# Overview of the 3rd Workshop on Asian Translation

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

**Chenchen Ding** and **Hideya Mino**
National Institute of Information and Communications Technology
{chenchen.ding, hideya.mino}@nict.go.jp

**Isao Goto**
NHK
goto.i-es@nhk.or.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 3rd workshop on Asian translation (WAT2016) including J↔E, J↔C scientific paper translation subtasks, C↔J, K↔J, E↔J patent translation subtasks, I↔E newswire subtasks and H↔E, H↔J mixed domain subtasks. For the WAT2016, 15 institutions participated in the shared tasks. About 500 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) and WAT2015 (Nakazawa et al., 2015), WAT2016 brings together machine translation researchers and users to try, evaluate, share and discuss brand-new ideas of machine translation. We are working toward the practical use of machine translation among all Asian countries.

For the 3rd WAT, we adopt new translation subtasks with English-Japanese patent description, Indonesian-English news description and Hindi-English and Hindi-Japanese mixed domain corpus in addition to the subtasks that were conducted in WAT2015. Furthermore, we invited research papers on topics related to the machine translation, especially for Asian languages. The submissions of the research papers were peer reviewed by at least 2 program committee members and the program committee accepted 7 papers that cover wide variety of topics such as neural machine translation, simultaneous interpretation, southeast Asian languages and so on.

WAT is unique for the following reasons:

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

- Domain and language pairs
  WAT is the world's first workshop that uses scientific papers as the domain, and Chinese ↔ Japanese, Korean ↔ Japanese and Indonesian ↔ English as 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 and manually. For human evaluation, WAT uses pairwise evaluation as the first-stage evaluation. Also, JPO adequacy evaluation is conducted for the selected submissions according to the pairwise evaluation results.





*Proceedings of the 3rd Workshop on Asian Translation*,
pages 1–46, Osaka, Japan, December 11-17 2016.

| 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 for ASPEC.

## 2 Dataset

WAT uses the Asian Scientific Paper Excerpt Corpus (ASPEC) [1], JPO Patent Corpus (JPC) [2], BPPT Corpus [3] and IIT Bombay English-Hindi Corpus (IITB Corpus) [4] as the dataset.

### 2.1 ASPEC

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 for each corpus are described in Table1.

#### 2.1.1 ASPEC-JE

The training data for ASPEC-JE was constructed by the NICT from approximately 2 million Japanese-English scientific paper abstracts owned by the JST. Because the abstracts are comparable corpora, the sentence correspondences are found automatically using the method from (Utiyama and Isahara, 2007). Each sentence pair is accompanied by a similarity score and the field symbol. The similarity scores are calculated by the method from (Utiyama and Isahara, 2007). The field symbols are single letters A-Z and show the scientific field for each document[5]. The correspondence between the symbols and field names, along with the frequency and occurrence ratios for the training data, are given in the README file from ASPEC-JE.

The development, development-test and test data were extracted from parallel sentences from the 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.

#### 2.1.2 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 with 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.

### 2.2 JPC

JPC was constructed by the Japan Patent Office (JPO). It consists of a Chinese-Japanese patent description corpus (JPC-CJ), Korean-Japanese patent description corpus (JPC-KJ) and English-Japanese patent description corpus (JPC-EJ) with four sections, which are Chemistry, Electricity, Mechanical engineering, and Physics, based on International Patent Classification (IPC). Each corpus is separated into

[1]http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[2]http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html
[3]http://orchid.kuee.kyoto-u.ac.jp/WAT/bppt-corpus/index.html
[4]http://www.cfilt.iitb.ac.in/iitb_parallel/index.html
[5]http://opac.jst.go.jp/bunrui/index.html

| LangPair | Train | Dev | DevTest | Test |
|---|---|---|---|---|
| JPC-CJ | 1,000,000 | 2,000 | 2,000 | 2,000 |
| JPC-KJ | 1,000,000 | 2,000 | 2,000 | 2,000 |
| JPC-EJ | 1,000,000 | 2,000 | 2,000 | 2,000 |

Table 2: Statistics for JPC.

| LangPair | Train | Dev | DevTest | Test |
|---|---|---|---|---|
| BPPT-IE | 50,000 | 400 | 400 | 400 |

Table 3: Statistics for BPPT Corpus.

training, development, development-test and test data, which are sentence pairs. This corpus was used for patent subtasks C↔J, K↔J and E↔J. The statistics for each corpus are described in Table2.

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 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 to be 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 BPPT Corpus

BPPT Corpus was constructed by Badan Pengkajian dan Penerapan Teknologi (BPPT). This corpus consists of a Indonesian-English news corpus (BPPT-IE) with five sections, which are Finance, International, Science and Technology, National, and Sports. These data come from Antara News Agency. This corpus was used for newswire subtasks I↔E. The statistics for each corpus are described in Table3.

### 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 existing sources and corpora developed at the Center for Indian Language Technology, IIT Bombay over the years. This corpus was 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 openly available corpora (IITB-JH) [6]. Most sentence pairs in IITB-JH come from the Bible corpus. The statistics for each corpus are described in Table4.

## 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 for human evaluation. A phrase-based statistical machine translation (SMT) system was adopted as the specific baseline system at WAT 2016, which is the same system as that at WAT 2014 and WAT 2015.

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

[6] http://lotus.kuee.kyoto-u.ac.jp/WAT/Hindi-corpus/WAT2016-Ja-Hi.zip

| LangPair | Train | Dev | Test | Monolingual Corpus (Hindi) |
|---|---|---|---|---|
| IITB-EH | 1,492,827 | 520 | 2,507 | 45,075,279 |
| IITB-JH | 152,692 | 1,566 | 2,000 | - |

Table 4: Statistics for IITB Corpus.

SMT system, a tree-to-string syntax-based SMT system, seven 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 for translating using the systems were published on the WAT web page[7]. 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 5.

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.

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

| | | | ASPEC | | | | JPC | | | | | | IITB | | BPPT | | pivot | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| System ID | System | Type | JE | EJ | JC | CJ | JE | EJ | JC | CJ | JK | KJ | HE | EH | IE | EI | HJ | JH |
| 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 | | | ✓ | ✓ | | | | ✓ | | | | | | | | |
| RBMT X | J Soul 9 (Commercial system) | RBMT | | | | | | | | | ✓ | ✓ | | | | | | |
| RBMT X | Korai 2011 (Commercial system) | RBMT | | | | | | | | | ✓ | ✓ | | | | | | |
| Online X | Google translate (July and August, 2016 or August, 2015) | (SMT) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Online X | Bing translator (July and August, 2016 or August and September, 2015) | (SMT) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 5: 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[8] for Japanese segmentation.
- Stanford Word Segmenter version 2014-01-04[9] (Chinese Penn Treebank (CTB) model) for Chinese segmentation.
- The Moses toolkit for English and Indonesian tokenization.
- Mecab-ko[10] for Korean segmentation.
- Indic NLP Library[11] 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.

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

### 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 calculated automatic evaluation scores for the translation results by applying 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* distributed with the Moses toolkit (Koehn et al., 2007); RIBES scores were calculated using *RIBES.py* version 1.02.4 [12]; AMFM scores were calculated using scripts created by technical collaborators of WAT2016. All scores for each task were calculated using one 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 [13] and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0 [14]. For Chinese segmentation we used two different tools: 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 [15] (Tseng, 2005). For Korean segmentation we used mecab-ko [16]. For English and Indonesian segmentations we used tokenizer.perl [17] in the Moses toolkit. For Hindi segmentation we used Indic NLP Library [18].

Detailed procedures for the automatic evaluation are shown on the WAT2016 evaluation web page [19].

### 4.2 Automatic Evaluation System

The participants submit translation results via an automatic evaluation system deployed on the WAT2016 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 ($J \leftrightarrow E$, $J \leftrightarrow C$);
  - Patents subtask ($C \leftrightarrow J$, $K \leftrightarrow J$, $E \leftrightarrow J$);
  - Newswire subtask ($I \leftrightarrow E$)
  - Mixed domain subtask ($H \leftrightarrow E$, $H \leftrightarrow J$)
- Method (SMT, RBMT, SMT and RBMT, EBMT, NMT, Other);

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

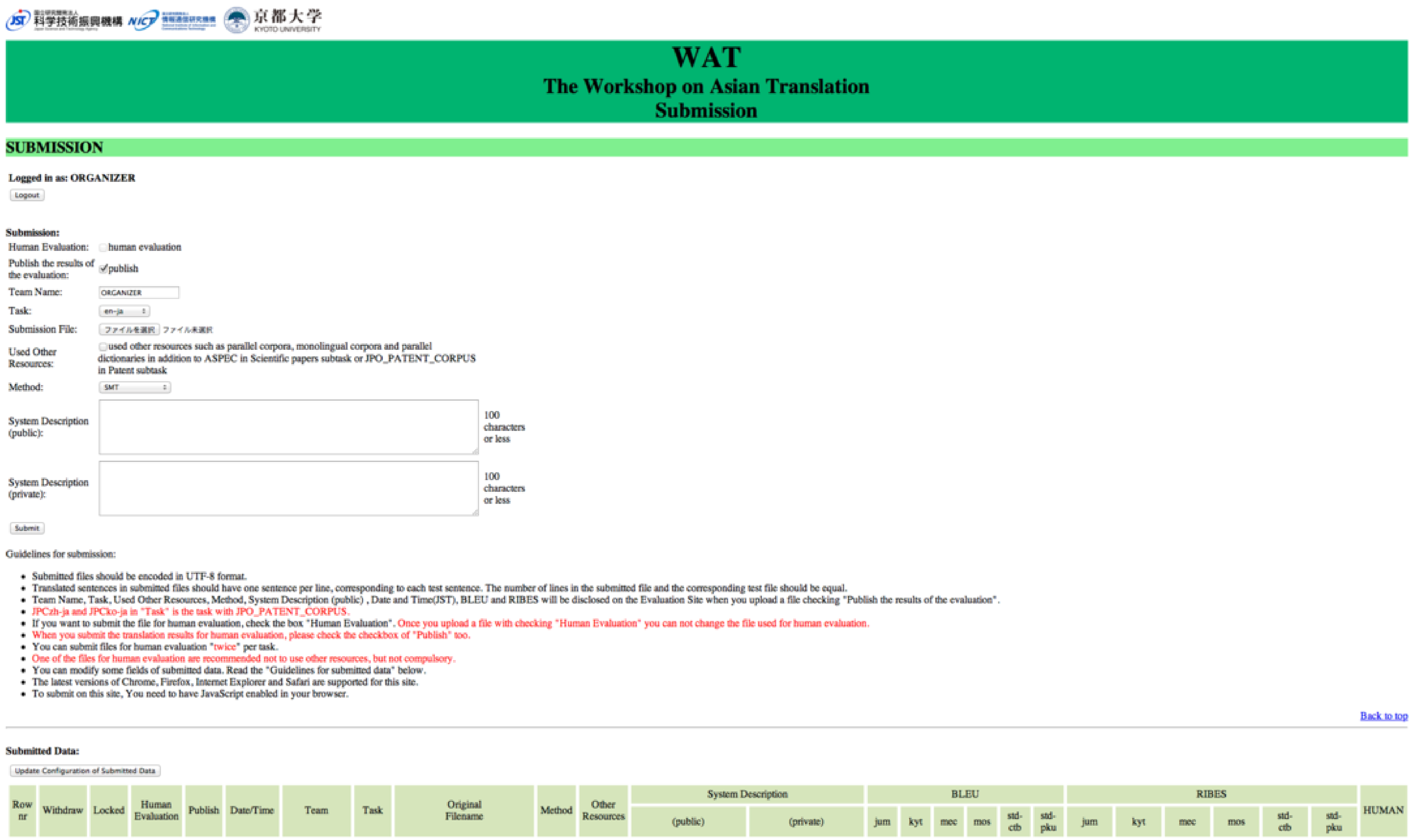


Figure 1: The submission web page for participants

- Use of other resources in addition to ASPEC / JPC / BPPT Corpus / IITB Corpus;
- Permission to publish the automatic evaluation scores on the WAT2016 web page.

The server for the system stores all submitted information, including translation results and scores, although participants can confirm only the information that they uploaded. Information about translation results that participants permit to be published is disclosed on the web page. In addition to submitting translation results for automatic evaluation, participants submit the results for human evaluation using the same web interface. This automatic evaluation system will remain available even after WAT2016. Anybody can register to use the system on the registration web page [20].

## 5 Human Evaluation

In WAT2016, 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[21].

#### 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.

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

[21] It was called HUMAN score in WAT2014 and Crowd score in WAT2015.

| | |
|---|---|
| 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 6: The JPO adequacy criterion

#### 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.

#### 5.2.2 Evaluation Criterion

Table 6 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 7 shows the list of participants for WAT2016. This includes not only Japanese organizations, but also some organizations from outside Japan. 15 teams submitted one or more translation results to the automatic evaluation server or human evaluation.

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

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

| Team ID | Organization | ASPEC | | | | JPC | | | | | | BPPT | | IITBC | | pivot | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | JE | EJ | JC | CJ | JE | EJ | JC | CJ | JK | KJ | IE | EI | HE | EH | HJ | JH |
| NAIST (Neubig, 2016) | Nara Institute of Science and Technology | ✓ | | | | | | | | | | | | | | | |
| Kyoto-U (Cromieres et al., 2016) | Kyoto University | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | |
| TMU (Yamagishi et al., 2016) | Tokyo Metropolitan University | ✓ | | | | | | | | | | | | | | | |
| bjtu_nlp (Li et al., 2016) | Beijing Jiaotong University | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | |
| Sense (Tan, 2016) | Saarland University | | | | | | | | | | | ✓ | ✓ | | | | |
| NICT-2 (Imamura and Sumita, 2016) | National Institute of Information and Communication Technology | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | |
| WASUIPS (Yang and Lepage, 2016) | Waseda University | | | | | | | | ✓ | | | | | | | | |
| EHR (Ehara, 2016) | Ehara NLP Research Laboratory | | ✓ | | ✓ | | | | ✓ | | ✓ | | | | ✓ | ✓ | |
| ntt (Sudoh and Nagata, 2016) | NTT Communication Science Laboratories | | | | | | | | ✓ | | | | | | | | |
| TOKYOMT (Shu and Miura, 2016) | Weblio, Inc. | | ✓ | | | | | | | | | | | | | | |
| IITB-EN-ID (Singh et al., 2016) | Indian Institute of Technology Bombay | | | | | | | | | | | ✓ | ✓ | | | | |
| JAPIO (Kinoshita et al., 2016) | Japan Patent Information Organization | | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | | | | | |
| IITP-MT (Sen et al., 2016) | Indian Institute of Technology Patna | | | | | | | | | | | | | | ✓ | | |
| UT-KAY (Hashimoto et al., 2016) | University of Tokyo | | | | ✓ | | | | | | | | | | | | |
| UT-AKY (Eriguchi et al., 2016) | University of Tokyo | | ✓ | | | | | | | | | | | | | | |

Table 7: List of participants who submitted translation results to WAT2016 and their participation in each subtasks.

## 7 Evaluation Results

In this section, the evaluation results for WAT2016 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2016 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 BPPT subtasks and Figures 13 and 14 show those of IITB 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 8. The weights for the weighted $\kappa$ (Cohen, 1968) is defined as $|Evaluation1 - Evaluation2|/4$.

From the evaluation results, the following can be observed:

- Neural network based translation models work very well also for Asian languages.
- None of the automatic evaluation measures perfectly correlate to the human evaluation result (JPO adequacy).
- The JPO adequacy evaluation result of IITB E→H shows an interesting tendency: the system which achieved the best average score has the lowest ratio of the perfect translations and vice versa.

### 7.2 Statistical Significance Testing of Pairwise Evaluation between Submissions

Tables 9, 10, 11 and 12 show the results of statistical significance testing of ASPEC subtasks, Tables 13, 14, 15, 16 and 17 show those of JPC subtasks, 18 shows those of BPPT subtasks and 19 shows those of JPC 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' $\kappa$ (Fleiss and others, 1971) values. The results are shown in Table 20. We can see that the $\kappa$ values are larger for X $\rightarrow$ J translations than for J $\rightarrow$ 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.

### 7.3 Chronological Evaluation

Figure 15 shows the chronological evaluation results of 4 subtasks of ASPEC and 2 subtasks of JPC. The Kyoto-U (2016) (Cromieres et al., 2016), ntt (2016) (Sudoh and Nagata, 2016) and naver (2015) (Lee et al., 2015) are NMT systems, the NAIST (2015) (Neubig et al., 2015) is a forest-to-string SMT system, Kyoto-U (2015) (Richardson et al., 2015) is a dependency tree-to-tree EBMT system and JAPIO (2016) (Kinoshita et al., 2016) system is a phrase-based SMT system.

What we can see is that in ASPEC-JE and EJ, the overall quality is improved from the last year, but the ratio of grade 5 is decreased. This is because the NMT systems can output much fluent translations

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

but the adequacy is worse. As for ASPEC-JC and CJ, the quality is very much improved. Literatures (Junczys-Dowmunt et al., 2016) say that Chinese receives the biggest benefits from NMT.

The translation quality of JPC-CJ does not so much varied from the last year, but that of JPC-KJ is much worse. Unfortunately, the best systems participated last year did not participate this year, so it is not directly comparable.

## 8 Submitted Data

The number of published automatic evaluation results for the 15 teams exceeded 400 before the start of WAT2016, and 63 translation results for pairwise evaluation were submitted by 14 teams. Furthermore, we selected maximum 3 translation results from each subtask 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 WAT2016. We had 15 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 include newspaper translation tasks for Japanese, Chinese and English where the context information is important to achieve high translation quality, so it is a challenging task.

We would also be very happy to include other languages if the resources are available.

## Appendix A Submissions

Tables 21 to 36 summarize all the submissions listed in the automatic evaluation server at the time of the WAT2016 workshop (12th, December, 2016). The OTHER RESOURCES column shows the use of resources such as parallel corpora, monolingual corpora and parallel dictionaries in addition to ASPEC, JPC, BPPT Corpus, IITB Corpus.

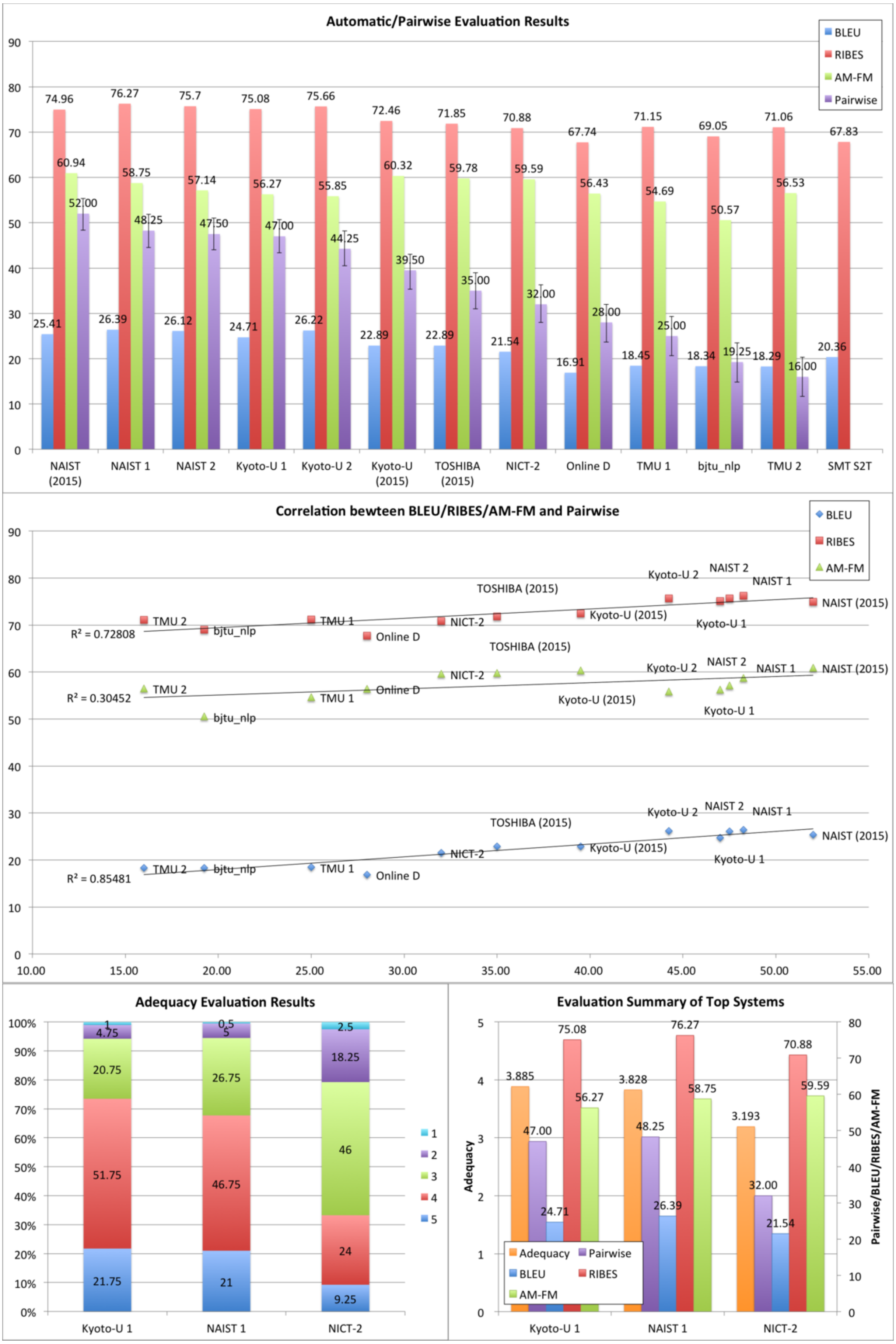


Figure 2: Official evaluation results of ASPEC-JE.


Figure 3: Official evaluation results of ASPEC-EJ.


Figure 4: Official evaluation results of ASPEC-JC.

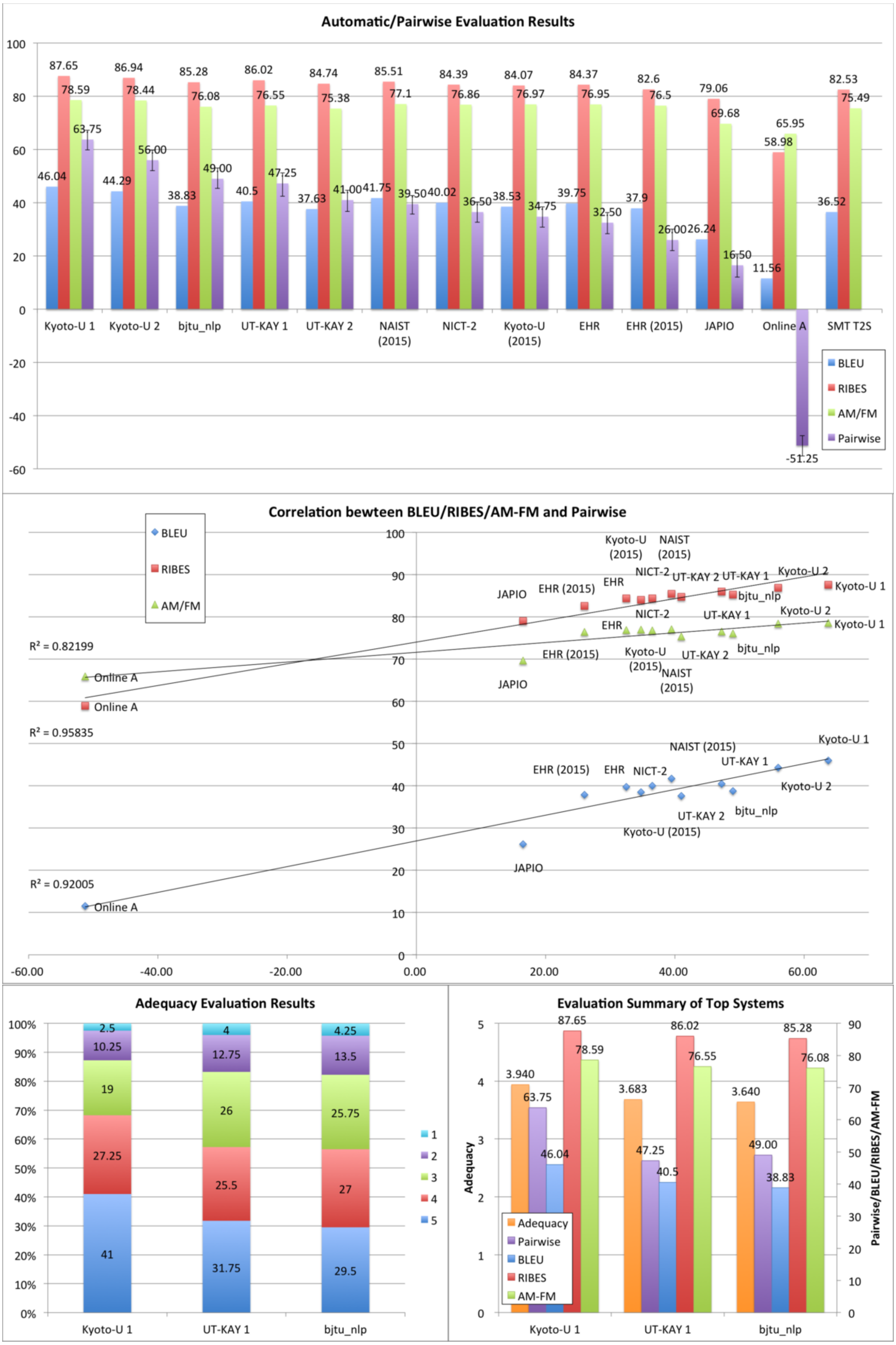


Figure 5: Official evaluation results of ASPEC-CJ.


Figure 6: Official evaluation results of JPC-JE.

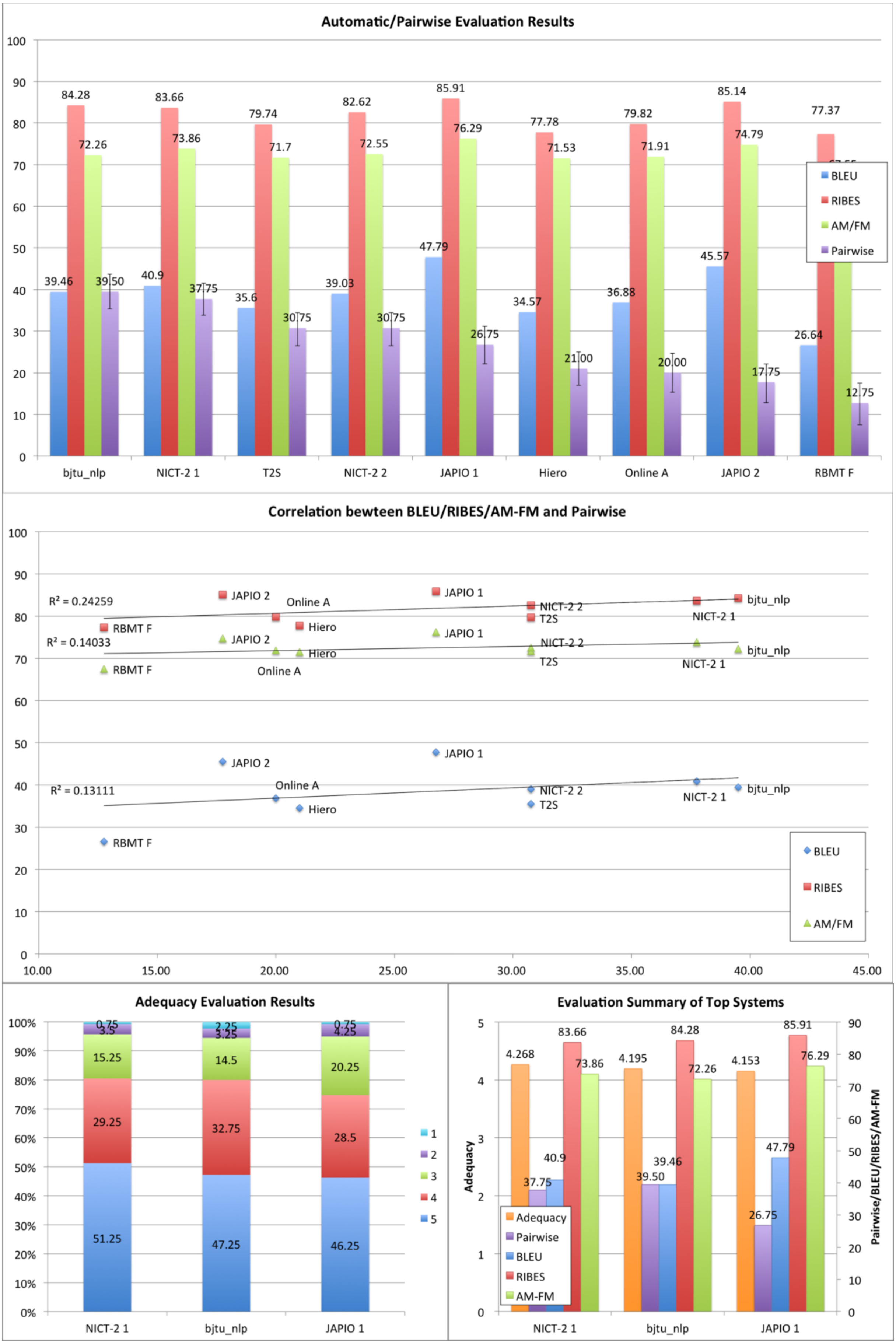


Figure 7: Official evaluation results of JPC-EJ.


Figure 8: Official evaluation results of JPC-JC.

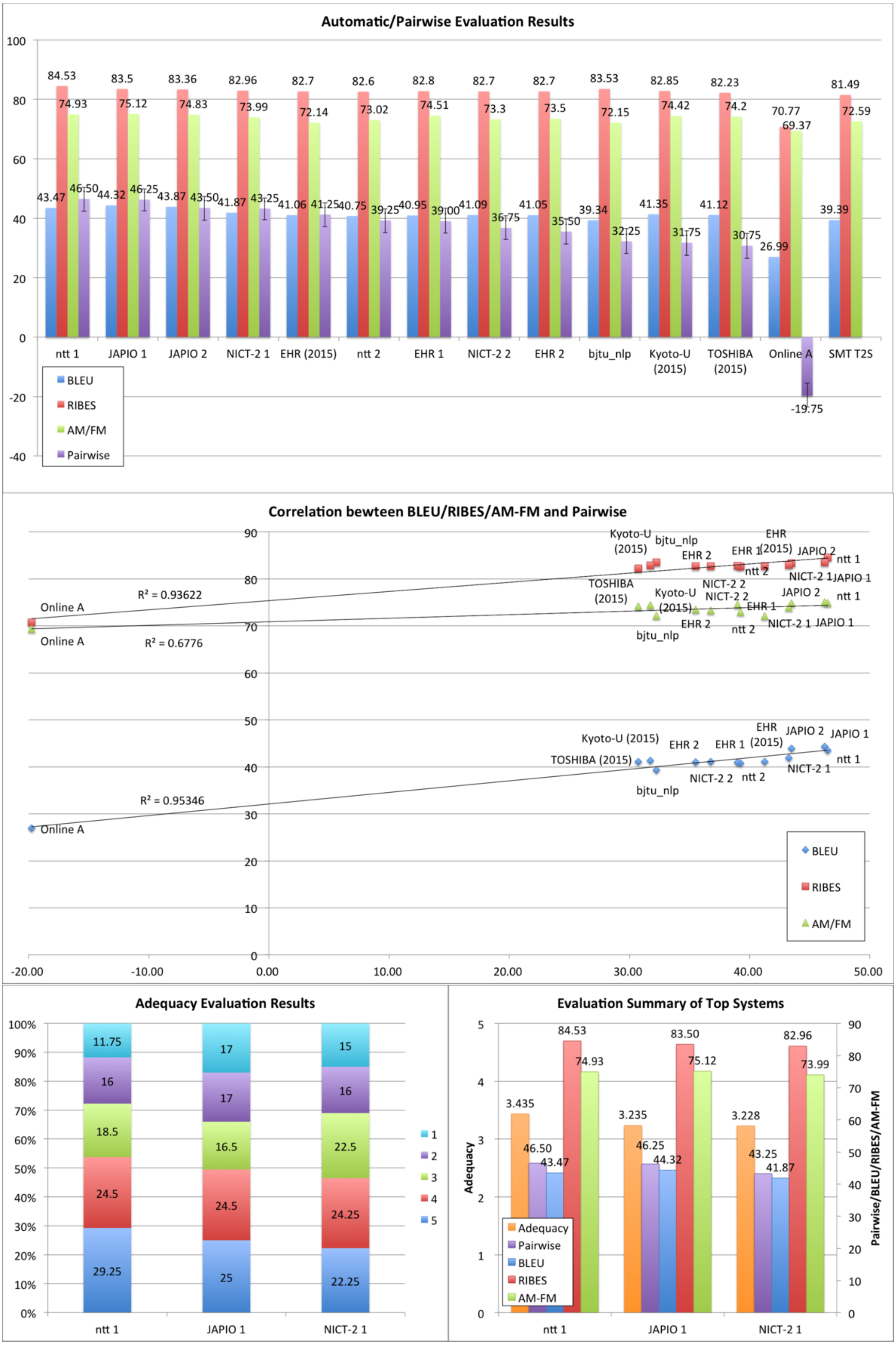


Figure 9: Official evaluation results of JPC-CJ.

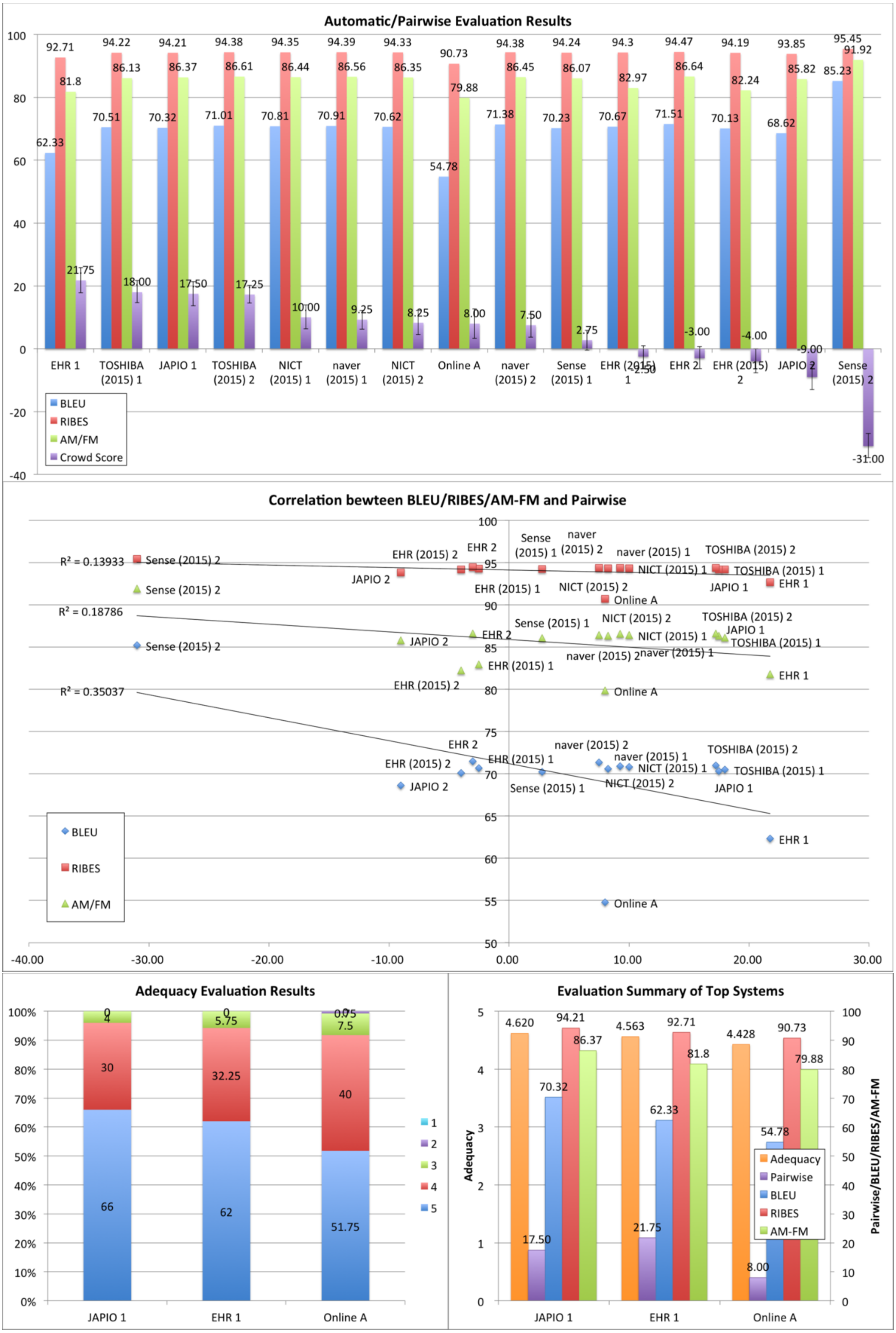


Figure 10: Official evaluation results of JPC-KJ.


Figure 11: Official evaluation results of BPPT-IE.


Figure 12: Official evaluation results of BPPT-EI.


Figure 13: Official evaluation results of IITB-EH.

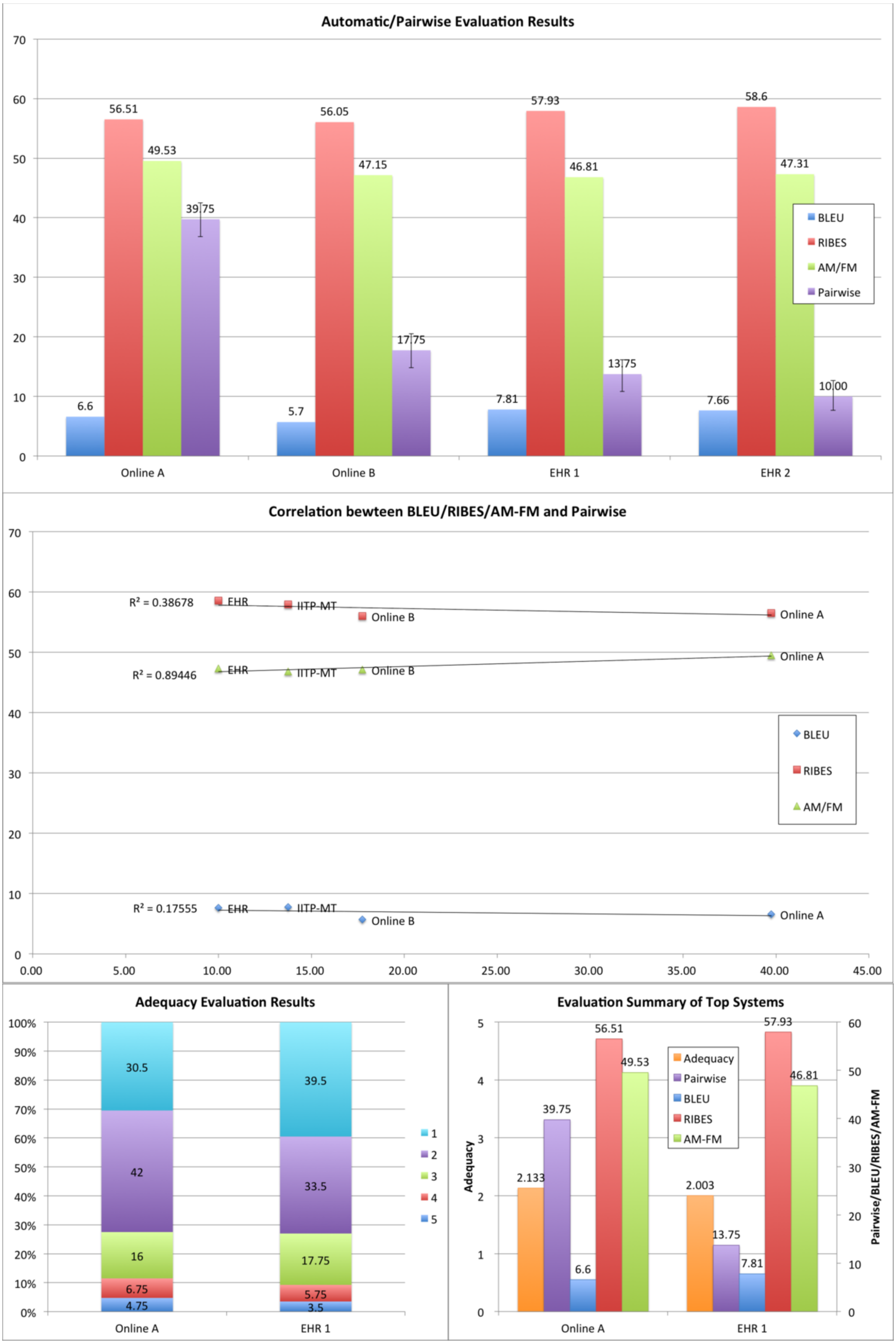


Figure 14: Official evaluation results of IITB-HJ.

| SYSTEM ID | Annotator A average | Annotator A variance | Annotator B average | Annotator B variance | all average | $\kappa$ | weighted $\kappa$ |
|---|---|---|---|---|---|---|---|
| ASPEC-JE | | | | | | | |
| Kyoto-U 1 | 3.760 | 0.682 | 4.010 | 0.670 | 3.885 | 0.205 | 0.313 |
| NAIST 1 | 3.705 | 0.728 | 3.950 | 0.628 | 3.828 | 0.257 | 0.356 |
| NICT-2 | 3.025 | 0.914 | 3.360 | 0.740 | 3.193 | 0.199 | 0.369 |
| ASPEC-EJ | | | | | | | |
| Kyoto-U 1 | 3.970 | 0.759 | 4.065 | 0.851 | 4.018 | 0.346 | 0.494 |
| bjtu_nlp | 3.800 | 0.980 | 3.625 | 1.364 | 3.713 | 0.299 | 0.509 |
| NICT-2 | 3.745 | 0.820 | 3.670 | 0.931 | 3.708 | 0.299 | 0.486 |
| Online A | 3.600 | 0.770 | 3.590 | 0.862 | 3.595 | 0.273 | 0.450 |
| ASPEC-JC | | | | | | | |
| Kyoto-U 1 | 3.995 | 1.095 | 3.755 | 1.145 | 3.875 | 0.203 | 0.362 |
| bjtu_nlp | 3.920 | 1.054 | 3.340 | 1.244 | 3.630 | 0.154 | 0.290 |
| NICT-2 | 2.940 | 1.846 | 2.850 | 1.368 | 2.895 | 0.237 | 0.477 |
| ASPEC-CJ | | | | | | | |
| Kyoto-U 1 | 4.245 | 1.045 | 3.635 | 1.232 | 3.940 | 0.234 | 0.341 |
| UT-KAY 1 | 3.995 | 1.355 | 3.370 | 1.143 | 3.683 | 0.152 | 0.348 |
| bjtu_nlp | 3.950 | 1.278 | 3.330 | 1.221 | 3.640 | 0.179 | 0.401 |
| JPC-JE | | | | | | | |
| bjtu_nlp | 4.085 | 0.798 | 4.505 | 0.580 | 4.295 | 0.254 | 0.393 |
| Online A | 3.910 | 0.652 | 4.300 | 0.830 | 4.105 | 0.166 | 0.336 |
| NICT-2 1 | 3.705 | 1.118 | 4.155 | 1.011 | 3.930 | 0.277 | 0.458 |
| JPC-EJ | | | | | | | |
| NICT-2 1 | 4.025 | 0.914 | 4.510 | 0.570 | 4.268 | 0.234 | 0.412 |
| bjtu_nlp | 3.920 | 0.924 | 4.470 | 0.749 | 4.195 | 0.151 | 0.340 |
| JAPIO 1 | 4.055 | 0.932 | 4.250 | 0.808 | 4.153 | 0.407 | 0.562 |
| JPC-JC | | | | | | | |
| bjtu_nlp | 3.485 | 1.720 | 3.015 | 1.755 | 3.250 | 0.274 | 0.507 |
| NICT-2 1 | 3.230 | 1.867 | 2.935 | 1.791 | 3.083 | 0.307 | 0.492 |
| S2T | 2.745 | 2.000 | 2.680 | 1.838 | 2.713 | 0.305 | 0.534 |
| JPC-CJ | | | | | | | |
| ntt 1 | 3.605 | 1.889 | 3.265 | 1.765 | 3.435 | 0.263 | 0.519 |
| JAPIO 1 | 3.385 | 1.947 | 3.085 | 2.088 | 3.235 | 0.365 | 0.592 |
| NICT-2 1 | 3.410 | 1.732 | 3.045 | 1.883 | 3.228 | 0.322 | 0.518 |
| JPC-KJ | | | | | | | |
| JAPIO 1 | 4.580 | 0.324 | 4.660 | 0.304 | 4.620 | 0.328 | 0.357 |
| EHR 1 | 4.510 | 0.380 | 4.615 | 0.337 | 4.563 | 0.424 | 0.478 |
| Online A | 4.380 | 0.466 | 4.475 | 0.409 | 4.428 | 0.517 | 0.574 |
| BPPT-IE | | | | | | | |
| Online A | 2.675 | 0.489 | 3.375 | 1.564 | 3.025 | 0.048 | 0.187 |
| Sense 1 | 2.685 | 0.826 | 2.420 | 1.294 | 2.553 | 0.242 | 0.408 |
| IITB-EN-ID | 2.485 | 0.870 | 2.345 | 1.216 | 2.415 | 0.139 | 0.324 |
| BPPT-EI | | | | | | | |
| Online A | 2.890 | 1.778 | 3.375 | 1.874 | 3.133 | 0.163 | 0.446 |
| Sense 1 | 2.395 | 1.059 | 2.450 | 1.328 | 2.423 | 0.305 | 0.494 |
| IITB-EN-ID | 2.185 | 1.241 | 2.360 | 1.130 | 2.273 | 0.246 | 0.477 |
| IITB-EH | | | | | | | |
| Online A | 3.200 | 1.330 | 3.525 | 1.189 | 3.363 | 0.103 | 0.155 |
| EHR | 2.590 | 1.372 | 1.900 | 0.520 | 2.245 | 0.136 | 0.263 |
| IITP-MT | 2.350 | 1.198 | 1.780 | 0.362 | 2.065 | 0.066 | 0.164 |
| IITB-HJ | | | | | | | |
| Online A | 1.955 | 1.563 | 2.310 | 0.664 | 2.133 | 0.120 | 0.287 |
| EHR 1 | 1.530 | 1.049 | 2.475 | 0.739 | 2.003 | 0.055 | 0.194 |

Table 8: JPO adequacy evaluation results in detail.

| | NAIST 1 | NAIST 2 | Kyoto-U 1 | Kyoto-U 2 | Kyoto-U (2015) | TOSHIBA (2015) | NICT-2 | Online D | TMU 1 | bjtu_nlp | TMU 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NAIST (2015) | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST 1 | | - | - | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST 2 | | | - | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 1 | | | | > | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 2 | | | | | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U (2015) | | | | | | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| TOSHIBA (2015) | | | | | | | > | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT-2 | | | | | | | | ≫ | ⋙ | ⋙ | ⋙ |
| Online D | | | | | | | | | - | ⋙ | ⋙ |
| TMU 1 | | | | | | | | | | ≫ | ⋙ |
| bjtu_nlp | | | | | | | | | | | - |

Table 9: Statistical significance testing of the ASPEC-JE Pairwise scores.

| | Kyoto-U | naver (2015) | Online A | WEBLIO_MT (2015) | NICT-2 | bjtu_nlp | EHR | UT-AKY 1 | TOKYOMT 1 | TOKYOMT 2 | UT-AKY 2 | JAPIO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NAIST (2015) | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U | | - | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| naver (2015) | | | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Online A | | | | > | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| WEBLIO_MT (2015) | | | | | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT-2 | | | | | | - | - | ≫ | ⋙ | ⋙ | ⋙ | ⋙ |
| bjtu_nlp | | | | | | | - | > | ⋙ | ⋙ | ⋙ | ⋙ |
| EHR | | | | | | | | - | ⋙ | ⋙ | ⋙ | ⋙ |
| UT-AKY 1 | | | | | | | | | ⋙ | ⋙ | ⋙ | ⋙ |
| TOKYOMT 1 | | | | | | | | | | - | ⋙ | ⋙ |
| TOKYOMT 2 | | | | | | | | | | | ⋙ | ⋙ |
| UT-AKY 2 | | | | | | | | | | | | ⋙ |

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

| | NAIST (2015) | bjtu_nlp | Kyoto-U (2015) | Kyoto-U 2 | NICT-2 | TOSHIBA (2015) | Online D |
|---|---|---|---|---|---|---|---|
| Kyoto-U 1 | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST (2015) | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| bjtu_nlp | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U (2015) | | | | - | ⋙ | ⋙ | ⋙ |
| Kyoto-U 2 | | | | | ⋙ | ⋙ | ⋙ |
| NICT-2 | | | | | | ≫ | ⋙ |
| TOSHIBA (2015) | | | | | | | ⋙ |

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

| | Kyoto-U 2 | bjtu_nlp | UT-KAY 1 | UT-KAY 2 | NAIST (2015) | NICT-2 | Kyoto-U (2015) | EHR | EHR (2015) | JAPIO | Online A |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Kyoto-U 1 | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 2 | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| bjtu_nlp | | | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| UT-KAY 1 | | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| UT-KAY 2 | | | | | - | ≫ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST (2015) | | | | | | > | ≫ | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT-2 | | | | | | | - | ≫ | ⋙ | ⋙ | ⋙ |
| Kyoto-U (2015) | | | | | | | | - | ⋙ | ⋙ | ⋙ |
| EHR | | | | | | | | | ⋙ | ⋙ | ⋙ |
| EHR (2015) | | | | | | | | | | ⋙ | ⋙ |
| JAPIO | | | | | | | | | | | ⋙ |

Table 12: Statistical significance testing of the ASPEC-CJ Pairwise scores.

| | Online A | NICT-2 1 | NICT-2 2 | RBMT A | S2T | SMT Hiero |
|---|---|---|---|---|---|---|
| bjtu_nlp | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Online A | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT-2 1 | | | - | - | - | ⋙ |
| NICT-2 2 | | | | - | - | ⋙ |
| RBMT A | | | | | - | ⋙ |
| SMT S2T | | | | | | ⋙ |

Table 13: Statistical significance testing of the JPC-JE Pairwise scores.

| | NICT-2 1 | SMT T2S | NICT-2 2 | JAPIO 1 | SMT Hiero | Online A | JAPIO 2 | RBMT F |
|---|---|---|---|---|---|---|---|---|
| bjtu_nlp | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT-2 1 | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| SMT T2S | | | - | > | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT-2 2 | | | | > | ⋙ | ⋙ | ⋙ | ⋙ |
| JAPIO 1 | | | | | ⋙ | ⋙ | ⋙ | ⋙ |
| SMT Hiero | | | | | | - | > | ⋙ |
| Online A | | | | | | | - | ⋙ |
| JAPIO 2 | | | | | | | | ≫ |

Table 14: Statistical significance testing of the JPC-EJ Pairwise scores.

| | SMT Hiero | SMT S2T | bjtu_nlp | NICT-2 2 | Online A | RBMT C |
|---|---|---|---|---|---|---|
| NICT-2 1 | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| SMT Hiero | | - | ≫ | ⋙ | ⋙ | ⋙ |
| SMT S2T | | | ≫ | ⋙ | ⋙ | ⋙ |
| bjtu_nlp | | | | ⋙ | ⋙ | ⋙ |
| NICT-2 2 | | | | | ⋙ | ⋙ |
| Online A | | | | | | ⋙ |

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

| | JAPIO 1 | JAPIO 2 | NICT-2 1 | EHR (2015) | ntt 2 | EHR 1 | NICT-2 2 | EHR 2 | bjtu_nlp | Kyoto-U (2015) | TOSHIBA (2015) | Online A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ntt 1 | - | > | > | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| JAPIO 1 | | $\gg$ | > | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| JAPIO 2 | | | - | - | $\gg$ | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| NICT-2 1 | | | | - | $\gg$ | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| EHR (2015) | | | | | - | - | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| ntt 2 | | | | | | - | - | > | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| EHR 1 | | | | | | | - | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| NICT-2 2 | | | | | | | | - | $\gg$ | $\gg$ | $\ggg$ | $\ggg$ |
| EHR 2 | | | | | | | | | > | > | $\gg$ | $\ggg$ |
| bjtu_nlp | | | | | | | | | | - | - | $\ggg$ |
| Kyoto-U (2015) | | | | | | | | | | | - | $\ggg$ |
| TOSHIBA (2015) | | | | | | | | | | | | $\ggg$ |

Table 16: Statistical significance testing of the JPC-CJ Pairwise scores.

| | TOSHIBA (2015) 1 | JAPIO 1 | TOSHIBA (2015) 2 | NICT (2015) 1 | naver (2015) 1 | NICT (2015) 2 | Online A | naver (2015) 2 | Sense (2015) 1 | EHR (2015) 1 | EHR 2 | EHR (2015) 2 | JAPIO 2 | Sense (2015) 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EHR 1 | $\gg$ | $\gg$ | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| TOSHIBA (2015) 1 | | - | - | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| JAPIO 1 | | | - | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| TOSHIBA (2015) 2 | | | | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| NICT (2015) 1 | | | | | - | - | - | - | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| naver (2015) 1 | | | | | | - | - | - | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| NICT (2015) 2 | | | | | | | - | - | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| Online A | | | | | | | | - | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| naver (2015) 2 | | | | | | | | | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| Sense (2015) 1 | | | | | | | | | | $\gg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| EHR (2015) 1 | | | | | | | | | | | - | - | $\ggg$ | $\ggg$ |
| EHR 2 | | | | | | | | | | | | - | $\ggg$ | $\ggg$ |
| EHR (2015) 2 | | | | | | | | | | | | | $\gg$ | $\ggg$ |
| JAPIO 2 | | | | | | | | | | | | | | $\ggg$ |

Table 17: Statistical significance testing of the JPC-KJ Pairwise scores.

| | Online B | SMT S2T | Sense 1 | SMT Hiero | Sense 2 | IITB-EN-ID |
|---|---|---|---|---|---|---|
| Online A | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| Online B | | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| SMT S2T | | | - | $\gg$ | $\gg$ | $\ggg$ |
| Sense 1 | | | | > | > | $\ggg$ |
| SMT Hiero | | | | | - | $\ggg$ |
| Sense 2 | | | | | | $\ggg$ |

| | Online B | Sense 1 | Sense 2 | SMT T2S | IITB-EN-ID | SMT Hiero |
|---|---|---|---|---|---|---|
| Online A | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| Online B | | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ | $\ggg$ |
| Sense 1 | | | > | $\ggg$ | $\ggg$ | $\ggg$ |
| Sense 2 | | | | $\gg$ | $\ggg$ | $\ggg$ |
| SMT T2S | | | | | - | $\ggg$ |
| IITB-EN-ID | | | | | | $\ggg$ |

Table 18: Statistical significance testing of the BPPT-IE (left) and BPPT-EI (right) Pairwise scores.

| | Online B | IITP-MT | EHR |
|---|---|---|---|
| Online A | ⋙ | ⋙ | ⋙ |
| Online B | | ⋙ | ⋙ |
| IITP-MT | | | ≫ |

| | Online B | EHR 1 | EHR 2 |
|---|---|---|---|
| Online A | ⋙ | ⋙ | ⋙ |
| Online B | | ⋙ | ⋙ |
| EHR 1 | | | ⋙ |

Table 19: Statistical significance testing of the IITB-EH (left) and IITB-HJ (right) Pairwise scores.

| ASPEC-JE | |
|---|---|
| SYSTEM ID | $\kappa$ |
| NAIST (2015) | 0.078 |
| NAIST 1 | 0.081 |
| NAIST 2 | 0.091 |
| Kyoto-U 1 | 0.106 |
| Kyoto-U 2 | 0.148 |
| Kyoto-U (2015) | 0.066 |
| TOSHIBA (2015) | 0.068 |
| NICT-2 | 0.106 |
| Online D | 0.081 |
| TMU 1 | 0.060 |
| bjtu_nlp | 0.146 |
| TMU 2 | 0.072 |
| ave. | 0.092 |

| ASPEC-EJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| NAIST (2015) | 0.239 |
| Kyoto-U | 0.215 |
| naver (2015) | 0.187 |
| Online A | 0.181 |
| WEBLIO MT (2015) | 0.193 |
| NICT-2 | 0.177 |
| bjtu_nlp | 0.247 |
| EHR | 0.195 |
| UT-AKY 1 | 0.204 |
| TOKYOMT 1 | 0.189 |
| TOKYOMT 2 | 0.200 |
| UT-AKY 2 | 0.201 |
| JAPIO | 0.183 |
| ave | 0.201 |

| ASPEC-JC | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Kyoto-U 1 | 0.177 |
| NAIST (2015) | 0.221 |
| bjtu_nlp | 0.187 |
| Kyoto-U (2015) | 0.197 |
| Kyoto-U 2 | 0.251 |
| NICT-2 | 0.190 |
| TOSHIBA (2015) | 0.214 |
| Online D | 0.180 |
| ave. | 0.202 |

| ASPEC-CJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Kyoto-U 1 | 0.195 |
| Kyoto-U 2 | 0.151 |
| bjtu_nlp | 0.168 |
| UT-KAY 1 | 0.172 |
| UT-KAY 2 | 0.156 |
| NAIST (2015) | 0.089 |
| NICT-2 | 0.168 |
| Kyoto-U (2015) | 0.144 |
| EHR | 0.152 |
| EHR (2015) | 0.190 |
| JAPIO | 0.185 |
| Online A | 0.207 |
| ave. | 0.165 |

| JPC-JE | |
|---|---|
| SYSTEM ID | $\kappa$ |
| bjtu_nlp | 0.256 |
| Online A | 0.242 |
| NICT-2 1 | 0.280 |
| NICT-2 2 | 0.293 |
| RBMT A | 0.179 |
| S2T | 0.296 |
| Hiero | 0.324 |
| ave. | 0.267 |

| JPC-EJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| bjtu_nlp | 0.339 |
| NICT-2 1 | 0.367 |
| T2S | 0.378 |
| NICT-2 2 | 0.346 |
| JAPIO 1 | 0.323 |
| Hiero | 0.383 |
| Online A | 0.403 |
| JAPIO 2 | 0.336 |
| RBMT F | 0.323 |
| ave. | 0.355 |

| JPC-JC | |
|---|---|
| SYSTEM ID | $\kappa$ |
| NICT-2 1 | 0.076 |
| Hiero | 0.127 |
| S2T | 0.133 |
| bjtu_nlp | 0.085 |
| NICT-2 2 | 0.068 |
| Online A | 0.055 |
| RBMT C | 0.116 |
| ave. | 0.094 |

| JPC-CJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| ntt 1 | 0.169 |
| JAPIO 1 | 0.121 |
| JAPIO 2 | 0.160 |
| NICT-2 1 | 0.150 |
| EHR (2015) | 0.123 |
| ntt 2 | 0.114 |
| EHR 1 | 0.155 |
| NICT-2 2 | 0.151 |
| EHR 2 | 0.150 |
| bjtu_nlp | 0.200 |
| Kyoto-U (2015) | 0.096 |
| TOSHIBA (2015) | 0.131 |
| Online A | 0.116 |
| ave. | 0.141 |

| JPC-KJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| EHR 1 | 0.256 |
| TOSHIBA (2015) 1 | 0.221 |
| JAPIO 1 | 0.228 |
| TOSHIBA (2015) 2 | 0.176 |
| NICT (2015) 1 | 0.351 |
| naver (2015) 1 | 0.469 |
| NICT (2015) 2 | 0.345 |
| Online A | 0.232 |
| naver (2015) 2 | 0.299 |
| Sense (2015) 1 | 0.522 |
| EHR (2015) 1 | 0.363 |
| EHR 2 | 0.399 |
| EHR (2015) 2 | 0.373 |
| JAPIO 2 | 0.260 |
| Sense (2015) 2 | 0.329 |
| ave. | 0.322 |

| BPPT-IE | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Online A | -0.083 |
| Online B | -0.051 |
| S2T | 0.025 |
| Sense 1 | 0.145 |
| Hiero | 0.057 |
| Sense 2 | 0.102 |
| IITB-EN-ID | 0.063 |
| ave. | 0.037 |

| BPPT-EI | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Online A | 0.094 |
| Online B | 0.063 |
| Sense 1 | 0.135 |
| Sense 2 | 0.160 |
| T2S | 0.089 |
| IITB-EN-ID | 0.115 |
| Hiero | 0.165 |
| ave. | 0.117 |

| IITB-EH | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Online A | 0.141 |
| Online B | 0.110 |
| IITP-MT | 0.215 |
| EHR | 0.196 |
| ave. | 0.166 |

| IITB-HJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Online A | 0.285 |
| Online B | 0.488 |
| EHR 1 | 0.452 |
| EHR 2 | 0.510 |
| ave. | 0.434 |

Table 20: The Fleiss' kappa values for the pairwise evaluation results.


Figure 15: The chronological evaluation results of JPO adequacy evaluation.

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| SMT Hiero | 2 | SMT | NO | 18.72 | 0.651066 | 0.588880 | —— | Hierarchical Phrase-based SMT |
| SMT Phrase | 6 | SMT | NO | 18.45 | 0.645137 | 0.590950 | —— | Phrase-based SMT |
| SMT S2T | 877 | SMT | NO | 20.36 | 0.678253 | 0.593410 | +7.00 | String-to-Tree SMT |
| RBMT D | 887 | Other | YES | 15.29 | 0.683378 | 0.551690 | +16.75 | RBMT D |
| RBMT E | 76 | Other | YES | 14.82 | 0.663851 | 0.561620 | —— | RBMT E |
| RBMT F | 79 | Other | YES | 13.86 | 0.661387 | 0.556840 | —— | RBMT F |
| Online C (2014) | 87 | Other | YES | 10.64 | 0.624827 | 0.466480 | —— | Online C (2014) |
| Online D (2014) | 35 | Other | YES | 15.08 | 0.643588 | 0.564170 | —— | Online D (2014) |
| Online D (2015) | 775 | Other | YES | 16.85 | 0.676609 | 0.562270 | +0.25 | Online D (2015) |
| Online D | 1042 | Other | YES | 16.91 | 0.677412 | 0.564270 | +28.00 | Online D (2016) |
| NAIST 1 | 1122 | SMT | NO | 26.39 | 0.762712 | 0.587450 | +48.25 | Neural MT w/ Lexicon and MinRisk Training 4 Ensemble |
| NAIST 2 | 1247 | SMT | NO | 26.12 | 0.756956 | 0.571360 | +47.50 | Neural MT w/ Lexicon 6 Ensemble |
| Kyoto-U 1 | 1182 | NMT | NO | 26.22 | 0.756601 | 0.558540 | +44.25 | Ensemble of 4 single-layer model (30k voc) |
| Kyoto-U 2 | 1246 | NMT | NO | 24.71 | 0.750802 | 0.562650 | +47.00 | voc src:200k voc tgt: 52k + BPE 2-layer self-ensembling |
| TMU 1 | 1222 | NMT | NO | 18.29 | 0.710613 | 0.565270 | +16.00 | 2016 our proposed method to control output voice |
| TMU 2 | 1234 | NMT | NO | 18.45 | 0.711542 | 0.546880 | +25.00 | 6 ensemble |
| BJTU-nlp 1 | 1168 | NMT | NO | 18.34 | 0.690455 | 0.505730 | +19.25 | RNN Encoder-Decoder with attention mechanism, single model |
| NICT-2 1 | 1104 | SMT | YES | 21.54 | 0.708808 | 0.595930 | —— | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) + Google 5-gram LM |

Table 21: ASPEC-JE submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 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 | —— | Phrase-based SMT |
| 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 | Hierarchical Phrase-based SMT |
| 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 | Tree-to-String SMT |
| RBMT A | 68 | Other | YES | 12.86 | 14.43 | 13.16 | 0.670167 | 0.676464 | 0.678934 | 0.626940 | 0.626940 | 0.626940 | —— | RBMT A |
| RBMT B | 883 | Other | YES | 13.18 | 14.85 | 13.48 | 0.671958 | 0.680748 | 0.682683 | 0.622930 | 0.622930 | 0.622930 | +9.75 | RBMT B |
| RBMT C | 95 | Other | YES | 12.19 | 13.32 | 12.14 | 0.668372 | 0.672645 | 0.676018 | 0.594380 | 0.594380 | 0.594380 | —— | RBMT C |
| Online A (2014) | 34 | Other | YES | 19.66 | 21.63 | 20.17 | 0.718019 | 0.723486 | 0.725848 | 0.695420 | 0.695420 | 0.695420 | —— | Online A (2014) |
| 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 | Online A (2015) |
| 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) |
| Online B (2014) | 91 | Other | YES | 17.04 | 18.67 | 17.36 | 0.687797 | 0.693390 | 0.698126 | 0.643070 | 0.643070 | 0.643070 | —— | Online B (2014) |
| 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 B (2015) |
| Kyoto-U 1 | 1172 | NMT | NO | 36.19 | 38.20 | 36.78 | 0.819836 | 0.823878 | 0.828956 | 0.738700 | 0.738700 | 0.738700 | +55.25 | BPE tgt/src: 52k 2-layer lstm self-ensemble of 3 |
| EHR 1 | 1140 | SMT | NO | 31.32 | 33.58 | 32.28 | 0.759914 | 0.771427 | 0.775023 | 0.746720 | 0.746720 | 0.746720 | +39.00 | PBSMT with preordering (DL=6) |
| BJTU-nlp 1 | 1143 | NMT | NO | 31.18 | 33.47 | 31.80 | 0.780510 | 0.787497 | 0.791088 | 0.704340 | 0.704340 | 0.704340 | +39.50 | RNN Encoder-Decoder with attention mechanism, single model |
| TOKYOMT 1 | 1131 | NMT | NO | 30.21 | 33.38 | 31.24 | 0.809691 | 0.817258 | 0.819951 | 0.705210 | 0.705210 | 0.705210 | +29.75 | char 1 , ens 2 , version 1 |
| TOKYOMT 2 | 1217 | NMT | NO | 32.03 | 34.77 | 32.98 | 0.808189 | 0.814452 | 0.818130 | 0.720810 | 0.720810 | 0.720810 | +30.50 | Combination of NMT and T2S |
| JAPIO 1 | 1165 | SMT | YES | 20.52 | 22.56 | 21.05 | 0.723467 | 0.728584 | 0.731474 | 0.660790 | 0.660790 | 0.660790 | +4.25 | Phrase-based SMT with Preordering + JAPIO corpus + rule-based posteditor |
| NICT-2 1 | 1097 | SMT | YES | 34.67 | 36.86 | 35.37 | 0.784335 | 0.790993 | 0.793409 | 0.753080 | 0.753080 | 0.753080 | +41.25 | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) + Google 5-gram LM |
| UT-AKY 1 | 1224 | NMT | NO | 30.14 | 33.20 | 31.09 | 0.806025 | 0.814490 | 0.815836 | 0.708140 | 0.708140 | 0.708140 | +21.75 | tree-to-seq NMT model (character-based decoder) |
| UT-AKY 2 | 1228 | NMT | NO | 33.57 | 36.95 | 34.65 | 0.816984 | 0.824456 | 0.827647 | 0.731440 | 0.731440 | 0.731440 | +36.25 | tree-to-seq NMT model (word-based decoder) |

Table 22: ASPEC-EJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | kytea | stanford (ctb) | stanford (pku) | kytea | stanford (ctb) | stanford (pku) | kytea | stanford (ctb) | stanford (pku) | | |
| SMT Phrase | 7 | SMT | NO | 27.96 | 28.01 | 27.68 | 0.788961 | 0.790263 | 0.790937 | 0.749450 | 0.749450 | 0.749450 | —— | Phrase-based SMT |
| SMT Hiero | 3 | SMT | NO | 27.71 | 27.70 | 27.35 | 0.809128 | 0.809561 | 0.811394 | 0.745100 | 0.745100 | 0.745100 | —— | Hierarchical Phrase-based SMT |
| 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 | String-to-Tree SMT |
| RBMT B | 886 | Other | YES | 17.86 | 17.75 | 17.49 | 0.744818 | 0.745885 | 0.743794 | 0.667960 | 0.667960 | 0.667960 | -11.00 | RBMT B |
| RBMT C | 244 | Other | NO | 9.62 | 9.96 | 9.59 | 0.642278 | 0.648758 | 0.645385 | 0.594900 | 0.594900 | 0.594900 | —— | RBMT C |
| Online C (2014) | 216 | Other | YES | 7.26 | 7.01 | 6.72 | 0.612808 | 0.613075 | 0.611563 | 0.587820 | 0.587820 | 0.587820 | —— | Online C (2014) |
| 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 C (2015) |
| Online D (2014) | 37 | Other | YES | 9.37 | 8.93 | 8.84 | 0.606905 | 0.606328 | 0.604149 | 0.625430 | 0.625430 | 0.625430 | —— | Online D (2014) |
| 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 | Online D (2015) |
| 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) |
| Kyoto-U 1 | 1071 | NMT | NO | 31.98 | 32.08 | 31.72 | 0.837579 | 0.839354 | 0.835932 | 0.763290 | 0.763290 | 0.763290 | +58.75 | 2 layer lstm dropout 0.5 200k source voc unk replaced |
| Kyoto-U 2 | 1109 | EBMT | NO | 30.27 | 29.94 | 29.92 | 0.813114 | 0.813581 | 0.813054 | 0.764230 | 0.764230 | 0.764230 | +30.75 | KyotoEBMT 2016 w/o reranking |
| BJTU-nlp 1 | 1120 | NMT | NO | 30.57 | 30.49 | 30.31 | 0.829679 | 0.829113 | 0.827637 | 0.754690 | 0.754690 | 0.754690 | +46.25 | RNN Encoder-Decoder with attention mechanism, single model |
| NICT-2 1 | 1105 | SMT | YES | 30.00 | 29.97 | 29.78 | 0.820891 | 0.820069 | 0.821090 | 0.759670 | 0.759670 | 0.759670 | +24.00 | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) |

Table 23: ASPEC-JC submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | juman | kytea | mecab | juman | kytea | mecab | juman | kytea | mecab | | |
| SMT Phrase | 8 | SMT | NO | 34.65 | 35.16 | 34.77 | 0.772498 | 0.766384 | 0.771005 | 0.753010 | 0.753010 | 0.753010 | —— | Phrase-based SMT |
| SMT Hiero | 4 | SMT | NO | 35.43 | 35.91 | 35.64 | 0.810406 | 0.798726 | 0.807665 | 0.750950 | 0.750950 | 0.750950 | —— | Hierarchical Phrase-based SMT |
| 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 | Tree-to-String SMT |
| RBMT A | 885 | Other | YES | 9.37 | 9.87 | 9.35 | 0.666277 | 0.652402 | 0.661730 | 0.626070 | 0.626070 | 0.626070 | -28.00 | RBMT A |
| RBMT D | 242 | Other | NO | 8.39 | 8.70 | 8.30 | 0.641189 | 0.626400 | 0.633319 | 0.586790 | 0.586790 | 0.586790 | —— | RBMT D |
| Online A (2014) | 36 | Other | YES | 11.63 | 13.21 | 11.87 | 0.595925 | 0.598172 | 0.598573 | 0.658060 | 0.658060 | 0.658060 | —— | Online A (2014) |
| 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 | Online A (2015) |
| 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) |
| Online B (2014) | 215 | Other | YES | 10.48 | 11.26 | 10.47 | 0.600733 | 0.596006 | 0.600706 | 0.636930 | 0.636930 | 0.636930 | —— | Online B (2014) |
| 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 B (2015) |
| Kyoto-U 1 | 1255 | NMT | NO | 44.29 | 45.05 | 44.32 | 0.869360 | 0.864748 | 0.869913 | 0.784380 | 0.784380 | 0.784380 | +56.00 | src: 200k tgt: 50k 2-layers self-ensembling |
| Kyoto-U 2 | 1256 | NMT | NO | 46.04 | 46.70 | 46.05 | 0.876531 | 0.872904 | 0.876946 | 0.785910 | 0.785910 | 0.785910 | +63.75 | voc: 30k ensemble of 3 independent model + reverse rescoring |
| EHR 1 | 1063 | SMT | YES | 39.75 | 39.85 | 39.40 | 0.843723 | 0.836156 | 0.841952 | 0.769490 | 0.769490 | 0.769490 | +32.50 | LM-based merging of outputs of preordered word-based PB-SMT(DL=6) and preordered character-based PBSMT(DL=6). |
| BJTU-nlp 1 | 1138 | NMT | NO | 38.83 | 39.25 | 38.68 | 0.852818 | 0.846301 | 0.852298 | 0.760840 | 0.760840 | 0.760840 | +49.00 | RNN Encoder-Decoder with attention mechanism, single model |
| JAPIO 1 | 1208 | SMT | YES | 26.24 | 27.87 | 26.37 | 0.790553 | 0.780637 | 0.785917 | 0.696770 | 0.696770 | 0.696770 | +16.50 | Phrase-based SMT with Preordering + JAPIO corpus + rule-based posteditor |
| NICT-2 1 | 1099 | SMT | YES | 40.02 | 40.45 | 40.29 | 0.843941 | 0.837707 | 0.842513 | 0.768580 | 0.768580 | 0.768580 | +36.50 | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) + Google 5-gram LM |
| UT-KAY 1 | 1220 | NMT | NO | 37.63 | 39.07 | 37.82 | 0.847407 | 0.842055 | 0.848040 | 0.753820 | 0.753820 | 0.753820 | +41.00 | An end-to-end NMT with 512 dimensional single-layer LSTMs, UNK replacement, and domain adaptation |
| UT-KAY 2 | 1221 | NMT | NO | 40.50 | 41.81 | 40.67 | 0.860214 | 0.854690 | 0.860449 | 0.765530 | 0.765530 | 0.765530 | +47.25 | Ensemble of our NMT models with and without domain adaptation |

Table 24: ASPEC-CJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| SMT Phrase | 977 | SMT | NO | 30.80 | 0.730056 | 0.664830 | —— | Phrase-based SMT |
| SMT Hiero | 979 | SMT | NO | 32.23 | 0.763030 | 0.672500 | +8.75 | Hierarchical Phrase-based SMT |
| SMT S2T | 980 | SMT | NO | 34.40 | 0.793483 | 0.672760 | +23.00 | String-to-Tree SMT |
| RBMT A | 1090 | Other | YES | 21.57 | 0.750381 | 0.521230 | +23.75 | RBMT A |
| RBMT B | 1095 | Other | YES | 18.38 | 0.710992 | 0.518110 | —— | RBMT B |
| RBMT C | 1088 | Other | YES | 21.00 | 0.755017 | 0.519210 | —— | RBMT C |
| Online A (2016) | 1035 | Other | YES | 35.77 | 0.803661 | 0.673950 | +32.25 | Online A (2016) |
| Online B (2016) | 1051 | Other | YES | 16.00 | 0.688004 | 0.486450 | —— | Online B (2016) |
| BJTU-nlp 1 | 1149 | NMT | NO | 41.62 | 0.851975 | 0.690750 | +41.50 | RNN Encoder-Decoder with attention mechanism, single model |
| NICT-2 1 | 1080 | SMT | NO | 35.68 | 0.824398 | 0.667540 | +25.00 | Phrase-based SMT with Preordering + Domain Adaptation |
| NICT-2 2 | 1103 | SMT | YES | 36.06 | 0.825420 | 0.672890 | +24.25 | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) + Google 5-gram LM |

Table 25: JPC-JE submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 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 | —— | Phrase-based SMT |
| 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 | Hierarchical Phrase-based SMT |
| 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 | Tree-to-String SMT |
| RBMT D | 1085 | Other | YES | 23.02 | 24.90 | 23.45 | 0.761224 | 0.757341 | 0.760325 | 0.647730 | 0.647730 | 0.647730 | —— | RBMT D |
| RBMT E | 1087 | Other | YES | 21.35 | 23.17 | 21.53 | 0.743484 | 0.741985 | 0.742300 | 0.646930 | 0.646930 | 0.646930 | —— | RBMT E |
| RBMT F | 1086 | Other | YES | 26.64 | 28.48 | 26.84 | 0.773673 | 0.769244 | 0.773344 | 0.675470 | 0.675470 | 0.675470 | +12.75 | RBMT F |
| 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 A (2016) |
| Online B (2016) | 1073 | Other | YES | 21.57 | 22.62 | 21.65 | 0.743083 | 0.735203 | 0.740962 | 0.659950 | 0.659950 | 0.659950 | —— | Online B (2016) |
| BJTU-nlp 1 | 1112 | NMT | NO | 39.46 | 41.16 | 39.45 | 0.842762 | 0.840148 | 0.842669 | 0.722560 | 0.722560 | 0.722560 | +39.50 | RNN Encoder-Decoder with attention mechanism, single model |
| JAPIO 1 | 1141 | SMT | YES | 45.57 | 46.40 | 45.74 | 0.851376 | 0.848580 | 0.849513 | 0.747910 | 0.747910 | 0.747910 | +17.75 | Phrase-based SMT with Preordering + JAPIO corpus |
| JAPIO 2 | 1156 | SMT | YES | 47.79 | 48.57 | 47.92 | 0.859139 | 0.856392 | 0.857422 | 0.762850 | 0.762850 | 0.762850 | +26.75 | Phrase-based SMT with Preordering + JPC/JAPIO corpora |
| NICT-2 1 | 1078 | SMT | NO | 39.03 | 40.74 | 38.98 | 0.826228 | 0.823582 | 0.824428 | 0.725540 | 0.725540 | 0.725540 | +30.75 | Phrase-based SMT with Preordering + Domain Adaptation |
| NICT-2 2 | 1098 | SMT | YES | 40.90 | 42.51 | 40.66 | 0.836556 | 0.832401 | 0.832622 | 0.738630 | 0.738630 | 0.738630 | +37.75 | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) + Google 5-gram LM |

Table 26: JPC-EJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 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 | —— | Phrase-based SMT |
| 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 | Hierarchical Phrase-based SMT |
| 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 | String-to-Tree SMT |
| RBMT C | 1118 | Other | YES | 12.35 | 13.72 | 13.17 | 0.688240 | 0.708681 | 0.700210 | 0.475430 | 0.475430 | 0.475430 | -41.25 | RBMT C |
| Online A | 1038 | Other | YES | 23.02 | 23.57 | 23.29 | 0.754241 | 0.760672 | 0.760148 | 0.702350 | 0.702350 | 0.702350 | -23.00 | Online A (2016) |
| Online B | 1069 | Other | YES | 9.42 | 9.59 | 8.79 | 0.642026 | 0.651070 | 0.643520 | 0.527180 | 0.527180 | 0.527180 | —— | Online B (2016) |
| BJTU-nlp 1 | 1150 | NMT | NO | 31.49 | 32.79 | 32.51 | 0.816577 | 0.822978 | 0.820820 | 0.701490 | 0.701490 | 0.701490 | -1.00 | RNN Encoder-Decoder with attention mechanism, single model |
| NICT-2 1 | 1081 | SMT | NO | 33.35 | 34.64 | 33.81 | 0.808513 | 0.817996 | 0.815322 | 0.723270 | 0.723270 | 0.723270 | -11.00 | Phrase-based SMT with Preordering + Domain Adaptation |
| NICT-2 2 | 1106 | SMT | YES | 33.40 | 34.64 | 33.83 | 0.811788 | 0.820320 | 0.818701 | 0.731520 | 0.731520 | 0.731520 | +14.00 | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) |

Table 27: JPC-JC submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | juman | kytea | mecab | juman | kytea | mecab | juman | kytea | mecab | | |
| SMT Phrase | 431 | SMT | NO | 38.34 | 38.51 | 38.22 | 0.782019 | 0.778921 | 0.781456 | 0.723110 | 0.723110 | 0.723110 | —— | Phrase-based SMT |
| SMT Hiero | 430 | SMT | NO | 39.22 | 39.52 | 39.14 | 0.806058 | 0.802059 | 0.804523 | 0.729370 | 0.729370 | 0.729370 | —— | Hierarchical Phrase-based SMT |
| 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 | Tree-to-String SMT |
| RBMT A | 759 | Other | NO | 10.49 | 10.72 | 10.35 | 0.674060 | 0.664098 | 0.667349 | 0.557130 | 0.557130 | 0.557130 | -39.25 | RBMT A |
| RBMT B | 760 | Other | NO | 7.94 | 8.07 | 7.73 | 0.596200 | 0.581837 | 0.586941 | 0.502100 | 0.502100 | 0.502100 | —— | RBMT B |
| 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 A (2015) |
| 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) |
| Online B (2015) | 648 | Other | YES | 12.33 | 12.72 | 12.44 | 0.648996 | 0.641255 | 0.648742 | 0.588380 | 0.588380 | 0.588380 | —— | Online B (2015) |
| EHR 1 | 1007 | SMT | YES | 40.95 | 41.20 | 40.51 | 0.828040 | 0.824502 | 0.826864 | 0.745080 | 0.745080 | 0.745080 | +39.00 | Combination of word-based PBSMT and character-based PBSMT with DL=6. |
| EHR 2 | 1009 | SMT and RBMT | YES | 41.05 | 41.05 | 40.52 | 0.827048 | 0.821940 | 0.824852 | 0.735010 | 0.735010 | 0.735010 | +35.50 | Combination of word-based PBSMT, character-based PBSMT and RBMT+PBSPE with DL=6. |
| ntt 1 | 1193 | SMT | NO | 40.75 | 41.05 | 40.68 | 0.825985 | 0.822125 | 0.824840 | 0.730190 | 0.730190 | 0.730190 | +39.25 | PBMT with pre-ordering on dependency structures |
| ntt 2 | 1200 | NMT | NO | 43.47 | 44.27 | 43.53 | 0.845271 | 0.843105 | 0.844968 | 0.749270 | 0.749270 | 0.749270 | +46.50 | NMT with pre-ordering and attention over bidirectional LSTMs (pre-ordering module is the same as the PBMT submission) |
| BJTU-nlp 1 | 1128 | NMT | NO | 39.34 | 39.72 | 39.30 | 0.835314 | 0.830505 | 0.833216 | 0.721460 | 0.721460 | 0.721460 | +32.25 | RNN Encoder-Decoder with attention mechanism, single model |
| JAPIO 1 | 1180 | SMT | YES | 43.87 | 44.47 | 43.66 | 0.833586 | 0.829360 | 0.831534 | 0.748330 | 0.748330 | 0.748330 | +43.50 | Phrase-based SMT with Preordering + JAPIO corpus |
| JAPIO 2 | 1192 | SMT | YES | 44.32 | 45.12 | 44.09 | 0.834959 | 0.830164 | 0.832955 | 0.751200 | 0.751200 | 0.751200 | +46.25 | Phrase-based SMT with Preordering + JAPIO corpus |
| NICT-2 1 | 1079 | SMT | NO | 41.09 | 41.27 | 41.24 | 0.827009 | 0.822664 | 0.825323 | 0.733020 | 0.733020 | 0.733020 | +36.75 | Phrase-based SMT with Preordering + Domain Adaptation |
| NICT-2 2 | 1100 | SMT | YES | 41.87 | 42.39 | 42.13 | 0.829640 | 0.826744 | 0.828107 | 0.739890 | 0.739890 | 0.739890 | +43.25 | Phrase-based SMT with Preordering + Domain Adaptation (JPC and ASPEC) + Google 5-gram LM |

Table 28: JPC-CJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| SMT Phrase | 1020 | SMT | NO | 67.09 | 0.933825 | 0.844950 | —— | Phrase-based SMT |
| SMT Hiero | 1021 | SMT | NO | 66.52 | 0.932391 | 0.844550 | -3.50 | Hierarchical Phrase-based SMT |
| RBMT C | 1083 | Other | YES | 43.26 | 0.872746 | 0.766520 | —— | RBMT C |
| RBMT D | 1089 | Other | YES | 45.59 | 0.877411 | 0.765530 | -53.25 | RBMT D |
| Online A | 1037 | Other | YES | 48.75 | 0.898976 | 0.791320 | -21.00 | Online A (2016) |
| Online B | 1068 | Other | YES | 28.21 | 0.827843 | 0.692980 | —— | Online B (2016) |

Table 29: JPC-JK submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 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 | —— | Phrase-based SMT |
| 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 | Hierarchical Phrase-based SMT |
| RBMT A | 653 | Other | YES | 42.00 | 43.97 | 42.45 | 0.876396 | 0.873734 | 0.875146 | 0.712020 | 0.712020 | 0.712020 | -7.25 | RBMT A |
| RBMT B | 654 | Other | YES | 34.74 | 37.51 | 35.54 | 0.845712 | 0.849014 | 0.846228 | 0.643150 | 0.643150 | 0.643150 | —— | RBMT B |
| 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 | Online A (2015) |
| 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) |
| 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 B (2015) |
| EHR 1 | 1005 | SMT | YES | 71.51 | 72.32 | 71.77 | 0.944651 | 0.943514 | 0.944606 | 0.866370 | 0.866370 | 0.866370 | -3.00 | Combination of word-based PBSMT and character-based PBSMT with DL=0. Parentheses surrounding number in Korean sentences are deleted. |
| EHR 2 | 1006 | SMT | YES | 62.33 | 64.17 | 62.75 | 0.927065 | 0.927215 | 0.927017 | 0.818030 | 0.818030 | 0.818030 | +21.75 | Combination of word-based PBSMT and character-based PBSMT with DL=0. Parentheses in Korean side and not in Japanese side are added to Japanese for training and dev sets. |
| JAPIO 1 | 1206 | SMT | YES | 68.62 | 69.49 | 68.90 | 0.938474 | 0.937066 | 0.938230 | 0.858190 | 0.858190 | 0.858190 | -9.00 | Phrase-based SMT + JAPIO corpus + rule-based posteditor |
| JAPIO 2 | 1209 | SMT | YES | 70.32 | 71.07 | 70.52 | 0.942137 | 0.940544 | 0.941746 | 0.863660 | 0.863660 | 0.863660 | +17.50 | Phrase-based SMT + JPC/JAPIO corpora + rule-based posteditor |

Table 30: JPC-KJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| SMT Phrase | 971 | SMT | NO | 24.57 | 0.779545 | 0.578310 | 0.00 | Phrase-based SMT |
| SMT Hiero | 981 | SMT | NO | 23.62 | 0.776309 | 0.575450 | -8.25 | Hierarchical Phrase-based SMT |
| SMT S2T | 982 | SMT | NO | 22.90 | 0.780436 | 0.577210 | -3.25 | String-to-Tree SMT |
| Online A | 1033 | Other | YES | 28.11 | 0.797852 | 0.607290 | +49.25 | Online A |
| Online B | 1052 | Other | YES | 19.69 | 0.770690 | 0.578920 | +34.50 | Online B |
| Sense 1 | 1171 | SMT | NO | 25.62 | 0.782761 | 0.564500 | -5.00 | Baseline-C50-PBMT |
| Sense 2 | 1173 | SMT | NO | 25.97 | 0.787768 | 0.570710 | -8.25 | Clustercat-PBMT |

Table 31: BPPT-IE submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| SMT Phrase | 972 | SMT | NO | 23.95 | 0.808362 | 0.559800 | 0.00 | Phrase-based SMT |
| SMT Hiero | 983 | SMT | NO | 22.64 | 0.796701 | 0.568660 | -17.00 | Hierarchical Phrase-based SMT |
| SMT T2S | 984 | SMT | NO | 23.65 | 0.792346 | 0.572520 | -7.75 | Tree-to-String SMT |
| Online A | 1034 | Other | YES | 24.20 | 0.819504 | 0.554720 | +35.75 | Online A |
| Online B | 1050 | Other | YES | 18.09 | 0.789499 | 0.514430 | +10.50 | Online B |
| Sense 1 | 1170 | SMT | NO | 25.16 | 0.807097 | 0.568780 | +1.25 | Baseline-C50-PBMT |
| Sense 2 | 1174 | SMT | NO | 25.31 | 0.808484 | 0.571890 | -2.75 | Clustercat-C50-PBMT |
| ITTB-EN-ID 1 | 1239 | SMT | NO | 22.35 | 0.808943 | 0.555970 | -9.25 | BLNLM |

Table 32: BPPT-EI submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| Online A | 1031 | Other | YES | 21.37 | 0.714537 | 0.621100 | +44.75 | Online A (2016) |
| Online B | 1048 | Other | YES | 15.58 | 0.683214 | 0.590520 | +14.00 | Online B (2016) |
| SMT Phrase | 1054 | SMT | NO | 10.32 | 0.638090 | 0.574850 | 0.00 | Phrase-based SMT |

Table 33: IITB-HE submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| SMT Phrase | 1252 | SMT | NO | 10.790000 | 0.651166 | 0.660860 | —— | Phrase-based SMT |
| Online A | 1032 | Other | YES | 18.720000 | 0.716788 | 0.670660 | +57.25 | Online A (2016) |
| Online B | 1047 | Other | YES | 16.970000 | 0.691298 | 0.668450 | +42.50 | Online B (2016) |
| EHR 1 | 1166 | SMT | NO | 11.750000 | 0.671866 | 0.650750 | 0.00 | PBSMT with preordering (DL=6) |
| IITP-MT 1 | 1185 | SMT | YES | 13.710000 | 0.688913 | 0.657330 | +4.75 | IITP-MT System1 |

Table 34: IITB-EH submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | AMFM | | | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | juman | kytea | mecab | juman | kytea | mecab | juman | kytea | mecab | | |
| SMT Phrase | 1251 | SMT | NO | 2.05 | 4.17 | 2.42 | 0.440122 | 0.496402 | 0.461763 | 0.360910 | 0.360910 | 0.360910 | —— | Phrase-based SMT |
| Online A | 1064 | Other | YES | 6.60 | 10.42 | 7.47 | 0.565109 | 0.597863 | 0.576725 | 0.495270 | 0.495270 | 0.495270 | +39.75 | Online A (2016) |
| Online B | 1065 | Other | YES | 5.70 | 8.91 | 6.38 | 0.560486 | 0.589558 | 0.571670 | 0.471450 | 0.471450 | 0.471450 | +17.75 | Online B (2016) |
| EHR 1 | 1167 | SMT | YES | 7.81 | 10.12 | 8.11 | 0.579285 | 0.617098 | 0.588723 | 0.468140 | 0.468140 | 0.468140 | +13.75 | PBSMT with phrase table pivoting and pivot language (en) reordering. User dictionary and TED based LM are used. |
| EHR 2 | 1179 | SMT | YES | 7.66 | 9.80 | 7.95 | 0.585953 | 0.618106 | 0.597490 | 0.473120 | 0.473120 | 0.473120 | +10.00 | PBSMT with sentence level pivoting and pivot language (en) reordering. User dictionary and TED based LM are used. |

Table 35: IITB-HJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | AMFM | Pair | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|
| SMT Phrase | 1253 | SMT | NO | 1.590000 | 0.399448 | 0.467980 | —— | Phrase-based SMT |
| Online B | 1066 | Other | YES | 4.210000 | 0.488631 | 0.528220 | +51.75 | Online B (2016) |
| Online A | 1067 | Other | YES | 4.430000 | 0.495349 | 0.525690 | +54.50 | Online A (2016) |

Table 36: IITB-JH 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, March.

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

Fabien Cromieres, Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2016. Kyoto university participation to wat 2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 166–174, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Terumasa Ehara. 2016. Translation systems and experimental results of the ehr group for wat2016 tasks. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 111–118, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Character-based decoding in tree-to-sequence attention-based neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 175–183, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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

Kazuma Hashimoto, Akiko Eriguchi, and Yoshimasa Tsuruoka. 2016. Domain adaptation and attention-based unknown word replacement in chinese-to-japanese neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 75–83, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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.

Kenji Imamura and Eiichiro Sumita. 2016. Nict-2 translation system for wat2016: Applying domain adaptation to phrase-based statistical machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 126–132, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions.

Satoshi Kinoshita, Tadaaki Oshio, Tomoharu Mitsuhashi, and Terumasa Ehara. 2016. Translation using japio patent corpora: Japio at wat2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 133–138, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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.

Hyoung-Gyu Lee, JaeSong Lee, Jun-Seok Kim, and Chang-Ki Lee. 2015. NAVER Machine Translation System for WAT 2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 69–73, Kyoto, Japan, October.

Shaotong Li, JinAn Xu, Yufeng Chen, and Yujie Zhang. 2016. System description of bjtu_nlp neural machine translation system. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 104–110, Osaka, Japan, December. 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, October.

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, October.

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, Makoto Morishita, and Satoshi Nakamura, 2015. *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, chapter Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015, pages 35–41. Workshop on Asian Translation.

Graham Neubig. 2016. Lexicons and minimum risk training for neural machine translation: Naist-cmu at wat2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 119–125, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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, Raj Dabre, Fabien Cromières, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. KyotoEBMT System Description for the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 54–60, Kyoto, Japan, October.

Sukanta Sen, Debajyoty Banik, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Iitp english-hindi machine translation system at wat 2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 216–222, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Raphael Shu and Akiva Miura. 2016. Residual stacking of rnns for neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 223–229, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Sandhya Singh, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2016. Iit bombay ' s english-indonesian submission at wat: Integrating neural language models with smt. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 68–74, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Katsuhito Sudoh and Masaaki Nagata. 2016. Chinese-to-japanese patent machine translation based on syntactic pre-ordering for wat 2016. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 211–215, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Liling Tan. 2016. Faster and lighter phrase-based machine translation baseline. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 184–193, Osaka, Japan, December. The COLING 2016 Organizing Committee.

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.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in japanese-to-english neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Wei Yang and Yves Lepage. 2016. Improving patent translation using bilingual term extraction and re-tokenization for chinese–japanese. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 194–202, Osaka, Japan, December. The COLING 2016 Organizing Committee.