

# Overview of the 6th Workshop on Asian Translation

Toshiaki Nakazawa  
The University of Tokyo  
nakazawa@logos.t.u-tokyo.ac.jp

Nobushige Doi  
Japan Exchange Group  
n-doi@jpx.co.jp

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

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

Win Pa Pa  
University of Computer Study, Yangon  
winpapa@ucsy.edu.mm

Anoop Kunchukuttan  
Microsoft AI and Research  
anoop.kunchukuttan@microsoft.com

Shantipriya Parida  
Idiap Research Institute  
shantipriya.parida@idiap.ch

Ondřej Bojar  
Charles University, MFF, ÚFAL  
bojar@ufal.mff.cuni.cz

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

## Abstract

This paper presents the results of the shared tasks from the 6th workshop on Asian translation (WAT2019) including Ja↔En, Ja↔Zh scientific paper translation subtasks, Ja↔En, Ja↔Ko, Ja↔En patent translation subtasks, Hi↔En, My↔En, Km↔En, Ta↔En mixed domain subtasks, Ru↔Ja news commentary translation task, and En→Hi multi-modal translation task. For the WAT2019, 25 teams participated in the shared tasks. We also received 10 research paper submissions out of which 7<sup>1</sup> were accepted. About 400 translation results were submitted to the automatic evaluation server, and selected submissions were manually evaluated.

## 1 Introduction

The Workshop on Asian Translation (WAT) is an open evaluation campaign focusing on Asian languages. Following the success of the previous workshops WAT2014-WAT2018 (Nakazawa et al., 2014, 2015, 2016, 2017, 2018), WAT2019 brings together machine

<sup>1</sup>One paper was withdrawn post acceptance and hence only 6 papers will be in the proceedings.

translation researchers and users to try, evaluate, share and discuss brand-new ideas for machine translation. We have been working toward practical use of machine translation among all Asian countries.

For the 6th WAT, we adopted new translation subtasks with Khmer↔English and Tamil↔English mixed domain corpora,<sup>2</sup> Russian↔Japanese news commentary corpus and English→Hindi multi-modal corpus<sup>3</sup> in addition to most of the subtasks of WAT2018.

WAT is a unique workshop on Asian language translation with the following characteristics:

- Open innovation platform  
Due to the fixed and open test data, we can repeatedly evaluate translation systems on the same dataset over years. WAT receives submissions at any time; i.e., there is no submission deadline of translation results w.r.t automatic evaluation of translation quality.

<sup>2</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/>

<sup>3</sup><https://ufal.mff.cuni.cz/hindi-visual-genome/wat-2019-multimodal-task>

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

Table 1: Statistics for ASPEC

- Domain and language pairs  
WAT is the world’s first workshop that targets scientific paper domain, and Chinese↔Japanese and Korean↔Japanese language pairs. In the future, we will add more Asian languages such as Vietnamese, Thai and so on.
- Evaluation method  
Evaluation is done both automatically and manually. Firstly, all submitted translation results are automatically evaluated using three metrics: BLEU, RIBES and AMFM. Among them, selected translation results are assessed by two kinds of human evaluation: pairwise evaluation and JPO adequacy evaluation.

## 2 Datasets

### 2.1 ASPEC

ASPEC was constructed by the Japan Science and Technology Agency (JST) in collaboration with the National Institute of Information and Communications Technology (NICT). The corpus consists of a Japanese-English scientific paper abstract corpus (ASPEC-JE), which is used for ja↔en subtasks, and a Japanese-Chinese scientific paper excerpt corpus (ASPEC-JC), which is used for ja↔zh subtasks. The statistics for each corpus are shown in Table 1.

#### 2.1.1 ASPEC-JE

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

Lang	Train	Dev	DevTest	Test-N
zh-ja	1,000,000	2,000	2,000	5,204
ko-ja	1,000,000	2,000	2,000	5,230
en-ja	1,000,000	2,000	2,000	5,668

Lang	Test-N1	Test-N2	Test-N3	Test-EP
zh-ja	2,000	3,000	204	1,151
ko-ja	2,000	3,000	230	–
en-ja	2,000	3,000	668	–

Table 2: Statistics for JPC

frequency and occurrence ratios for the training data, are described in the README file of ASPEC-JE.

The development, development-test and test data were extracted from parallel sentences from the Japanese-English paper abstracts that exclude the sentences in the training data. Each dataset consists of 400 documents and contains sentences in each field at the same rate. The document alignment was conducted automatically and only documents with a 1-to-1 alignment are included. It is therefore possible to restore the original documents. The format is the same as the training data except that there is no similarity score.

#### 2.1.2 ASPEC-JC

ASPEC-JC is a parallel corpus consisting of Japanese scientific papers, which come from the literature database and electronic journal site J-STAGE by JST, and their translation to Chinese with permission from the necessary academic associations. Abstracts and paragraph units are selected from the body text so as to contain the highest overall vocabulary coverage.

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

## 2.2 JPC

JPO Patent Corpus (JPC) for the patent tasks was constructed by the Japan Patent Office (JPO) in collaboration with NICT. The corpus consists of Chinese-Japanese, Korean-Japanese and English-Japanese patent descriptions whose International Patent Classi-

Disclosure Period	Train	Dev		DevTest		Test	
		Texts	Items	Texts	Items	Texts	Items
2016-01-01 to 2017-12-31	1,089,346 (614,817)	-	-	-	-	-	-
2018-01-01 to 2018-06-30	314,649 (218,495)	1,153 (1,148)	2,845 (2,650)	1,114 (1,111)	2,900 (2,671)	1,153 (1,135)	2,129 (1,763)

Table 3: Statistics for TDDC (The number of unique sentences)

fication (IPC) sections are chemistry, electricity, mechanical engineering, and physics.

At WAT2019, the patent tasks has two subtasks: normal subtask and expression pattern subtask. Both subtasks use common training, development and development-test data for each language pair. The normal subtask for three language pairs uses four test data with different characteristics:

- test-N: union of the following three sets;
- test-N1: patent documents from patent families published between 2011 and 2013;
- test-N2: patent documents from patent families published between 2016 and 2017; and
- test-N3: patent documents published between 2016 and 2017 where target sentences are manually created by translating source sentences.

The expression pattern subtask for zh→ja pair uses test-EP data. The test-EP data consists of sentences annotated with expression pattern categories: title of invention (TIT), abstract (ABS), scope of claim (CLM) or description (DES). The corpus statistics are shown in Table 2. Note that training, development, development-test and test-N1 data are the same as those used in WAT2017.

### 2.3 TDDC

Timely Disclosure Documents Corpus (TDDC) was constructed by Japan Exchange Group (JPX). The corpus was made by aligning the sentences manually from past Japanese and English timely disclosure documents in PDF format published by companies listed on Tokyo Stock Exchange (TSE). Timely Disclosure tasks focus on Japanese to English translation of sentences extracted from timely disclosure documents in order

to avoid mistranslations that would confuse investors.

TSE is one of the largest capital markets in the world that has over 3,600 companies listed as of the end of 2018. Companies are required to disclose material information including financial statements, corporate actions, and corporate governance policies to the public in a timely manner. These timely disclosure documents form an important basis for investment decisions, containing important figures (e.g., sales, profits, significant dates) and proper nouns (e.g., names of persons, places, companies, business and product). Since such information is critical for investors, mistranslations should be avoided and translations should be of a high quality.

The corpus consists of Japanese-English sentence pairs, document hashes, and sentence hashes. A document hash is a hash of the Document ID, which is a unique identifier of the source document. A sentence hash is a hash of the Document ID and the Sentence ID, which is a unique identifier of the sentence in each source document.

The corpus is partitioned into training, development, development-test, and test data. The training data is split into two (2) sets of data from different periods. The first data set was created based on documents disclosed from January 1, 2016 to December 31, 2017, and the second data set was based on documents from January 1, 2018 to June 30, 2018. The development, development-test, and test data set were extracted from timely disclosure documents disclosed from January 1, 2018 to June 30, 2018, excluding documents that were used to create the training data. The documents for the period were randomly selected, and the sentences were extracted from each randomly selected, discrete document set so that the sources extracted are not biased. Therefore, the set of source documents for training, development, development-test and

Lang	Train	Dev	DevTest	Test
en-ja	200,000	2,000	2,000	2,000

Table 4: Statistics for JJI Corpus

Lang	Train	Dev	Test	Mono
hi-en	1,492,827	520	2,507	–
hi-ja	152,692	1,566	2,000	–
hi	–	–	–	45,075,279

Table 5: Statistics for IITB Corpus. “Mono” indicates monolingual Hindi corpus.

test data are independent of each other. Furthermore, each data set of the development, development-test, and test is further split into two (2) sets of data: sentences that end with a Japanese period (。 : U+3002) are classified as ‘Texts’, which has various sentences, and others are classified as ‘Items’, which has many duplicates and similar expressions. The statistics for each corpus are shown in Table 3.

#### 2.4 JJI Corpus

JJI Corpus was constructed by Jiji Press Ltd. in collaboration with NICT. The corpus consists of news text that comes from Jiji Press news of various categories including politics, economy, nation, business, markets, sports and so on. The corpus is partitioned into training, development, development-test and test data, which consists of Japanese-English sentence pairs. The statistics for each corpus are shown in Table 4.

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

#### 2.5 IITB Corpus

IIT Bombay English-Hindi Corpus contains English-Hindi parallel corpus as well as monolingual Hindi corpus collected from a variety of sources and corpora. This corpus had been developed at the Center for Indian Language Technology, IIT Bombay over the years. The corpus is used for mixed domain tasks hi↔en. The statistics for the corpus are shown in Table 5.

#### 2.6 ALT and UCSY Corpus

The parallel data for Myanmar-English translation tasks at WAT2019 consists of two cor-

Corpus	Train	Dev	Test
ALT	18,088	1,000	1,018
UCSY	204,539	–	–
All	222,627	1,000	1,018

Table 6: Statistics for the data used in Myanmar-English translation tasks

pora, the ALT corpus and UCSY corpus.

- The ALT corpus is one part from the Asian Language Treebank (ALT) project (Riza et al., 2016), consisting of twenty thousand Myanmar-English parallel sentences from news articles.
- The UCSY corpus (Yi Mon Shwe Sin and Khin Mar Soe, 2018) is constructed by the NLP Lab, University of Computer Studies, Yangon (UCSY), Myanmar. The corpus consists of 200 thousand Myanmar-English parallel sentences collected from different domains, including news articles and textbooks.

The ALT corpus has been manually segmented into words (Ding et al., 2018, 2019), and the UCSY corpus is unsegmented. A script to tokenize the Myanmar data into writing units is released with the data. The automatic evaluation of Myanmar translation results is based on the tokenized writing units, regardless to the segmented words in the ALT data. However, participants can make a use of the segmentation in ALT data in their own manner.

The detailed composition of training, development, and test data of the Myanmar-English translation tasks are listed in Table 6. Notice that both of the corpora have been modified from the data used in WAT2018.

#### 2.7 ALT and ECCC Corpus

The parallel data for Khmer-English translation tasks at WAT2019 consists of two corpora, the ALT corpus and ECCC corpus.

- The ALT corpus is one part from the Asian Language Treebank (ALT) project (Riza et al., 2016), consisting of twenty thousand Khmer-English parallel sentences from news articles.

Corpus	Train	Dev	Test
ALT	18,088	1,000	1,018
ECCC	104,660	–	–
All	122,748	1,000	1,018

Table 7: Statistics for the data used in Khmer-English translation tasks

- The ECCC corpus consists of 100 thousand Khmer-English parallel sentences extracted from document pairs of Khmer-English bi-lingual records in Extraordinary Chambers in the Court of Cambodia, collected by National Institute of Posts, Telecoms & ICT, Cambodia.

The ALT corpus has been manually segmented into words (Ding et al., 2018), and the ECCC corpus is unsegmented. A script to tokenize the Khmer data into writing units is released with the data. The automatic evaluation of Khmer translation results is based on the tokenized writing units, regardless to the segmented words in the ALT data. However, participants can make a use of the segmentation in ALT data in their own manner.

The detailed composition of training, development, and test data of the Khmer-English translation tasks are listed in Table 7.

## 2.8 Multi-Modal Task Corpus

For English→Hindi multi-modal translation task we asked the participants to use the Hindi Visual Genome corpus (HVG, Parida et al., 2019a,b). The statistics of the corpus are given in Table 8. One “item” in the original HVG consists of an image with a rectangular region highlighting a part of the image, the original English caption of this region and the Hindi reference translation. Depending on the track (see 2.8.1 below), some of these item components are available as the source and some serve as the reference or play the role of a competing candidate solution.

HVG Training, D-Test and E-Test sections were accessible to the participants in advance. The participants were explicitly instructed not to consult E-Test in any way but strictly speaking, they could have used the reference translation (which would mean cheating from the evaluation point of view). C-Test was provided only for the task itself: the source side

Dataset	Items	Tokens	
		English	Hindi
Training Set	28,932	143,178	136,722
D-Test	998	4,922	4,695
E-Test (EV)	1,595	7,852	7,535
C-Test (CH)	1,400	8,185	8,665

Table 8: Data for the English→Hindi multi-modal translation task. One item consists of source English sentence, target Hindi sentence, and a rectangular region within an image. The total number of English and Hindi tokens in the dataset also listed. The abbreviations EV and CH are used in the official task names in WAT scoring tables.

was distributed to task participants and the target side was published only after output submission deadline.

Note that the original Visual Genome suffers from a considerable level of noise. Some observed English grammar errors are illustrated in Figure 1. We also took the chance and used our manual evaluation for validating the quality of the captions given the picture, see 8.4.1 below.

The multi-modal task includes three tracks as illustrated in Figure 1:

### 2.8.1 Multi-Modal Task Tracks

1. Text-Only Translation (labeled “TEXT” in WAT official tables): The participants are asked to translate short English captions (text) into Hindi. No visual information can be used. On the other hand, additional text resources are permitted (but they need to be specified in the corresponding system description paper).
2. Hindi Captioning (labeled “HI”): The participants are asked to generate captions in Hindi for the given rectangular region in an input image.
3. Multi-Modal Translation (labeled “MM”): Given an image, a rectangular region in it and an English caption for the rectangular region, the participants are asked to translate the English text into Hindi. Both textual and visual information can be used.



	Text-Only MT	Hindi Captioning	Multi-Modal MT
Image	–		
Source Text	the bird is stand on a tree branch	–	man stand on skateboard
System Output Gloss	चिड़िया एक पेड़ शाखा पर है Bird on a branch of tree	लाल और सफेद चिह्न Red and white sign	स्केटबोर्ड पर मनुष्य स्टेपिंग Man stepping on skateboard
Reference Solution Gloss	पक्षी एक पेड़ की शाखा पर खड़ा है A bird standing on the branch of a tree	सन्केत अंग्रेजी और विदेशी भाषा में लिखे गये हैं A sign is written in English and a foreign language	आदमी स्केटबोर्ड पर खड़ा है A man is standing on a skateboard

Figure 1: An illustration of the three tracks of WAT 2019 Multi-Modal Task. Note the grammatical errors in the English source. The correct sentences would be “The bird is standing on a tree branch.” and “A man is standing on a skateboard.”

Dataset	Sentences	English tokens	Tamil tokens
train	166,871	3,913,541	2,727,174
test	2,000	47,144	32,847
development	1,000	23,353	16,376
total	169,871	3,984,038	2,776,397

Domain	Sentences	English tokens	Tamil tokens
bible	26,792 (15.77%)	703,838	373,082
cinema	30,242 (17.80%)	445,230	298,419
news	112,837 (66.43%)	2,834,970	2,104,896
total	169,871	3,984,038	2,776,397

Table 9: Data for the Tamil↔English task.

## 2.9 EnTam Corpus

For Tamil↔English translation task we asked the participants to use the publicly available EnTam mixed domain corpus<sup>4</sup> (Ramasamy et al., 2012). This corpus contains training, development and test sentences mostly from the news-domain. The other domains are Bible and Cinema. The statistics of the corpus are given in Table 9.

## 2.10 JaRuNC Corpus

For the Russian↔Japanese task we asked participants to use the JaRuNC corpus<sup>5</sup> (Imankulova et al., 2019) which belongs to the news commentary domain. This dataset was manually aligned and cleaned and is trilingual. It can be used to evaluate Russian↔English

<sup>4</sup><http://ufal.mff.cuni.cz/~ramasamy/parallel/html/>

<sup>5</sup><https://github.com/aizhanti/JaRuNC>

Lang.pair	Partition	#sent.	#tokens	#types
Ja↔Ru	train	12,356	341k / 229k	22k / 42k
	development	486	16k / 11k	2.9k / 4.3k
	test	600	22k / 15k	3.5k / 5.6k
Ja↔En	train	47,082	1.27M / 1.01M	48k / 55k
	development	589	21k / 16k	3.5k / 3.8k
	test	600	22k / 17k	3.5k / 3.8k
Ru↔En	train	82,072	1.61M / 1.83M	144k / 74k
	development	313	7.8k / 8.4k	3.2k / 2.3k
	test	600	15k / 17k	5.6k / 3.8k

Table 10: In-Domain data for the Russian–Japanese task.

translation quality as well but this is beyond the scope of this years sub-task. Refer to Table 10 for the statistics of the in-domain parallel corpora. In addition we encouraged the participants to use out-of-domain parallel corpora from various sources such as KFTT,<sup>6</sup> JESC,<sup>7</sup> TED,<sup>8</sup> ASPEC,<sup>9</sup> UN,<sup>10</sup> Yandex<sup>11</sup> and Russian↔English news-commentary corpus.<sup>12</sup>

## 3 Baseline Systems

Human evaluations of most of WAT tasks were conducted as pairwise comparisons between the translation results for a specific baseline system and translation results for each partic-

<sup>6</sup><http://www.phontron.com/kfft/>

<sup>7</sup><https://datarepository.wolframcloud.com/resources/Japanese-English-Subtitle-Corpus>

<sup>8</sup><https://wit3.fbk.eu/>

<sup>9</sup><http://lotus.kuee.kyoto-u.ac.jp/ASPEC/>

<sup>10</sup><https://cms.unov.org/UNCORpus/>

<sup>11</sup><https://translate.yandex.ru/corpus?lang=en>

<sup>12</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/News-Commentary/news-commentary-v14.en-ru.filtered.tar.gz>

ipant’s system. That is, the specific baseline system was the standard for human evaluation. At WAT 2019, we adopted a neural machine translation (NMT) with attention mechanism as a baseline system.

The NMT 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.<sup>13</sup> We also have SMT baseline systems for the tasks that started at WAT 2017 or before 2017. The baseline systems are shown in Tables 11, 12, and 13. SMT baseline systems are described in the WAT 2017 overview paper (Nakazawa et al., 2017). 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 NMT baseline systems.

- All of the training data for each task were used for training except for the ASPEC Japanese–English task. For the ASPEC Japanese–English task, we only used train-1.txt, which consists of one million parallel sentence pairs with high similarity scores.
- All of the development data for each task was used for validation.

### 3.2 Tokenization

We used the following tools for tokenization.

#### 3.2.1 For ASPEC, JPC, TDDC, JIJI, ALT, UCSY, ECCC, and IITB

- Juman version 7.0<sup>14</sup> for Japanese segmentation.
- Stanford Word Segmenter version 2014-01-04<sup>15</sup> (Chinese Penn Treebank (CTB)

<sup>13</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/baseline/baselineSystems.html>

<sup>14</sup><http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

<sup>15</sup><http://nlp.stanford.edu/software/segmenter.shtml>

System ID	System	Type	ASPEC				JPC						
			ja-en	en-ja	ja-zh	zh-ja	ja-en	en-ja	ja-zh	zh-ja	ja-ko	ko-ja	
NMT	OpenNMT’s NMT with attention	NMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SMT Phrase	Moses’ Phrase-based SMT	SMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SMT Hiero	Moses’ Hierarchical Phrase-based SMT	SMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SMT S2T	Moses’ String-to-Tree Syntax-based SMT and Berkeley parser	SMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SMT T2S	Moses’ Tree-to-String Syntax-based SMT and Berkeley parser	SMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	The Honyaku V15 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	ATLAS V14 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	PAT-Transer 2009 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	PC-Transer V13 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	J-Beijing 7 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	Hohrai 2011 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	J Soul 9 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RBMT X	Korai 2011 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Online X	Google translate	Other	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Online X	Bing translator	Other	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AIAYN	Google’s implementation of “Attention Is All You Need”	NMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 11: Baseline Systems I

System ID	System	Type	JLI		ITB				ALT				
			ja-en	en-ja	hi-en	en-hi	hi-ja	ja-hi	my-en	en-my	km-en	en-km	
NMT	OpenNMT's NMT with attention	NMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SMT Phrase	Moses' Phrase-based SMT	SMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SMT Hiero	Moses' Hierarchical Phrase-based SMT	SMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SMT S2T	Moses' Tree-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	PC-Translator V13 (Commercial system)	RBMT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Online X	Google translate	Other	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Online X	Bing translator	Other	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 12: Baseline Systems II

System ID	System	Type	NewsCommentary		TDDC		EnTam		Multimodal	
			ru-ja	ja-ru	ja-en	en-ja	ta-en	en-ta	en-hi	hi-en
NMT	OpenNMT's NMT with attention	NMT	✓	✓	✓	✓	✓	✓	✓	✓
NMT T2T	Tensor2Tensor's Transformer	NMT	✓	✓	✓	✓	✓	✓	✓	✓
NMT OT	OpenNMT-py's Transformer	NMT	✓	✓	✓	✓	✓	✓	✓	✓
Online X	Azure Custom Translator	Other	✓	✓	✓	✓	✓	✓	✓	✓

Table 13: Baseline Systems III



model) for Chinese segmentation.

- The Moses toolkit for English and Indonesian tokenization.
- Mecab-ko<sup>16</sup> for Korean segmentation.
- Indic NLP Library<sup>17</sup> for Indic language segmentation.
- The tools included in the ALT corpus for Myanmar and Khmer segmentation.
- subword-nmt<sup>18</sup> for all languages.

When we built BPE-codes, we merged source and target sentences and we used 100,000 for -s option. We used 10 for vocabulary-threshold when subword-nmt applied BPE.

### 3.2.2 For EnTam, News Commentary

- The Moses toolkit for English and Russian only for the News Commentary data.
- Mecab<sup>19</sup> for Japanese segmentation.
- The EnTam corpus is not tokenized by any external toolkits.
- Both corpora are further processed by tensor2tensor’s internal pre/post-processing which includes sub-word segmentation.

### 3.2.3 For Multi-Modal Task

- Hindi Visual Genome comes untokenized and we did not use or recommend any specific external tokenizer.
- The standard OpenNMT-py sub-word segmentation was used for pre/post-processing for the baseline system and each participant used what they wanted.

## 3.3 Baseline NMT Methods

We used the following NMT with attention for most of the tasks. We used Transformer (Vaswani et al., 2017) (Tensor2Tensor)) for the News Commentary and English↔Tamil tasks and Transformer (OpenNMT-py) for the Multimodal task.

<sup>16</sup><https://bitbucket.org/eunjeon/mecab-ko/>

<sup>17</sup>[https://bitbucket.org/anoopk/indic\\_nlp\\_library](https://bitbucket.org/anoopk/indic_nlp_library)

<sup>18</sup><https://github.com/rsennrich/subword-nmt>

<sup>19</sup><https://taku910.github.io/mecab/>

### 3.3.1 NMT with Attention

We used OpenNMT (Klein et al., 2017) as the implementation of the baseline NMT systems of NMT with attention (System ID: NMT). We used the following OpenNMT configuration.

- encoder\_type = brnn
- brnn\_merge = concat
- src\_seq\_length = 150
- tgt\_seq\_length = 150
- src\_vocab\_size = 100000
- tgt\_vocab\_size = 100000
- src\_words\_min\_frequency = 1
- tgt\_words\_min\_frequency = 1

The default values were used for the other system parameters.

### 3.3.2 Transformer (Tensor2Tensor)

For the News Commentary and English↔Tamil tasks, we used tensor2tensor’s<sup>20</sup> implementation of the Transformer (Vaswani et al., 2017) and use default hyperparameter settings corresponding to the “base” model for all baseline models. The baseline for the News Commentary task is a multilingual model as described in Imankulova et al. (2019) which is trained using only the in-domain parallel corpora. We use the token trick proposed by (Johnson et al., 2017) to train the multilingual model. As for the English↔Tamil task, we train separate baseline models for each translation direction with 32,000 separate sub-word vocabularies.

### 3.3.3 Transformer (OpenNMT-py)

For the Multimodal task, we used the Transformer model (Vaswani et al., 2018) as implemented in OpenNMT-py (Klein et al., 2017) and used the “base” model with default parameters for the multi-modal task baseline. We have generated the vocabulary of 32k sub-word types jointly for both the source and target languages. The vocabulary is shared between the encoder and decoder.

## 4 Automatic Evaluation

### 4.1 Procedure for Calculating Automatic Evaluation Score

We evaluated translation results by three metrics: BLEU (Papineni et al., 2002), RIBES

<sup>20</sup><https://github.com/tensorflow/tensor2tensor>

(Isozaki et al., 2010) and AMFM (Banchs et al., 2015). BLEU scores were calculated using multi-bleu.perl in the Moses toolkit (Koehn et al., 2007). RIBES scores were calculated using RIBES.py version 1.02.4.<sup>21</sup> AMFM scores were calculated using scripts created by the technical collaborators listed in the WAT2019 web page.<sup>22</sup> All scores for each task were calculated using the corresponding reference translations.

Before the calculation of the automatic evaluation scores, the translation results were tokenized or segmented with tokenization/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<sup>23</sup> and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0.<sup>24</sup> For Chinese segmentation, we used two different tools: KyTea 0.4.6 with full SVM Model in MSR model and Stanford Word Segmenter (Tseng, 2005) version 2014-06-16 with Chinese Penn Treebank (CTB) and Peking University (PKU) model.<sup>25</sup> For Korean segmentation, we used mecab-ko.<sup>26</sup> For Myanmar and Khmer segmentations, we used myseg.py<sup>27</sup> and kmseg.py<sup>28</sup>. For English and Russian tokenizations, we used tokenizer.perl<sup>29</sup> in the Moses toolkit. For Hindi and Tamil tokenizations, we used Indic NLP Library.<sup>30</sup> The detailed procedures for the automatic evaluation are shown on the WAT2019 evaluation web page.<sup>31</sup>

<sup>21</sup><http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html>

<sup>22</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/>

<sup>23</sup><http://www.phontron.com/kytea/model.html>

<sup>24</sup><http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz>

<sup>25</sup><http://nlp.stanford.edu/software/segmenter.shtml>

<sup>26</sup><https://bitbucket.org/eunjeon/mecab-ko/>

<sup>27</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/wat2019.my-en.zip>

<sup>28</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/km-en-data/km-en.zip>

<sup>29</sup><https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl>

<sup>30</sup>[https://bitbucket.org/anoopk/indic\\_nlp\\_library](https://bitbucket.org/anoopk/indic_nlp_library)

<sup>31</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html>

## 4.2 Automatic Evaluation System

The automatic evaluation system receives translation results by participants and automatically gives evaluation scores to the uploaded results. As shown in Figure 2, the system requires participants to provide the following information for each submission:

- Human Evaluation: whether or not they submit the results for human evaluation;
- Publish the results of the evaluation: whether or not they permit to publish automatic evaluation scores on the WAT2019 web page.
- Task: the task you submit the results for;
- Used Other Resources: whether or not they used additional resources; and
- Method: the type of the method including SMT, RBMT, SMT and RBMT, EBMT, NMT and Other.

Evaluation scores of translation results that participants permit to be published are disclosed via the WAT2019 evaluation web page. Participants can also submit the results for human evaluation using the same web interface.

This automatic evaluation system will remain available even after WAT2019. Anybody can register an account for the system by the procedures described in the registration web page.<sup>32</sup>

## 4.3 Additional Automatic Scores in Multi-Modal Task

For the multi-modal task, several additional automatic metrics were run aside from the WAT evaluation server, namely: BLEU (now calculated by Moses scorer<sup>33</sup>), characTER (Wang et al., 2016), chrF3 (Popović, 2015), TER (Snover et al., 2006), WER, PER and CDER (Leusch et al., 2006). Except for chrF3 and characTER, we ran Moses tokenizer<sup>34</sup> on the candidate and reference before scoring. For all error metrics, i.e. metrics where better

<sup>32</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/registration/index.html>

<sup>33</sup><https://github.com/moses-smt/mosesdecoder/blob/master/mert/evaluator.cpp>

<sup>34</sup><https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl>

# 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:

Task:

Submission File:  選択されていません

Used Other Resources:  used other resources such as parallel corpora, monolingual corpora and parallel dictionaries in addition to official corpora

Method:

System Description (public):  100 characters or less

System Description (private):  100 characters or less

Guidelines for submission:

- System requirements:
  - The latest versions of Chrome, Firefox, Internet Explorer and Safari are supported for this site.
  - Before you submit files, you need to enable JavaScript in your browser.
- File format:
  - Submitted files should **NOT** be tokenized/segmented. Please check [the automatic evaluation procedures](#).
  - 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 that of the corresponding test file should be the same.
- Tasks:
  - en-ja, ja-en, zh-ja, ja-zh indicate the scientific paper tasks with ASPEC.
  - HINDENen-hi, HINDENhi-en, HINDENja-hi, and HINDENhi-ja indicate the mixed domain tasks with IITB Corpus.
  - JJIen-ja and JJIja-en are the newswire tasks with JIJI Corpus.
  - RECIPE{ALL,TTL,STE,ING}en-ja and RECIPE{ALL,TTL,STE,ING}ja-en indicate the recipe tasks with Recipe Corpus.
  - ALTen-my and ALTmy-en indicate the mixed domain tasks with UCSY and ALT Corpus.
  - INDICen-{bn,hi,ml,ta,te,ur,si} and INDIC{bn,hi,ml,ta,te,ur,si}-en indicate the Indic languages multilingual tasks with Indic Languages Multilingual Parallel Corpus.
  - JPC{N,N1,N2,N3,EP}zh-ja ,JPC{N,N1,N2,N3}ja-zh, JPC{N,N1,N2,N3}ko-ja, JPC{N,N1,N2,N3}ja-ko, JPC{N,N1,N2,N3}en-ja, and JPC{N,N1,N2,N3}ja-en indicate the patent tasks with JPO Patent Corpus. JPCN1{zh-ja,ja-zh,ko-ja,ja-ko,en-ja,ja-en} are the same tasks as JPC{zh-ja,ja-zh,ko-ja,ja-ko,en-ja,ja-en} in WAT2015-WAT2017. AMFM is not calculated for JPC{N,N2,N3} tasks.
- Human evaluation:
  - 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 cannot 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 **two files** for human evaluation per task.
  - One of the files for human evaluation is recommended not to use other resources, but it is not compulsory.
- Other:
  - Team Name, Task, Used Other Resources, Method, System Description (public) , Date and Time(JST), BLEU, RIBES and AMFM will be disclosed on the Evaluation Site when you upload a file checking "Publish the results of the evaluation".
  - You can modify some fields of submitted data. Read "Guidelines for submitted data" at the bottom of this page.

[Back to top](#)

Figure 2: The interface for translation results submission

scores are lower, we reverse the score by taking  $1 - x$  and indicate this by prepending “n” to the metric name. With this modification, higher scores always indicate a better translation result. Also, we multiply all metric scores by 100 for better readability.

## 5 Human Evaluation

In WAT2019, we conducted three kinds of human evaluations: pairwise evaluation (Section 5.1) and JPO adequacy evaluation (Section 5.2) for text-only language pairs and a pairwise variation of direct assessment (Section 5.3) for the multi-modal task.

## 5.1 Pairwise Evaluation

We conducted pairwise evaluation for participants’ systems submitted for human evaluation. The submitted translations were evaluated by a professional translation company and Pairwise scores were given to the submissions by comparing with baseline translations (described in Section 3).

### 5.1.1 Sentence Selection and Evaluation

For the pairwise evaluation, we randomly selected 400 sentences from the test set of each task. We used the same sentences as the last year for the continuous subtasks. Baseline and submitted translations were shown to annotators in random order with the input source sentence. The annotators were asked to judge which of the translations is better, or whether they are on par.

### 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  $j_i (i = 1, \dots, 5)$  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  $S = \sum j_i$ :

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

### 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:

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

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

### 5.1.4 Confidence Interval Estimation

There are several ways to estimate a confidence interval. We chose to use bootstrap resampling (Koehn, 2004) to estimate the 95%

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 14: The JPO adequacy criterion

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

We conducted JPO adequacy evaluation for the top two or three participants’ systems of pairwise evaluation for each subtask.<sup>35</sup> The evaluation was carried out by translation experts based on the JPO adequacy evaluation criterion, which is originally defined by JPO to assess the quality of translated patent documents.

### 5.2.1 Sentence Selection and Evaluation

For the JPO adequacy evaluation, the 200 test sentences were randomly selected from the 400 test sentences used for the pairwise evaluation. For each test sentence, input source sentence, translation by participants’ system, and reference translation were shown to the annotators. To guarantee the quality of the evaluation, each sentence was evaluated by two annotators. Note that the selected sentences are the same as those used in the previous workshops except for the new subtasks at WAT2019.

### 5.2.2 Evaluation Criterion

Table 14 shows the JPO adequacy criterion from 5 to 1. The evaluation is performed

<sup>35</sup>The number of systems varies depending on the subtasks.

Indicate to what extent each of these candidate translations expresses the meaning of the English source text (independently of the other candidate).

Sentence: 1  
 SRC Text:   
 CAND1 Text:   
 CAND1 Score: worst  best  
 CAND2 Text:   
 CAND2 Score: worst  best

Figure 3: Manual evaluation of text-only translation in the multi-modal task.

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 is described in the JPO document (in Japanese).<sup>36</sup>

### 5.3 Manual Evaluation for the Multi-Modal Task

The evaluations of the three tracks of the multi-modal task follow the Direct Assessment (DA, Graham et al., 2016) technique by asking annotators to assign a score from 0 to 100 to each candidate. The score is assigned using a slider with no numeric feedback, the scale is therefore effectively continuous. After a certain number of scored items, each of the annotators stabilizes in their predictions.

The collected DA scores can be either directly averaged for each system and track (denoted “Ave”), or first standardized per annotator and then averaged (“Ave Z”). The standardization removes the effect of individual differences in the range of scores assigned: the scores are scaled so that the average score of each annotator is 0 and the standard deviation is 1.

Our evaluation differs from the basic DA in the following respects: (1) we run the evaluation bilingually, i.e. we require the annotators to understand the source English sufficiently to be able to assess the adequacy of the Hindi translation, (2) we ask the annotators to score two distinct segments at once, while the original DA displays only one candidate at a time.

The main benefit of bilingual evaluation is that the reference is not needed for the evalu-

<sup>36</sup>[http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku\\_hyouka.htm](http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm)



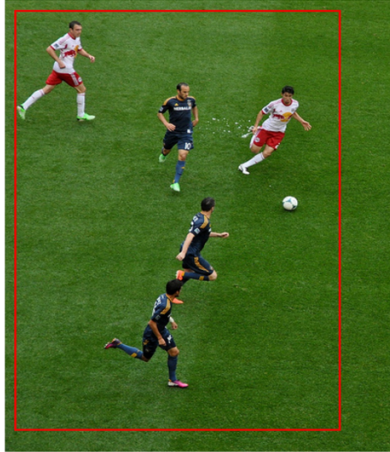
Sentence: 1  
 Is the English text (SRC) a good caption for the highlighted area of the image? :  Yes  No  
 SRC Text:   
 Indicate to what extent each of these candidate translations expresses the meaning of the English source text (independently of the other candidate).  
 CAND1 Text:   
 CAND1 Score: worst  best  
 CAND2 Text:   
 CAND2 Score: worst  best

Figure 4: Manual evaluation of multi-modal translation.

ation. Instead, the reference can be included among other candidates and the manual evaluation allows us to directly compare the performance of MT to human translators.

The dual judgment (scoring two candidates at once) was added experimentally. The advantage is saving some of the annotators’ time (they do not need to read the source or examine the picture again) and the chance to evaluate candidates also in terms of direct pairwise comparisons. In the history of WMT (Bojar et al., 2016), 5-way relative ranking was used for many years. With 5 candidates, the individual pairs may not be compared very precisely. With the single-candidate DA, pairwise comparisons cannot be used as the basis for system ranking. We believe that two candidates on one screen could be a good balance.

For the full statistical soundness, the judgments should be independent of each other. This is not the case in our dual scoring, even if we explicitly ask people to score the candidates independent of each other. The full independence is however not assured even in the original approach because annotators will remember their past judgments. This year, WMT even ran DA with document context available to the annotators by scoring all segments from a given document one after another in their natural order. We thus dare to pretend independence of judgments when interpreting DA scores.



Sentence: 1

Indicate how plausible these captions are for the highlighted area of the image. Judge each of the captions independently of the other. Each of the captions may be focusing on a different aspect of the area in the image.

CAND1 Text:

CAND1 Score: worst  best

CAND2 Text:

CAND2 Score: worst  best

Figure 5: Manual evaluation of Hindi captioning.

The user interface for our annotation for each of the tracks is illustrated in Figure 3, Figure 4, and Figure 5.

In the “text-only” evaluation, one English text (source) and two Hindi translations (candidate 1 and 2) are shown to the annotators. In the “multi-modal” evaluation, the annotators are shown both the image and the source English text. The first question is to validate if the source English text is a good caption for the indicated area. For two translation candidates, the annotators are asked to independently indicate to what extent the meaning is preserved. The “Hindi captioning” evaluation shows only the image and two Hindi candidates. The annotators are reminded that the two captions should be treated independently and that each of them can consider a very different aspect of the region.

## 6 Participants

Table 15 shows the participants in WAT2019. The table lists 25 organizations from various countries, including Japan, India, Myanmar, USA, Korea, China, France, and Switzerland.

About 400 translation results by 25 teams were submitted for automatic evaluation and about 30 translation results by 8 teams were submitted for pairwise evaluation. We selected about 50 translation results for JPO adequacy

evaluation. Table 16 shows tasks for which each team submitted results by the deadline.

## 7 Evaluation Results

In this section, the evaluation results for WAT2019 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2019 website.<sup>37</sup>

### 7.1 Official Evaluation Results

Figures 6, 7, 8 and 9 show the official evaluation results of ASPEC subtasks, Figures 10, 11, 12, 13, 14 and 15 show those of JPC subtasks, Figures 16 and 17 show those of JJI subtasks, Figures 18 and 19 show those of NCPD subtasks, Figures 20 and 21 show those of IITB subtasks, Figures 22, 23, 24 and 25 show those of ALT subtasks, Figures 26 and 27 show those of TDDC subtasks and Figures 28 and 29 show those of UFAL subtasks. Each figure contains the JPO adequacy evaluation result and evaluation summary of top systems.

The detailed automatic evaluation results are shown in Appendix A. The detailed JPO adequacy evaluation results for the selected submissions are shown in Tables 17 and 18. The weights for the weighted  $\kappa$  (Cohen, 1968) is defined as  $|Evaluation1 - Evaluation2|/4$ .

The automatic scores for the multi-modal task along with the WAT evaluation server BLEU scores are provided in Table 20. For each of the test sets (E-Test and C-Test), the scores are comparable across all the tracks (text-only, captioning or multi-modal translation) because of the underlying set of reference translations is the same. The scores for the captioning task will be however very low because captions generated independently of the English source caption are very likely to differ from the reference translation.

For multi-modal task, Table 19 shows the manual evaluation scores for all valid system submissions. As mentioned above, we used the reference translation as if it was one of the competing systems, see the rows “Reference” in the table. The annotation was fully anonymized, so the annotators had no chance of knowing if they are scoring human translation or MT output.

<sup>37</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/>

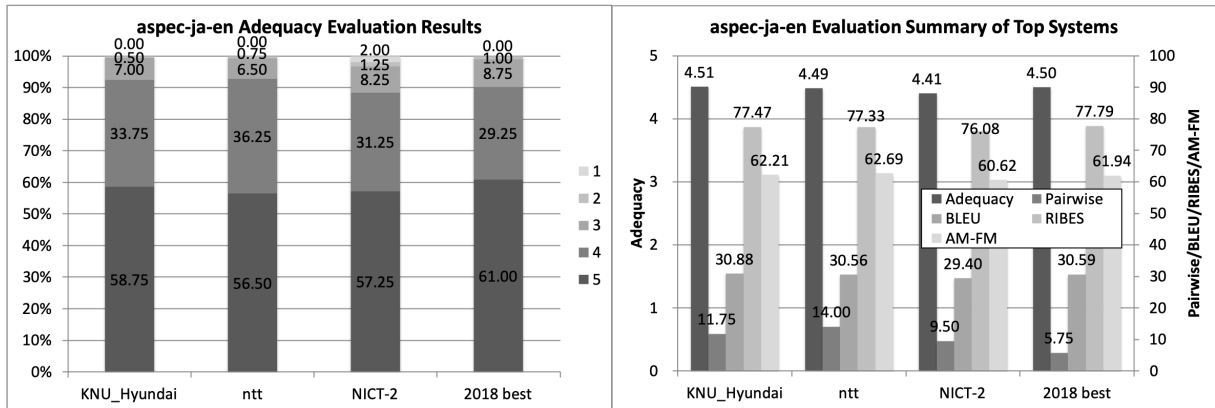


Figure 6: Official evaluation results of aspec-ja-en.

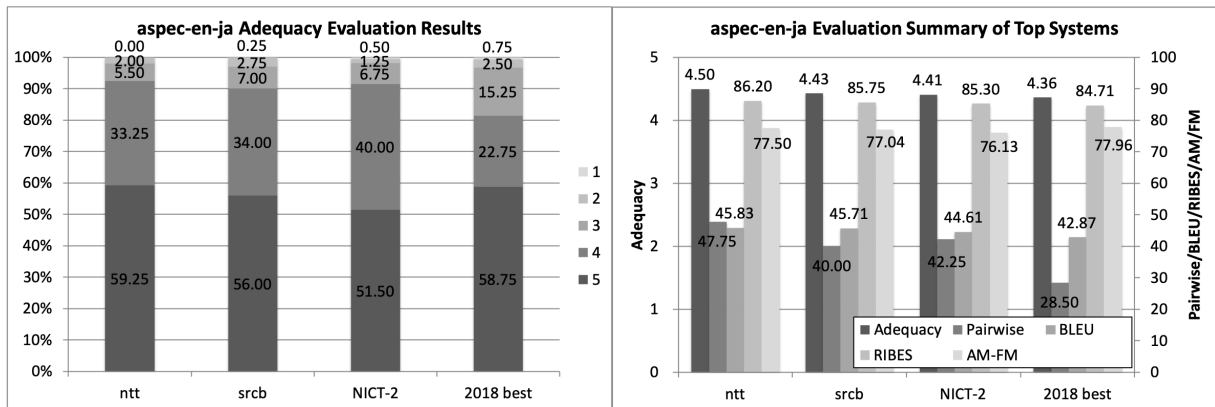


Figure 7: Official evaluation results of aspec-en-ja.

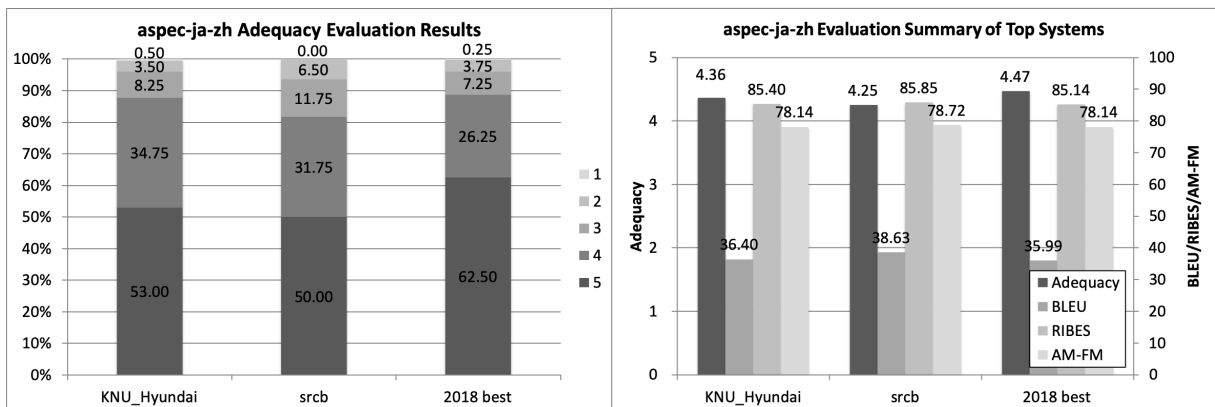


Figure 8: Official evaluation results of aspec-ja-zh.

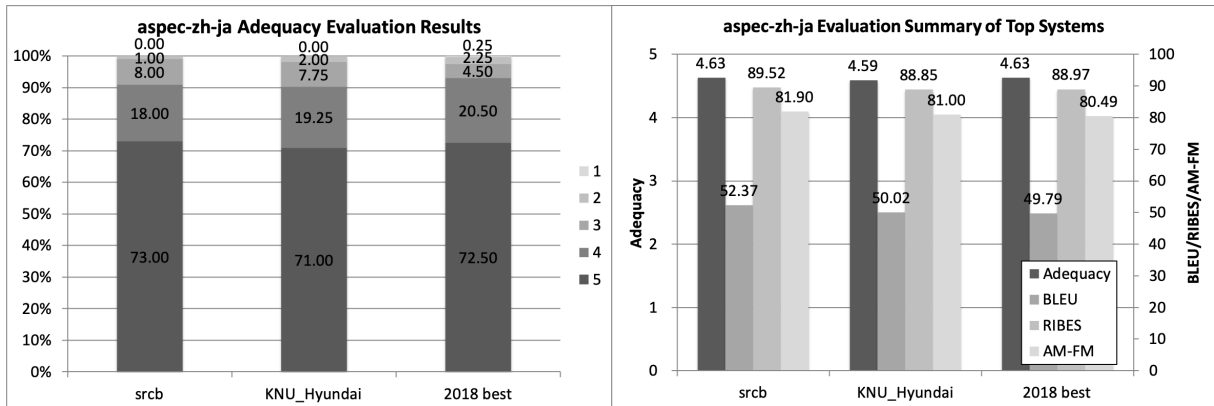


Figure 9: Official evaluation results of aspec-zh-ja.

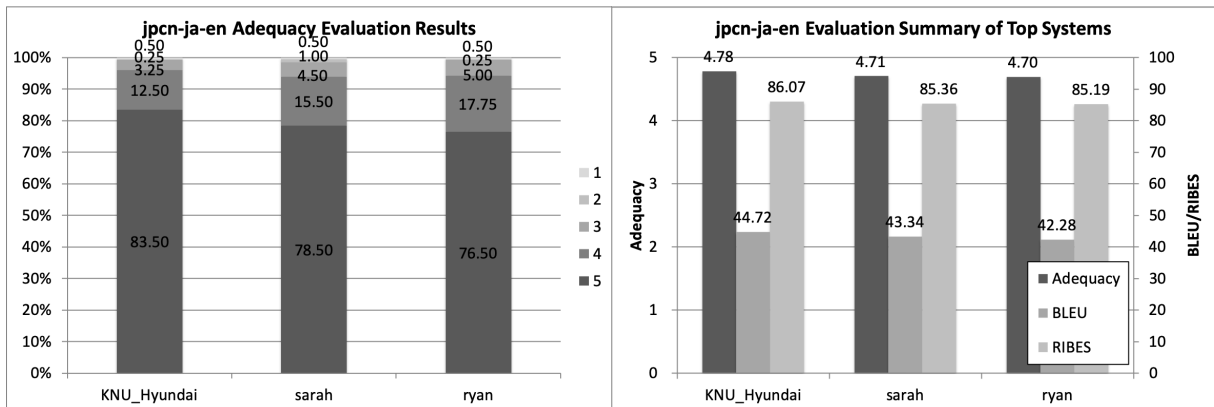


Figure 10: Official evaluation results of jpcn-ja-en.

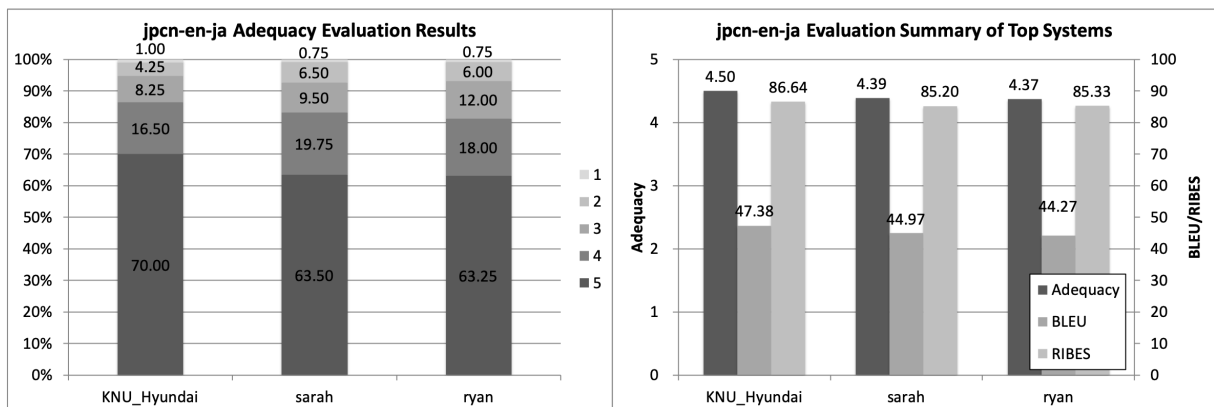


Figure 11: Official evaluation results of jpcn-en-ja.



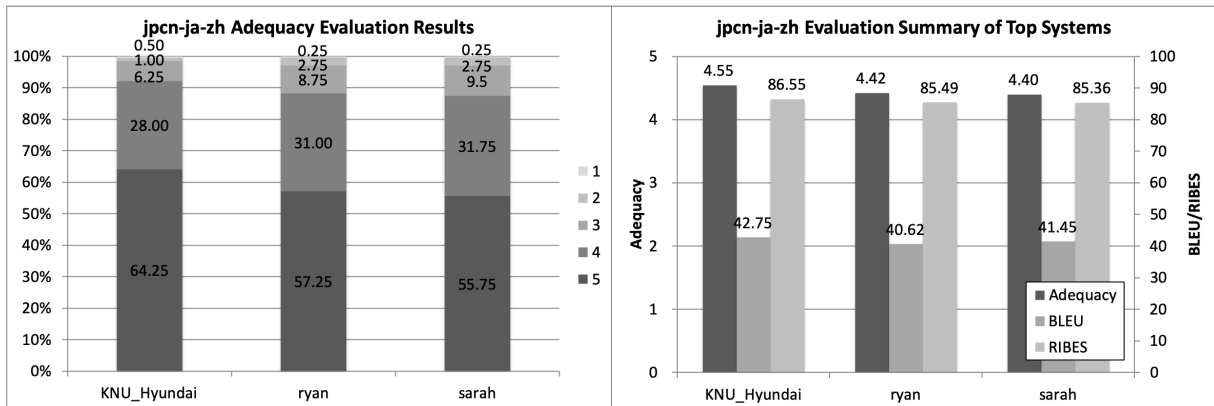


Figure 12: Official evaluation results of jpcn-ja-zh.

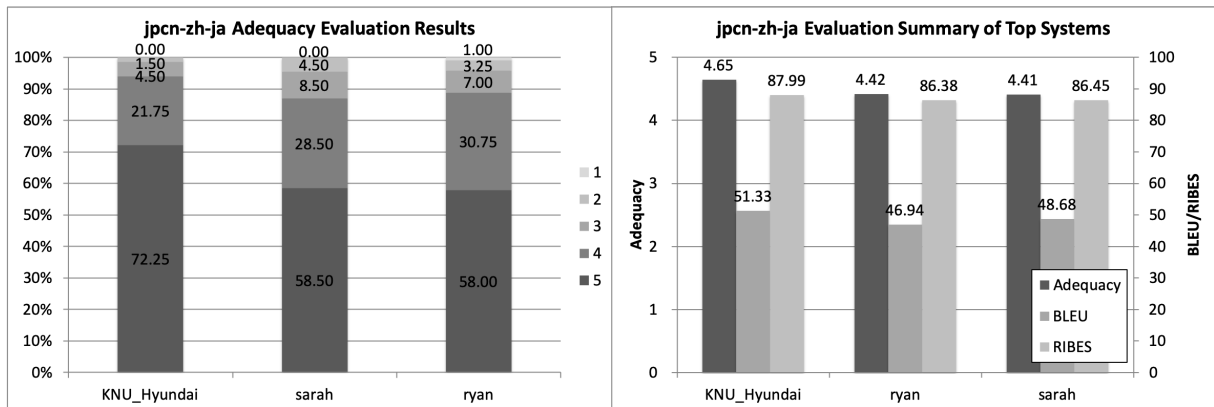


Figure 13: Official evaluation results of jpcn-zh-ja.

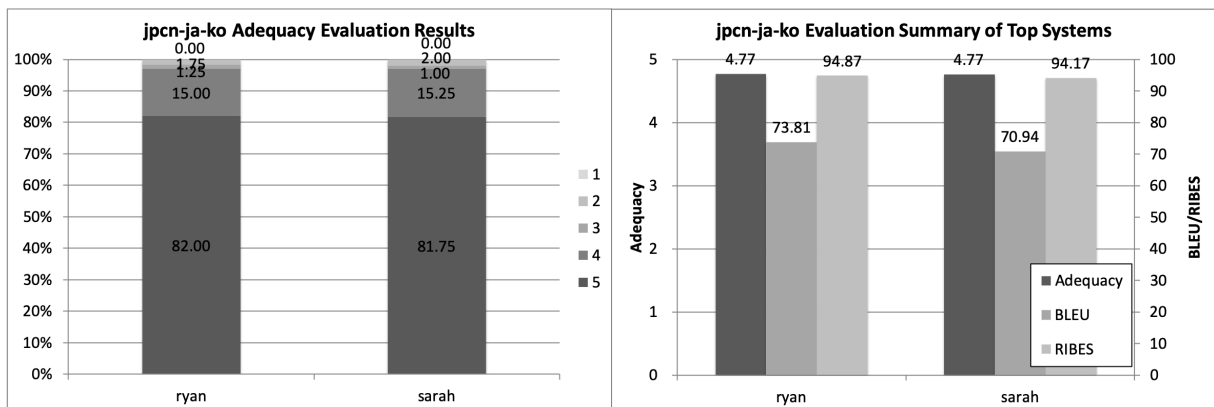


Figure 14: Official evaluation results of jpcn-ja-ko.

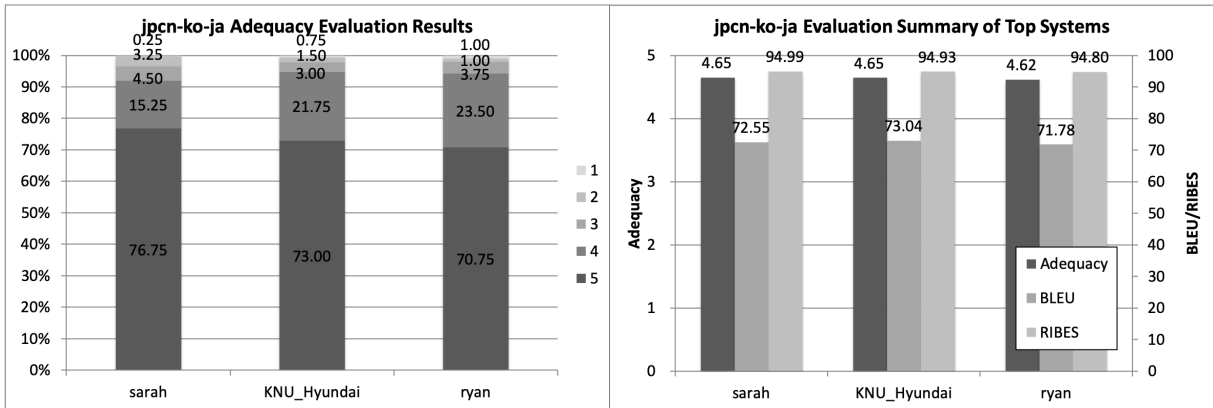


Figure 15: Official evaluation results of jpcn-ko-ja.

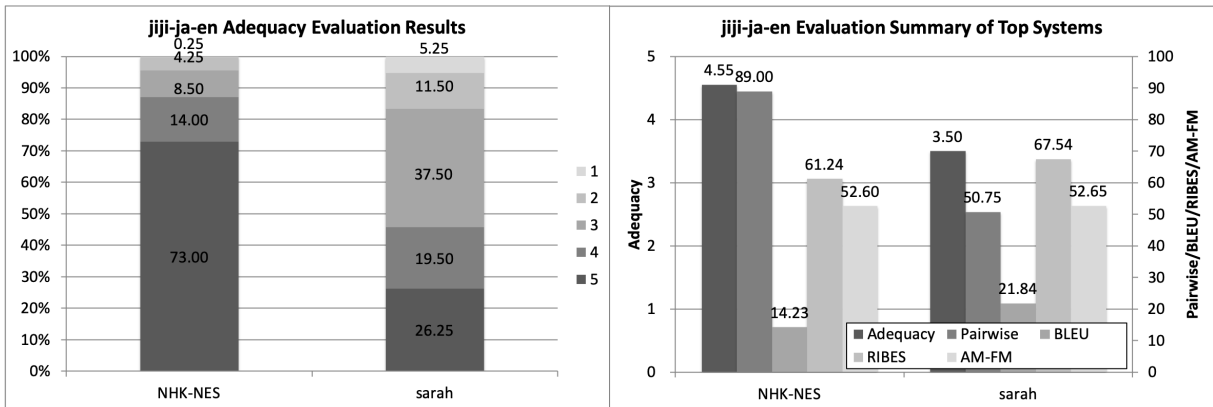


Figure 16: Official evaluation results of jiji-ja-en.

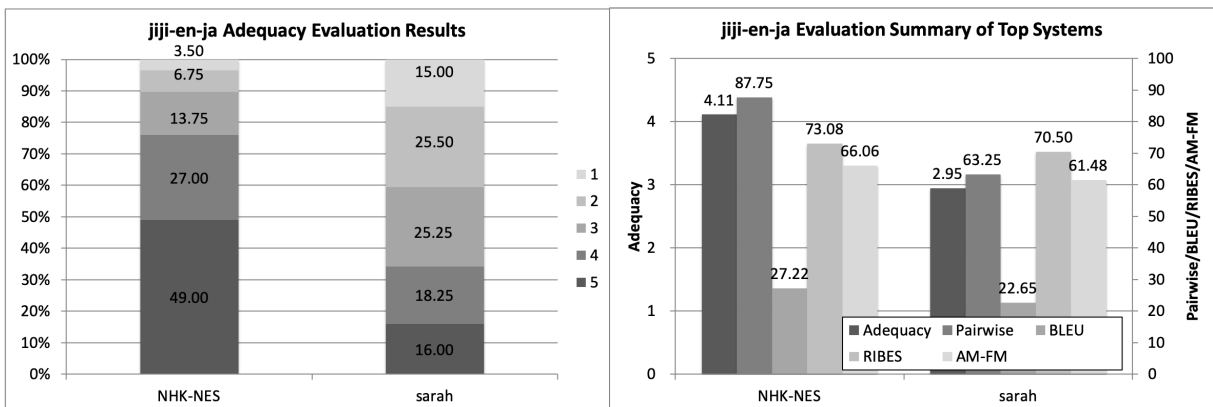


Figure 17: Official evaluation results of jiji-en-ja.

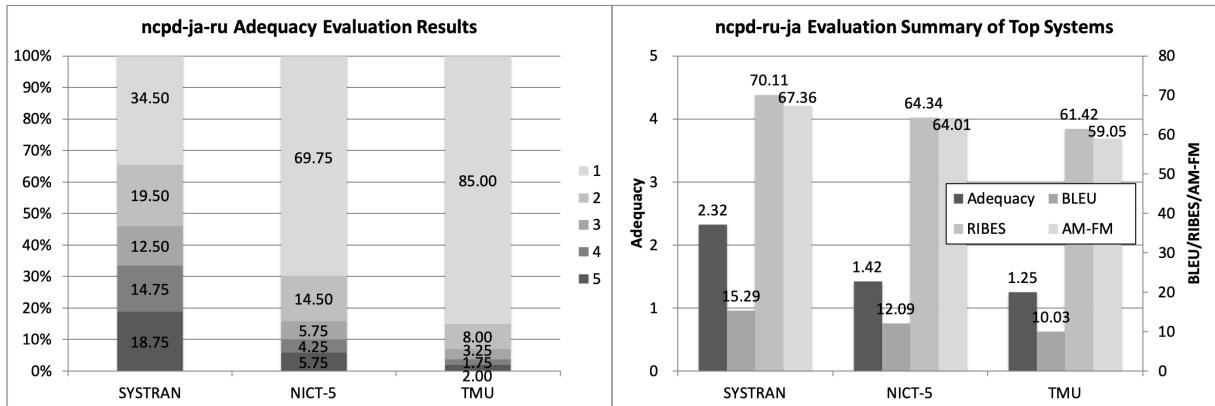


Figure 18: Official evaluation results of ncpd-ja-ru.

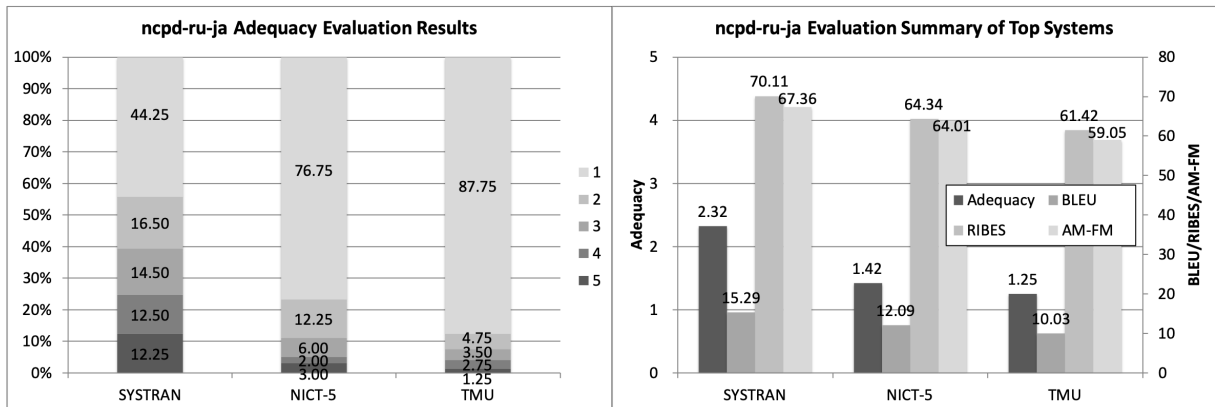


Figure 19: Official evaluation results of ncpd-ru-ja.

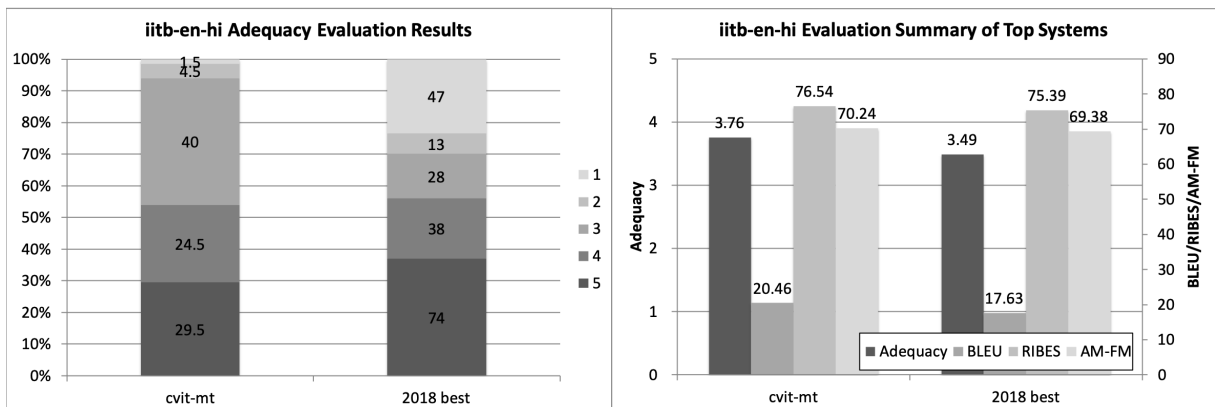


Figure 20: Official evaluation results of iitb-en-hi.

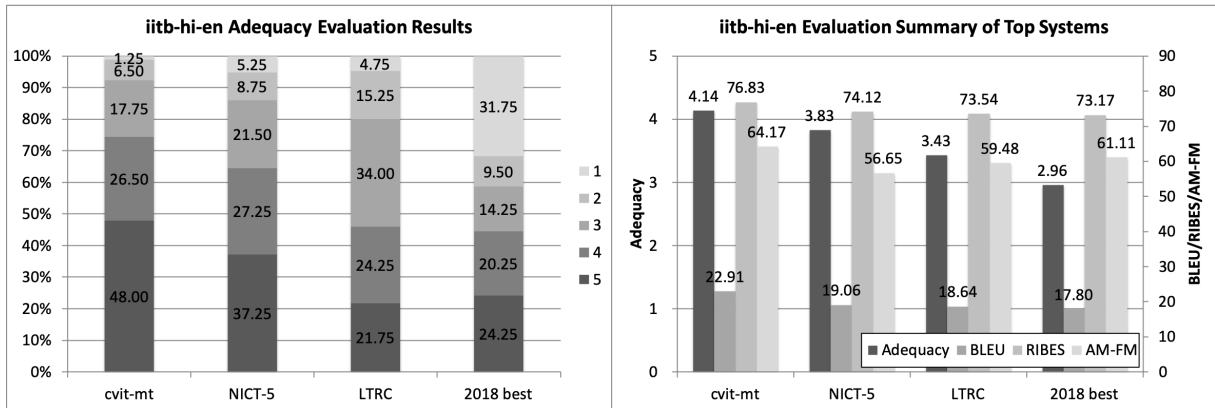


Figure 21: Official evaluation results of iitb-hi-en.

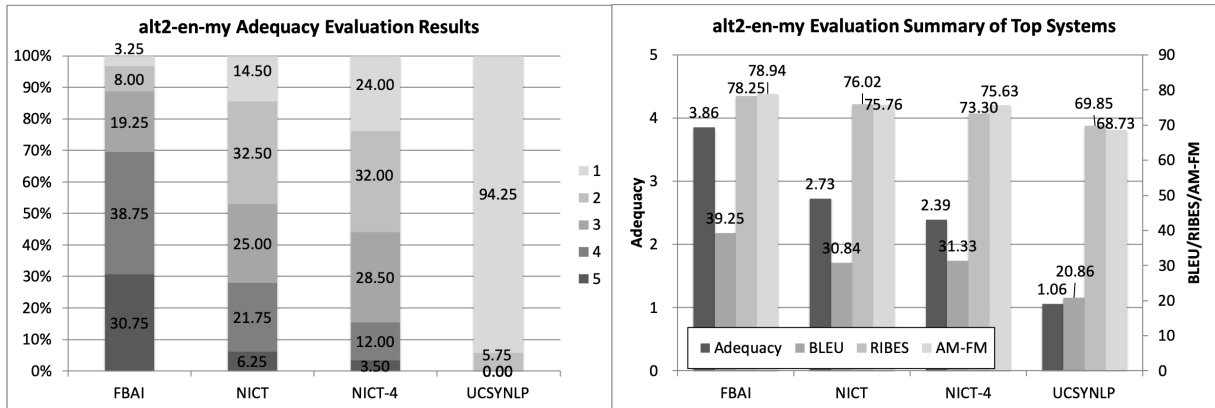


Figure 22: Official evaluation results of alt2-en-my.

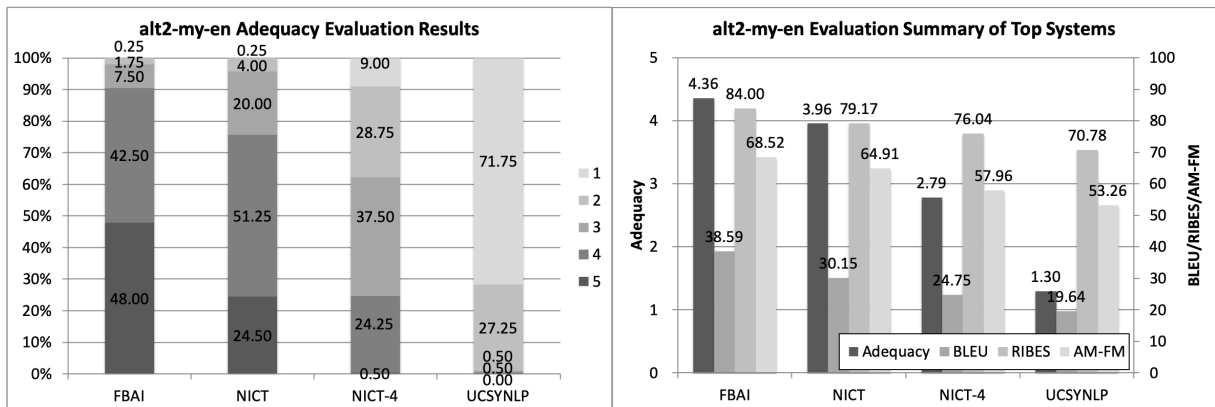


Figure 23: Official evaluation results of alt2-my-en.

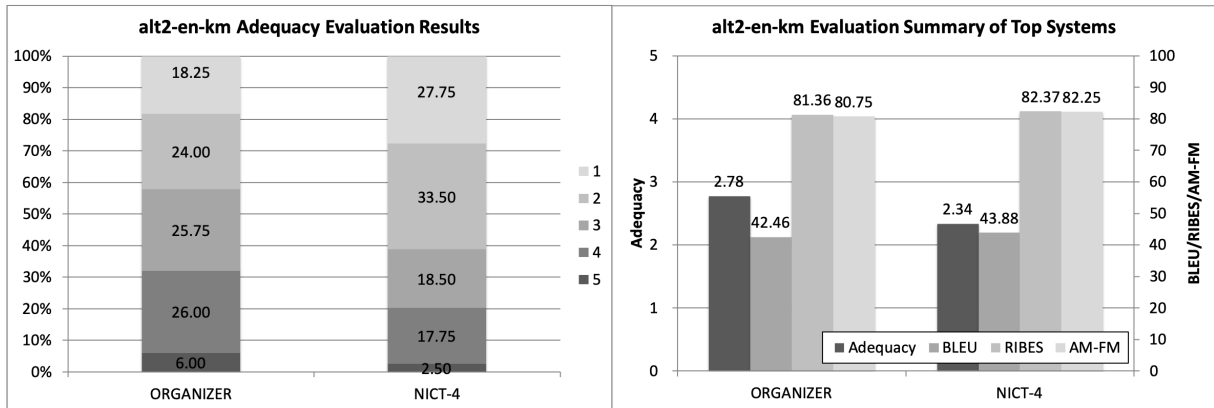


Figure 24: Official evaluation results of alt2-en-km.

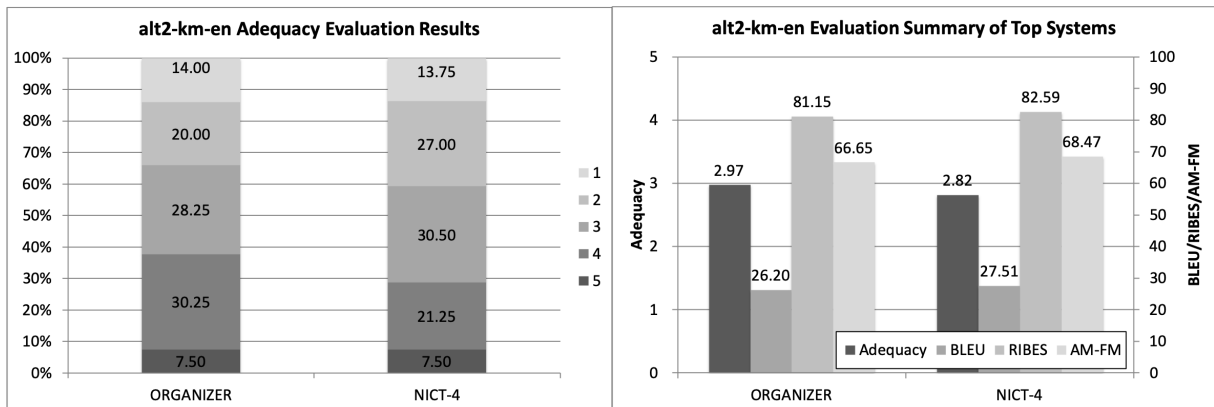


Figure 25: Official evaluation results of alt2-km-en.

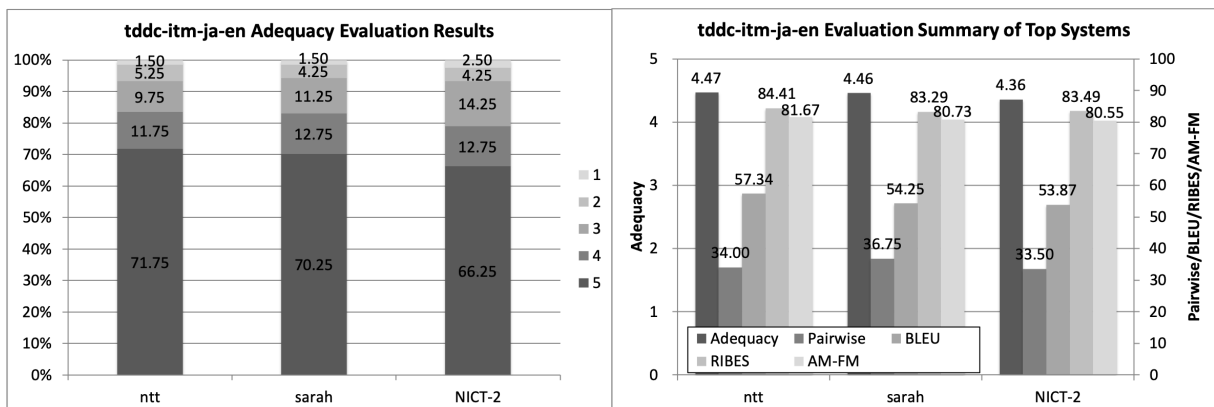


Figure 26: Official evaluation results of tddc-itm-ja-en.

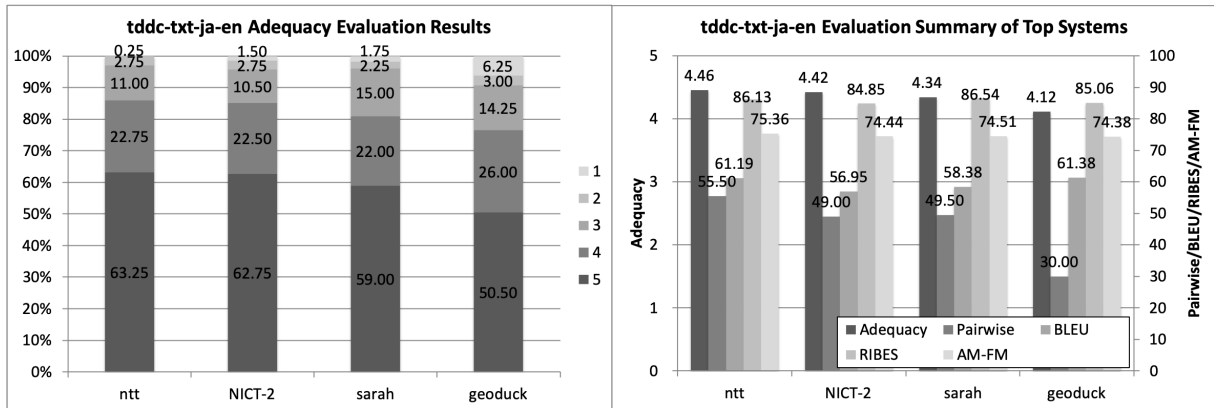


Figure 27: Official evaluation results of tddc-txt-ja-en.

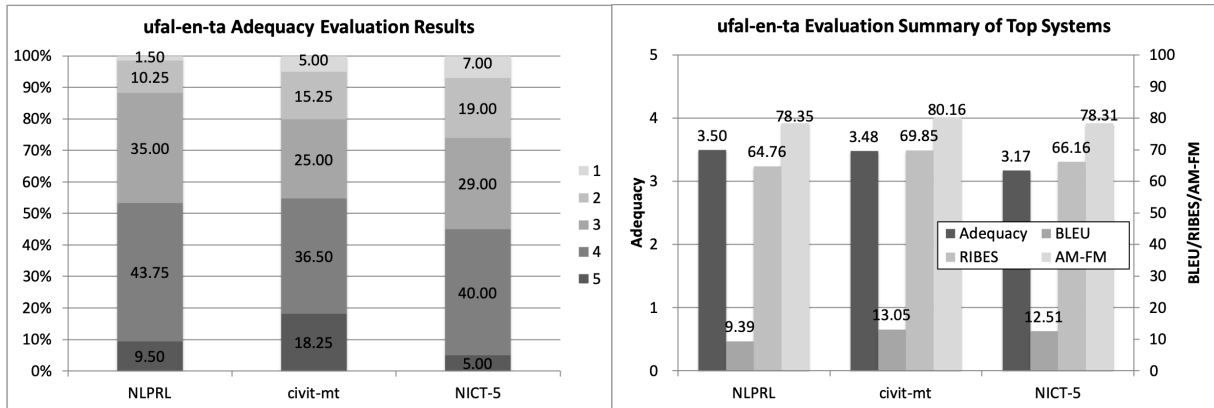


Figure 28: Official evaluation results of ufal-en-ta.

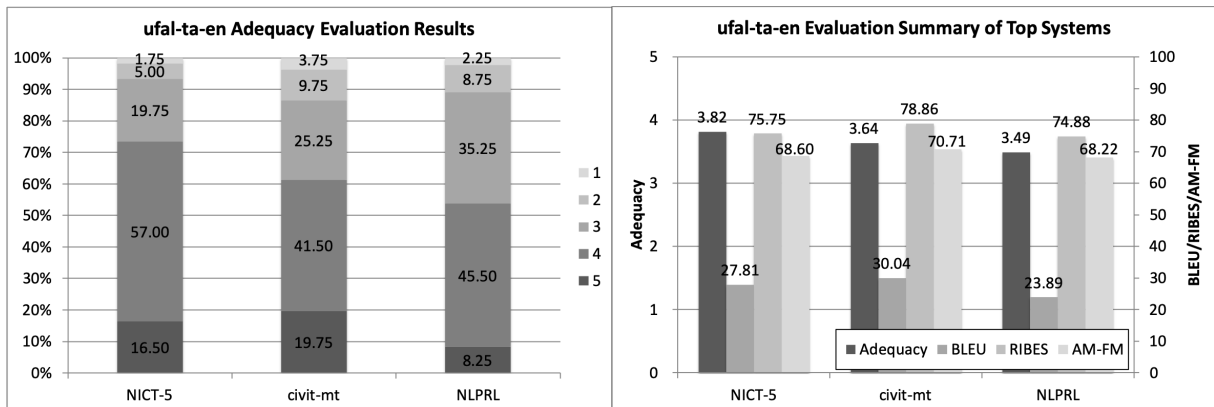


Figure 29: Official evaluation results of ufal-ta-en.

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NICT	NICT	Japan
NTT	NTT Corporation	Japan
NICT-2	NICT	Japan
NICT-4	NICT	Japan
NICT-5	NICT	Japan
UCSYNLP	University of Computer Studies, Yangon	Myanmar
UCSMNLP	University of Computer Studies, Mandalay	Myanmar
cvit	IIIT Hyderabad	India
srcb	RICOH Software Research Center Beijing Co.,Ltd	China
sarah	Rakuten Institute of Technology	Japan
683	National Institute of Technology Silchar	India
KNU_Hyundai	Kangwon National University	Korea
NITSNLP	National Institute of Technology Silchar	India
ryan	Kakao Brain	Korea
PUP-IND	Punjabi University Patiala	India
FBAI	Facebook AI Research	USA
AISTAI	National Institute of Advanced Industrial Science and Technology	Japan
SYSTRAN	SYSTRAN	France
NHK-NES	NHK & NHK Engineering System	Japan
geoduck	Microsoft Research	USA
LTRC-MT	IIIT Hyderabad	India
ykkd	The University of Tokyo	Japan
IDIAP	Idiap Research Institute	Switzerland
NLPRL	Indian Institute of Technology (BHU) Varanasi	India

Table 15: List of participants in WAT2019

Team ID	ASPEC				JPC						TDDC	JIJI		NCPD	
	EJ	JE	CJ	JC	EJ	JE	CJ	JC	Ko-J	J-Ko	JE	EJ	JE	RJ	JR
TMU															
NTT	✓	✓									✓				✓
NICT-2	✓	✓									✓				
NICT-5		✓												✓	✓
srcb	✓	✓	✓	✓						✓					
sarah					✓	✓	✓	✓	✓	✓	✓	✓	✓		
KNU_Hyundai	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓					
ryan					✓	✓	✓	✓	✓	✓					
AISTAI	✓														
SYSTRAN														✓	✓
NHK-NES												✓	✓		
geoduck											✓				
ykkd		✓													

Team ID	Mixed-domain tasks								Mutimodal task
	EM	ALT			IITB		UFAL (EnTam)		EV/CH EH
		ME	E-Kh	Kh-E	EH	HE	ET	TE	
NICT	✓	✓							
NICT-4	✓	✓	✓	✓					
NICT-5						✓	✓	✓	
UCSYNLP	✓	✓							
UCSMNLP	✓	✓							
cvit					✓	✓	✓	✓	
sarah	✓								
683									✓
NITSNLP									✓
PUP-IND									✓
FBAI	✓	✓							
LTRC-MT						✓			
IDIAP									✓
NLPRL							✓	✓	

Table 16: Submissions for each task by each team. E, J, C, Ko, R, M, Kh, H, and T denote English, Japanese, Chinese, Korean, Russian, Myanmar, Khmer, Hindi, and Tamil language, respectively.

Subtask	SYSTEM ID	DATA ID	Annotator A		Annotator B		all average	weighted	
			average	variance	average	variance		$\kappa$	$\kappa$
aspec-ja-en	KNU_Hyundai	3173	4.63	0.40	4.38	0.41	4.51	0.23	0.31
	ntt	3225	4.62	0.39	4.34	0.42	4.49	0.26	0.32
	NICT-2	3086	4.51	0.74	4.30	0.68	4.41	0.22	0.39
	2018 best	2474	4.37	0.49	4.63	0.44	4.50	0.15	0.25
aspec-en-ja	ntt	3236	4.60	0.51	4.39	0.43	4.50	0.07	0.15
	srcb	3212	4.54	0.62	4.32	0.52	4.43	0.17	0.31
	NICT-2	3182	4.54	0.53	4.28	0.46	4.41	0.22	0.31
	2018 best	2219	4.16	0.90	4.57	0.57	4.36	0.17	0.30
aspec-ja-zh	KNU_Hyundai	3170	4.44	0.47	4.29	0.85	4.36	0.15	0.15
	srcb	3208	4.14	0.74	4.37	0.86	4.25	0.16	0.26
	2018 best	2266	4.67	0.32	4.27	0.90	4.47	0.28	0.36
aspec-zh-ja	srcb	3210	4.80	0.24	4.46	0.61	4.63	0.27	0.31
	KNU_Hyundai	3179	4.76	0.26	4.42	0.71	4.59	0.14	0.16
	2018 best	2267	4.78	0.26	4.48	0.67	4.63	0.31	0.33
jpcn-ja-en	KNU_Hyundai	3188	4.73	0.40	4.83	0.22	4.78	0.36	0.46
	sarah	2927	4.63	0.53	4.78	0.29	4.71	0.44	0.55
	ryan	2962	4.62	0.50	4.77	0.27	4.70	0.33	0.38
jpcn-en-ja	KNU_Hyundai	3192	4.43	0.81	4.57	0.77	4.50	0.36	0.49
	sarah	2926	4.38	0.83	4.40	0.98	4.39	0.35	0.51
	ryan	2961	4.30	0.90	4.44	0.94	4.37	0.36	0.53
jpcn-ja-zh	KNU_Hyundai	3157	4.53	0.45	4.56	0.54	4.54	0.29	0.35
	ryan	2948	4.43	0.49	4.41	0.74	4.42	0.29	0.39
	sarah	2921	4.39	0.50	4.41	0.75	4.40	0.39	0.48
jpcn-zh-ja	KNU_Hyundai	3152	4.72	0.26	4.57	0.55	4.65	0.26	0.35
	ryan	2949	4.42	0.58	4.42	0.81	4.42	0.35	0.48
	sarah	2920	4.45	0.56	4.37	0.80	4.41	0.36	0.51
jpcn-ja-ko	ryan	2850	4.82	0.27	4.73	0.34	4.77	0.56	0.65
	sarah	2925	4.83	0.27	4.71	0.36	4.77	0.31	0.46
jpcn-ko-ja	sarah	2924	4.72	0.39	4.58	0.68	4.65	0.59	0.69
	KNU_Hyundai	2998	4.70	0.35	4.59	0.58	4.65	0.68	0.76
	ryan	2890	4.68	0.35	4.55	0.63	4.62	0.71	0.74
jiji-ja-en	NHK-NES	2884	4.50	0.68	4.61	0.72	4.55	0.26	0.38
	sarah	2793	3.27	1.13	3.73	1.39	3.50	0.23	0.39
jiji-en-ja	NHK-NES	2886	4.04	1.02	4.18	1.37	4.11	0.21	0.42
	sarah	2814	3.00	1.35	2.89	1.99	2.95	0.19	0.42

Table 17: JPO adequacy evaluation results in detail (1).



Subtask	SYSTEM ID	DATA ID	Annotator A		Annotator B		all average	weighted	
			average	variance	average	variance		$\kappa$	$\kappa$
ncpd-ja-ru	SYSTRAN	3076	2.65	2.56	2.62	2.12	2.64	0.26	0.47
	NICT-5	3026	1.70	1.39	1.54	1.19	1.62	0.29	0.52
	TMU	3095	1.32	0.65	1.23	0.57	1.28	0.30	0.43
ncpd-ru-ja	SYSTRAN	2912	2.23	2.16	2.42	1.99	2.32	0.24	0.48
	NICT-5	3027	1.41	0.95	1.44	0.74	1.42	0.34	0.57
	TMU	3097	1.21	0.50	1.29	0.65	1.25	0.36	0.56
iitb-en-hi	cvit-mt	2680	3.94	1.02	3.58	0.82	3.76	0.53	0.58
	2018 best	2362	3.58	2.71	3.40	2.52	3.49	0.52	0.74
iitb-hi-en	cvit-mt	2681	4.53	0.53	3.74	1.18	4.13	0.05	0.13
	NICT-5	2865	4.26	0.90	3.39	1.48	3.83	0.10	0.22
	LTRC	3119	3.92	0.91	2.94	1.15	3.43	0.05	0.16
	2018 best	2381	2.96	2.55	2.96	2.52	2.96	0.48	0.76
alt2-en-my	fbai	3203	4.36	0.67	3.36	1.02	3.86	0.02	0.19
	NICT	2818	2.74	1.34	2.71	1.25	2.73	0.97	0.98
	NICT-4	2979	2.40	1.21	2.38	1.13	2.39	0.97	0.98
	UCSYNLP	2858	1.05	0.05	1.06	0.06	1.06	0.59	0.59
alt2-my-en	fbai	3201	4.49	0.53	4.24	0.47	4.36	0.13	0.18
	NICT	2816	3.88	0.78	4.03	0.46	3.96	0.07	0.18
	NICT-4	2977	2.56	0.78	3.01	0.85	2.79	0.26	0.42
	UCSYNLP	3252	1.25	0.27	1.34	0.22	1.30	0.62	0.60
alt2-en-km	organizer	2898	2.54	1.54	3.00	1.19	2.77	0.24	0.49
	NICT-4	2929	2.43	1.35	2.25	1.20	2.34	0.67	0.80
alt2-km-en	organizer	2897	2.60	1.52	3.35	0.92	2.97	0.08	0.31
	NICT-4	2915	2.91	1.33	2.73	1.25	2.82	0.61	0.76
tddc-itm-ja-en	ntt	3002	4.48	0.83	4.46	1.05	4.47	0.17	0.30
	sarah	2807	4.52	0.77	4.40	1.04	4.46	0.29	0.47
	NICT-2	3081	4.42	0.94	4.30	1.19	4.36	0.40	0.52
tddc-txt-ja-en	ntt	3005	4.34	0.65	4.58	0.65	4.46	0.16	0.27
	NICT-2	3084	4.26	0.82	4.58	0.72	4.42	0.22	0.33
	sarah	2808	4.20	0.83	4.49	0.87	4.34	0.26	0.39
	geoduck	3200	4.07	0.99	4.17	1.64	4.12	0.30	0.48
ufal-en-ta	NLPRL018	3015	3.71	0.56	3.29	0.82	3.50	0.17	0.34
	cvit-mt	2830	3.46	1.10	3.50	1.34	3.48	0.84	0.90
	NICT-5	3046	3.15	0.92	3.19	1.16	3.17	0.62	0.74
ufal-ta-en	NICT-5	3054	3.88	0.46	3.75	0.90	3.81	0.39	0.48
	cvit-mt	2833	3.58	0.89	3.70	1.19	3.64	0.75	0.83
	NLPRL018	3014	3.67	0.47	3.31	0.91	3.49	0.10	0.22

Table 18: JPO adequacy evaluation results in detail (2).

	Team ID	Data ID	Ave	Ave Z
EV TEXT	IDIAP	2956	72.85	0.70
	Reference		71.34	0.66
	683	3285	68.89	0.57
	683	3286	61.64	0.36
	NITSNLP	3299	52.53	0.00
CH TEXT	Reference		79.23	0.94
	IDIAP	3277	60.81	0.25
	IDIAP	3267	60.17	0.25
	683	3284	45.69	-0.28
	683	3287	45.52	-0.24
EV MM	NITSNLP	3300	28.48	-0.81
	Reference		70.04	0.60
	683	3271	69.17	0.61
	PUP-IND	3296	62.42	0.35
	PUP-IND	3295	60.22	0.28
CH MM	NITSNLP	3288	58.98	0.25
	Reference		75.96	0.76
	683	3270	54.51	0.08
	NITSNLP	3298	48.45	-0.20
	PUP-IND	3281	48.06	-0.13
EV HI	PUP-IND	3280	47.06	-0.17
	Reference		68.80	0.52
	NITSNLP	3289	51.78	-0.05
CH HI	Reference		72.60	0.61
	NITSNLP	3297	44.46	-0.35
	683	3304	26.54	-0.94

Table 19: Manual evaluation result for WAT Multi-Modal Tasks.

	System	Run	BLEU	chrF3	nCDER	nCharacTER	nPER	nTER	nWER	BLEU <sub>w</sub>
EV TEXT	IDIAP	2956	52.18	58.81	62.18	57.95	69.32	56.87	55.07	41.32
	683	3285	48.29	54.66	58.18	54.12	65.34	52.52	51.00	38.19
	683	3286	33.47	40.37	45.36	00.11	50.54	43.11	42.13	25.34
	NITSNLP	3299	30.05	34.49	41.36	∧ 10.92	48.23	36.42	35.10	20.13
	IDIAP	3277	40.40	50.18	52.58	44.32	60.19	49.11	46.02	30.94
CH TEXT	IDIAP	3267	39.08	49.30	51.78	41.72	59.49	48.42	45.51	30.34
	683	3284	21.56	30.90	33.92	13.69	41.14	30.53	28.40	14.69
	683	3287	21.50	30.27	∧ 34.66	-65.00	38.98	∧ 32.91	∧ 31.47	∧ 15.85
	NITSNLP	3300	10.50	17.91	23.04	∧ -60.87	28.05	20.87	19.90	5.56
	683	3271	51.46	57.63	61.51	52.61	68.52	55.99	54.28	40.55
EV MM	PUP-IND	3296	39.67	47.76	51.98	46.84	59.50	43.47	41.92	28.27
	NITSNLP	3288	39.13	45.50	49.45	27.92	57.43	∧ 43.91	∧ 42.17	∧ 28.45
	PUP-IND	3295	38.50	45.35	∧ 50.33	∧ 41.40	∧ 58.82	41.84	40.65	27.39
	683	3270	28.62	37.86	41.60	20.10	48.64	38.38	36.44	20.37
	NITSNLP	3298	19.68	27.99	31.84	-24.40	38.61	29.38	27.16	12.58
CH MM	PUP-IND	3281	18.32	27.79	30.08	∧ 19.63	∧ 40.51	23.51	21.12	11.77
	PUP-IND	3280	16.15	25.78	28.57	06.31	37.34	23.38	∧ 21.28	10.19
	NITSNLP	3289	8.68	14.45	14.27	-15.81	22.51	06.85	06.19	2.59
EV HI	NITSNLP	3297	2.28	8.88	8.00	-50.33	12.97	06.05	05.62	0.00
	NITSNLP	3297	2.28	8.88	8.00	-50.33	12.97	06.05	05.62	0.00
	683	3304	1.07	8.63	6.65	-19.81	-32.82	-52.44	-52.59	0.00

Table 20: Multi-Modal Task automatic evaluation results. For each test set (EV and CH) and each track (TEXT, MM and HI), we sort the entries by our BLEU scores. The symbol “∧” in subsequent columns indicates fields where the other metric ranks candidates in a different order. BLEU<sub>w</sub> denotes the WAT official BLEU scores.

## 7.2 Statistical Significance Testing of Pairwise Evaluation between Submissions

Table 21 shows the results of statistical significance testing of aspec-ja-en subtasks, Table 22 shows that of JIJI subtasks, Table 23 shows that of TDDC 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 et al., 1971) values. The results are shown in Table 24. 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 for these language pairs are Japanese, and the evaluation of one's mother tongue is much easier than for other languages in general. The  $\kappa$  values for Hindi languages are relatively high. This might be because the overall translation quality of the Hindi languages are low, and the evaluators can easily distinguish better translations from worse ones.

## 8 Findings

In this section, we will show findings of some of the translation tasks.

### 8.1 TDDC

In the results of both the automatic evaluation and the human evaluation, every system translated most sentences correctly. According to the human evaluation of the subtasks of 'Items' and 'Texts', all evaluators rated more than 70% of all the pairs at 4 or 5. Most of

these high-rated pairs consist of typical terms and sentences from timely disclosure documents. This tasks focus on the accurate translation of figures, so the evaluation criteria confirmed there are no mistranslation in the typical sentences containing figures, such as unit of money and dates.

However, uncommon sentences used in timely disclosure documents tend to be mistranslated. For example, uncommon proper nouns tended to be omitted or mistranslated to other meaning words, besides sentences which has complex and uncommon structures, generally long sentences, caused errors at dependency of subordinate clauses.

In addition, some systems translated sentences without subjects into sentences with incorrect subjects. Japanese sentences often omit subjects and objects, which would normally be included in English. For example, a Japanese sentence, “当社普通株式 27,000 株を上限とする。” (Common shares of the Company, limited to a maximum of 27,000 shares), was translated to “(Unrelated company name) common stocks up to 27,000 shares” .

Moreover, there are some incorrect modifiers or determiners. In Japanese timely disclosure documents, there are many variable prefix for dates, such as “本” (this), “当” (this), “次” (next), and “前” (last). Some systems translated sentences containing these words with incorrect year. For example, a Japanese sentence contains “当第 3 四半期連結会計期間末” (the end of third quarter of this fiscal year) was translated to “the end of the third quarter of FY 2016” .

In summary, the causes of these mistranslations are considered as follows:

- It is difficult for the systems to translate long sentence and proper nouns which TDDC does not contain.
- Some source sentences are unclear due to lack of subjects and/or objects, so these are not suitable for English translation.
- TDDC contains not semantically balanced pairs and the systems might be affected strongly by either of source pair sentences.

On the other hand, some translations seem to be fitted to sentences of TDDC which are

	KNU_Hyundai (3173)		
	NICT-2 (3086)		
	srcb (3205)		
NTT (3225)	-	»	»
KNU_Hyundai (3173)	-	»	»
NICT-2 (3086)	-	»	»

	NICT-2 (3182)		
	srcb (3212)		
	AISTAI (3251)		
	KNU_Hyundai (3172)		
NTT (3236)	»	»	»
NICT-2 (3182)	-	»	»
srcb (3212)	-	»	»
AISTAI (3251)	-	»	»

Table 21: Statistical significance testing of the aspec-ja-en (left) and aspec-en-ja (right) Pairwise scores.

	NHK-NES (2883)		
	sarah (2793)		
	sarah (2813)		
NHK-NES (2884)	»	»	»
NHK-NES (2883)	»	»	»
sarah (2793)	»	»	»

	NHK-NES (2885)		
	sarah (2814)		
	sarah (2815)		
NHK-NES (2886)	»	»	»
NHK-NES (2885)	»	»	»
sarah (2814)	»	»	»

Table 22: Statistical significance testing of the jiji-ja-en (left) and jiji-en-ja (right) Pairwise scores.

	sarah (2807)				
	NTT (3002)				
	NICT-2 (3081)				
	sarah (2811)				
	geoduck (3197)				
	geoduck (3216)				
ORGANIZER (3264)	»	»	»	»	»
sarah (2807)	»	»	»	»	»
NTT (3002)	-	»	»	»	»
NICT-2 (3081)	-	»	»	»	»
sarah (2811)	-	»	»	»	»
geoduck (3197)	-	»	»	»	»

	NTT (3005)				
	sarah (2808)				
	NICT-2 (3084)				
	sarah (2812)				
	geoduck (3200)				
	geoduck (3217)				
ORGANIZER (3265)	-	»	»	»	»
NTT (3005)	»	»	»	»	»
sarah (2808)	-	»	»	»	»
NICT-2 (3084)	-	»	»	»	»
sarah (2812)	-	»	»	»	»
geoduck (3200)	-	»	»	»	»

Table 23: Statistical significance testing of the tddc-itm-ja-en (left) and tddc-txt-ja-en (right) Pairwise scores.

freely and omitted redundant expressions, but evaluators mark them as low scores, probably because they are not literal translations. This result implies that it is necessary to create another evaluation criterion, which evaluates the correctness of transmitting information to investors correctly.

## 8.2 English↔Tamil Task

We observed that most participants used transfer learning techniques such as fine-tuning and mixed fine-tuning for Tamil→English translation leading to reasonably high quality translations. However, English→Tamil translation is still poor and

the main reason is the lack of helping parallel corpora. We expect that utilization of large in-domain monolingual corpora for back-translation should help alleviate this problem. We will provide such corpora for next year’s task.

## 8.3 News Commentary Task

We only received 3 submissions for Russian↔Japanese translation and all submissions leveraged multilingualism and multi-step fine-tuning proposed by [Imankulova et al. \(2019\)](#) and showed that carefully choosing corpora and robust training can dramatically enhance the quality of NMT for language pairs

aspec-ja-en			aspec-en-ja		
SYSTEM	DATA	$\kappa$	SYSTEM	DATA	$\kappa$
NTT	3225	0.157	NTT	3236	0.298
NICT-2	3086	0.187	NICT-2	3182	0.319
srcb	3205	0.220	srcb	3212	0.305
KNU_Hyundai	3173	0.156	KNU_Hyundai	3172	0.302
ave.		0.180	AISTAI	3251	0.303
			ave.		0.305

jiji-ja-en			jiji-en-ja		
SYSTEM	DATA	$\kappa$	SYSTEM	DATA	$\kappa$
sarah	2793	0.146	sarah	2814	0.357
sarah	2813	0.158	sarah	2815	0.299
NHK-NES	2883	0.273	NHK-NES	2885	0.354
NHK-NES	2884	0.159	NHK-NES	2886	0.390
ave.		0.184	ave.		0.350

tddc-itm-ja-en			tddc-txt-ja-en		
SYSTEM	DATA	$\kappa$	SYSTEM	DATA	$\kappa$
ORGANIZER	3264	0.382	ORGANIZER	3265	0.135
NTT	3002	0.403	NTT	3005	0.163
NICT-2	3081	0.423	NICT-2	3084	0.175
sarah	2807	0.408	sarah	2808	0.163
sarah	2811	0.391	sarah	2812	0.167
geoduck	3197	0.404	geoduck	3200	0.172
geoduck	3216	0.493	geoduck	3217	0.321
ave.		0.415	ave.		0.185

Table 24: The Fleiss’ kappa values for the pairwise evaluation results.

Source Good?	C-Test	E-Test
Yes	1586 (78.7 %)	1348 (66.9 %)
No	20 (1.0 %)	46 (2.3 %)
No Answer	410 (20.3 %)	622 (30.9 %)
Total	2016 (100.0 %)	2016 (100.0 %)

Table 25: Appropriateness of source English captions in the 4032 assessments collected for the multi-modal track.

that have very small in-domain parallel corpora. For next year’s task we expect more submissions where participants will leverage additional larger helping monolingual as well as bilingual corpora.

## 8.4 Multi-Modal Task

### 8.4.1 Validation of Source English Captions

In the manual evaluation of multimodal track, our annotators saw both the picture and the source text (and the two scored candidates). We took this opportunity to double check the quality of the original HVG data. Prior to scoring the candidates, we asked our annotators to confirm that the source English text is a good caption for the indicated region of the image.

The results in Table 25 indicate that for a surprisingly high number of items we did not

receive any answer. This confirms that even non-anonymous annotators can easily provide sloppy evaluations. It is possible that part of these omissions can be attributed to our annotation interface which was showing all items on one page and relying on scrolling. Next time, we will show only one annotation item on each page and also consider highlighting unanswered questions. Strictly requiring an answer would not be always appropriate but we need to ensure that annotators are aware that they are skipping a question.

Luckily, the bad source captions are not a frequent case, amounting to 1 or 2% of assessed examples.

### 8.4.2 Relation to Human Translation

The bilingual style of evaluation of the multi-modal task allowed us to evaluate the reference translations as if they were yet another competing MT system. Table 19 thus lists also the “Reference”.

Across the tracks and test sets (EV vs. CH), humans surpass MT candidates. One single exception is IDIAP run 2956 winning in text-only translation of the E-Test, but this is not confirmed on the C-Test (CH). The score of the anonymized system 683 on E-Test in multi-

modal track (MM) has also almost reached human performance. These are not the first cases of MT performing on par with humans and we are happy to see this when targetting an Indian language.

### 8.4.3 Evaluating Captioning

While the automatic scores are comparable across tasks, the Hindi-only captioning (“HI”) must be considered separately. Without a source sentence, both humans and machines are very likely to come up with highly varying textual captions. The same image can be described in many different aspects. All our automatic metrics compare the candidate caption with the reference one generally on the basis of the presence of the same character sequences, words or n-grams. Candidates diverging from the reference will get a low score regardless of their actual quality.

The automatic evaluation score for the “Hindi caption” is very very low as compared to other sub-tasks (“text-only” and “multi-modal” translations) as can be seen in the Table 20. Even the human annotators couldn’t give any score for most of the segments submitted from the “Hindi caption” entries due to the wrong caption generation.

## 9 Conclusion and Future Perspective

This paper summarizes the shared tasks of WAT2019. We had 25 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 document-level evaluation using the new dataset with context for some translation subtasks and we would like to consider how to realize context-aware machine translation in WAT. Also, we are planning to do extrinsic evaluation of the translations.

### Acknowledgement

The multi-modal shared task was supported by the following grants at Idiap Research Institute and Charles University.

- At Idiap Research Institute, the work was supported by an innovation project (under an InnoSuisse grant) oriented to im-

prove the automatic speech recognition and natural language understanding technologies for German (Title: “SM2: Extracting Semantic Meaning from Spoken Material” funding application no. 29814.1 IP-ICT). The work was also supported by the EU H2020 project “Real-time network, text, and speaker analytics for combating organized crime” (ROXANNE), grant agreement: 833635.

- At Charles University, the work was supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation and “Progress” Q18+Q48 of Charles University, and using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16013/0001781).

Some of the human evaluations for the other tasks were supported by “Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation”, the Commissioned Research of National Institute of Information and Communications Technology (NICT), JAPAN.

### 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.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Névól, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of the First Conference on Machine Translation (WMT)*. Volume 2: Shared Task Papers, volume 2, pages 131–198, Stroudsburg, PA, USA. Association for Computational Linguistics, Association for Computational Linguistics.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled dis-

- agreement or partial credit. *Psychological Bulletin*, 70(4):213 – 220.
- Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):5.
- Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):17.
- J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. [Can machine translation systems be evaluated by the crowd alone](#). *Natural Language Engineering*, FirstView:1–28.
- Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. [Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation](#). In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139, Dublin, Ireland.
- 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.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. [Opennmt: Open-source toolkit for neural machine translation](#). In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session.
- T. Kudo. 2005. [Mecab : Yet another part-of-speech and morphological analyzer](#). <http://mecab.sourceforge.net/>.
- Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28.
- Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of EACL*, pages 241–248.
- Toshiaki Nakazawa, Chenchen Ding, Hideya MINO, Isao Goto, Graham Neubig, and Sadao Kurohashi. 2016. [Overview of the 3rd workshop on asian translation](#). In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 1–46, Osaka, Japan. The COLING 2016 Organizing Committee.
- Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. [Overview of the 4th workshop on asian translation](#). In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 1–54. Asian Federation of Natural Language Processing.
- Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi, and Eiichiro Sumita. 2014. Overview of the 1st Workshop on Asian Translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT2014)*, pages 1–19, Tokyo, Japan.
- Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 1–28, Kyoto, Japan.
- Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong

- Kong. Association for Computational Linguistics.
- 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.
- 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.
- Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019a. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation. arXiv preprint arXiv:1907.08948.
- Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019b. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *Computación y Sistemas*. In print. Presented at CICLing 2019, La Rochelle, France.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisboa, Portugal. Association for Computational Linguistics.
- Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtský. 2012. Morphological processing for english-tamil statistical machine translation. In Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MT-PIL-2012), pages 113–122.
- Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the asian language treebank. In In Proc. of O-COCOSDA, pages 1–6.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In In Proceedings of Association for Machine Translation in the Americas, pages 223–231.
- Huihsin Tseng. 2005. A conditional random field word segmenter. In In Fourth SIGHAN Workshop on Chinese Language Processing.
- Masao Utiyama and Hitoshi Isahara. 2007. A japanese-english patent parallel corpus. In MT summit XI, pages 475–482.
- Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. [Tensor2tensor for neural machine translation](#). In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers), pages 193–199. Association for Machine Translation in the Americas.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). CoRR, abs/1706.03762.
- Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In ACL 2016 First Conference on Machine Translation, pages 505–510, Berlin, Germany.
- Yi Mon Shwe Sin and Khin Mar Soe. 2018. Syllable-based myanmar-english neural machine translation. In In Proc. of ICCA, pages 228–233.

## Appendix A Submissions

Tables 26 to 31 summarize translation results submitted for WAT2019 pairwise evaluation. Type, RSRC, and Pair columns indicate type of method, use of other resources, and pairwise evaluation score, respectively. The tables also include results by the organizers’ baselines, which are listed in Table 13.



System	ID	Type	RSRC	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
NMT	1900	NMT	NO	36.37	38.48	37.15	0.824985	0.831183	0.833207	0.759910	0.759910	0.759910	—
NTT	3236	NMT	NO	45.83	47.63	46.57	0.861994	0.865640	0.868175	0.774950	0.774950	0.774950	47.75
NICT-2	3182	NMT	NO	44.61	46.59	45.66	0.852970	0.856755	0.859059	0.761340	0.761340	0.761340	42.25
srcb	3212	NMT	NO	45.71	47.55	46.29	0.857506	0.860642	0.862819	0.770440	0.770440	0.770440	40.00
AISTAI	3251	NMT	NO	42.64	44.17	43.34	0.849129	0.851400	0.856177	0.757590	0.757590	0.757590	36.75
KNU_Hyundai	3172	NMT	NO	44.08	45.88	44.78	0.857060	0.859885	0.863400	0.760050	0.760050	0.760050	36.00

Table 26: ASPEC en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM	Pair
NMT	1901	NMT	NO	26.91	0.764968	0.595370	—
NTT	3225	SMT	NO	30.56	0.773281	0.626880	14.000
KNU_Hyundai	3173	NMT	NO	30.88	0.774653	0.622070	11.750
srcb	3205	NMT	NO	30.92	0.778832	0.630150	6.500

Table 27: ASPEC ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM	Pair
NMT	3262	NMT	NO	38.55	0.758265	0.712040	-
sarah	2807	NMT	NO	54.25	0.832909	0.807250	36.750
NTT	3002	SMT	NO	57.34	0.844086	0.816660	34.000
NICT-2	3081	NMT	NO	53.87	0.834898	0.805460	33.500
sarah	2811	NMT	NO	52.77	0.823336	0.794050	29.250
geoduck	3197	Other	YES	54.27	0.828493	0.800360	27.250
geoduck	3216	Other	YES	46.21	0.703675	0.710630	-32.500

Table 28: TDDC ITM ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM	Pair
NMT	3263	NMT	NO	24.11	0.701005	0.583850	-
NTT	3005	NMT	NO	61.19	0.861346	0.753630	55.500
sarah	2808	NMT	NO	58.38	0.865364	0.745110	49.500
NICT-2	3084	NMT	NO	56.95	0.848530	0.744390	49.000
sarah	2812	NMT	NO	54.84	0.845921	0.732790	37.750
geoduck	3200	Other	YES	61.38	0.850551	0.743820	30.000
geoduck	3217	Other	YES	51.08	0.659190	0.623770	-28.250

Table 29: TDDC TXT ja-en submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
NMT	1904	NMT	NO	16.63	17.01	16.36	0.653766	0.657279	0.658830	0.514530	0.514530	0.514530	—
NHK-NES	2886	NMT	YES	27.22	28.75	27.63	0.730816	0.731653	0.734618	0.660570	0.660570	0.660570	87.750
NHK-NES	2885	NMT	YES	28.25	29.76	28.63	0.739036	0.741164	0.743977	0.654260	0.654260	0.654260	81.250
sarah	2814	NMT	NO	22.65	23.48	22.76	0.705044	0.707209	0.710111	0.614840	0.614840	0.614840	63.250
sarah	2815	NMT	NO	21.80	22.67	21.91	0.698609	0.699981	0.701739	0.613770	0.613770	0.613770	55.250

Table 30: JLIJ en-ja submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM			Pair
				juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab	
NMT	1905	NMT	NO	16.48	16.48	16.48	0.640558	0.640558	0.640558	0.459080	0.459080	0.459080	—
NHK-NES	2884	NMT	YES	14.23	14.23	14.23	0.612351	0.612351	0.612351	0.526010	0.526010	0.526010	89.000
NHK-NES	2883	NMT	YES	26.38	26.38	26.38	0.703808	0.703808	0.703808	0.554110	0.554110	0.554110	72.000
sarah	2793	NMT	NO	21.84	21.84	21.84	0.675386	0.675386	0.675386	0.526530	0.526530	0.526530	50.750
sarah	2813	NMT	NO	21.34	21.34	21.34	0.676723	0.676723	0.676723	0.524530	0.524530	0.524530	44.750

Table 31: JLIJ ja-en submissions