	1729	4.54	0.56	4.28	0.49	4.41	0.33	0.43
	AIAYN	1737	4.38	0.83	4.21	0.76	4.30	0.36	0.52
	NICT-2	1479	4.43	0.73	4.16	0.69	4.29	0.35	0.48
	Kyoto-U	1731	4.37	0.84	4.15	0.74	4.26	0.39	0.54
	NAIST-NICT	1507	4.36	0.69	4.06	0.57	4.21	0.26	0.36
	2016 best	1172	3.97	0.76	4.07	0.85	4.02	0.35	0.49
ASPEC-JC	NICT-2	1483	4.25	0.73	3.71	0.98	3.98	0.10	0.18
	Kyoto-U	1722	4.25	0.79	3.64	1.07	3.95	0.12	0.23
	AIAYN	1738	4.26	0.69	3.54	1.03	3.90	0.17	0.27
	2016 best	1071	4.00	1.09	3.76	1.14	3.88	0.20	0.36
ASPEC-CJ	NICT-2	1481	4.63	0.47	3.99	0.98	4.31	0.17	0.23
	Kyoto-U	1720	4.62	0.56	3.97	0.94	4.30	0.16	0.22
	AIAYN	1740	4.59	0.61	3.96	1.04	4.27	0.14	0.23
	2016 best	1256	4.25	1.04	3.64	1.23	3.94	0.23	0.34
JPC-JE	JAPIO	1574	4.80	0.26	4.78	0.51	4.79	0.34	0.42
	u-tkb	1472	4.24	1.26	4.08	2.27	4.16	0.43	0.64
	CUNI	1666	4.12	1.49	3.99	2.35	4.05	0.40	0.63
	2016 best	1149	4.09	0.80	4.51	0.58	4.30	0.25	0.39
JPC-EJ	JAPIO	1454	4.74	0.45	4.76	0.38	4.75	0.32	0.48
	EHR	1407	4.64	0.61	4.61	0.65	4.63	0.42	0.60
	u-tkb	1470	4.39	1.07	4.42	0.99	4.40	0.43	0.61
	2016 best	1098	4.03	0.91	4.51	0.57	4.27	0.23	0.41
JPC-JC	u-tkb	1465	3.99	1.12	4.19	0.94	4.09	0.22	0.32
	2016 best	1150	3.49	1.72	3.02	1.75	3.25	0.27	0.51
JPC-CJ	JAPIO	1484	4.41	0.68	4.51	0.64	4.46	0.26	0.34
	EHR	1414	4.27	0.92	4.35	1.03	4.31	0.33	0.48
	u-tkb	1468	3.84	1.16	4.04	1.36	3.94	0.23	0.43
	2016 best	1200	3.61	1.89	3.27	1.76	3.44	0.26	0.52
JPC-KJ	JAPIO	1448	4.82	0.24	4.87	0.11	4.84	0.55	0.55
	EHR	1417	4.76	0.30	4.86	0.23	4.81	0.35	0.47
	2016 best	1209	4.58	0.32	4.66	0.30	4.62	0.33	0.36
IITBC-HE	XMUNLP	1511	3.43	1.64	3.60	1.74	3.51	0.22	0.45
	IITB-MTG	1726	2.14	1.45	2.45	1.87	2.29	0.30	0.51
IITBC-EH	XMUNLP	1576	3.95	1.18	3.76	1.85	3.86	0.17	0.36
	IITB-MTG	1725	2.78	1.74	2.58	1.87	2.68	0.15	0.38
	2016 best	1032	3.20	1.33	3.53	1.19	3.36	0.10	0.16
JIJI-JE	Online A	1523	3.03	1.60	3.28	2.24	3.15	0.15	0.37
	NTT	1599	1.87	1.25	2.23	1.69	2.05	0.26	0.46
	XMUNLP	1442	1.91	1.26	2.19	1.56	2.05	0.24	0.44
JIJI-EJ	Online A	1518	3.31	1.92	3.78	2.06	3.54	0.23	0.50
	NTT	1679	1.78	1.18	2.28	1.97	2.03	0.29	0.52
	XMUNLP	1443	1.72	1.02	2.20	1.70	1.96	0.33	0.51
RECIPE-TTL-JE	XMUNLP	1637	3.90	1.98	3.62	1.57	3.76	0.30	0.56
	Online A	1534	3.52	2.04	3.16	2.07	3.34	0.36	0.60
RECIPE-TTL-EJ	Online B	1533	4.56	0.55	3.84	2.02	4.20	0.25	0.35
	XMUNLP	1636	4.54	0.62	3.62	2.40	4.08	0.26	0.34
RECIPE-ING-JE	XMULNP	1635	4.68	0.65	4.58	0.85	4.63	0.47	0.67
	Online A	1544	4.29	1.44	4.23	1.57	4.26	0.55	0.76
RECIPE-ING-EJ	XMUNLP	1634	4.71	0.43	4.43	1.03	4.57	0.40	0.53
	Online A	1542	4.50	0.95	4.54	0.91	4.52	0.50	0.65
RECIPE-STE-JE	XMUNLP	1632	4.61	0.76	3.98	0.96	4.29	0.13	0.28
	Online A	1551	3.34	1.54	2.69	1.21	3.01	0.14	0.36
RECIPE-STE-EJ	XMUNLP	1633	4.75	0.36	4.04	1.33	4.39	0.12	0.21
	Online A	1549	4.18	0.42	3.16	1.52	3.67	0.11	0.17

Table 9: JPO adequacy evaluation results in detail.

	NTT (1681)	ORGANIZER (1736)	NTT (1616)	Kyoto-U (1733)	NICT-2 (1480)	NICT-2 (1476)	CUNI (1665)	ORGANIZER (1333)	TMU (1703)	TMU (1695)
Kyoto-U (1717)	-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NTT (1681)		$>$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
ORGANIZER (1736)			-	-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NTT (1616)				-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
Kyoto-U (1733)					$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NICT-2 (1480)						-	$\gg$	$\gg$	$\gg$	$\gg$
NICT-2 (1476)							$\gg$	$\gg$	$\gg$	$\gg$
CUNI (1665)								$>$	$\gg$	$\gg$
ORGANIZER (1333)									-	$\gg$
TMU (1703)										$\gg$

Table 10: Statistical significance testing of the ASPEC-JE Pairwise scores.

	NICT-2 (1479)	ORGANIZER (1334)	NTT (1684)	NAIST-NICT (1507)	ORGANIZER (1737)	Kyoto-U (1731)	UT-IIS (1710)	NAIST-NICT (1506)	NICT-2 (1475)	TMU (1709)	TMU (1704)
NTT (1729)	-	-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NICT-2 (1479)		-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
ORGANIZER (1334)			-	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NTT (1684)				$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
NAIST-NICT (1507)					-	-	$>$	$\gg$	$\gg$	$\gg$	$\gg$
ORGANIZER (1737)						-	$>$	$\gg$	$\gg$	$\gg$	$\gg$
Kyoto-U (1731)							$>$	$\gg$	$\gg$	$\gg$	$\gg$
UT-IIS (1710)								$\gg$	$\gg$	$\gg$	$\gg$
NAIST-NICT (1506)									-	$\gg$	$\gg$
NICT-2 (1475)										$\gg$	$\gg$
TMU (1709)											$\gg$

Table 11: Statistical significance testing of the ASPEC-EJ Pairwise scores.

	Kyoto-U (1642)	ORGANIZER (1738)	NICT-2 (1483)	NICT-2 (1478)	ORGANIZER (1336)	TMU (1743)
Kyoto-U (1722)	-	$>$	$\gg$	$\gg$	$\gg$	$\gg$
Kyoto-U (1642)		-	$>$	$\gg$	$\gg$	$\gg$
ORGANIZER (1738)			-	$\gg$	$\gg$	$\gg$
NICT-2 (1483)				$>$	$\gg$	$\gg$
NICT-2 (1478)					$\gg$	$\gg$
ORGANIZER (1336)						$\gg$

	Kyoto-U (1577)	NICT-2 (1481)	ORGANIZER (1740)	NICT-2 (1477)	ORGANIZER (1342)
Kyoto-U (1720)	$\gg$	$\gg$	$\gg$	$\gg$	$\gg$
Kyoto-U (1577)		-	-	-	$\gg$
NICT-2 (1481)			-	-	$\gg$
ORGANIZER (1740)				-	$\gg$
NICT-2 (1477)					$\gg$

Table 12: Statistical significance testing of the ASPEC-JC (left) and ASPEC-CJ (right) Pairwise scores.

	JAPIO (1574)	JAPIO (1578)	CUNI (1666)	u-tkb (1472)
ORGANIZER (1338)	≫	≫	≫	≫
JAPIO (1574)		-	≫	≫
JAPIO (1578)			≫	≫
CUNI (1666)				≫

	EHR (1406)	JAPIO (1454)	u-tkb (1470)	ORGANIZER (1339)	JAPIO (1462)
EHR (1407)	-	≫	≫	≫	≫
EHR (1406)		-	≫	≫	≫
JAPIO (1454)			≫	≫	≫
u-tkb (1470)				-	≫
ORGANIZER (1339)					≫

Table 13: Statistical significance testing of the JPC-JE (left) and JPC-EJ (right) Pairwise scores.

	u-tkb (1465)
ORGANIZER (1340)	≫

	EHR (1414)	EHR (1408)	JAPIO (1447)	u-tkb (1468)	ORGANIZER (1341)
JAPIO (1484)	≫	≫	≫	≫	≫
EHR (1414)		-	≫	≫	≫
EHR (1408)			≫	≫	≫
JAPIO (1447)				≫	≫
u-tkb (1468)					-

Table 14: Statistical significance testing of the JPC-JC (left) and JPC-CJ (right) Pairwise scores.

	JAPIO (1450)	EHR (1417)	EHR (1416)	ORGANIZER (1344)
JAPIO (1448)	-	≫	≫	≫
JAPIO (1450)		≫	≫	≫
EHR (1417)			≫	≫
EHR (1416)				≫

Table 15: Statistical significance testing of the JPC-KJ Pairwise scores.

	IITB-MTG (1726)
XMUNLP (1511)	≫

	IITB-MTG (1725)
XMUNLP (1576)	≫

Table 16: Statistical significance testing of the IITBC-HE (left) and IITBC-EH (right) Pairwise scores.

	ORGANIZER (1526)	NTT (1599)	NTT (1677)	XMUNLP (1442)	ORGANIZER (1396)	NICT-2 (1474)	NICT-2 (1473)	CUNI (1668)
ORGANIZER (1523)	⋙	⋙	⋙	⋙	⋙	⋙	⋙	⋙
ORGANIZER (1526)		⋙	⋙	⋙	⋙	⋙	⋙	⋙
NTT (1599)			⋙	⋙	⋙	⋙	⋙	⋙
NTT (1677)				⋙	⋙	⋙	⋙	⋙
XMUNLP (1442)					⋙	⋙	⋙	⋙
ORGANIZER (1396)						>	⋙	⋙
NICT-2 (1474)							⋙	⋙
NICT-2 (1473)								⋙

	ORGANIZER (1514)	NTT (1679)	NTT (1603)	XMUNLP (1443)	ORGANIZER (1395)
ORGANIZER (1518)	⋙	⋙	⋙	⋙	⋙
ORGANIZER (1514)		⋙	⋙	⋙	⋙
NTT (1679)			⋙	⋙	⋙
NTT (1603)				-	⋙
XMUNLP (1443)					-

Table 17: Statistical significance testing of the JIJI-JE (left) and JIJI-EJ (right) Pairwise scores.

	XMUNLP (1637)	ORGANIZER (1531)
ORGANIZER (1534)	-	⋙
XMUNLP (1637)		⋙

	ORGANIZER (1533)	ORGANIZER (1528)
XMUNLP (1636)	⋙	⋙
ORGANIZER (1533)		⋙

Table 18: Statistical significance testing of the RECIPE-TTL-JE (left) and RECIPE-TTL-EJ (right) Pairwise scores.

	ORGANIZER (1544)	ORGANIZER (1539)
XMUNLP (1635)	⋙	⋙
ORGANIZER (1544)		⋙

	ORGANIZER (1542)	ORGANIZER (1537)
XMUNLP (1634)	⋙	⋙
ORGANIZER (1542)		⋙

Table 19: Statistical significance testing of the RECIPE-ING-JE (left) and RECIPE-ING-EJ (right) Pairwise scores.

	ORGANIZER (1551)	ORGANIZER (1548)
XMUNLP (1632)	⋙	⋙
ORGANIZER (1551)		⋙

	ORGANIZER (1549)	ORGANIZER (1546)
XMUNLP (1633)	⋙	⋙
ORGANIZER (1549)		⋙

Table 20: Statistical significance testing of the RECIPE-STE-JE (left) and RECIPE-STE-EJ (right) Pairwise scores.

ASPEC-JE		
SYSTEM	DATA	κ
Online D	1333	0.230
AIAYN	1736	0.204
Kyoto-U	1717	0.217
Kyoto-U	1733	0.204
TMU	1695	0.188
TMU	1703	0.191
NTT	1616	0.201
NTT	1681	0.173
NICT-2	1476	0.274
NICT-2	1480	0.257
CUNI	1665	0.241
ave.		0.216

ASPEC-EJ		
SYSTEM	DATA	κ
Online A	1334	0.290
AIAYN	1737	0.338
Kyoto-U	1731	0.321
TMU	1704	0.269
TMU	1709	0.260
NTT	1684	0.353
NTT	1729	0.341
NICT-2	1475	0.315
NICT-2	1479	0.395
UT-IIS	1710	0.305
NAIST-NICT	1506	0.301
NAIST-NICT	1507	0.339
ave.		0.319

ASPEC-JC		
SYSTEM	DATA	κ
Online D	1336	0.189
AIAYN	1738	0.183
Kyoto-U	1642	0.159
Kyoto-U	1722	0.128
TMU	1743	0.171
NICT-2	1478	0.222
NICT-2	1483	0.194
ave.		0.178

ASPEC-CJ		
SYSTEM	DATA	κ
Online A	1342	0.215
AIAYN	1740	0.310
Kyoto-U	1577	0.284
Kyoto-U	1720	0.254
NICT-2	1477	0.191
NICT-2	1481	0.279
ave.		0.255

JPC-JE		
SYSTEM	DATA	κ
Online A	1338	0.424
JAPIO	1574	0.280
JAPIO	1578	0.296
CUNI	1666	0.249
u-tkb	1472	0.380
ave.		0.326

JPC-EJ		
SYSTEM	DATA	κ
Online A	1339	0.410
EHR	1406	0.364
EHR	1407	0.385
JAPIO	1454	0.409
JAPIO	1462	0.280
u-tkb	1470	0.349
ave.		0.366

JPC-JC		
SYSTEM	DATA	κ
Online A	1340	0.185
u-tkb	1465	0.176
ave.		0.180

JPC-CJ		
SYSTEM	DATA	κ
Online A	1341	0.194
EHR	1408	0.201
EHR	1414	0.170
JAPIO	1447	0.257
JAPIO	1484	0.247
u-tkb	1468	0.172
ave.		0.207

JPC-KJ		
SYSTEM	DATA	κ
Online A	1344	0.257
EHR	1416	0.413
EHR	1417	0.459
JAPIO	1448	0.224
JAPIO	1450	0.235
ave.		0.317

IITBC-HE		
SYSTEM	DATA	κ
XMUNLP	1511	0.376
IITB-MTG	1726	0.626
ave.		0.501

IITBC-EH		
SYSTEM	DATA	κ
XMUNLP	1576	0.269
IITB-MTG	1725	0.371
ave.		0.320

JIJI-JE		
SYSTEM	DATA	κ
Hiero	1396	0.117
Online A	1523	0.035
RBMT B	1526	0.004
NTT	1599	0.095
NTT	1677	0.077
NICT-2	1473	0.078
NICT-2	1474	0.064
XMUNLP	1442	0.070
CUNI	1668	0.060
ave.		0.067

JIJI-EJ		
SYSTEM	DATA	κ
Hiero	1395	0.104
RBMT A	1514	0.167
Online A	1518	0.179
NTT	1603	0.189
NTT	1679	0.155
XMUNLP	1443	0.151
ave.		0.157

RECIPE-TTL-JE		
SYSTEM	DATA	κ
RBMT B	1531	0.305
Online A	1534	0.333
XMUNLP	1637	0.366
ave.		0.334

RECIPE-TTL-EJ		
SYSTEM	DATA	κ
RBMT B	1528	0.340
Online B	1533	0.356
XMUNLP	1636	0.341
ave.		0.345

RECIPE-STE-JE		
SYSTEM	DATA	κ
RBMT B	1548	0.290
Online A	1551	0.289
XMUNLP	1632	0.261
ave.		0.280

RECIPE-STE-EJ		
SYSTEM	DATA	κ
RBMT B	1546	0.108
Online A	1549	0.138
XMUNLP	1633	0.162
ave.		0.136

RECIPE-ING-JE		
SYSTEM	DATA	κ
RBMT B	1539	0.537
Online B	1544	0.551
XMUNLP	1635	0.614
ave.		0.567

RECIPE-ING-EJ		
SYSTEM	DATA	κ
RBMT B	1537	0.665
Online B	1542	0.515
XMUNLP	1634	0.618
ave.		0.599

Table 21: The Fleiss' kappa values for the pairwise evaluation results.

SYSTEM ID	ID	METHOD	OTHER	BLEU	RIBES	AMFM	Pair
SMT Hiero	2	SMT	NO	18.72	0.651066	0.588880	+7.75
SMT Phrase	6	SMT	NO	18.45	0.645137	0.590950	—
SMT S2T	9	SMT	NO	20.36	0.678253	0.593410	+25.50
Online D (2014)	35	Other	YES	15.08	0.643588	0.564170	+13.75
RBMT E	76	Other	YES	14.82	0.663851	0.561620	—
RBMT F	79	Other	YES	13.86	0.661387	0.556840	—
Online C (2014)	87	Other	YES	10.64	0.624827	0.466480	—
RBMT D (2014)	96	Other	YES	15.29	0.683378	0.551690	+23.00
Online D (2015)	775	Other	YES	16.85	0.676609	0.562270	+0.25
SMT S2T	877	SMT	NO	20.36	0.678253	0.593410	+7.00
RBMT D (2015)	887	Other	YES	15.29	0.683378	0.551690	+16.75
Online C (2015)	892	Other	YES	10.29	0.622564	0.453370	—
Online D (2016)	1042	Other	YES	16.91	0.677412	0.564270	+28.00
Online D (2016/11)	1333	NMT	YES	22.04	0.733483	0.584390	+63.00
AIAYN	1736	NMT	NO	28.06	0.767577	0.595580	+75.25
Kyoto-U 1	1717	NMT	NO	27.53	0.761403	0.585540	+77.75
Kyoto-U 2	1733	NMT	NO	27.66	0.765464	0.591160	+74.50
TMU 1	1695	NMT	NO	21.00	0.725284	0.585710	+56.75
TMU 2	1703	NMT	NO	23.03	0.741175	0.595260	+61.00
NTT 1	1616	NMT	NO	27.43	0.764831	0.597620	+75.00
NTT 2	1681	NMT	NO	28.36	0.768880	0.597860	+77.25
NICT-2 1	1476	NMT	NO	24.79	0.747335	0.574810	+68.75
NICT-2 2	1480	NMT	NO	26.76	0.741329	0.578150	+69.75
CUNI 1	1665	NMT	NO	23.43	0.741699	0.583780	+66.00

Table 22: ASPEC-JE submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
SMT Phrase	5	SMT	NO	27.48	29.80	28.27	0.683735	0.691926	0.695390	0.736380	0.736380	0.736380	—
SMT T2S	12	SMT	NO	31.05	33.44	32.10	0.748883	0.758031	0.760516	0.744370	0.744370	0.744370	+34.25
Online A (2014)	34	Other	YES	19.66	21.63	20.17	0.718019	0.723486	0.725848	0.695420	0.695420	0.695420	+42.50
RBMT B (2014)	66	Other	YES	13.18	14.85	13.48	0.671958	0.680748	0.682683	0.622930	0.622930	0.622930	+0.75
RBMT A	68	Other	YES	12.86	14.43	13.16	0.670167	0.676464	0.678934	0.626940	0.626940	0.626940	—
Online B (2014)	91	Other	YES	17.04	18.67	17.36	0.687797	0.693390	0.698126	0.643070	0.643070	0.643070	—
RBMT C	95	Other	YES	12.19	13.32	12.14	0.668372	0.672645	0.676018	0.594380	0.594380	0.594380	—
SMT Hiero	367	SMT	NO	30.19	32.56	30.94	0.734705	0.746978	0.747722	0.743900	0.743900	0.743900	+31.50
Online A (2015)	774	Other	YES	18.22	19.77	18.46	0.705882	0.713960	0.718150	0.677200	0.677200	0.677200	+34.25
SMT T2S	875	SMT	NO	31.05	33.44	32.10	0.748883	0.758031	0.760516	0.744370	0.744370	0.744370	+30.00
RBMT B (2015)	883	Other	YES	13.18	14.85	13.48	0.671958	0.680748	0.682683	0.622930	0.622930	0.622930	+9.75
Online B (2015)	889	Other	YES	17.80	19.52	18.11	0.693359	0.701966	0.703859	0.646160	0.646160	0.646160	—
Online A (2016)	1041	Other	YES	18.28	19.81	18.51	0.706639	0.715222	0.718559	0.677020	0.677020	0.677020	+49.75
Online A (2016/11)	1334	NMT	YES	26.19	28.22	26.68	0.776787	0.780217	0.782674	0.727040	0.727040	0.727040	+74.00
AIAYN	1737	NMT	NO	40.79	42.55	41.50	0.844896	0.847559	0.851471	0.768630	0.768630	0.768630	+69.75

Table 23: ASPEC-EJ submissions (Organizer)

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
Kyoto-U 1	1731	NMT	NO	38.72	40.65	39.37	0.832472	0.835870	0.839646	0.754220	0.754220	0.754220	+69.75
TMU 1	1704	NMT	NO	32.65	35.05	33.72	0.802262	0.809649	0.811057	0.740620	0.740620	0.740620	+50.75
TMU 2	1709	NMT	NO	34.05	36.69	35.32	0.812926	0.818443	0.821563	0.744890	0.744890	0.744890	+56.50
NTT 1	1684	NMT	NO	39.80	42.27	40.47	0.835806	0.839981	0.844326	0.757740	0.757740	0.757740	+72.25
NTT 2	1729	NMT	NO	40.32	42.80	40.95	0.838594	0.841769	0.846486	0.762170	0.762170	0.762170	+75.75
NICT-2 1	1475	NMT	NO	36.85	38.94	37.87	0.826791	0.834448	0.835255	0.759570	0.759570	0.759570	+62.00
NICT-2 2	1479	NMT	NO	40.17	42.25	41.17	0.842206	0.848170	0.849929	0.765580	0.765580	0.765580	+74.75
UT-IIS 1	1710	NMT	NO	36.26	38.93	37.06	0.827891	0.832054	0.836394	0.746910	0.746910	0.746910	+68.00
NAIST-NICT 1	1506	NMT	NO	36.47	38.54	37.30	0.821989	0.827225	0.830116	0.763310	0.763310	0.763310	+63.50
NAIST-NICT 2	1507	NMT	NO	38.25	40.29	39.05	0.834492	0.839321	0.842337	0.770480	0.770480	0.770480	+70.00

Table 24: ASPEC-EJ submissions (Participants)

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				kytea	stanford (ctb)	stanford (pku)	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	0.745100	0.745100	0.745100	+3.75
SMT Phrase	7	SMT	NO	27.96	28.01	27.68	0.788961	0.790263	0.790937	0.749450	0.749450	0.749450	—
SMT S2T	10	SMT	NO	28.65	28.65	28.35	0.807606	0.809457	0.808417	0.755230	0.755230	0.755230	+14.00
Online D (2014)	37	Other	YES	9.37	8.93	8.84	0.606905	0.606328	0.604149	0.625430	0.625430	0.625430	-14.50
Online C (2014)	216	Other	YES	7.26	7.01	6.72	0.612808	0.613075	0.611563	0.587820	0.587820	0.587820	—
RBMT B (2014)	243	RBMT	NO	17.86	17.75	17.49	0.744818	0.745885	0.743794	0.667960	0.667960	0.667960	-20.00
RBMT C	244	RBMT	NO	9.62	9.96	9.59	0.642278	0.648758	0.645385	0.594900	0.594900	0.594900	—
Online D (2015)	777	Other	YES	10.73	10.33	10.08	0.660484	0.660847	0.660482	0.634090	0.634090	0.634090	-14.75
SMT S2T	881	SMT	NO	28.65	28.65	28.35	0.807606	0.809457	0.808417	0.755230	0.755230	0.755230	+7.75
RBMT B (2015)	886	Other	YES	17.86	17.75	17.49	0.744818	0.745885	0.743794	0.667960	0.667960	0.667960	-11.00
Online C (2015)	891	Other	YES	7.44	7.05	6.75	0.611964	0.615048	0.612158	0.566060	0.566060	0.566060	—
Online D (2016)	1045	Other	YES	11.16	10.72	10.54	0.665185	0.667382	0.666953	0.639440	0.639440	0.639440	-26.00
Online D (2016/11)	1336	NMT	YES	15.94	15.68	15.38	0.728453	0.728270	0.728284	0.673730	0.673730	0.673730	+17.75
AIAYN	1738	NMT	NO	34.97	34.96	34.72	0.850199	0.850052	0.848394	0.787250	0.787250	0.787250	+70.50
Kyoto-U 1	1642	NMT	NO	35.67	35.30	35.40	0.849464	0.848107	0.848318	0.779400	0.779400	0.779400	+71.50
Kyoto-U 2	1722	NMT	NO	35.31	35.37	35.06	0.850103	0.849168	0.847879	0.785420	0.785420	0.785420	+72.50
TMU 1	1743	NMT	NO	22.92	22.86	22.74	0.798681	0.798736	0.797969	0.700030	0.700030	0.700030	+4.25
NICT-2 1	1478	NMT	NO	33.72	33.64	33.60	0.847223	0.846578	0.846158	0.779870	0.779870	0.779870	+67.25
NICT-2 2	1483	NMT	NO	35.23	35.23	35.14	0.852084	0.851893	0.851548	0.785820	0.785820	0.785820	+69.50

Table 25: ASPEC-JC submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
SMT Hiero	4	SMT	NO	35.43	35.91	35.64	0.810406	0.798726	0.807665	0.750950	0.750950	0.750950	+4.75
SMT Phrase	8	SMT	NO	34.65	35.16	34.77	0.772498	0.766384	0.771005	0.753010	0.753010	0.753010	—
SMT T2S	13	SMT	NO	36.52	37.07	36.64	0.825292	0.820490	0.825025	0.754870	0.754870	0.754870	+16.00
Online A (2014)	36	Other	YES	11.63	13.21	11.87	0.595925	0.598172	0.598573	0.658060	0.658060	0.658060	-21.75
Online B (2014)	215	Other	YES	10.48	11.26	10.47	0.600733	0.596006	0.600706	0.636930	0.636930	0.636930	—
RBMT A (2014)	239	RBMT	NO	9.37	9.87	9.35	0.666277	0.652402	0.661730	0.626070	0.626070	0.626070	-37.75
RBMT D	242	RBMT	NO	8.39	8.70	8.30	0.641189	0.626400	0.633319	0.586790	0.586790	0.586790	—
Online A (2015)	776	Other	YES	11.53	12.82	11.68	0.588285	0.590393	0.592887	0.649860	0.649860	0.649860	-19.00
SMT T2S	879	SMT	NO	36.52	37.07	36.64	0.825292	0.820490	0.825025	0.754870	0.754870	0.754870	+17.25
RBMT A (2015)	885	Other	YES	9.37	9.87	9.35	0.666277	0.652402	0.661730	0.626070	0.626070	0.626070	-28.00
Online B (2015)	890	Other	YES	10.41	11.03	10.36	0.597355	0.592841	0.597298	0.628290	0.628290	0.628290	—
Online A (2016)	1043	Other	YES	11.56	12.87	11.69	0.589802	0.589397	0.593361	0.659540	0.659540	0.659540	-51.25
Online A (2016/11)	1342	NMT	YES	18.75	20.64	19.04	0.719022	0.717173	0.720095	0.692820	0.692820	0.692820	+22.50
AIAYN	1740	NMT	NO	46.87	47.30	47.00	0.880815	0.875511	0.880368	0.798110	0.798110	0.798110	+78.50
Kyoto-U 1	1577	NMT	NO	48.43	48.84	48.51	0.883457	0.878964	0.884137	0.799520	0.799520	0.799520	+79.50
Kyoto-U 2	1720	NMT	NO	48.34	48.76	48.40	0.884210	0.880069	0.884745	0.799840	0.799840	0.799840	+82.75
NICT-2 1	1477	NMT	NO	44.26	44.90	44.50	0.871438	0.868359	0.871736	0.788940	0.788940	0.788940	+78.00
NICT-2 2	1481	NMT	NO	46.84	47.51	47.27	0.882356	0.878580	0.882195	0.799680	0.799680	0.799680	+79.00

Table 26: ASPEC-CJ submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU	RIBES	AMFM	Pair
SMT Phrase	977	SMT	NO	30.80	0.730056	0.664830	—
SMT Hiero	979	SMT	NO	32.23	0.763030	0.672500	+8.75
SMT S2T	980	SMT	NO	34.40	0.793483	0.672760	+23.00
Online A (2016)	1035	Other	YES	35.77	0.803661	0.673950	+32.25
Online B (2016)	1051	Other	YES	16.00	0.688004	0.486450	—
RBMT C (2016)	1088	Other	YES	21.00	0.755017	0.519210	—
RBMT A (2016)	1090	Other	YES	21.57	0.750381	0.521230	+23.75
RBMT B (2016)	1095	Other	YES	18.38	0.710992	0.518110	—
Online A (2016/11)	1338	NMT	YES	49.35	0.878342	0.722590	+71.50
JAPIO 1	1574	NMT	YES	49.00	0.878298	0.724710	+68.50
JAPIO 2	1578	NMT	YES	48.08	0.873093	0.715560	+67.00
CUNI 1	1666	SMT	NO	38.29	0.837425	0.681520	+58.00
u-tkb 1	1472	NMT	NO	37.31	0.841136	0.697290	+51.50

Table 27: JPC-JE submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
SMT Phrase	973	SMT	NO	32.36	34.26	32.52	0.728539	0.728281	0.729077	0.711900	0.711900	0.711900	—
SMT Hiero	974	SMT	NO	34.57	36.61	34.79	0.777759	0.778657	0.779049	0.715300	0.715300	0.715300	+21.00
SMT T2S	975	SMT	NO	35.60	37.65	35.82	0.797353	0.796783	0.798025	0.717030	0.717030	0.717030	+30.75
Online A (2016)	1036	Other	YES	36.88	37.89	36.83	0.798168	0.792471	0.796308	0.719110	0.719110	0.719110	+20.00
Online B (2016)	1073	Other	YES	21.57	22.62	21.65	0.743083	0.735203	0.740962	0.659950	0.659950	0.659950	—
RBMT D (2016)	1085	Other	YES	23.02	24.90	23.45	0.761224	0.757341	0.760325	0.647730	0.647730	0.647730	—
RBMT F (2016)	1086	Other	YES	26.64	28.48	26.84	0.773673	0.769244	0.773344	0.675470	0.675470	0.675470	+12.75
RBMT E (2016)	1087	Other	YES	21.35	23.17	21.53	0.743484	0.741985	0.742300	0.646930	0.646930	0.646930	—
Online A (2016/11)	1339	NMT	YES	50.60	51.65	50.83	0.879382	0.877336	0.878316	0.770480	0.770480	0.770480	+48.50
EHR 1	1406	NMT	NO	44.44	45.59	44.15	0.860998	0.858466	0.860659	0.747050	0.747050	0.747050	+58.25
EHR 2	1407	NMT	NO	44.63	45.94	44.53	0.866722	0.864256	0.866205	0.747770	0.747770	0.747770	+60.00
JAPIO 1	1454	NMT	YES	50.27	51.23	50.17	0.886403	0.883481	0.885747	0.776790	0.776790	0.776790	+56.25
JAPIO 2	1462	SMT	YES	51.79	52.23	51.75	0.864038	0.861596	0.862200	0.781150	0.781150	0.781150	+41.00
u-tkb 1	1470	NMT	NO	38.91	41.12	39.11	0.845815	0.846888	0.845551	0.734010	0.734010	0.734010	+49.50

Table 28: JPC-EJ submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				kytea	stanford (ctb)	stanford (pku)	kytea	stanford (ctb)	stanford (pku)	kytea	stanford (ctb)	stanford (pku)	
SMT Phrase	966	SMT	NO	30.60	32.03	31.25	0.787321	0.797888	0.794388	0.710940	0.710940	0.710940	—
SMT Hiero	967	SMT	NO	30.26	31.57	30.91	0.788415	0.799118	0.796685	0.718360	0.718360	0.718360	+4.75
SMT S2T	968	SMT	NO	31.05	32.35	31.70	0.793846	0.802805	0.800848	0.720030	0.720030	0.720030	+4.25
Online A (2016)	1038	Other	YES	23.02	23.57	23.29	0.754241	0.760672	0.760148	0.702350	0.702350	0.702350	-23.00
Online B (2016)	1069	Other	YES	9.42	9.59	8.79	0.642026	0.651070	0.643520	0.527180	0.527180	0.527180	—
RBMT C (2016)	1118	Other	YES	12.35	13.72	13.17	0.688240	0.708681	0.700210	0.475430	0.475430	0.475430	-41.25
Online A (2016/11)	1340	NMT	YES	33.04	33.92	33.34	0.824829	0.829122	0.829067	0.735470	0.735470	0.735470	+32.50
u-tkb 1	1465	NMT	NO	31.80	33.19	32.83	0.819791	0.826055	0.825025	0.706720	0.706720	0.706720	+21.75

Table 29: JPC-JC submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
SMT Hiero	430	SMT	NO	39.22	39.52	39.14	0.806058	0.802059	0.804523	0.729370	0.729370	0.729370	—
SMT Phrase	431	SMT	NO	38.34	38.51	38.22	0.782019	0.778921	0.781456	0.723110	0.723110	0.723110	—
SMT T2S	432	SMT	NO	39.39	39.90	39.39	0.814919	0.811350	0.813595	0.725920	0.725920	0.725920	+20.75
Online A (2015)	647	Other	YES	26.80	27.81	26.89	0.712242	0.707264	0.711273	0.693840	0.693840	0.693840	-7.00
Online B (2015)	648	Other	YES	12.33	12.72	12.44	0.648996	0.641255	0.648742	0.588380	0.588380	0.588380	—
RBMT A (2015)	759	RBMT	NO	10.49	10.72	10.35	0.674060	0.664098	0.667349	0.557130	0.557130	0.557130	-39.25
RBMT B	760	RBMT	NO	7.94	8.07	7.73	0.596200	0.581837	0.586941	0.502100	0.502100	0.502100	—
Online A (2016)	1040	Other	YES	26.99	27.91	27.02	0.707739	0.702718	0.706707	0.693720	0.693720	0.693720	-19.75
Online A (2016/11)	1341	NMT	YES	42.66	43.76	42.95	0.845858	0.844918	0.845794	0.747240	0.747240	0.747240	+54.25
EHR 1	1408	NMT	NO	47.08	47.44	46.83	0.859070	0.856376	0.858888	0.756350	0.756350	0.756350	+68.25
EHR 2	1414	NMT	NO	46.52	47.17	46.35	0.859619	0.856784	0.858353	0.761370	0.761370	0.761370	+69.75
JAPIO 1	1447	SMT	YES	50.52	51.25	50.57	0.847793	0.843774	0.846081	0.774660	0.774660	0.774660	+60.50
JAPIO 2	1484	NMT	YES	50.06	50.51	50.00	0.875398	0.873390	0.874822	0.779420	0.779420	0.779420	+80.25
u-tkb 1	1468	NMT	NO	38.79	40.47	38.99	0.832144	0.833610	0.831209	0.729580	0.729580	0.729580	+55.50

Table 30: JPC-CJ submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
SMT Phrase	438	SMT	NO	69.22	70.36	69.73	0.941302	0.939729	0.940756	0.856220	0.856220	0.856220	—
SMT Hiero	439	SMT	NO	67.41	68.65	68.00	0.937162	0.935903	0.936570	0.850560	0.850560	0.850560	+2.75
Online B (2015)	651	Other	YES	36.41	38.72	37.01	0.851745	0.852263	0.851945	0.728750	0.728750	0.728750	—
Online A (2015)	652	Other	YES	55.05	56.84	55.46	0.909152	0.909385	0.908838	0.800460	0.800460	0.800460	+38.75
RBMT A (2015)	653	Other	YES	42.00	43.97	42.45	0.876396	0.873734	0.875146	0.712020	0.712020	0.712020	-7.25
RBMT B	654	Other	YES	34.74	37.51	35.54	0.845712	0.849014	0.846228	0.643150	0.643150	0.643150	—
Online A (2015)	963	Other	YES	55.05	56.84	55.46	0.909152	0.909385	0.908838	0.800610	0.800610	0.800610	—
RBMT A (2015)	964	Other	YES	42.00	43.97	42.45	0.876396	0.873734	0.875146	0.712700	0.712700	0.712700	—
Online A (2016)	1039	Other	YES	54.78	56.68	55.14	0.907320	0.907652	0.906743	0.798750	0.798750	0.798750	+8.00
Online A (2016/11)	1344	NMT	NO	44.42	45.14	44.72	0.857642	0.854158	0.857083	0.783850	0.783850	0.783850	-55.75
EHR 1	1416	NMT	NO	71.52	72.34	71.82	0.944516	0.942940	0.944219	0.866060	0.866060	0.866060	+6.25
EHR 2	1417	NMT	NO	71.36	72.26	71.65	0.946126	0.944812	0.945888	0.871110	0.871110	0.871110	+11.25
JAPIO 1	1448	SMT	YES	73.00	73.71	73.23	0.946880	0.945754	0.946645	0.872510	0.872510	0.872510	+48.75
JAPIO 2	1450	SMT	YES	73.00	73.73	73.25	0.946985	0.945841	0.946745	0.873200	0.873200	0.873200	+48.50

Table 31: JPC-KJ submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU	RIBES	AMFM	Pair
Online A (2016)	1031	Other	YES	21.37	0.714537	0.621100	+44.75
Online B (2016)	1048	Other	YES	15.58	0.683214	0.590520	+14.00
SMT Phrase	1054	SMT	NO	10.32	0.638090	0.574850	0.00
XMUNLP 1	1511	NMT	NO	22.44	0.750921	0.629530	+68.25
IITB-MTG 1	1726	NMT	NO	11.55	0.682902	0.557040	+21.00

Table 32: IITB-HE submissions

SYSTEM ID	ID	METHOD	OTHER RESOURCES	BLEU	RIBES	AMFM	Pair
Online A (2016)	1032	Other	YES	18.720000	0.716788	0.670660	+57.25
Online B (2016)	1047	Other	YES	16.970000	0.691298	0.668450	+42.50
SMT Phrase	1252	SMT	NO	10.790000	0.651166	0.660860	—
XMUNLP 1	1576	NMT	NO	21.390000	0.749660	0.688770	+64.50
IITB-MTG 1	1725	NMT	NO	12.230000	0.688606	0.624780	+28.75

Table 33: IITB-EH submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU	RIBES	AMFM	Pair
SMT Phrase	1394	SMT	NO	15.11	0.554550	0.475740	—
SMT Hiero	1396	SMT	NO	15.67	0.558225	0.470610	+10.25
SMT S2T	1398	SMT	NO	14.54	0.556728	0.477170	—
ONLINE-A 1	1523	NMT	NO	8.19	0.529844	0.450850	+70.00
RBMT-A	1525	RBMT	NO	4.36	0.472312	0.391050	—
RBMT-B	1526	RBMT	NO	4.67	0.475760	0.385600	+51.75
NTT 1	1599	NMT	NO	19.44	0.638841	0.476200	+32.00
NTT 2	1677	NMT	NO	20.90	0.648931	0.474360	+26.75
NICT-2 1	1473	NMT	NO	16.52	0.642379	0.459000	+0.25
NICT-2 2	1474	NMT	NO	18.19	0.632638	0.453420	+7.25
XMUNLP 1	1442	NMT	NO	17.95	0.637059	0.465710	+20.75
CUNI 1	1668	SMT	NO	10.67	0.564797	0.432700	-24.00

Table 34: JIJI-JE submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
SMT Phrase	1393	SMT	NO	15.77	16.65	15.76	0.580284	0.584892	0.585437	0.545240	0.545240	0.545240	—
SMT Hiero	1395	SMT	NO	16.22	16.95	16.22	0.594923	0.601505	0.602516	0.550260	0.550260	0.550260	+10.25
SMT T2S	1397	SMT	NO	14.95	15.38	14.79	0.594072	0.597791	0.599530	0.530370	0.530370	0.530370	—
RBMT-A	1514	RBMT	NO	5.31	6.68	5.69	0.505227	0.515050	0.513580	0.473940	0.473940	0.473940	+31.25
RBMT-B	1515	RBMT	NO	4.72	5.98	4.97	0.518416	0.531603	0.532079	0.487320	0.487320	0.487320	—
ONLINE-A 1	1518	NMT	NO	11.29	13.12	11.84	0.597473	0.605532	0.603374	0.533120	0.533120	0.533120	+69.75
NTT 1	1603	NMT	NO	19.13	20.47	19.41	0.668517	0.670920	0.676594	0.536970	0.536970	0.536970	+14.50
NTT 2	1679	NMT	NO	20.37	21.82	20.68	0.680598	0.684048	0.688863	0.537800	0.537800	0.537800	+17.75
XMUNLP 1	1443	NMT	NO	19.61	20.72	20.14	0.684120	0.688497	0.691056	0.546360	0.546360	0.546360	+11.75

Table 35: JIJI-EJ submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU	RIBES	AMFM	Pair
RBMT-A	1538	RBMT	NO	11.14	0.484800	0.613950	—
RBMT-B	1539	RBMT	NO	12.52	0.520800	0.607190	-24.75
ONLINE-B 1	1544	NMT	NO	20.33	0.563419	0.656630	-15.50
SMT Phrase	1571	SMT	NO	44.42	0.830105	0.859040	—
XMUNLP 1	1635	NMT	NO	46.98	0.831261	0.854970	+3.50

Table 36: RECIPEING-JE submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
RBMT-A	1536	RBMT	NO	4.52	4.74	4.29	0.365913	0.371584	0.355798	0.426550	0.426550	0.426550	—
RBMT-B	1537	RBMT	NO	5.44	4.95	5.07	0.385269	0.363344	0.372793	0.445370	0.445370	0.445370	-48.50
ONLINE-B 1	1542	NMT	NO	15.57	14.89	14.85	0.548026	0.544581	0.537147	0.618010	0.618010	0.618010	-24.25
SMT Phrase	1570	SMT	NO	31.39	30.61	29.60	0.749305	0.740283	0.741249	0.775770	0.775770	0.775770	—
XMUNLP 1	1634	NMT	NO	34.88	34.26	33.19	0.747521	0.742770	0.739909	0.778530	0.778530	0.778530	-3.75

Table 37: RECIPEING-EJ submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU	RIBES	AMFM	Pair
RBMT-A	1547	RBMT	NO	5.37	0.546642	0.315930	—
RBMT-B	1548	RBMT	NO	5.82	0.565086	0.268580	-60.25
ONLINE-A 1	1551	NMT	NO	11.04	0.670484	0.415880	+2.75
SMT Phrase	1569	SMT	NO	22.84	0.705506	0.595290	—
XMUNLP 1	1632	NMT	NO	28.03	0.784235	0.598050	+40.50

Table 38: RECIPESTE-JE submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
RBMT-A	1545	RBMT	NO	3.06	4.55	3.43	0.518467	0.533326	0.521713	0.368710	0.368710	0.368710	—
RBMT-B	1546	RBMT	NO	3.30	5.19	4.00	0.525116	0.532919	0.529113	0.441180	0.441180	0.441180	-65.50
ONLINE-A 1	1549	NMT	NO	8.14	11.23	8.97	0.623865	0.635888	0.629012	0.526680	0.526680	0.526680	-29.25
SMT Phrase	1568	SMT	NO	17.60	21.43	18.53	0.694179	0.698499	0.695400	0.626610	0.626610	0.626610	—
XMUNLP 1	1633	NMT	NO	22.55	26.87	24.00	0.776539	0.776469	0.775689	0.645050	0.645050	0.645050	+45.50

Table 39: RECIPESTE-EJ submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU	RIBES	AMFM	Pair
RBMT-A	1530	RBMT	NO	0.53	0.086378	0.433380	—
RBMT-B	1531	RBMT	NO	0.56	0.100899	0.445400	-25.75
ONLINE-A 1	1534	NMT	NO	2.19	0.199338	0.509470	+10.25
SMT Phrase	1567	SMT	NO	9.72	0.451707	0.571230	—
XMUNLP 1	1637	NMT	NO	15.57	0.526993	0.542690	+10.25

Table 40: RECIPETTL-JE submissions

SYSTEM ID	ID	METHOD	OTHER	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
RBMT-A	1527	RBMT	NO	0.00	0.00	0.00	0.128134	0.137884	0.122371	0.187540	0.187540	0.187540	—
RBMT-B	1528	RBMT	NO	2.57	2.30	2.08	0.299133	0.300578	0.270482	0.402830	0.402830	0.402830	-50.25
ONLINE-B 1	1533	NMT	NO	16.16	15.85	15.40	0.573771	0.559142	0.532130	0.590440	0.590440	0.590440	+3.75
SMT Phrase	1566	SMT	NO	17.16	16.23	16.57	0.600503	0.576617	0.548811	0.571650	0.571650	0.571650	—
XMUNLP 1	1636	NMT	NO	19.41	18.87	18.78	0.592087	0.573466	0.558997	0.584980	0.584980	0.584980	+23.75

Table 41: RECIPETTL-EJ submissions

References

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(3):472–482.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213 – 220.

Fabien Cromieres, Raj Dabre, Toshiaki Nakazawa, and Sadao Kurohashi. 2017. Kyoto university participation to wat 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 146–153, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Terumasa Ehara. 2017. Smt reranked nmt. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 119–126, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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. 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.

Kenji Imamura and Eiichiro Sumita. 2017. Ensemble and reranking: Using multiple models in the nict-2 neural machine translation system at wat2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 127–134, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Satoshi Kinoshita, Tadaaki Oshio, and Tomoharu Mitsuhashi. 2017. Comparison of smt and nmt trained with large patent corpora: Japio at wat2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 140–145, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tom Kocmi, Dušan Variš, and Ondřej Bojar. 2017. Cuni nmt system for wat 2017 translation tasks. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 154–159, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

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.

Zi Long, Ryuichiro Kimura, Takehito Utsuro, Tomoharu Mitsuhashi, and Mikio Yamamoto. 2017. Patent nmt integrated with large vocabulary phrase translation by smt at wat 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 110–118, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yukio Matsumura and Mamoru Komachi. 2017. Tokyo metropolitan university neural machine translation system for wat 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 160–166, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. Ntt neural machine translation systems at wat 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 89–94, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Toshiaki Nakazawa, Chenchen Ding, Hideya MINO, Isao Goto, Graham Neubig, and Sadao

Kurohashi. 2016. Overview of the 3rd workshop on asian translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 1–46, Osaka, Japan. The COLING 2016 Organizing Committee.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st Workshop on Asian Translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*, pages 1–19, Tokyo, Japan.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 1–28, Kyoto, Japan.

Masato Neishi, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 99–109, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Yusuke Oda, Katsuhito Sudoh, Satoshi Nakamura, Masao Utiyama, and Eiichiro Sumita. 2017. A simple and strong baseline: Naist-nict neural machine translation system for wat2017 english-japanese translation task. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 135–139, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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. Association for Computational Linguistics.

Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2017. Comparing recurrent and convolutional architectures for english-hindi neural machine translation. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 167–170, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Boli Wang, Zhixing Tan, Jinming Hu, Yidong Chen, and xiaodong shi. 2017. Xmu neural machine translation systems for wat 2017. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 95–98, Taipei, Taiwan. Asian Federation of Natural Language Processing.