# Overview of the 2nd Workshop on Asian Translation

**Toshiaki Nakazawa**
Japan Science and Technology Agency

**Hideya Mino**
National Institute of Information and Communications Technology

**Isao Goto**
NHK

nakazawa@pa.jst.jp hideya.mino@nict.go.jp goto.i-es@nhk.or.jp

**Graham Neubig**
Nara Institute of Science and Technology

**Sadao Kurohashi**
Kyoto University

**Eiichiro Sumita**
National Institute of Information and Communications Technology

neubig@is.naist.jp kuro@i.kyoto-u.ac.jp eiichiro.sumita@nict.go.jp

## Abstract

This paper presents the results of the shared tasks from the 2nd workshop on Asian translation (WAT2015) including J↔E, J↔C scientific paper translation subtasks and C→J, K→J patent translation subtasks. For the WAT2015, 12 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 workshop WAT2014(Nakazawa et al., 2014), WAT2015 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 2nd WAT, we adopt new translation subtasks "Chinese-to-Japanese and Korean-to-Japanese patent translation" in addition to the subtasks that were conducted in WAT2014.

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 and Korean→Japanese as language pairs. In the future, we will add more Asian languages, such as Vietnamese, Indonesian, Thai, Burmese and so on.
- Evaluation method
  Evaluation is done both automatically and manually. For human evaluation, WAT uses crowdsourcing, which is low cost and allows multiple evaluations, as the first-stage evaluation. Also, JPO adequacy evaluation is conducted for the selected submissions according to the crowdsourcing evaluation results.

## 2 Dataset

WAT uses the Asian Scientific Paper Excerpt Corpus (ASPEC)[1] and JPO Patent Corpus (JPC) [2] 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.

[1] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html



Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 1–28,
Kyoto, Japan, 16th October 2015.

| 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.

#### 2.1.1 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[3]. 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.

[3]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 |

Table 2: Statistics for JPC.

### 2.2 JPC

JPC was constructed by the Japan Patent Office (JPO). It consists of a Chinese-Japanese patent description corpus (JPC-CJ) and Korean-Japanese patent description corpus (JPC-KJ) with four sections, which are Chemistry, Electricity, Mechanical engineering, and Physics, based on International Patent Classification (IPC). Each corpus is separated into training, development, development-test and test data, which are sentence pairs. This corpus was used for patent subtasks C→J and K→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 was 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 developent, 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, the training data was extracted from sentence pairs of Korean-Japanese patent documents published in 2011 and 2012. The development, development-test and test data for JPC-CJ and JPC-KJ were respectively made with 100 patent documents published in 2013.

## 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 2015, which is the same system as that at WAT 2014.

In addition to the results for the baseline phrase-based SMT system, we produced results for the baseline systems that consisted of a hierarchical phrase-based SMT system, a string-to-tree syntax-based SMT system, a tree-to-string syntax-based SMT system, 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[4]. 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 3.

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.

### 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 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[5] for Japanese segmentation.
- Stanford Word Segmenter version 2014-01-04[6] (Chinese Penn Treebank (CTB) model) for Chinese segmentation.
- The Moses toolkit for English tokenization.
- Mecab-ko[7] for Korean 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 $= 20$ except for KJ and distortion-limit $= 0$ for KJ
- 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.

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

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

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

[7] https://bitbucket.org/eunjeon/mecab-ko/

| System ID | System | Type | ASPEC | | | | JPC | |
|---|---|---|---|---|---|---|---|---|
| | | | JE | EJ | JC | CJ | CJ | KJ |
| 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 (August, 2015) | (SMT) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Online X | Bing translator (August and September, 2015) | (SMT) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 3: Baseline Systems

### 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 two popular metrics: BLEU (Papineni et al., 2002) and RIBES (Isozaki et al., 2010). 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 [8]. 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 [9] and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0 [10]. 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 [11] (Tseng, 2005). For Korean segmentation we used mecab-ko [12]. For English segmentation we used tokenizer.perl [13] in the Moses toolkit.

Detailed procedures for the automatic evaluation are shown on the WAT2015 evaluation web page [14].

### 4.2 Automatic Evaluation System

The participants submit translation results via an automatic evaluation system deployed on the WAT2015 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 ($C \rightarrow J$, $K \rightarrow J$);
- Method (SMT, RBMT, SMT and RBMT, EBMT, Other);

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

# WAT
## The Workshop on Asian Translation
## Submission

## SUBMISSION

**Logged in as: ORGANIZER**

Logout

**Submission:**

Human Evaluation: human evaluation

Publish the results of the evaluation: publish

Team Name: ORGANIZER

Task: en-ja

Submission File: ファイルを選択 ファイル未選択

Used Other Resources: used other resources such as parallel corpora, monolingual corpora and parallel dictionaries in addition to ASPEC in Scientific papers subtask or JPO_PATENT_CORPUS in Patent subtask

Method: SMT

System Description (public): 100 characters or less

System Description (private): 100 characters or less

Submit

Guidelines for submission:

- Submitted files should be encoded in UTF-8 format.
- Translated sentences in submitted files should have one sentence per line, corresponding to each test sentence. The number of lines in the submitted file and the corresponding test file should be equal.
- Team Name, Task, Used Other Resources, Method, System Description (public) , Date and Time(JST), BLEU and RIBES will be disclosed on the Evaluation Site when you upload a file checking "Publish the results of the evaluation".
- JPCzh-ja and JPCko-ja in "Task" is the task with JPO_PATENT_CORPUS.
- If you want to submit the file for human evaluation, check the box "Human Evaluation". Once you upload a file with checking "Human Evaluation" you can not change the file used for human evaluation.
- When you submit the translation results for human evaluation, please check the checkbox of "Publish" too.
- You can submit files for human evaluation "twice" per task.
- One of the files for human evaluation are recommended not to use other resources, but not compulsory.
- You can modify some fields of submitted data. Read the "Guidelines for submitted data" below.
- The latest versions of Chrome, Firefox, Internet Explorer and Safari are supported for this site.
- To submit on this site, You need to have JavaScript enabled in your browser.

Back to top

**Submitted Data:**

Update Configuration of Submitted Data

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

Figure 1: The submission web page for participants

- Use of other resources in addition to ASPEC or JPC;
- Permission to publish the automatic evaluation scores on the WAT2015 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 WAT2015. Anybody can register to use the system on the registration web page [15].

## 5 Human Evaluation

In WAT2015, we conducted 2 kinds of human evaluations: *pairwise crowdsourcing evaluation* and *JPO adequacy evaluation*.

### 5.1 Pairwise Crowdsourcing Evaluation

The pairwise crowdsourcing evaluation is the same as the last year. We used Lancers as the crowdsourcing platform. There are two reasons of choosing Lancers. One is that we can set the category of the crowdsourcing task ('Translation' in this case). We can reach the appropriate workers by choosing the appropriate categories. The other reason is that we can assign the task to identity-verified workers. This function guaranteed the quality of the workers. These two advantages ensure a high evaluation quality.

We used the same sentences as the last year for the pairwise crowdsourcing evaluation. We randomly chose *documents* from the Test set from the ASPEC data, for a total of 400 sentence pairs for JE and JC. We excluded documents containing sentences longer than 100 Japanese characters. Each submission is compared with the baseline translation (Phrase-based SMT, described in Section 3) and given a *Crowd* score[16].

#### 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 workers, and the workers 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. Figure 2 illustrates the evaluation.

#### 5.1.2 Voting

The crowdsourcing workers are not specialists, and thus the quality of the judgments are not necessarily precise. To guarantee the quality of the evaluations, each sentence is evaluated by 5 different workers and the final decision is made depending on the 5 judgements[17]. 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 Crowd 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 Crowd score can be calculated by the following formula:

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

From the definition, the Crowd 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 Crowd score of the selected sentences
2. iterate the previous step 1000 times and get 1000 Crowd scores
3. sort the 1000 scores and estimate the 95% confidence interval by discarding the top 25 scores and the bottom 25 scores

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

[16] It was called HUMAN score in WAT2014.

[17] We used 3 judgements in WAT2014.

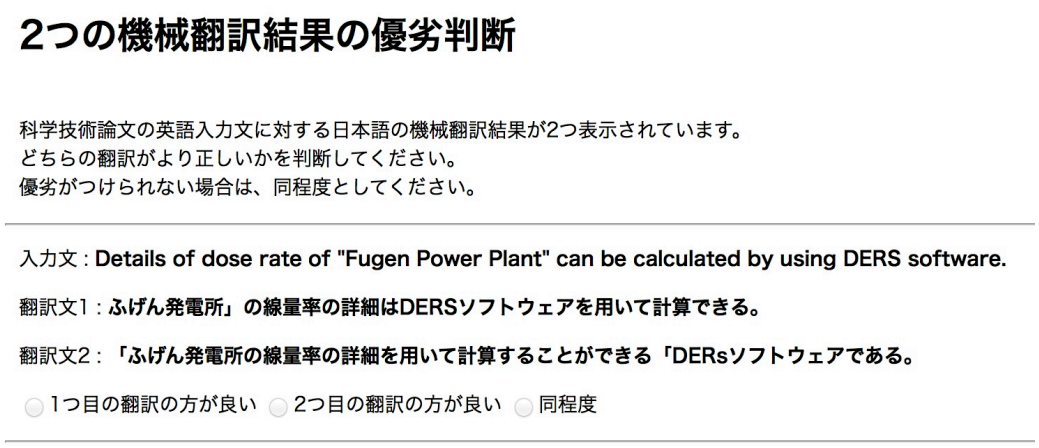


Figure 2: Illustration of the crowdsourcing evaluation. The workers are asked to judge which translation is better, or the same.

#### 5.1.5 Cost

A major benefit of using crowdsourcing is that it reduces the cost of evaluations. In WAT2015, one judgment costs 5 JPY. The evaluation of a submission requires 5 (judgments) $\times$ 400 (sentence pairs) = 2,000 judgments and costs 5 $\times$ 2,000 = 10,000 JPY. The time for the evaluation differs depending on the translation direction. On average, one evaluation takes a couple of days.

### 5.2 JPO Adequacy Evaluation

The participants' systems, which achieved the top 3 highest scores among the pairwise crowdsourcing evaluation results of each subtask, 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. In addition to the top 3 systems, the Sense 1 system of the JPC-KJ subtask, which was the lower score on the pairwise crowdsourcing evaluation despite the high score on the automatic evaluation, was evaluated exceptionally. 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 | 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 4: The JPO adequacy criterion

#### 5.2.2 Evaluation Criterion

Table 4 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) [18].

## 6 Participants List

Table 5 shows the list of participants for WAT2015. This includes not only Japanese organizations, but also some organizations from outside Japan. 12 teams submitted one or more translation results to the both automatic evaluation server and human evaluation.

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

| Team ID | Organization | ASPEC | | | | JPC | |
|---|---|---|---|---|---|---|---|
| | | JE | EJ | JC | CJ | CJ | KJ |
| NAIST (Neubig et al., 2015) | Nara Institute of Science and Technology | ✓ | ✓ | ✓ | ✓ | | |
| Kyoto-U (Richardson et al., 2015) | Kyoto University | ✓ | ✓ | ✓ | ✓ | ✓ | |
| WEBLIO_MT (Zhu, 2015) | Weblio, Inc. | | ✓ | | | | |
| TMU (Matsuo et al., 2015) | Tokyo Metropolitan University | ✓ | | | | | |
| BJTUNLP (Shan et al., 2015) | Beijing Jiaotong University | | | | ✓ | | |
| Sense (Tan et al., 2015) | Saarland University & Nanyang Technological University | ✓ | ✓ | | | | ✓ |
| NICT (Ding et al., 2015) | National Institute of Information and Communication Technology | ✓ | | | | | ✓ |
| TOSHIBA (Sonoh and Kinoshita, 2015) | Toshiba Corporation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| WASUIPS (Yang et al., 2015) | Waseda University | | | | | ✓ | |
| naver (Lee et al., 2015) | NAVER Corporation | | ✓ | | | | ✓ |
| EHR (Ehara, 2015) | Ehara NLP Research Laboratory | | ✓ | | ✓ | ✓ | ✓ |
| ntt (Sudoh and Nagata, 2015) | NTT Communication Science Laboratories | | | | | ✓ | |

Table 5: List of participants who submitted translation results to WAT2015 and their participation in each subtasks.

## 7 Evaluation Results

In this section, the evaluation results for WAT2015 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2015 website[19].

### 7.1 Official Automatic Evaluation Results

Figures 3 and 4 show the official automatic evaluation results for the representative submissions and baseline systems. The automatic evaluation results for all the submissions are shown in Section Appendix A.

### 7.2 Official Crowdsourcing Evaluation Results

#### Crowd Score

Figure 5 shows the official crowdsourcing evaluation results. The error bars in the figures show the 95% confidence interval (see Section 5.1.4). Note that overlapping error bars between two submissions do not necessarily mean that there is no significant difference. If an error bar crosses the x-axis (Crowd score = 0), it means that there is no significant difference between the submission and the baseline (SMT Phrase).

#### Statistical Significance Testing between Submissions

Tables 6, 7, 8, 9, 10 and 11 show the results of statistical significance testing for the JE, EJ, JC and CJ translations respectively where all the pairs of submissions are tested. $\ggg$, $\gg$ and $>$ mean that the system in the row is *better* than the system in the column 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 crowdsourcing evaluation sentences, and calculate the Crowd scores on the selected sentences for both systems
2. iterate the previous step 1000 times and count the number of wins ($W$), losses ($L$) and ties ($T$)
3. calculate $p = \frac{L}{W+L}$

#### Inter-annotator Agreement

To assess the reliability of agreement between the crowdsourcing workers, we calculated the Fleiss' $\kappa$ (Fleiss and others, 1971) values. The results are shown in Table 12. We can see that the $\kappa$ values are larger for X $\rightarrow$ J translations than for J $\rightarrow$ X translations. This may be because we used a Japanese crowdsourcing service for the evaluation and so the majority of the crowdsourcing workers are Japanese, and the evaluation of one's mother tongue is much easier than for other languages in general. Also, K $\rightarrow$ J evaluations seem to be more consistent than the other directions. This may be because the quality of K $\rightarrow$ J translations are much better than the other directions.

#### Correlation between Automatic and Crowdsourcing Evaluations

Figure 6 and 7 show the correlations between the automatic evaluation measures (BLEU/RIBES) and the Crowd score.

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

| | Kyoto-U 2 | TOSHIBA 2 | TOSHIBA 1 | RBMT D | Kyoto-U 1 | NICT 1 | NAIST 2 | SMT S2T | NICT 2 | Online D | Sense 1 | Sense 2 | TMU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NAIST 1 | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 2 | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| TOSHIBA 2 | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| TOSHIBA 1 | | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| RBMT D | | | | | - | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 1 | | | | | | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT 1 | | | | | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST 2 | | | | | | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| SMT S2T | | | | | | | | | - | ⋙ | ⋙ | ⋙ | ⋙ |
| NICT 2 | | | | | | | | | | ⋙ | ⋙ | ⋙ | ⋙ |
| Online D | | | | | | | | | | | ⋙ | ⋙ | ⋙ |
| Sense 1 | | | | | | | | | | | | ⋙ | ⋙ |
| Sense 2 | | | | | | | | | | | | | ⋙ |

Table 6: Statistical significance testing of the ASPEC-JE Crowd scores.

| | WEBLIO_MT 1 | naver 2 | Kyoto-U 2 | NAIST 2 | naver 1 | WEBLIO MT 2 | Kyoto-U 1 | TOSHIBA | Online A | EHR | SMT T2S | RBMT B | Sense 2 | Sense 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NAIST 1 | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| WEBLIO_MT 1 | | - | > | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| naver 2 | | | - | > | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 2 | | | | - | > | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST 2 | | | | | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| naver 1 | | | | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| WEBLIO_MT 2 | | | | | | | > | > | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 1 | | | | | | | | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| TOSHIBA | | | | | | | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Online A | | | | | | | | | | - | ⋙ | ⋙ | ⋙ | ⋙ |
| EHR | | | | | | | | | | | - | ⋙ | ⋙ | ⋙ |
| SMT T2S | | | | | | | | | | | | ⋙ | ⋙ | ⋙ |
| RBMT B | | | | | | | | | | | | | ⋙ | ⋙ |
| Sense 2 | | | | | | | | | | | | | | ⋙ |

Table 7: Statistical significance testing of the ASPEC-EJ Crowd scores.

### 7.3 Chronological Evaluation

Figure 8 shows the chronological evaluation results of ASPEC. The x-axis indicates the BLEU score and the y-axis indicates the Crowd score. Note that the first 3 annotations among 5 by the crowdsourcing were used, and the decision for each sentence is made by the same criteria in WAT2014 for calculating the Crowd score of WAT2015 submissions.

### 7.4 Official JPO Adequacy Evaluation Results

Table 13 and Figure 9 show the JPO Adequacy Evaluation results for the selected submissions. The weights for the weighted $\kappa$ (Cohen, 1968) is defined as $|Evaluation1 - Evaluation2|/4$.

As described in Section 5.2, we selected top 3 teams per subtask for the JPO adequacy evaluation according to the pairwise crowdsourcing evaluation results. However, for JPC K $\rightarrow$ J, we exceptionally selected 4 teams including the top 3 teams and the Sense team. This is because the Sense team achieved notably better automatic evaluation scores, but got the worst crowdsourcing evaluation result, and we thought this is a very interesting case.

Figure 10 shows the summary of automatic and human evaluations for the selected submissions. We can see that all of Crowd, BLEU and RIBES scores partially correlate to the JPO adequacy score, but none of them perfectly correlates. Especially, for JPC-KJ, both BLEU and RIBES failed to correctly evaluate the quality of Sense team.

| | Kyoto-U 1 | Kyoto-U 2 | SMT S2T | NAIST 1 | NAIST 2 | TOSHIBA 1 | RBMT B | Online D |
|---|---|---|---|---|---|---|---|---|
| TOSHIBA 2 | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 1 | | > | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 2 | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| SMT S2T | | | | - | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST 1 | | | | | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST 2 | | | | | | - | ⋙ | ⋙ |
| TOSHIBA 1 | | | | | | | ⋙ | ⋙ |
| RBMT B | | | | | | | | > |

Table 8: Statistical significance testing of the ASPEC-JC Crowd scores.

| | NAIST 2 | EHR | Kyoto-U 2 | TOSHIBA 1 | SMT T2S | Kyoto-U 1 | BJTUNLP | TOSHIBA 2 | Online A | RBMT A |
|---|---|---|---|---|---|---|---|---|---|---|
| NAIST 1 | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| NAIST 2 | | - | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| EHR | | | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 2 | | | | - | - | - | ⋙ | ⋙ | ⋙ | ⋙ |
| TOSHIBA 1 | | | | | - | - | ⋙ | ⋙ | ⋙ | ⋙ |
| SMT T2S | | | | | | - | ⋙ | ⋙ | ⋙ | ⋙ |
| Kyoto-U 1 | | | | | | | ⋙ | ⋙ | ⋙ | ⋙ |
| BJTUNLP | | | | | | | | ⋙ | ⋙ | ⋙ |
| TOSHIBA 2 | | | | | | | | | ⋙ | ⋙ |
| Online A | | | | | | | | | | ⋙ |

Table 9: Statistical significance testing of the ASPEC-CJ Crowd scores.

From the evaluation results, the following can be observed (see Figure 10):

- Neural Network based re-ranking is effective (NAIST, Kyoto-U, naver).
- The top SMT outperformed RBMT for CJ and KJ patent translation.
- K→J patent translation achieved high scores for automatic and human evaluations.
- A new problem of automatic evaluation was found in the KJ evaluation.

## 8 Submitted Data

The number of published automatic evaluation results for the twelve teams exceeded 400 before the start of WAT2015, and 56 translation results for pairwise crowsdourcing evaluation were submitted by twelve teams. Furthermore, we selected 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 WAT2015 machine translation evaluation campaign. 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 conduct context-aware MT evaluations. The test data for WAT are prepared using the paragraphs as the unit, while almost all other evaluation campaigns use the sentences as the unit. Therefore, it is suitable to investigate the importance of context in translation.

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

## Appendix A Submissions

Tables 14, 15, 16, 17, 18, 19, summarize all the submissions listed in the automatic evaluation server at the time of the WAT2015 workshop (16th, October, 2015). The OTHER RESOURCES column shows the use of resources such as parallel corpora, monolingual corpora and

| | TOSHIBA 2 | EHR 1 | SMT T2S | ntt 1 | TOSHIBA 1 | Kyoto-U 1 | EHR 2 | ntt 2 | Online A | WASUIPS | RBMT A |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Kyoto-U 2 | - | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| TOSHIBA 2 | | - | > | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| EHR 1 | | | - | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| SMT T2S | | | | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| ntt 1 | | | | | - | - | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| TOSHIBA 1 | | | | | | - | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| Kyoto-U 1 | | | | | | | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| EHR 2 | | | | | | | | - | ⩢ | ⩢ | ⩢ |
| ntt 2 | | | | | | | | | ⩢ | ⩢ | ⩢ |
| Online A | | | | | | | | | | ⩢ | ⩢ |
| WASUIPS | | | | | | | | | | | ⩢ |

Table 10: Statistical significance testing of the JPC-CJ Crowd scores.

| | naver 2 | NICT 2 | EHR 1 | NICT 1 | naver 1 | EHR 2 | TOSHIBA 2 | Sense 2 | TOSHIBA 1 | SMT Hiero | RBMT A | Sense 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Online A | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| naver 2 | | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| NICT 2 | | | - | > | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| EHR 1 | | | | - | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| NICT 1 | | | | | - | - | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| naver 1 | | | | | | - | > | ⩢ | ⩢ | ⩢ | ⩢ | ⩢ |
| EHR 2 | | | | | | | - | > | ⩢ | > | ⩢ | ⩢ |
| TOSHIBA 2 | | | | | | | | - | - | - | ⩢ | ⩢ |
| Sense 2 | | | | | | | | | - | - | ⩢ | ⩢ |
| TOSHIBA 1 | | | | | | | | | | - | ⩢ | ⩢ |
| SMT Hiero | | | | | | | | | | | ⩢ | ⩢ |
| RBMT A | | | | | | | | | | | | ⩢ |

Table 11: Statistical significance testing of the JPC-KJ Crowd scores.

parallel dictionaries in addition to ASPEC or JPC.

| ASPEC-JE | |
|---|---|
| SYSTEM ID | $\kappa$ |
| NAIST 1 | 0.104 |
| Kyoto-U 2 | 0.070 |
| TOSHIBA 2 | 0.076 |
| TOSHIBA 1 | 0.080 |
| RBMT D | 0.070 |
| Kyoto-U 1 | 0.112 |
| NICT 1 | 0.118 |
| NAIST 2 | 0.109 |
| SMT S2T | 0.046 |
| NICT 2 | 0.078 |
| Online D | 0.044 |
| Sense 1 | 0.077 |
| Sense 2 | 0.073 |
| TMU | 0.065 |
| ave. | 0.080 |

| ASPEC-EJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| NAIST 1 | 0.243 |
| WEBLIO_MT 1 | 0.216 |
| naver 2 | 0.170 |
| Kyoto-U 2 | 0.219 |
| NAIST 2 | 0.187 |
| naver 1 | 0.193 |
| WEBLIO_MT 2 | 0.170 |
| Kyoto-U 1 | 0.208 |
| TOSHIBA | 0.221 |
| Online A | 0.170 |
| EHR | 0.196 |
| SMT T2S | 0.206 |
| RBMT B | 0.171 |
| Sense 2 | 0.200 |
| Sense 1 | 0.140 |
| ave. | 0.194 |

| ASPEC-JC | |
|---|---|
| SYSTEM ID | $\kappa$ |
| TOSHIBA 2 | 0.159 |
| Kyoto-U 1 | 0.022 |
| Kyoto-U 2 | -0.042 |
| SMT S2T | -0.013 |
| NAIST 1 | 0.004 |
| NAIST 2 | 0.021 |
| TOSHIBA 1 | 0.072 |
| RBMT B | -0.031 |
| Online D | 0.019 |
| ave. | 0.023 |

| ASPEC-CJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| NAIST 1 | 0.116 |
| NAIST 2 | 0.066 |
| EHR | 0.155 |
| Kyoto-U 2 | 0.078 |
| TOSHIBA 1 | 0.171 |
| SMT T2S | 0.169 |
| Kyoto-U 1 | 0.043 |
| BJTUNLP | 0.115 |
| TOSHIBA 2 | 0.095 |
| Online A | 0.076 |
| RBMT A | 0.094 |
| ave. | 0.107 |

| JPC-CJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Kyoto-U 2 | 0.087 |
| TOSHIBA 2 | 0.117 |
| EHR 1 | 0.111 |
| SMT T2S | 0.109 |
| ntt 1 | 0.114 |
| TOSHIBA 1 | 0.105 |
| Kyoto-U 1 | 0.057 |
| EHR 2 | 0.148 |
| ntt 2 | 0.042 |
| Online A | 0.027 |
| WASUIPS | 0.044 |
| RBMT A | 0.054 |
| ave. | 0.084 |

| JPC-KJ | |
|---|---|
| SYSTEM ID | $\kappa$ |
| Online A | 0.142 |
| naver 2 | 0.248 |
| NICT 2 | 0.317 |
| EHR 1 | 0.221 |
| NICT 1 | 0.295 |
| naver 1 | 0.388 |
| EHR 2 | 0.213 |
| TOSHIBA 2 | 0.305 |
| Sense 2 | 0.323 |
| TOSHIBA 1 | 0.304 |
| SMT Hiero | 0.246 |
| RBMT A | 0.125 |
| Sense 1 | 0.200 |
| ave. | 0.255 |

Table 12: The Fleiss' kappa values for the crowdsourcing evaluation results.

| ASPEC-JE SYSTEM ID | Annotator A average | Annotator B average | all average | $\kappa$ | weighted $\kappa$ |
|---|---|---|---|---|---|
| NAIST 1 | 3.605 | 4.055 | 3.830 | 0.250 | 0.439 |
| Kyoto-U 2 | 3.495 | 3.805 | 3.650 | 0.296 | 0.475 |
| TOSHIBA 2 | 3.450 | 3.755 | 3.600 | 0.269 | 0.459 |

| ASPEC-EJ SYSTEM ID | Annotator A average | Annotator B average | all average | $\kappa$ | weighted $\kappa$ |
|---|---|---|---|---|---|
| NAIST 1 | 3.865 | 4.220 | 4.043 | 0.377 | 0.535 |
| naver 2 | 3.760 | 4.240 | 4.000 | 0.371 | 0.544 |
| WEBLIO_MT 1 | 3.585 | 4.040 | 3.813 | 0.356 | 0.535 |

| ASPEC-JC SYSTEM ID | Annotator A average | Annotator B average | all average | $\kappa$ | weighted $\kappa$ |
|---|---|---|---|---|---|
| NAIST 1 | 3.600 | 2.740 | 3.170 | 0.151 | 0.296 |
| Kyoto-U 1 | 3.330 | 2.400 | 2.865 | 0.162 | 0.319 |
| TOSHIBA 2 | 3.220 | 2.275 | 2.748 | 0.112 | 0.287 |

| ASPEC-CJ SYSTEM ID | Annotator A average | Annotator B average | all average | $\kappa$ | weighted $\kappa$ |
|---|---|---|---|---|---|
| NAIST 1 | 3.970 | 3.785 | 3.878 | 0.247 | 0.417 |
| Kyoto-U 2 | 3.785 | 3.700 | 3.743 | 0.292 | 0.433 |
| EHR | 3.250 | 3.245 | 3.248 | 0.199 | 0.418 |

| JPC-CJ SYSTEM ID | Annotator A average | Annotator B average | all average | $\kappa$ | weighted $\kappa$ |
|---|---|---|---|---|---|
| Kyoto-U 2 | 3.570 | 3.245 | 3.408 | 0.287 | 0.513 |
| EHR 1 | 3.410 | 3.225 | 3.318 | 0.351 | 0.531 |
| TOSHIBA 2 | 3.405 | 3.095 | 3.250 | 0.349 | 0.557 |

| JPC-KJ SYSTEM ID | Annotator A average | Annotator B average | all average | $\kappa$ | weighted $\kappa$ |
|---|---|---|---|---|---|
| naver 2 | 4.900 | 4.655 | 4.778 | 0.329 | 0.333 |
| NICT 2 | 4.905 | 4.610 | 4.758 | 0.186 | 0.218 |
| Online A | 4.735 | 4.315 | 4.525 | 0.183 | 0.300 |
| Sense 1 | 4.445 | 4.195 | 4.320 | 0.367 | 0.528 |

Table 13: JPO adequacy evaluation results.

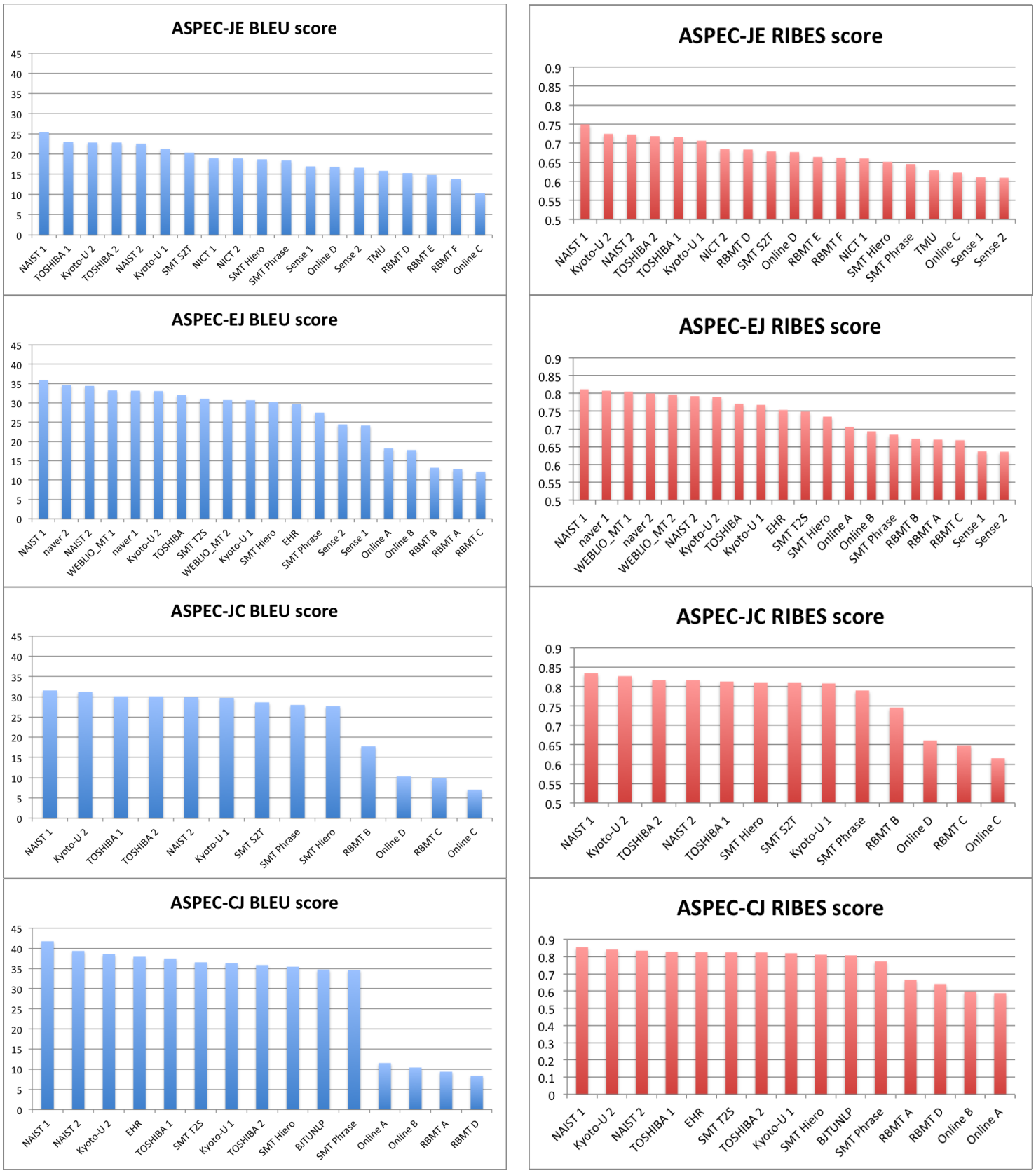


Figure 3: Official automatic evaluation results of ASPEC.

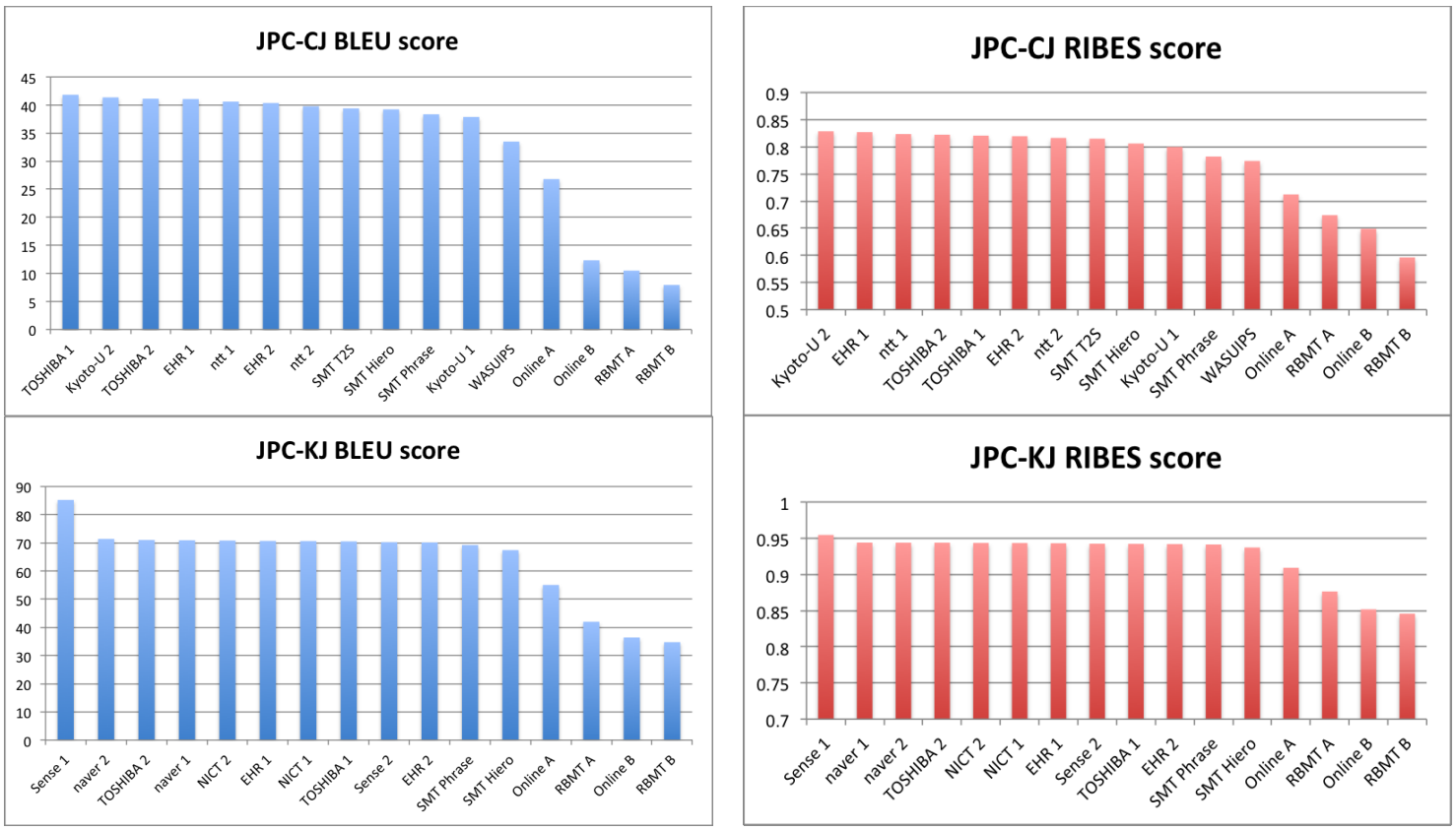


Figure 4: Official automatic evaluation results of JPC.

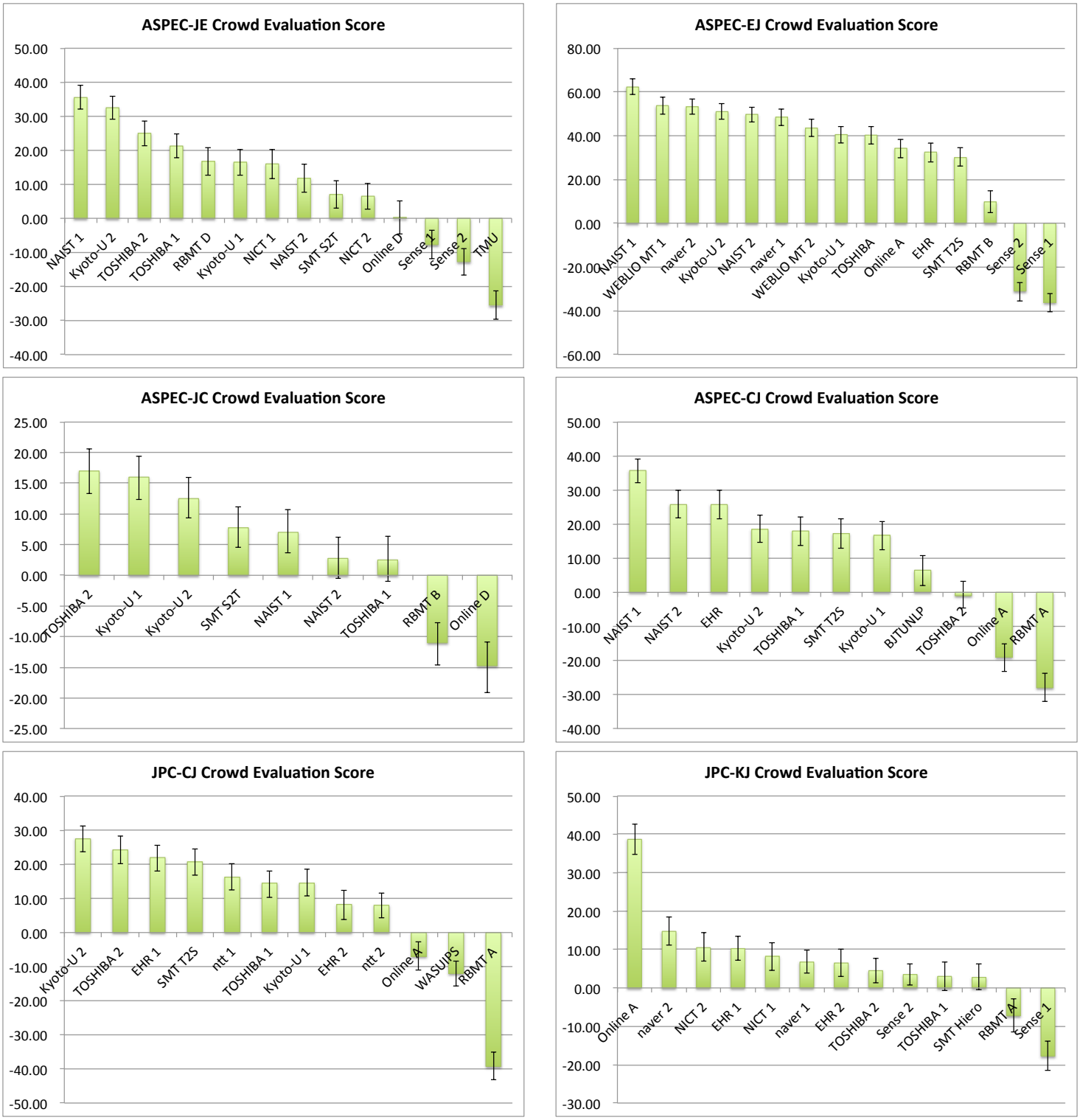


Figure 5: Official pairwise crowdsourcing evaluation results.

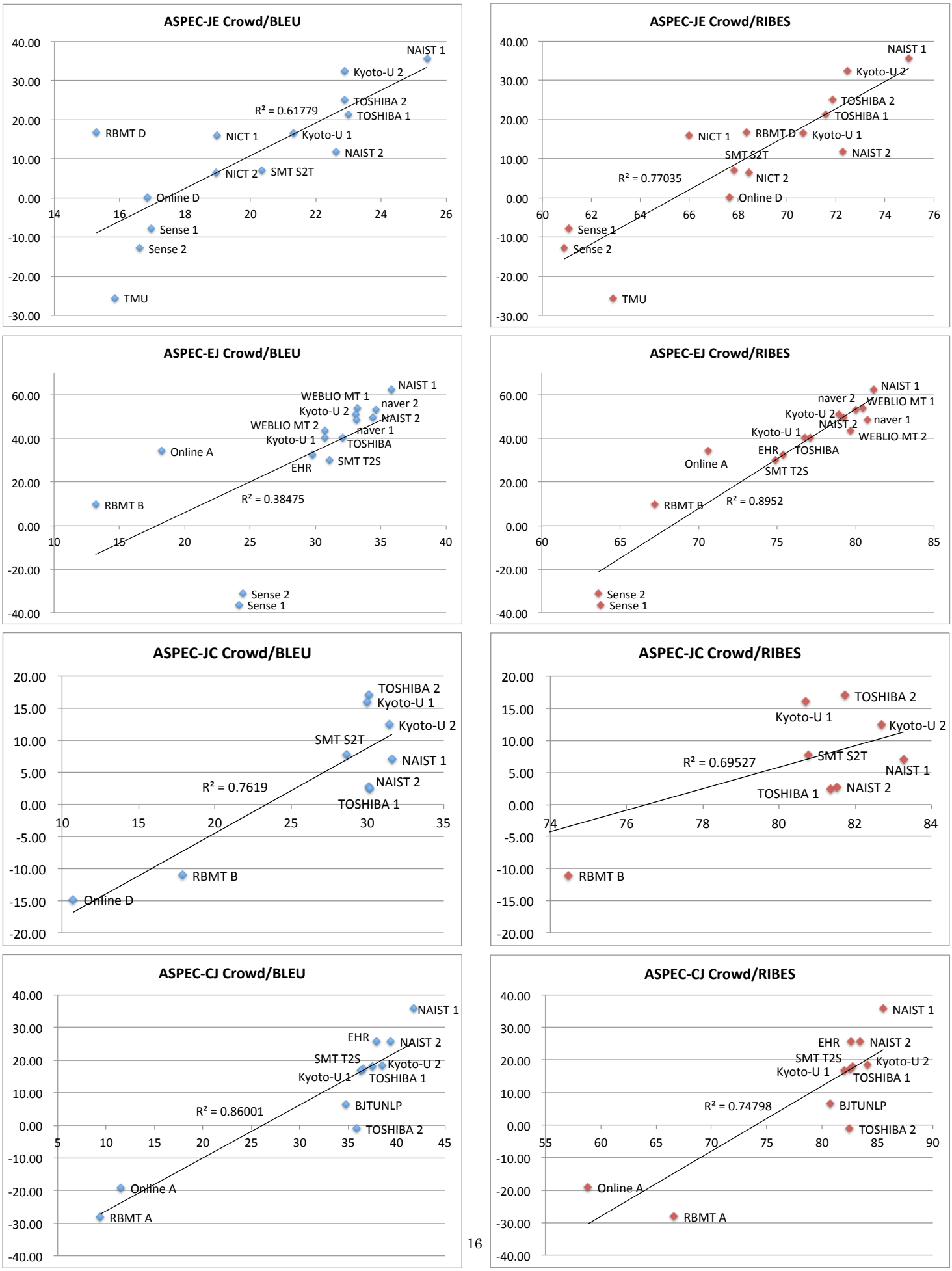




Figure 6: The correlations between the BLEU/RIBES and Crowd scores of ASPEC.

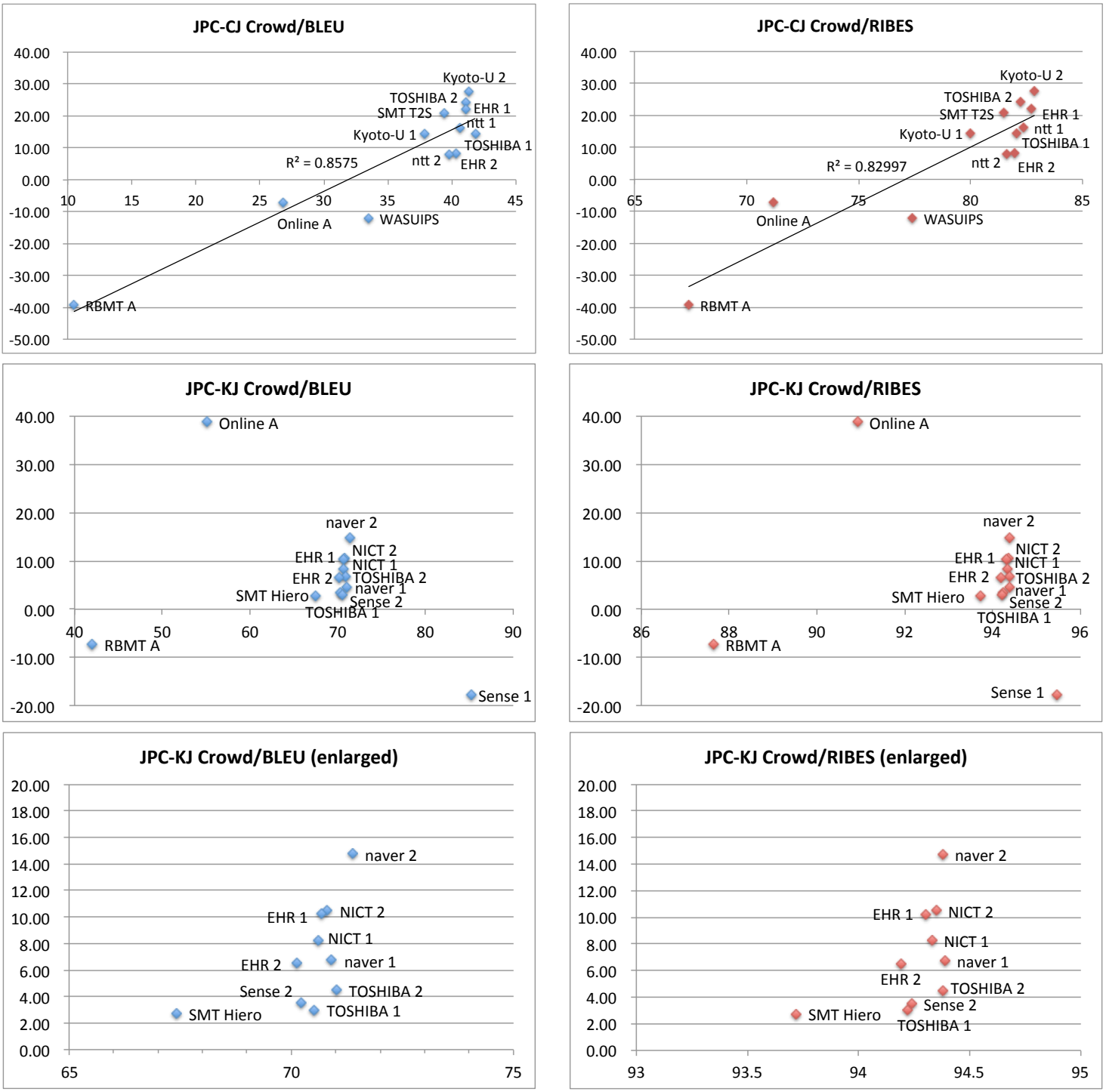


Figure 7: The correlations between the BLEU/RIBES and Crowd scores of JPC.

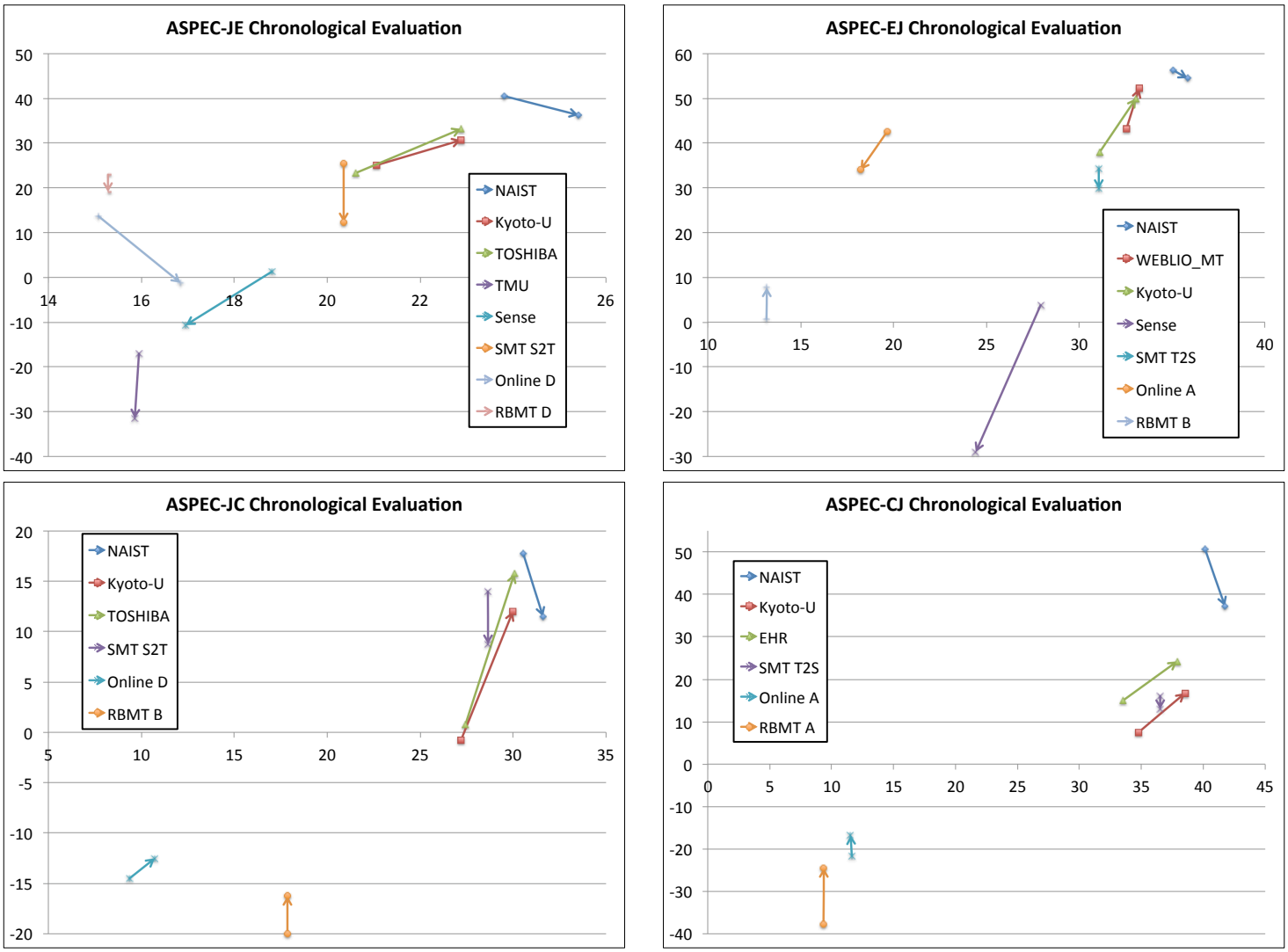


Figure 8: The chronological evaluation results of ASPEC. (The x-axis indicates the BLEU score and the y-axis indicates the Crowd score.)

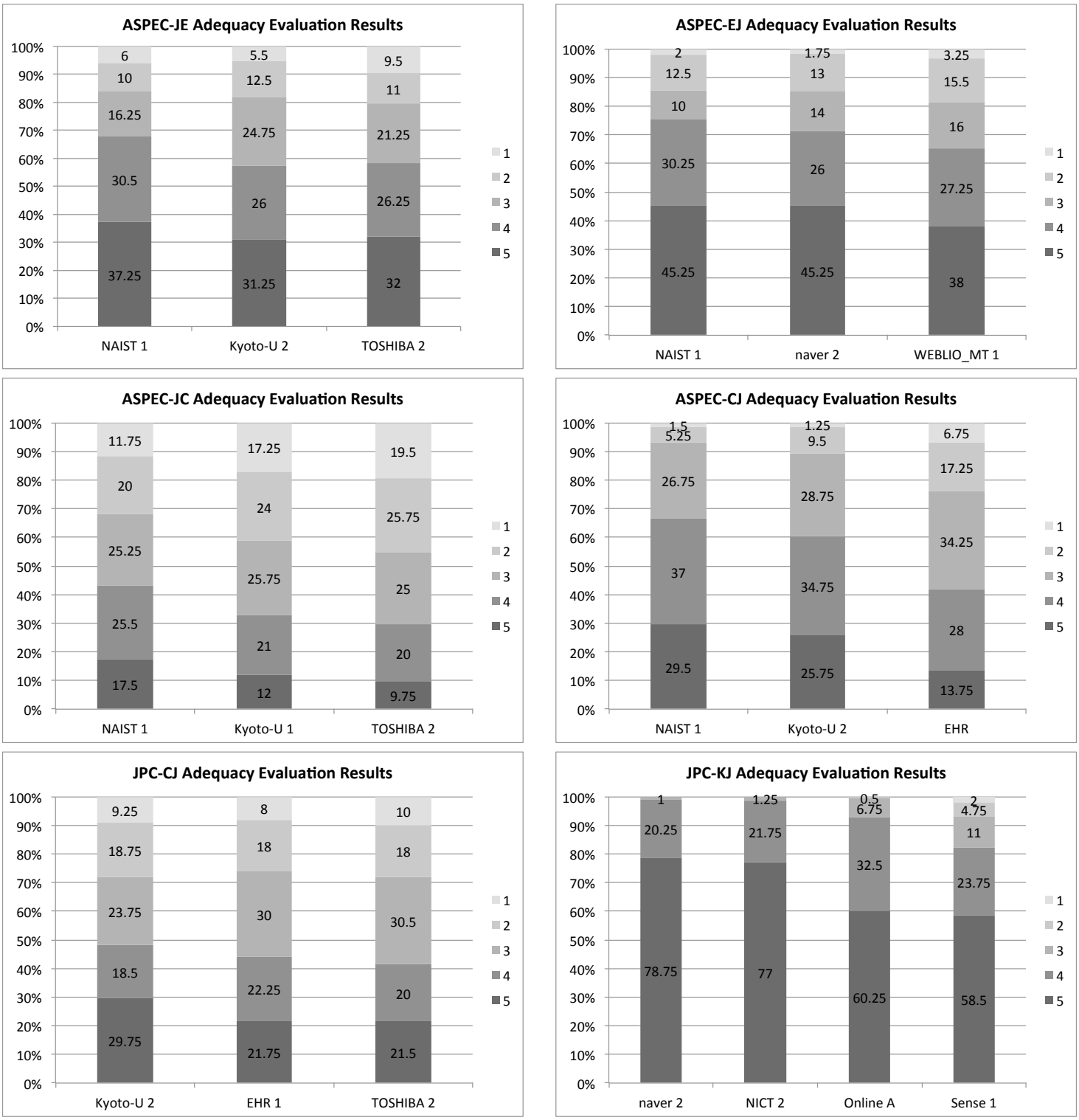


Figure 9: Distribution of JPO adequacy evaluations.

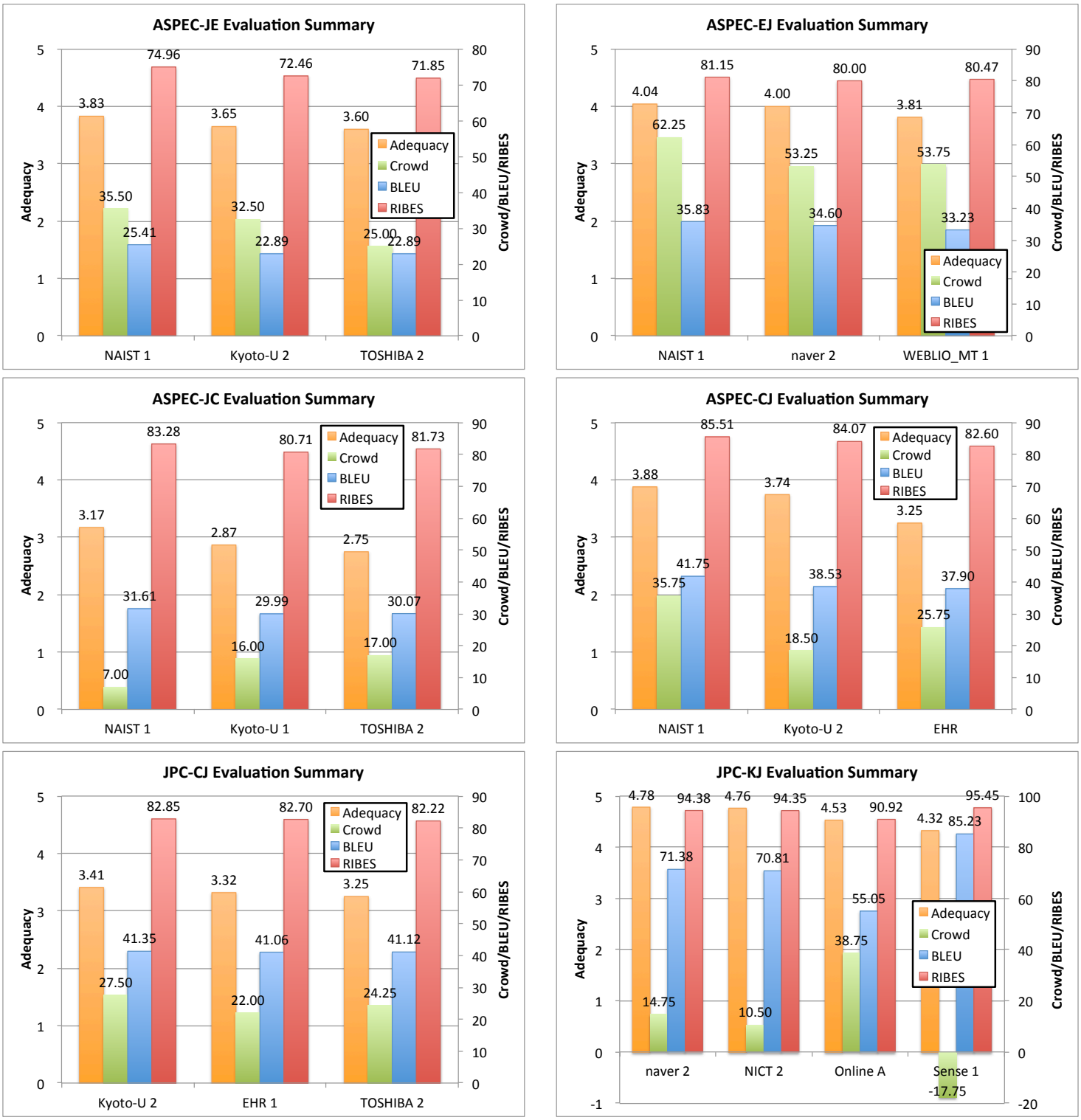


Figure 10: Summary of automatic and human evaluations.

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | RIBES | Crowd | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|
| SMT Hiero | 2 | SMT | NO | 18.72 | 0.651066 | +7.75 | Hierarchical Phrase-based SMT (2014) |
| SMT Phrase | 6 | SMT | NO | 18.45 | 0.645137 | —— | Phrase-based SMT |
| SMT S2T | 9 | SMT | NO | 20.36 | 0.678253 | +25.50 | String-to-Tree SMT (2014) |
| RBMT E | 76 | Other | YES | 14.82 | 0.663851 | —— | RBMT E |
| RBMT F | 79 | Other | YES | 13.86 | 0.661387 | —— | RBMT F |
| Online D | 775 | Other | YES | 16.85 | 0.676609 | +0.25 | Online D (2015) |
| SMT S2T | 877 | SMT | NO | 20.36 | 0.678253 | +7.00 | String-to-Tree SMT (2015) |
| RBMT D | 887 | Other | YES | 15.29 | 0.683378 | +16.75 | RBMT D (2015) |
| Online C | 892 | Other | YES | 10.29 | 0.622564 | —— | Online C (2015) |
| NAIST 1 | 655 | SMT | NO | 25.41 | 0.749573 | +35.50 | Travatar System with NeuralMT Reranking and Parser Self Training |
| NAIST 2 | 766 | SMT | NO | 22.62 | 0.722798 | +11.75 | Travatar System with Parser Self Training |
| Kyoto-U 1 | 796 | EBMT | NO | 21.31 | 0.706480 | +16.50 | KyotoEBMT system without reranking |
| Kyoto-U 2 | 829 | EBMT | NO | 22.89 | 0.724555 | +32.50 | KyotoEBMT system with bilingual RNNLM reranking |
| TMU | 847 | SMT | NO | 15.85 | 0.628897 | -25.50 | PBSMT with dependency based phrase segmentation |
| Sense 1 | 860 | SMT | YES | 16.96 | 0.610775 | -7.75 | Passive JSTx1 |
| Sense 2 | 861 | SMT | YES | 16.61 | 0.609008 | -12.75 | Pervasive JSTx1 |
| NICT 1 | 488 | SMT | NO | 18.98 | 0.659883 | +16.00 | our baseline (DL=6) + dependency-based pre-reordering [Ding+ 2015] |
| NICT 2 | 492 | SMT | NO | 18.96 | 0.684485 | +6.50 | our baseline (DL=9) + reverse pre-reordering [Katz-Brown & Collins 2008] |
| TOSHIBA 1 | 506 | SMT and RBMT | YES | 23.00 | 0.715795 | +21.25 | System combination SMT and RBMT(SPE) with RNNLM language model |
| TOSHIBA 2 | 529 | SMT and RBMT | YES | 22.89 | 0.718540 | +25.00 | RBMT with SPE(Statistical Post Editing) system |

Table 14: ASPEC-JE submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | Crowd | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | juman | kytea | mecab | juman | kytea | mecab | | |
| SMT Phrase | 5 | SMT | NO | 27.48 | 29.80 | 28.27 | 0.683735 | 0.691926 | 0.695390 | — | Phrase-based SMT |
| SMT T2S | 12 | SMT | NO | 31.05 | 33.44 | 32.10 | 0.748883 | 0.758031 | 0.760516 | +34.25 | Tree-to-String SMT (2014) |
| RBMT A | 68 | Other | YES | 12.86 | 14.43 | 13.16 | 0.670167 | 0.676464 | 0.678934 | — | RBMT A |
| RBMT C | 95 | Other | YES | 12.19 | 13.32 | 12.14 | 0.668372 | 0.672645 | 0.676018 | — | RBMT C |
| SMT Hiero | 367 | SMT | NO | 30.19 | 32.56 | 30.94 | 0.734705 | 0.746978 | 0.747722 | +31.50 | Hierarchical Phrase-based SMT (2014) |
| Online A | 774 | Other | YES | 18.22 | 19.77 | 18.46 | 0.705882 | 0.713960 | 0.718150 | +34.25 | Online A (2015) |
| SMT T2S | 875 | SMT | NO | 31.05 | 33.44 | 32.10 | 0.748883 | 0.758031 | 0.760516 | +30.00 | Tree-to-String SMT (2015) |
| RBMT B | 883 | Other | YES | 13.18 | 14.85 | 13.48 | 0.671958 | 0.680748 | 0.682683 | +9.75 | RBMT B (2015) |
| Online B | 889 | Other | YES | 17.80 | 19.52 | 18.11 | 0.693359 | 0.701966 | 0.703859 | — | Online B (2015) |
| NAIST 1 | 761 | SMT | NO | 35.83 | 38.17 | 36.61 | 0.811479 | 0.813827 | 0.820337 | +62.25 | Travatar System with NeuralMT Reranking |
| NAIST 2 | 763 | SMT | NO | 34.38 | 36.58 | 35.16 | 0.792447 | 0.796489 | 0.802228 | +49.75 | Travatar System Baseline |
| Kyoto-U 1 | 805 | EBMT | NO | 30.69 | 33.25 | 31.71 | 0.767778 | 0.776672 | 0.778358 | +40.50 | KyotoEBMT system without reranking |
| Kyoto-U 2 | 832 | EBMT | NO | 33.06 | 35.57 | 33.99 | 0.789514 | 0.797182 | 0.799979 | +51.00 | KyotoEBMT system with bilingual RNNLM reranking |
| WEBLIO_MT 1 | 786 | SMT | NO | 33.23 | 36.21 | 34.05 | 0.804722 | 0.809065 | 0.814337 | +53.75 | NMT, LSTM Search, 5 ensembles, beam size 20, UNK replacing, System Combination with NMT score (Pick top-1k results from NMT) |
| WEBLIO_MT 2 | 813 | Other | NO | 30.72 | 34.19 | 31.57 | 0.796863 | 0.802666 | 0.807186 | +43.50 | NMT, LSTM Search, Beam Size 20, Ensemble of 2 models, UNK replacing |
| Sense 1 | 700 | SMT | NO | 24.13 | 26.24 | 24.96 | 0.637378 | 0.642789 | 0.647831 | -36.25 | Baseline-dictmt |
| Sense 2 | 715 | SMT | YES | 24.43 | 26.58 | 25.36 | 0.635933 | 0.641517 | 0.646682 | -31.00 | Passive JSTx3 |
| TOSHIBA | 524 | SMT and RBMT | YES | 32.06 | 34.17 | 32.76 | 0.770989 | 0.778570 | 0.780467 | +40.25 | RBMT with SPE(Statistical Post Editing) system |
| naver 1 | 836 | SMT | NO | 33.14 | 35.75 | 33.93 | 0.807280 | 0.811487 | 0.817343 | +48.50 | NMT only |
| naver 2 | 837 | SMT | NO | 34.60 | 36.14 | 35.30 | 0.799966 | 0.803154 | 0.808787 | +53.25 | SMT t2s + Spell correction + NMT reranking |
| EHR | 742 | SMT | NO | 29.78 | 32.36 | 30.71 | 0.753576 | 0.766044 | 0.768105 | +32.50 | Phrase based SMT with preordering. |

Table 15: ASPEC-EJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | Crowd | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | kytea | stanford (ctb) | stanford (pku) | kytea | stanford (ctb) | stanford (pku) | | |
| SMT Hiero | 3 | SMT | NO | 27.71 | 27.70 | 27.35 | 0.809128 | 0.809561 | 0.811394 | +3.75 | Hierarchical Phrase-based SMT (2014) |
| SMT Phrase | 7 | SMT | NO | 27.96 | 28.01 | 27.68 | 0.788961 | 0.790263 | 0.790937 | — | Phrase-based SMT |
| SMT S2T | 10 | SMT | NO | 28.65 | 28.65 | 28.35 | 0.807606 | 0.809457 | 0.808417 | +14.00 | String-to-Tree SMT (2014) |
| RBMT C | 244 | RBMT | NO | 9.62 | 9.96 | 9.59 | 0.642278 | 0.648758 | 0.645385 | — | RBMT C |
| Online D | 777 | Other | YES | 10.73 | 10.33 | 10.08 | 0.660484 | 0.660847 | 0.660482 | -14.75 | Online D (2015) |
| SMT S2T | 881 | SMT | NO | 28.65 | 28.65 | 28.35 | 0.807606 | 0.809457 | 0.808417 | +7.75 | String-to-Tree SMT (2015) |
| RBMT B | 886 | Other | YES | 17.86 | 17.75 | 17.49 | 0.744818 | 0.745885 | 0.743794 | -11.00 | RBMT B (2015) |
| Online C | 891 | Other | YES | 7.44 | 7.05 | 6.75 | 0.611964 | 0.615048 | 0.612158 | — | Online C (2015) |
| NAIST 1 | 838 | SMT | NO | 31.61 | 31.59 | 31.42 | 0.832765 | 0.834245 | 0.833721 | +7.00 | Travatar System with NeuralMT Reranking |
| NAIST 2 | 839 | SMT | NO | 30.06 | 29.92 | 29.73 | 0.815084 | 0.816624 | 0.816462 | +2.75 | Travatar System Baseline |
| Kyoto-U 1 | 778 | EBMT | NO | 29.99 | 29.76 | 29.81 | 0.807083 | 0.808275 | 0.808010 | +16.00 | KyotoEBMT system without reranking |
| Kyoto-U 2 | 793 | EBMT | NO | 31.40 | 31.26 | 31.23 | 0.826986 | 0.826919 | 0.827190 | +12.50 | KyotoEBMT system with bilingual RNNLM reranking |
| TOSHIBA 1 | 505 | SMT and RBMT | YES | 30.17 | 30.15 | 29.89 | 0.813490 | 0.813233 | 0.813441 | +2.50 | SPE(Statistical Post Editing) System |
| TOSHIBA 2 | 676 | SMT and RBMT | YES | 30.07 | 30.14 | 29.83 | 0.817294 | 0.816984 | 0.816981 | +17.00 | System combination SMT and RBMT(SPE) with RNNLM language model + post-processing |

Table 16: ASPEC-JC submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | Crowd | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | juman | kytea | mecab | juman | kytea | mecab | | |
| SMT Hiero | 4 | SMT | NO | 35.43 | 35.91 | 35.64 | 0.810406 | 0.798726 | 0.807665 | +4.75 | Hierarchical Phrase-based SMT (2014) |
| SMT Phrase | 8 | SMT | NO | 34.65 | 35.16 | 34.77 | 0.772498 | 0.766384 | 0.771005 | —— | Phrase-based SMT |
| SMT T2S | 13 | SMT | NO | 36.52 | 37.07 | 36.64 | 0.825292 | 0.820490 | 0.825025 | +16.00 | Tree-to-String SMT (2014) |
| RBMT D | 242 | RBMT | NO | 8.39 | 8.70 | 8.30 | 0.641189 | 0.626400 | 0.633319 | —— | RBMT D |
| Online A | 776 | Other | YES | 11.53 | 12.82 | 11.68 | 0.588285 | 0.590393 | 0.592887 | -19.00 | Online A (2015) |
| SMT T2S | 879 | SMT | NO | 36.52 | 37.07 | 36.64 | 0.825292 | 0.820490 | 0.825025 | +17.25 | Tree-to-String SMT (2015) |
| RBMT A | 885 | Other | YES | 9.37 | 9.87 | 9.35 | 0.666277 | 0.652402 | 0.661730 | -28.00 | RBMT A (2015) |
| Online B | 890 | Other | YES | 10.41 | 11.03 | 10.36 | 0.597355 | 0.592841 | 0.597298 | —— | Online B (2015) |
| NAIST 1 | 834 | SMT | NO | 41.75 | 42.95 | 41.93 | 0.855089 | 0.847746 | 0.854587 | +35.75 | Travatar System with NeuralMT Reranking |
| NAIST 2 | 835 | SMT | NO | 39.36 | 40.51 | 39.47 | 0.834388 | 0.827148 | 0.834130 | +25.75 | Travatar System Baseline |
| Kyoto-U 1 | 844 | EBMT | NO | 36.30 | 37.22 | 36.44 | 0.819743 | 0.814581 | 0.818794 | +16.75 | KyotoEBMT system without reranking |
| Kyoto-U 2 | 845 | EBMT | NO | 38.53 | 39.41 | 38.66 | 0.840681 | 0.834451 | 0.839063 | +18.50 | KyotoEBMT system with bilingual RNNLM reranking |
| BJTUNLP | 862 | SMT | NO | 34.72 | 34.87 | 34.79 | 0.807012 | 0.792488 | 0.802430 | +6.50 | a dependency-to-string model for SMT |
| TOSHIBA 1 | 508 | SMT and RBMT | YES | 37.47 | 37.44 | 37.34 | 0.827291 | 0.817395 | 0.825472 | +18.00 | System combination SMT and RBMT(SPE) with RNNLM language model |
| TOSHIBA 2 | 525 | SMT and RBMT | YES | 35.85 | 36.02 | 35.73 | 0.824740 | 0.815388 | 0.822423 | -1.00 | RBMT with SPE(Statistical Post Editing) system |
| EHR | 720 | SMT and RBMT | YES | 37.90 | 38.68 | 37.98 | 0.826003 | 0.818620 | 0.824806 | +25.75 | System combination of RBMT with user dictionary plus SPE and phrase based SMT with preordering. Candidate selection by language model score. |

Table 17: ASPEC-CJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | Crowd | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | juman | kytea | mecab | juman | kytea | mecab | | |
| SMT Hiero | 430 | SMT | NO | 39.22 | 39.52 | 39.14 | 0.806058 | 0.802059 | 0.804523 | — | Hierarchical Phrase-based SMT |
| SMT Phrase | 431 | SMT | NO | 38.34 | 38.51 | 38.22 | 0.782019 | 0.778921 | 0.781456 | — | Phrase-based SMT |
| SMT T2S | 432 | SMT | NO | 39.39 | 39.90 | 39.39 | 0.814919 | 0.811350 | 0.813595 | +20.75 | Tree-to-String SMT (2015) |
| Online A | 647 | Other | YES | 26.80 | 27.81 | 26.89 | 0.712242 | 0.707264 | 0.711273 | -7.00 | Online A (2015) |
| Online B | 648 | Other | YES | 12.33 | 12.72 | 12.44 | 0.648996 | 0.641255 | 0.648742 | — | Online B (2015) |
| RBMT A | 759 | RBMT | NO | 10.49 | 10.72 | 10.35 | 0.674060 | 0.664098 | 0.667349 | -39.25 | RBMT A (2015) |
| RBMT B | 760 | RBMT | NO | 7.94 | 8.07 | 7.73 | 0.596200 | 0.581837 | 0.586941 | — | RBMT B |
| Kyoto-U 1 | 781 | EBMT | NO | 37.87 | 38.62 | 37.71 | 0.799730 | 0.797700 | 0.798979 | +14.50 | Baseline w/o reranking |
| Kyoto-U 2 | 864 | EBMT | NO | 41.35 | 41.92 | 41.16 | 0.828543 | 0.824199 | 0.827230 | +27.50 | KyotoEBMT system with bilingual RNNLM reranking (only character-base model) |
| TOSHIBA 1 | 504 | SMT and RBMT | YES | 41.82 | 41.90 | 41.60 | 0.820568 | 0.813536 | 0.817614 | +14.50 | Combination of phrase-based SMT and SPE systems. |
| TOSHIBA 2 | 526 | SMT and RBMT | YES | 41.12 | 40.87 | 40.59 | 0.822268 | 0.814249 | 0.818981 | +24.25 | RBMT with SPE(Statistical Post Editing) system |
| WASUIPS | 853 | SMT | NO | 33.48 | 34.55 | 33.55 | 0.773985 | 0.771099 | 0.772202 | -12.00 | Combining sampling-based alignment and bilingual hierarchical sub-sentential alignment methods. |
| EHR 1 | 671 | SMT and RBMT | YES | 41.06 | 42.24 | 41.15 | 0.826987 | 0.821983 | 0.825056 | +22.00 | System combination of RBMT with user dictionary plus SPE and phrase based SMT with preordering. Candidate selection by language model score. |
| EHR 2 | 828 | SMT and RBMT | YES | 40.35 | 40.16 | 39.92 | 0.819516 | 0.812982 | 0.816743 | +8.25 | RBMT with user dictionary plus SPE |
| ntt 1 | 736 | SMT | NO | 40.60 | 41.10 | 40.63 | 0.823436 | 0.820252 | 0.822026 | +16.25 | A pre-ordering-based PBMT with patent-tuned dependency parsing and phrase table smoothing. |
| ntt 2 | 811 | SMT | NO | 39.77 | 40.08 | 39.88 | 0.816288 | 0.811911 | 0.815543 | +8.00 | A pre-ordering-based PBMT with patent-tuned dependency parsing, learning-based pre-ordering, and phrase table smoothing. |

Table 18: JPC-CJ submissions

| SYSTEM ID | ID | METHOD | OTHER RESOURCES | BLEU | | | RIBES | | | Crowd | SYSTEM DESCRIPTION |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | juman | kytea | mecab | juman | kytea | mecab | | |
| SMT Phrase | 438 | SMT | NO | 69.22 | 70.36 | 69.73 | 0.941302 | 0.939729 | 0.940756 | — | Phrase-based SMT |
| SMT Hiero | 439 | SMT | NO | 67.41 | 68.65 | 68.00 | 0.937162 | 0.935903 | 0.936570 | +2.75 | Hierarchical Phrase-based SMT (2015) |
| Online B | 651 | Other | YES | 36.41 | 38.72 | 37.01 | 0.851745 | 0.852263 | 0.851945 | — | Online B (2015) |
| Online A | 652 | Other | YES | 55.05 | 56.84 | 55.46 | 0.909152 | 0.909385 | 0.908838 | +38.75 | Online A (2015) |
| RBMT A | 653 | Other | YES | 42.00 | 43.97 | 42.45 | 0.876396 | 0.873734 | 0.875146 | -7.25 | RBMT A (2015) |
| RBMT B | 654 | Other | YES | 34.74 | 37.51 | 35.54 | 0.845712 | 0.849014 | 0.846228 | — | RBMT B |
| Sense 1 | 657 | SMT | NO | 85.23 | 85.20 | 85.23 | 0.954506 | 0.954448 | 0.954435 | -17.75 | Unicode2String with devtest MERT run2 |
| Sense 2 | 833 | SMT | NO | 70.23 | 71.11 | 70.51 | 0.942415 | 0.940887 | 0.941687 | +3.50 | Baseline with only train.ja/ko |
| NICT 1 | 501 | SMT | NO | 70.62 | 71.52 | 70.92 | 0.943348 | 0.942402 | 0.942748 | +8.25 | our baseline: character-based / PB SMT in MOSES (DL=0, max-phrase-len=9, no lex-reordering) / SRILM (9-gram) |
| NICT 2 | 513 | SMT | NO | 70.81 | 71.70 | 71.11 | 0.943463 | 0.942519 | 0.942904 | +10.50 | our baseline + post-processing of bracket balancing |
| TOSHIBA 1 | 554 | SMT and RBMT | YES | 70.51 | 70.84 | 70.71 | 0.942183 | 0.939545 | 0.941471 | +3.00 | System combination SMT, SPE and RBMT with RNNLM + post-processing |
| TOSHIBA 2 | 568 | SMT | YES | 71.01 | 71.44 | 71.26 | 0.943794 | 0.941287 | 0.943181 | +4.50 | Phrase-based SMT with RNNLM reranking + post-processing |
| naver 1 | 630 | SMT | NO | 70.91 | 71.76 | 71.18 | 0.943928 | 0.942800 | 0.943376 | +6.75 | combined two phrase-based systems, post-processing |
| naver 2 | 816 | SMT | NO | 71.38 | 72.27 | 71.68 | 0.943814 | 0.942805 | 0.943584 | +14.75 | SMT PB + NMT reranking |
| EHR 1 | 500 | SMT and RBMT | YES | 70.67 | 71.52 | 70.93 | 0.943040 | 0.941893 | 0.942771 | +10.25 | System combination of RBMT with user dictionary plus SPE and phrase based SMT. Candidate selection by language model score. |
| EHR 2 | 831 | SMT and RBMT | YES | 70.13 | 70.86 | 70.35 | 0.941887 | 0.940341 | 0.941517 | +6.50 | RBMT with user dictionary plus SPE |

Table 19: JPC-KJ submissions

## References


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

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2015. NICT at WAT 2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 42–47, Kyoto, Japan, October.

Terumasa Ehara. 2015. System Combination of RBMT plus SPE and Preordering plus SMT. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 29–34, Kyoto, Japan, October.

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

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

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 152–159.

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

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

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

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

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

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.

Junki Matsuo, Kenichi Ohwada, and Mamoru Komachi. 2015. Source Phrase Segmentation and Translation for Japanese-English Translation Using Dependency Structure. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 99–104, Kyoto, Japan, October.

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.

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. Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 35–41, Kyoto, Japan, October.

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.

Hua Shan, Yujie Zhang, Lu Bai, and Te Luo. 2015. A Dependency-to-String Model for Chinese-Japanese SMT System. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 82–86, Kyoto, Japan, October.

Satoshi Sonoh and Satoshi Kinoshita. 2015. Toshiba MT System Description for the WAT2015 Workshop. In *Proceedings of the 2nd Workshop on Asian*

*Translation (WAT2015)*, pages 48–53, Kyoto, Japan, October.

Katsuhito Sudoh and Masaaki Nagata. 2015. Chinese-to-Japanese Patent Machine Translation based on Syntactic Pre-ordering forWAT 2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 95–98, Kyoto, Japan, October.

Liling Tan, Jon Dehdari, and Josef van Genabith. 2015. An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan, October.

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.

Wei Yang, Zhongwen Zhao, Baosong Yang, and Yves Lepage. 2015. Sampling-based Alignment and Hierarchical Sub-sentential Alignment in Chinese-Japanese Translation of Patents. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 87–94, Kyoto, Japan, October.

Zhongyuan Zhu. 2015. Evaluating Neural Machine Translation in English-Japanese Task. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 61–68, Kyoto, Japan, October.