

Overview of the 7th Workshop on Asian Translation

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

Hideki Nakayama
The University of Tokyo
nakayama@ci.i.u-tokyo.ac.jp

Chenchen Ding and Raj Dabre and Shohei Higashiyama
National Institute of
Information and Communications Technology
{chenchen.ding, raj.dabre, shohei.higashiyama}@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 7th workshop on Asian translation (WAT2020). For the WAT2020, 20 teams participated in the shared tasks and 14 teams submitted their translation results for the human evaluation. We also received 12 research paper submissions out of which 7 were accepted. About 500 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-WAT2019 (Nakazawa et al., 2019), WAT2020 brings together machine 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 7th WAT, we included the following new tasks:

- Hindi / Thai / Malay / Indonesian ↔ English IT and Wikinews tasks

- Odia ↔ English task
- Bengali / Hindi / Malayalam / Tamil / Telugu / Marathi / Gujarati ↔ English Indic multilingual tasks
- English ↔ Japanese multimodal tasks
- English ↔ Japanese document-level translation tasks

All the tasks are explained in Section 2.

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.
- 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

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

more Asian languages such as Vietnamese, Lao 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 Tasks

2.1 ASPEC Task

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 \leftrightarrow en subtasks, and a Japanese-Chinese scientific paper excerpt corpus (ASPEC-JC), which is used for ja \leftrightarrow 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 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 doc-

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

uments 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 Task

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

At WAT2020, 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 Newswire (JJI) Task

The Japanese ↔ English newswire task uses JJI Corpus which was constructed by Jiji Press Ltd. in collaboration with NICT and NHK. 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. In addition to the test set (test set I) that has been provided from WAT 2017, we added a new test set (test set II) with document-level context at WAT 2020. These test sets are as follows.

Test set I : A pair of test and reference sentences. The references were automatically extracted from English newswire sentences and not manually checked. There are no context data.

Test set II : New test set added at WAT 2020. A pair of test and reference sentences and context data that are articles including test sentences. The references were automatically extracted from English newswire sentences and manually selected. Therefore, the quality of the references of test set II is better than that of test set I.

The statistics of JJI Corpus are shown in Table 3.

The definition of data use is shown in Table 4.

Participants submit the translation results of one or more of the test data.

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

Training		0.2 M sentence pairs
Test set I	Test	2,000 sentence pairs
	DevTest	2,000 sentence pairs
	Dev	2,000 sentence pairs
Test set II	Test-2	1,912 sentence pairs
	Dev-2	497 sentence pairs
	Context for Test-2	567 article pairs
	Context for Dev-2	135 article pairs

Table 3: Statistics for JJI Corpus

2.4 Mixed-domain Task

2.4.1 ALT and UCSY Corpus

The parallel data for Myanmar-English translation tasks at WAT2020 consists of two corpora, 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 5. Notice that both of the corpora have been modified from the data used in WAT2018.

2.4.2 ALT and ECCC Corpus

The parallel data for Khmer-English translation tasks at WAT2020 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.,

Task	Use	Content
Japanese to English	Training	Training, DevTest, Dev, Dev-2, context for Dev2
	Test set I	To be translated Reference
	Test set II	Test-2 Context Reference
English to Japanese	Training	Training, DevTest, Dev, Dev-2, context for Dev2
	Test set I	To be translated Reference
	Test set II	To be translated Context in English for Test-2 Reference

Table 4: Definition of data use in the Japanese ↔ English newswire task

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

Table 5: Statistics for the data used in Myanmar-English translation tasks

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

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

2016), consisting of twenty thousand Khmer-English parallel sentences from news articles.

- The ECCC corpus consists of 100 thousand Khmer-English parallel sentences extracted from document pairs of Khmer-English bilingual 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 6.

Split	Domain	Language Pair			
		Hi	Id	Ms	Th
Train	ALT	18,088			
	IT	254,242	158,472	506,739	74,497
Dev	ALT	1,000			
	IT	2,016	2,023	2,050	2,049
Test	ALT	1,018			
	IT	2,073	2,037	2,050	2,050

Table 7: The NICT-SAP task corpora splits. The corpora belong to two domains: wikinews (ALT) and software documentation (IT). The Wikinews corpora are N-way parallel.

2.5 NICT-SAP Task

This year, we created a new task for joint multi-domain multilingual neural machine translation involving 4 low-resource Asian languages: Thai (Th), Hindi (Hi), Malay (Ms), Indonesian (Id). English (En) is the source or the target language for the translation directions being evaluated. The purpose of this task was to test the feasibility of multi-domain multilingual solutions for extremely low-resource language pairs and domains. Naturally the solutions could be one-to-many, many-to-one or many-to-many NMT models. The domains in question are Wikinews and IT (specifically, Software Documentation). The total number of evaluation directions are 16 (8 for each domain). There is very little clean and publicly available data for these domains and language pairs and thus we encouraged participants to not only utilize the small Asian Language Treebank (ALT) parallel corpora (Thu et al., 2016) but also the parallel corpora from OPUS¹. The ALT dataset contains 18,088, 1,000 and 1,018 training, development and testing sentences. As for corpora for the IT domain we only provided evaluation (dev and test sets)

¹<http://opus.nlpl.eu/>

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

corpora² (Buschbeck and Exel, 2020) and encouraged participants to consider GNOME, UBUNTU and KDE corpora from OPUS. In Table 7 we give statistics of the aforementioned corpora which we used for the organizer’s baselines. Note that we do not list³ all available corpora here and participants were not restricted from using any corpora as long as they are freely available.

2.6 News Commentary Task

For the Russian↔Japanese task we asked participants to use the JaRuNC corpus⁴ (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 translation quality as well but this is beyond the scope of this years sub-task. Refer to Table 8 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,⁵ JESC,⁶ TED,⁷ ASPEC,⁸ UN,⁹ Yandex¹⁰ and Russian↔English news-commentary corpus¹¹. This year we also encouraged participants to use any corpora from WMT 2020¹² involving Japanese, Russian and English as long as it did not belong to the news commentary domain to prevent any test set sentences from being intentionally seen during training.

²Software Domain Evaluation Splits

³<http://lotus.kuee.kyoto-u.ac.jp/WAT/NICT-SAP-Task>

⁴<https://github.com/aizhanti/JaRuNC>

⁵<http://www.phontron.com/kftt/>

⁶<https://datarepository.wolframcloud.com/resources/Japanese-English-Subtitle-Corpus>

⁷<https://wit3.fbk.eu/>

⁸<http://lotus.kuee.kyoto-u.ac.jp/ASPEC/>

⁹<https://cms.unov.org/UNCorpus/>

¹⁰<https://translate.yandex.ru/corpus?lang=en>

¹¹<http://lotus.kuee.kyoto-u.ac.jp/WAT/News-Commentary/news-commentary-v14.en-ru.filtered.tar.gz>

¹²<http://www.statmt.org/wmt20/translation-task.html>

2.7 Indic Multilingual Task

In 2018, we had organized an Indic languages task (Nakazawa et al., 2018) but due to lack of reliable evaluation corpora we discontinued it in WAT 2019. However, in 2020, high quality publicly available evaluation (and training) corpora became available which motivated us to relaunch the task. The Indic task involves mixed domain corpora for evaluation consisting of various articles composed by Indian Prime Minister. The languages involved are Hindi (Hi), Marathi (Mr), Tamil (Ta), Telugu (Te), Gujarati (Gu), Malayalam (Ml), Bengali (Bg) and English (En). English is either the source or the target language during evaluation leading to a total of 14 translation directions. The objective of this task, like the Indic languages task in 2018, was to evaluate the performance of multilingual NMT models. The desired solution could be one-to-many, many-to-one or many-to-many NMT models. We provided a filtered version of the PM India dataset¹³ and further encouraged the use of the CVIT-PIB dataset¹⁴. Our organizer’s baselines used the PMI and PIB corpora for training. Detailed statistics for the aforementioned corpora can be found in Table 9. We also listed additional sources of corpora for participants to use. See Appendix B for details.

2.8 UFAL (EnOdia) Task

This task introduced this year at WAT2020 and the first Odia↔English machine translation shared task running in any conference. For Odia↔English translation task we asked the participants to use OdiEnCorp 2.0 (Parida et al., 2020).¹⁵ The statistics of the corpus are given in Table 10.

2.9 English→Hindi Multi-Modal Task

For English→Hindi multi-modal translation task we asked the participants to use the updated version 1.1 of Hindi Visual Genome corpus (HVG, Parida et al., 2019a,b).¹⁶ The update consisted in correcting primarily the Hindi, i.e. the target side of the corpus.

The statistics of HVG 1.1 are given in Table 11. One “item” in HVG consists of an image with a

¹³<http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/cvit-pmindia-mono-bi.zip>

¹⁴http://preon.iiit.ac.in/~jerin/resources/datasets/pib_v0.2.tar

¹⁵<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3211>

¹⁶<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3267>

Split	Language						
	Bn	Gu	Hi	MI	Mr	Ta	Te
Train	74,593	73,504	247,926	61,678	112,429	122,337	41,741
Dev	2,000						
Test	3,522	4,463	3,169	2,886	3,760	3,637	3,049

Table 9: The Indic task corpora splits. The training corpora statistics are the result of combining the PIB and PMI corpora. While the number of development set sentences are the same, they are not N-way parallel as in the case of the Wikinews corpora.

Dataset	Items	Tokens	
		English	Odia
Train	69,370	1.34M	1.16M
Dev	13,544	158,188	140,726
Test	14,344	186,320	165,274

Table 10: Statistics of OdiEnCorp 2.0 Corpus.

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 2.9.1 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.

Since HVG 1.0 was used already in WAT 2019, all the data were publicly available before WAT 2020. We instructed the participants to use only the Training and D-Test sections and avoid using E-Test and C-Test which are the official test sets for the task this year.

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

2.9.1 English→Hindi 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.

Dataset	Items	Tokens	
		English	Hindi
Training Set	28,930	143,164	145,448
D-Test	998	4,922	4,978
E-Test (EV)	1,595	7,853	7,852
C-Test (CH)	1,400	8,186	8,639

Table 11: Statistics of Hindi Visual Genome 1.1 used for the English→Hindi Multi-Modal translation task. One item consists of a 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.

Data	Images	Sentences/Tokens	
		English	Japanese
Train	29,783	59,566/1.03M	59,562*/0.99M
Dev	1,000	2,000/34,670	2,000/33,022
Test	1,000	1,000/10,876	1,000/17,731

Table 12: Statistics of the dataset used for Japanese↔English multi-modal tasks. Here we use the MeCab tokenizer to count Japanese tokens. *Four of the original English sentences are actually broken so we did not provide their translations.

2.10 Japanese↔English Multi-Modal Tasks

The goal of Japanese↔English multi-modal task¹⁷ is to improve translation performance with the help of another modality (images) associated with input sentences. For both English→Japanese and Japanese→English tasks, we use the Flickr30k Entities Japanese (F30kEnt-Jp) dataset (Nakayama et al., 2020). This is an extended dataset of the Flickr30k¹⁸ and Flickr30k Entities¹⁹ datasets where manual Japanese translations are added. Notably, it has the annotations of many-to-many phrase-to-region correspondences in both English and Japanese captions, which are expected to strongly supervise multimodal grounding and provide new research directions.

We summarize the statistics of our dataset in Ta-

¹⁷<https://nlab-mpg.github.io/wat2020-mmt-jp/>

¹⁸<http://shannon.cs.illinois.edu/DenotationGraph/>

¹⁹<http://bryanplummer.com/Flickr30kEntities/>



	Text-Only MT	Hindi Captioning	Multi-Modal MT
Image	—		
Source Text	man in the middle of tennis court	—	man in a white hat on the tennis court
System Output Gloss	टेनिस कोर्ट के बीच में आदमी Man in the middle of a tennis court	आकाश में सफेद बादल White cloud on the sky	टेनिस कोर्ट पर व्यक्ति Man on the tennis court
Reference Solution	टेनिस कोर्ट के बीच में मनुष्य	फोटो पर तारीख की मोहर।	टेनिस कोर्ट पर एक सफेद टोपी में आदमी
Gloss	Man in the middle of a tennis court	Date stamp on the photo	Man in a white cap on the tennis court

Figure 1: An illustration of the three tracks of WAT 2020 English→Hindi Multi-Modal Task. Note the missing articles in the English source. The correct sentence would be “A man in the middle of a tennis court”. The system outputs are not correct for the Hindi captioning and multimodal with respect to the reference solution, although the output of the captioning system is understandable.

ble 12. We use the same splits of training, validation and test data specified in Flickr30k Entities. For the training and the validation data, we use the F30kEnt-Jp version 1.0 which are publicly available.²⁰ While the original Flickr30k has five English sentences for each image, our Japanese set has the translations of the first two sentences of each. Therefore, we have two parallel sentences for each image. For the test data, we use the Japanese sentences not included in the version 1.0 dataset (i.e., one of the other three sentences for each image) which are not publicly available at the time of WAT 2020. Note that phrase-to-region annotation is not included in the test data.

There are two settings of submission: with and without resource constraints. In the constrained setting, external resources such as additional data and pre-trained models (with external data) are not allowed to use, except for pre-trained convolutional neural networks (for visual analysis) and basic linguistic tools such as taggers, parsers, and morphological analyzers. As the baseline system to compute the Pairwise score, we implement the text-only model in (Nishihara et al., 2020) under the constrained setting.

2.11 Document-level Translation Task

In WAT2020, we set up 2 document-level translation tasks: ParaNatCom and BSD.

²⁰<https://github.com/nlab-mpg/Flickr30kEnt-JP>

2.11.1 Document-level Scientific Paper Translation

Traditional ASPEC translation tasks are sentence-level and the translation quality of them seem to be saturated. We think it’s high time to move on to document-level evaluation. For the first year, we use ParaNatCom²¹ (Parallel English-Japanese abstract corpus made from Nature Communications articles) for the development and test sets of the Document-level Scientific Paper Translation sub-task. We cannot provide document-level training corpus, but you can use ASPEC and any other extra resources.

2.11.2 Document-level Business Scene Dialogue Translation

There are a lot of ready-to-use parallel corpora for training machine translation systems, however, most of them are in written languages such as web crawl, news-commentary, patents, scientific papers and so on. Even though some of the parallel corpora are in spoken language, they are mostly spoken by only one person (TED talks) or contain a lot of noise (OpenSubtitle). Most of other MT evaluation campaigns adopt the written language, monologue or noisy dialogue parallel corpora for their translation tasks. Traditional ASPEC translation tasks are sentence-level and the translation quality of them seem to be saturated. We think it’s high time to move on to document-level evaluation.

²¹<http://www2.nict.go.jp/astrec-att/member/mutiyama/paranacom/>

Lang	Train	Dev	Test	Mono
hi-en	1,609,682	520	2,507	–
hi	–	–	–	45,075,279

Table 13: Statistics for IITB Corpus. “Mono” indicates monolingual Hindi corpus.

For the first year, WAT uses BSD Corpus²² (The Business Scene Dialogue corpus) for the dataset including training, development and test data. Participants of this task must get a copy of BSD corpus by themselves.

2.12 IITB Hindi–English task

In this task we use IIT Bombay English-Hindi Corpus (Kunchukuttan et al., 2018) which contains English-Hindi parallel corpus as well as monolingual Hindi corpus collected from a variety of sources and corpora (Bojar et al., 2014). 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 13.

3 Participants

Table 14 shows the participants in WAT2020. The table lists 14 organizations from various countries, including Japan, India, Singapore, China, Ireland, and Switzerland.

493 translation results by 20 teams were submitted for automatic evaluation and about 121 translation results by 14 teams were submitted for the human evaluation. Table 15 shows tasks for which each team submitted results by the deadline. The human evaluation was conducted only for the tasks with the check marks in “human eval” line.

4 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 participant’s system. That is, the specific baseline system was the standard for human evaluation. At WAT 2020, 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

²²<https://github.com/tsuruoka-lab/BS>

systems were published on the WAT web page.²³ We also have SMT baseline systems for the tasks that started at WAT 2017 or before 2017. The baseline systems are shown in Tables 16, 17, and 18. 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.

4.1 Tokenization

We used the following tools for tokenization.

4.1.1 For ASPEC, JPC, TDDC, JIJI, ALT, UCSY, ECCC, and IITB

- Juman version 7.0²⁴ for Japanese segmentation.
- Stanford Word Segmenter version 2014-01-04²⁵ (Chinese Penn Treebank (CTB) model) for Chinese segmentation.
- The Moses toolkit for English and Indonesian tokenization.
- Mecab-ko²⁶ for Korean segmentation.
- Indic NLP Library²⁷ (Kunchukuttan, 2020) for Indic language segmentation.
- The tools included in the ALT corpus for Myanmar and Khmer segmentation.
- subword-nmt²⁸ 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.

4.1.2 For News Commentary

- The Moses toolkit for English and Russian only for the News Commentary data.

²³<http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/baseline/baselineSystems.html>

²⁴<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

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

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

²⁷https://github.com/anoopkunchukuttan/indic_nlp_library

²⁸<https://github.com/rsennrich/subword-nmt>

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NICT-5	NICT	Japan
*cvit-mt	IIIT Hyderabad	India
NHK-NES	NHK & NHK Engineering System	Japan
ODIANLP	Idiap Research Institute	Switzerland
Kyoto-U+ECNU	Kyoto University and East China Normal University	Japan and China
goku20	Rakuten Institute of Technology Singapore, Rakuten Asia.	Singapore
WT	Wipro	India
HW-TSC	Huawei Translation Services Center	China
ut-mrt	The University of Tokyo	Japan
*iiitsc	Indian Institute of Information Technology	India
adapt-dcu	Dublin City University	Ireland
DEEPNLP	TATA CONSULTANCY SERVICES	India
CNLP-NITS	National Institute of Technology Silchar	India

Table 14: List of participants who submitted translations for the human evaluation in WAT2020 (Note: teams with '*' marks did not submit their system description papers, therefore the evaluation results are UNOFFICIAL according to our policy)

Team ID	ASPEC		JPC						JJI	Multimodal			En-Ja		
	CJ	JC	EJ	JE	CJ	JC	KJ	JK		JE	TX	HI	MM	EJ	JE
TMU							✓							✓	✓
NHK-NES									✓						
ODIANLP										✓	✓				
Kyoto-U+ECNU	✓	✓													
goku20			✓	✓	✓	✓	✓	✓						✓	✓
*HW-TSC															
*iiitsc										✓					
CNLP-NITS										✓		✓			
human eval	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Team ID	Indic Multilingual													
	En-Bn	Bn-En	En-Hi	Hi-En	En-Ml	Ml-En	En-Ta	Ta-En	En-Te	Te-En	En-Gu	Gu-En	En-Mr	Mr-En
NICT-5		✓		✓		✓		✓		✓		✓		✓
*cvit-mt	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ODIANLP	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
HW-TSC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
human eval	✓	✓	✓	✓										

Team ID	UFAL		EnOdia		BSD		Hinden		Multilingual Multi-domain (IT/Wikileaks)							
	En-Od	Od-En	EJ	JE	En-Hi	Hi-En	En-Hi	Hi-En	En-Th	Th-En	En-Ms	Ms-En	En-In	In-En		
NICT-5							✓/✓	✓/✓	✓/✓	-/✓	✓/✓	✓/✓	✓/✓	✓/✓		
*cvit-mt	✓	✓														
ODIANLP	✓	✓														
goku20			✓	✓												
WT					✓	✓										
ut-mrt			✓	✓												
adapt-dcu			✓	✓												
DEEPNLP			✓	✓												
human eval	✓	✓	✓	✓	✓	✓										

Table 15: Submissions for each task by each team. E, J, C, and K denote English, Japanese, Chinese, and Korean respectively. The human evaluation was conducted only for the tasks with the check marks in "human eval" line.

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 16: Baseline Systems I

System ID	System	Type	JJI		ALT	
			ja-en	en-ja	my↔en	km↔en
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	PC-Transer V13 (Commercial system)	RBMT	✓	✓	✓	✓
Online X	Google translate	Other	✓	✓	✓	✓
Online X	Bing translator	Other	✓	✓	✓	✓

Table 17: Baseline Systems II

System ID	System	Type	NewsCommentary ru↔ja	IT&Wikinews {hi,id,ms,th}↔en	od↔en	Indic {bn,hi,gu,m,mr,ta,te}-en	Multimodal	
							en-hi	ja↔en
NMT	OpenNMT's NMT with attention	NMT						
NMT T2T	Tensor2Tensor's Transformer	NMT	✓	✓	✓	✓	✓	✓
NMT OT	OpenNMT-py's Transformer	NMT						
MNMT	Multimodal NMT	NMT						✓

Table 18: Baseline Systems III

- Mecab²⁹ for Japanese segmentation.
- Corpora are further processed by tensor2tensor’s internal pre/post-processing which includes sub-word segmentation.

4.1.3 Indic and NICT-SAP Tasks

- For the Indic task we did not perform any explicit tokenization of the raw data.
- For the NICT-SAP task we only character segmented the Thai corpora as it was the only language for which character level BLEU was to be computed. Other languages corpora were not preprocessed in any way.
- Any subword segmentation or tokenization was handled by the internal mechanisms of tensor2tensor.

4.1.4 For English→Hindi Multi-Modal and UFAL EnOdia Tasks

- Hindi Visual Genome 1.1 and OdiEnCorp 2.0 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.

4.1.5 For English↔Japanese Multi-Modal Tasks

- For English sentences, we applied lowercase, punctuation normalization, and the Moses tokenizer.
- For Japanese sentences, we used KyTea for word segmentation.

4.2 Baseline NMT Methods

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

4.2.1 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.

²⁹<https://taku910.github.io/mecab/>

- 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.

We used the following data for training the NMT baseline systems of NMT with attention.

- All of the training data mentioned in Section 2 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.

4.2.2 Transformer (Tensor2Tensor)

For the News Commentary task, we used tensor2tensor’s³⁰ implementation of the Transformer (Vaswani et al., 2017) and used 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 Indic and NICT-SAP tasks, we used tensor2tensor to train many-to-one and one-to-many models where the latter were trained with the aforementioned token trick. We used default hyperparameter settings corresponding to the “big” model. Since the NICT-SAP task involves two domains for evaluation (Wikinews and IT) we used a modification of the token trick technique for domain adaptation to distinguish between corpora for different domains. In our case we used tokens such as *2alt* and *2it* to indicate whether the sentences belonged to the Wikinews or IT domain, respectively. For both tasks we used 32,000 separate sub-word vocabularies. We trained our models on 1 GPU till convergence on the development set BLEU scores, averaged the last 10 checkpoints

³⁰<https://github.com/tensorflow/tensor2tensor>

(separated by 1000 batches) and performed decoding with a beam of size 4 and a length penalty of 0.6.

4.2.3 Transformer (OpenNMT-py)

For the English→Hindi Multimodal and UFAL EnOdia tasks, 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.

4.2.4 Multimodal Transformer with Supervised Attention

As the baselines for the English↔Japanese Multimodal tasks, we implement five models described in (Nishihara et al., 2020): (i) the text-only Transformer NMT model, (ii) the multimodal Transformer NMT model (MNMT) which incorporates the ResNet50 convolutional neural network to extract visual features, (iii) the MNMT model with the supervised visual attention mechanism (MNMT+SVA), (iv) the MNMT model with the supervised cross-lingual attention mechanisms (MNMT+SCA), and (v) the MNMT model with the both supervised attentions (MNMT+SVA+SCA). For training with the supervised attention mechanisms, we utilize the phrase-to-phrase and phrase-to-region annotations available in the training dataset.

5 Automatic Evaluation

5.1 Procedure for Calculating Automatic Evaluation Score

We evaluated translation results by three metrics: BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and AMFM (Banchs et al., 2015). BLEU scores were calculated using `multi-bleu.perl` in the Moses toolkit (Koehn et al., 2007). RIBES scores were calculated using `RIBES.py` version 1.02.4.³¹ AMFM scores were calculated using scripts created by the technical collaborators listed in the WAT2020 web page.³² All scores for each task were calculated using the corresponding reference translations.

³¹<http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html>

³²lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/

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 model³³ and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0.³⁴ 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.³⁵ For Korean segmentation, we used `mecab-ko`.³⁶ For Myanmar and Khmer segmentations, we used `myseg.py`³⁷ and `kmseg.py`.³⁸ For English and Russian tokenizations, we used `tokenizer.perl`³⁹ in the Moses toolkit. For Indonesian and Malay tokenizations, we used `tokenizer.perl` as same as the English tokenization. For Thai tokenization, we segmented the whole character separately. For Bengali, Gujarati, Hindi, Marathi, Malayalam, Odia, Tamil, and Telugu tokenizations, we used Indic NLP Library⁴⁰ (Kunchukuttan, 2020). The detailed procedures for the automatic evaluation are shown on the WAT2020 evaluation web page.⁴¹

5.2 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;

³³<http://www.phontron.com/kytea/model.html>

³⁴<http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz>

³⁵<http://nlp.stanford.edu/software/segmenter.shtml>

³⁶<https://bitbucket.org/eunjeon/mecab-ko/>

³⁷<http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/wat2020.my-en.zip>

³⁸<http://lotus.kuee.kyoto-u.ac.jp/WAT/km-en-data/km-en.zip>

³⁹<https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl>

⁴⁰https://github.com/anoopkunchukuttan/indic_nlp_library

⁴¹<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:

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

- Publish the results of the evaluation: whether or not they permit to publish automatic evaluation scores on the WAT2020 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 includ-

ing SMT, RBMT, SMT and RBMT, EBMT, NMT and Other.

Evaluation scores of translation results that participants permit to be published are disclosed via the WAT2020 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 WAT2020. Anybody can reg-

ister an account for the system by the procedures described in the registration web page.⁴²

5.3 Additional Automatic Scores in Multi-Modal and UFAL EnOdia Tasks

For the multi-modal task and UFAL EnOdia task, several additional automatic metrics were run aside from the WAT evaluation server, namely: BLEU (this time calculated by Moses scorer⁴³), 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 tokenizer⁴⁴ on the candidate and reference before scoring. For all error metrics, i.e. metrics where better 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.

These additional scores document again, that BLEU implementations (and the underlying tokenization schemes) heavily vary in their outcomes. The scores are thus comparable only within each of the metric variation, even if it is supposed to be the same “BLEU”. In Table 22, we highlight with a special symbol whenever the ranking in one of the metrics differs from the top-to-bottom sorting of the scores. Last year, a number of these metric, including our BLEU vs. official WAT BLEU (BLEU_w in Table 22) lead to varying rankings. This year, the system differences are probably sufficiently big in the optics of these metrics that only nCharacTER in the E-Test text-only (“EV TEXT”) scoring differs.

6 Human Evaluation

In WAT2020, we conducted 2 kinds of human evaluations: *pairwise evaluation* (only for JaEn multi-modal translation task, Section 6.1) and *JPO adequacy evaluation* (other than HiEn multi-modal translation task, Section 6.2) and *a pairwise variation of direct assessment* (Section 6.4) for the HiEn multi-modal task.

⁴²<http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/registration/index.html>

⁴³<https://github.com/moses-smt/mosesdecoder/blob/master/mert/evaluator.cpp>

⁴⁴<https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl>

6.1 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 4).

6.1.1 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.

6.1.2 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 $S = \sum j_i$:

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

6.1.3 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:

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

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

6.2 JPO Adequacy Evaluation

We conducted JPO adequacy evaluation for the top two or three participants’ systems of pairwise evaluation for each subtask.⁴⁵ The evaluation was carried out by translation experts based on the JPO

⁴⁵The number of systems varies depending on the subtasks.

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

Table 19: The JPO adequacy criterion

adequacy evaluation criterion, which is originally defined by JPO to assess the quality of translated patent documents.

6.2.1 Sentence Selection and Evaluation

For the JPO adequacy evaluation, the 200 test sentences were randomly selected from the test sentences. For the Newswire (JJI) task test set II, articles were randomly selected from the context of test set II until the number of the test sentences that were contained in the selected articles became 200.

For each test sentence, input source sentence, translation by participants’ system, and reference translation were shown to the annotators. For the Newswire (JJI) task test set II, input source sentences were shown in articles, which means that not only input source sentences but also their context were shown to the evaluators. The evaluators considered the context of the input sentences to evaluate the translations. To guarantee the quality of the evaluation, each sentence was evaluated by two annotators. Note that the selected sentences are basically the same as those used in the previous workshop (WAT2019).

6.2.2 Evaluation Criterion

Table 19 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 is described in the JPO document (in Japanese).⁴⁶

⁴⁶http://www.jpo.go.jp/shiryuu/toushin/chousa/tokkyohonyaku_hyouka.htm

6.3 Manual Evaluation for the UFAL (EnOdia) Task

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

The interpretation of these judgements is carried out as described in the following Section 6.4.

6.4 Manual Evaluation for the English→Hindi Multi-Modal Task

The evaluations of the three tracks of the multi-modal task and also the UFAL EnOdia 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, it is assumed that each of the annotators stabilizes in their scoring criteria.

The collected DA scores can be either directly averaged for each system and track (denoted “Ave”), or first standardized per annotator across all annotation tasks 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 individual annotator across all tasks he or she annotated is 0 and the standard deviation is 1.

Our evaluation differs from the basic DA in the following respects: (1) we run the evaluation bilinearly, 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 evaluation. 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 last year. 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 better stabilize their score assignments by seeing another candidate output. This essentially direct pairwise comparison could be useful especially for systems very

Data :ENOD_ANNNOTATOR_4

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 English to Odia translation task.

Data :ODEN_ANNNOTATOR_0

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 4: Manual evaluation of Odia to English translation task.

close in their performance.⁴⁷

The user interface for our annotation for each of the tracks is illustrated in Figures 5 to 7. By default, the position of the slider appears to be at the “worst” score but technically, the user interface is capable of distinguishing if the annotator has touched the slider at all or not. The default score is -1 while the lowest score that the annotator can assign is 0.

Table 20 provides an overview of the usage of sliders in annotation for the 6 Hindi annotators (H*, each scoring 1256 outputs) and 6 Odia annotators (O*, each scoring 576 outputs). The Hindi annotation was carried out first and we observed that the default value was left untouched (“Unscored”) rather often, in up to 45.6% of cases for the annotator H2. Given this large proportion, we decided to consider two interpretations of the default score -1: we either disregard these scorings, assuming that the annotator forgot about that particular slider, or we interpret the scoring as “Worst”, i.e. merging the “Unscored” and “Min” cases. For the

	Value Chosen [% of Cases]				Total Cases
	Unscored	Min	Other	Max	
H0	25.8	0.2	52.8	21.3	1256 (100.0%)
H1	13.5	0.0	41.4	45.1	1256 (100.0%)
H2	45.6	0.3	23.4	30.7	1256 (100.0%)
H3	18.6	1.9	78.9	0.6	1256 (100.0%)
H4	21.6	0.2	58.5	19.7	1256 (100.0%)
H5	11.7	0.1	58.8	29.4	1256 (100.0%)
O0	0.2	0.7	90.8	8.3	576 (100.0%)
O1	0.3	0.5	87.5	11.6	576 (100.0%)
O2	0.0	0.3	90.8	8.9	576 (100.0%)
O3	0.5	8.5	67.7	23.3	576 (100.0%)
O4	0.0	8.5	70.0	21.5	576 (100.0%)
O5	0.0	10.9	65.3	23.8	576 (100.0%)

Table 20: Usage of DA sliders in English→Hindi Multi-Modal and UFAL EnOdia Tasks.

Odia task, we urged the annotators to touch every slider. The low “Unscored” rates indicate that this reminder helped and only very few items were forgotten.

It may seem surprising that many annotators used the very highest score (“Max”). This is possible because the sentences are often short and simple and also because human translations are included in the scoring. The big differences in the usage of the extreme values however justify the need for score standardization.


In the “text-only” evaluation, one English text (source) and two Hindi translations (candidate 1

⁴⁷For the full statistical soundness of the subsequent interpretation of DA scores, the judgments should be *independent* of each other. We explicitly ask our annotators to judge the candidates independently but the dependence cannot be denied. Whether the violation of the independence assumption is offset by the benefit of obtaining more stable judgements is yet to be analyzed

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



Sentence: 5
 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 6: Manual evaluation of multi-modal translation.



Sentence: 8
 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 7: Manual evaluation of Hindi captioning.

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.

7 Evaluation Results

In this section, the evaluation results for WAT2020 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2020 website.⁴⁸

⁴⁸<http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/>

7.1 Official Evaluation Results

Figures 8 and 9 show the official evaluation results of ASPEC subtasks, Figures 10, 11, 12, 13, 14 and 15 show those of JPC subtasks, Figure 16 shows that of JJI-c subtask, Figures 17 and 18 show those of MMT subtasks, Figures 19, 20, 21 and 22 show those of Indic Multilingual subtasks, Figures 23 and 24 show those of BSD subtasks and Figures 25 and 26 show those of Hinden 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 Table 21. The weights for the weighted κ (Cohen, 1968) is defined as $|Evaluation1 - Evaluation2|/4$.

The automatic scores for the multi-modal and UFAL EnOdia tasks along with the WAT evaluation server BLEU scores are provided in Table 22. For each of the test sets of the multi-modal task (E-Test, 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 23 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.

The UFAL EnOdia task has its official manual scores listed in Table 24.

8 Findings

8.1 ASPEC Task

There is only one team (Kyoto-U+ECNU) who participated ASPEC task this year. Kyoto-U+ECNU team participated Japanese ↔ Chinese translation subtasks. They achieved the state-of-the-art automatic evaluation scores, however, the human evaluation scores are below those of last year’s (see Figure 8 and 9). Strictly speaking, the human evaluation scores of this year and last year are not directly comparable because the evaluators might be different.

They trained a lot of different NMT models which exploit 1) different training data (out-of-domain external data, back/forward-translation of Japanese-side of ASPEC-JE), 2) different S2S frameworks (LSTM, ConvS2S, Transformer and Lightconv) and 3) different model capacities (different hyperparameter settings). They also tried to use mBART. Among all the models, the ones which exploit data augmentation by back/forward-translation performed the best. They also tried to combine various NMT models trained above. The BLEU score improves a lot (about 0.5 to 1.5 points) by adding first 2 or 3 models, however, the impact is getting smaller (about 0.1 to 0.2 points improvement) after that.

From the results, we can say that using external resources and combining various models both improves the automatic evaluation scores because of the generalization effect or improvement of fluency, but they might have a bad effect on lexical choice of technical term translations which directly

affect the human evaluation scores.

8.2 JPC Task

Two teams participated in the JPC task; goku20 submitted their systems for all language pairs ($J \leftrightarrow E$, $J \leftrightarrow C$, and $J \leftrightarrow K$) and TMU submitted their systems for the $K \rightarrow J$ pair. goku20 used baseline Transformer models and mBART models, which were pre-trained on large-scale monolingual corpus in 25 languages, fine-tuned on the JPO corpus. TMU used baseline Transformer models, enhanced models with Hanja loss to obtain close embeddings of Sino-Korean (Hanja) and Sino-Japanese (Kanji) words, and domain adaptation models fine-tuned for each domain on source test sentences and target sentences translated by their domain-specific model. Domain indicates section based on IPC: chemistry, electricity, mechanical engineering, or physics.

We discuss results on test-N data as follows. For $J \rightarrow C$ and $J \rightarrow E$, goku20’s ensemble Transformer models achieved higher BLEU scores than the best systems in previous years’ WAT except for systems using additional resources. According to goku20’s experiments, single mBART models outperformed single Transformer models for several language pairs but ensemble mBART models underperformed ensemble Transformer models for all language pairs. For $K \rightarrow J$, TMU’s domain adaptation model achieved the highest BLEU score, while their Hanja loss+domain adaptation models (and their unofficial baseline Transformer model) achieved similar scores. According to TMU’s experiments, single Hanja loss model slightly improved single Transformer model but no difference was observed between ensemble versions of both types of models. As for JPO adequacy evaluation, goku20 submitted ensemble mBART models for all language pairs and TMU submitted ensemble domain adaptation model and ensemble domain adaptation + Hanja loss model for $K \rightarrow J$. All systems by both teams achieved high adequacy scores close to or better than 4.5.

Thus, evaluation results in this year demonstrated high translation accuracy by Transformer models similarly to WAT2019’s results. Although both teams reported improvements by their additional techniques, their enhanced models performed similarly or worse than strong Transformer baselines if ensemble models were used.

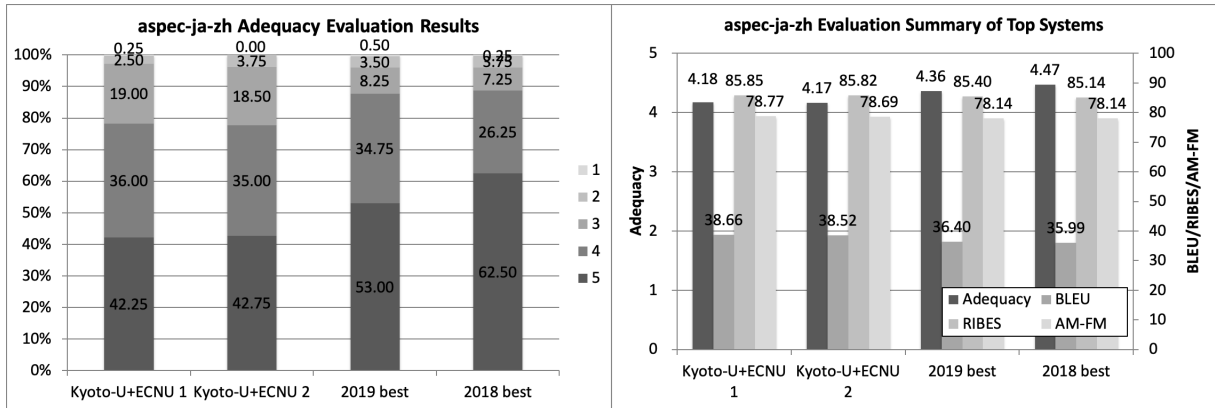


Figure 8: Official evaluation results of aspec-ja-zh.

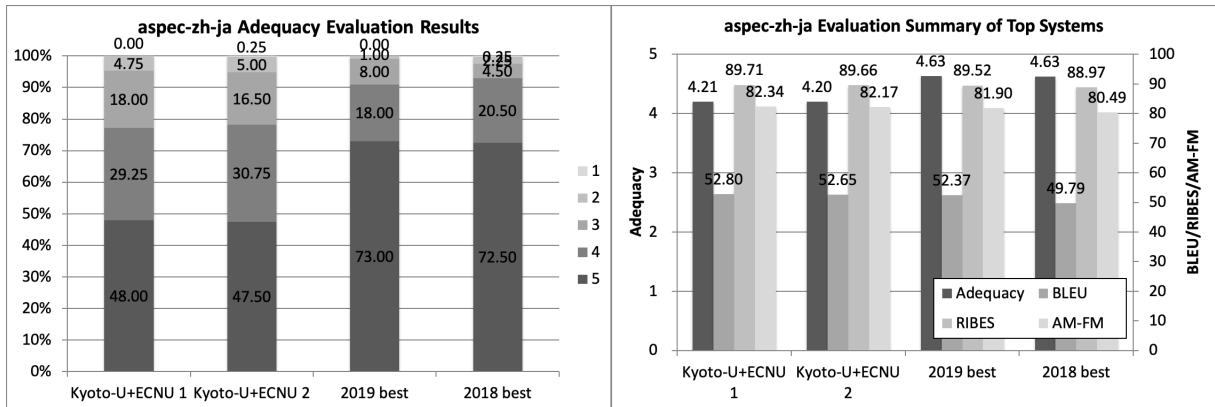


Figure 9: Official evaluation results of aspec-zh-ja.

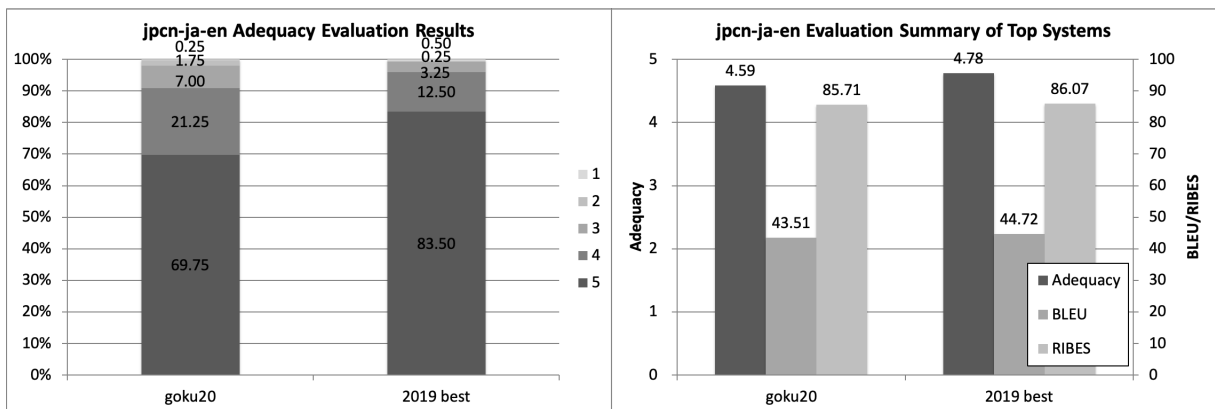


Figure 10: Official evaluation results of jpcn-ja-en.

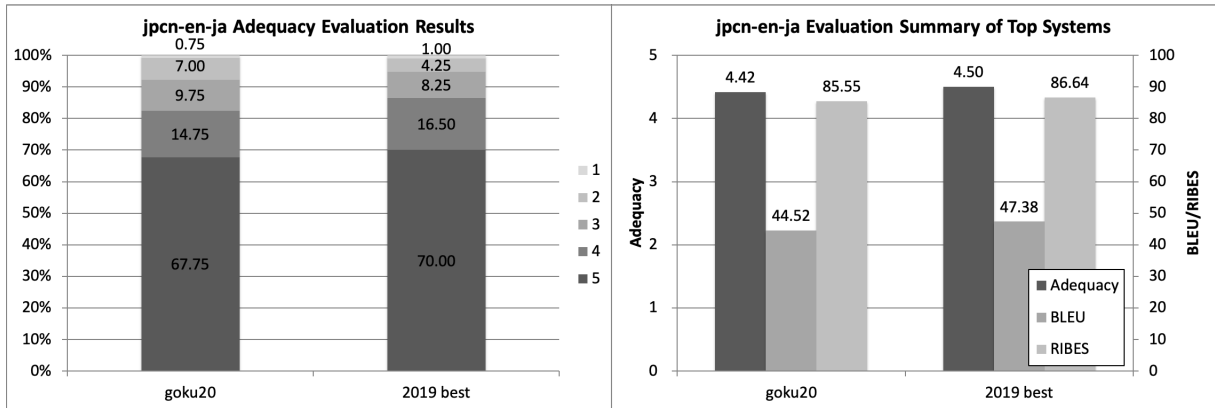


Figure 11: Official evaluation results of jpcn-en-ja.

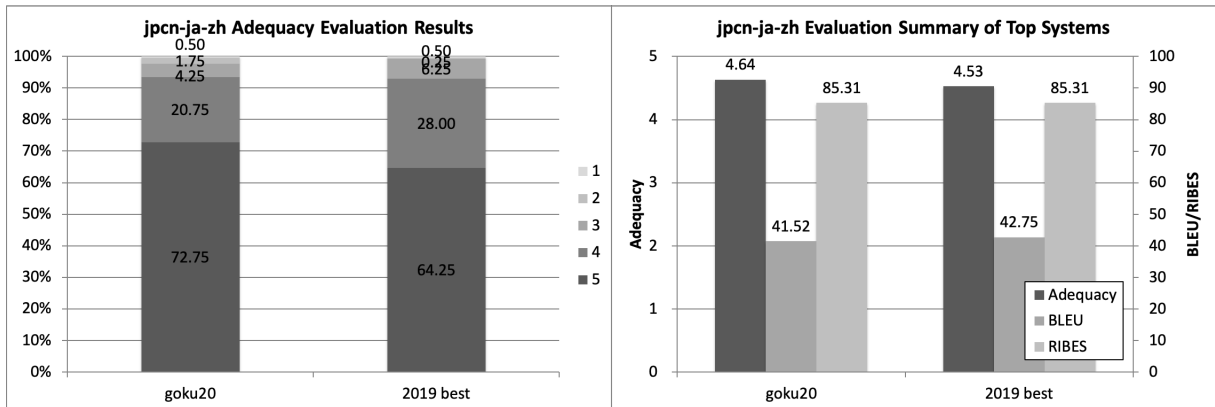


Figure 12: Official evaluation results of jpcn-ja-zh.

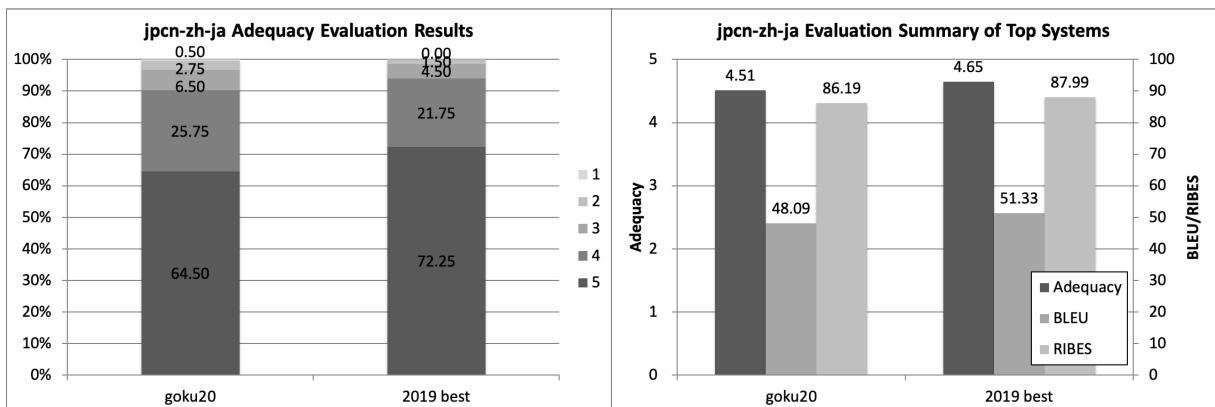


Figure 13: Official evaluation results of jpcn-zh-ja.

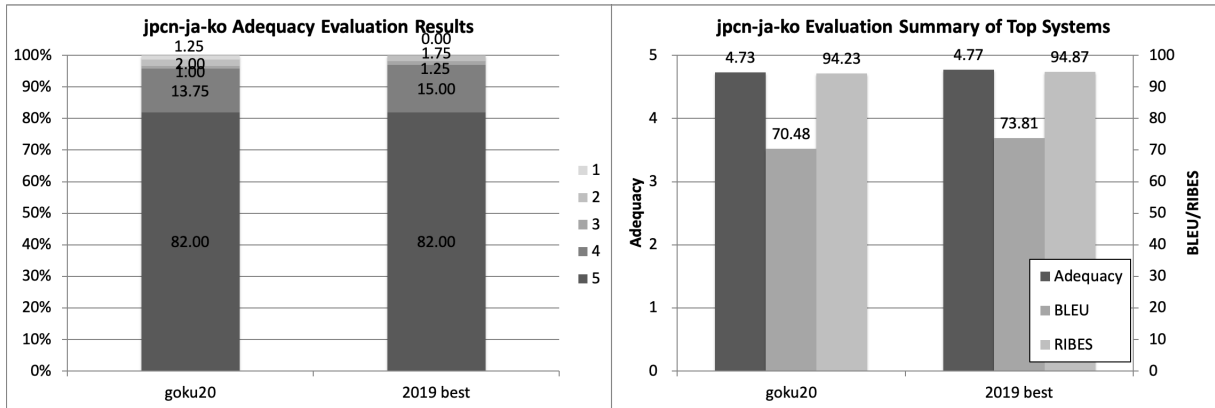


Figure 14: Official evaluation results of jpcn-ja-ko.

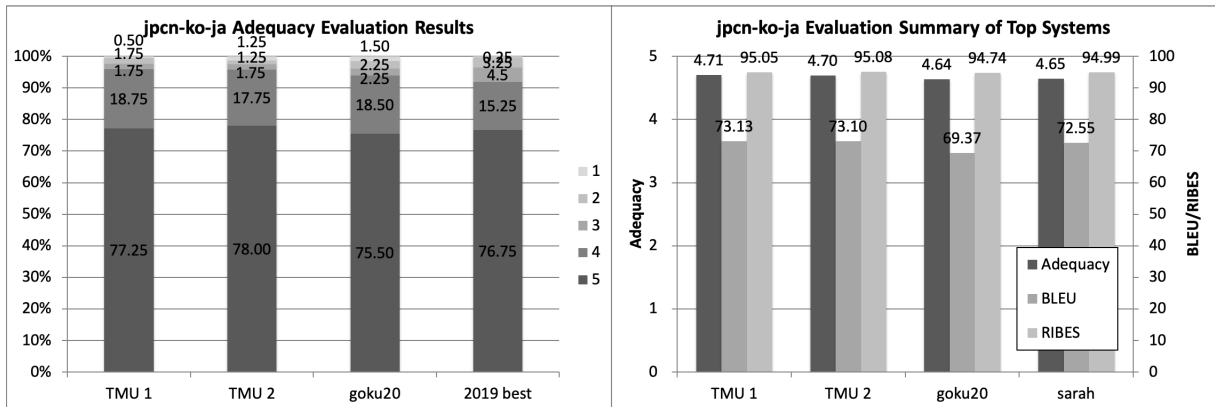


Figure 15: Official evaluation results of jpcn-ko-ja.

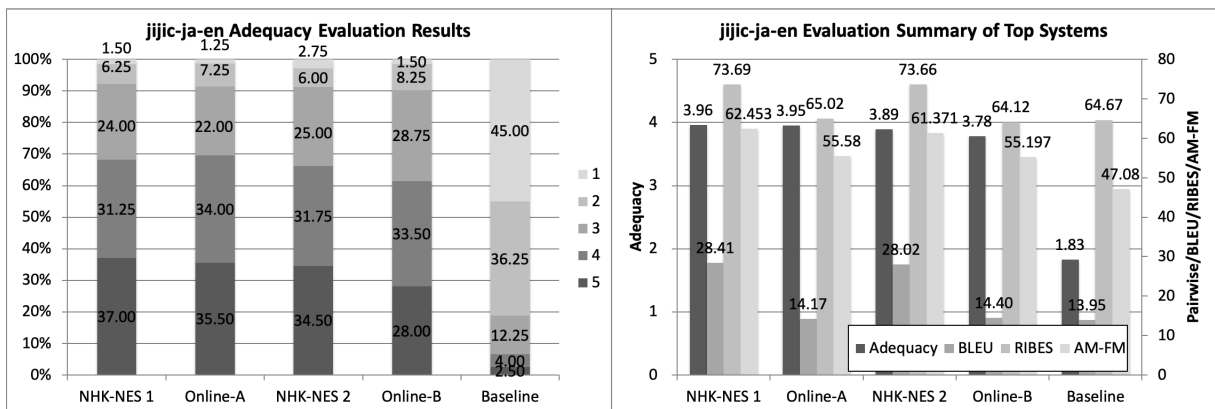


Figure 16: Official evaluation results of jijic-ja-en.

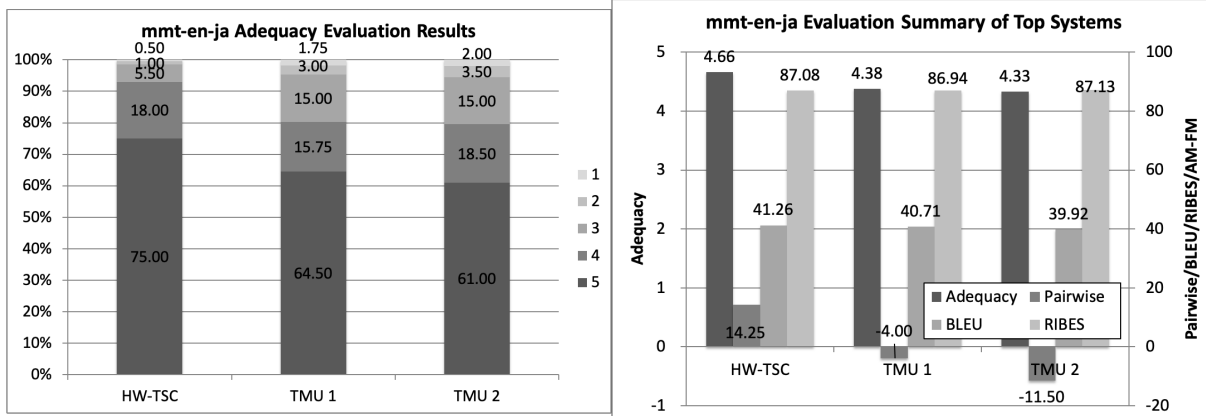


Figure 17: Official evaluation results of mmt-en-ja.

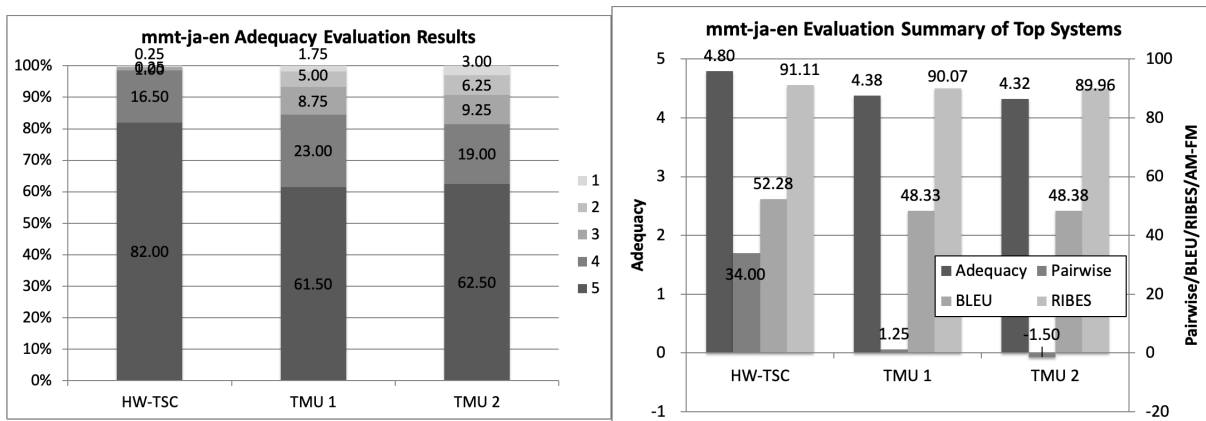


Figure 18: Official evaluation results of mmt-ja-en.

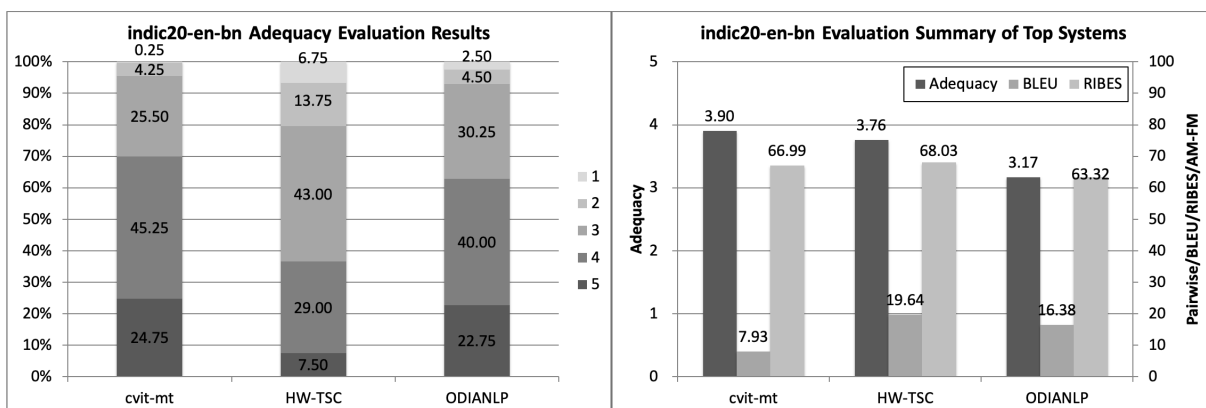


Figure 19: Official evaluation results of indic20-en-bn.

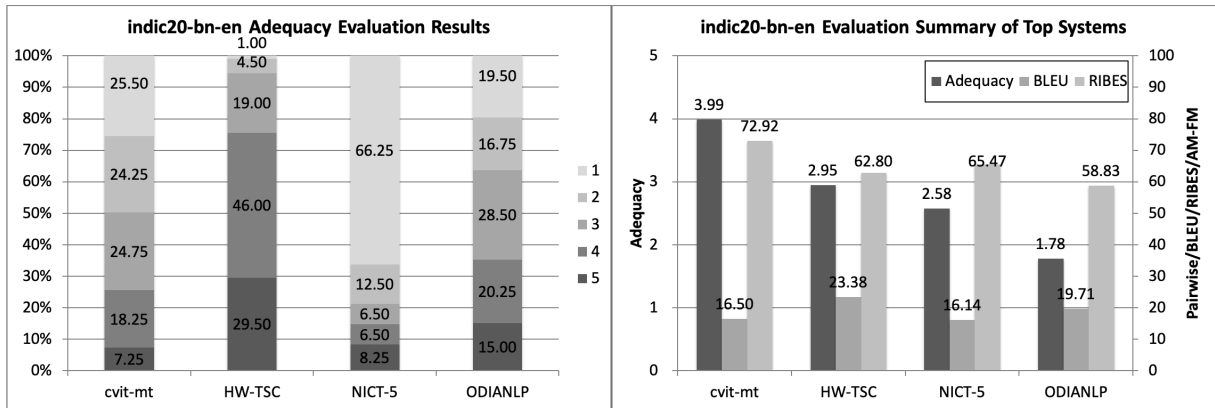


Figure 20: Official evaluation results of indic20-bn-en.

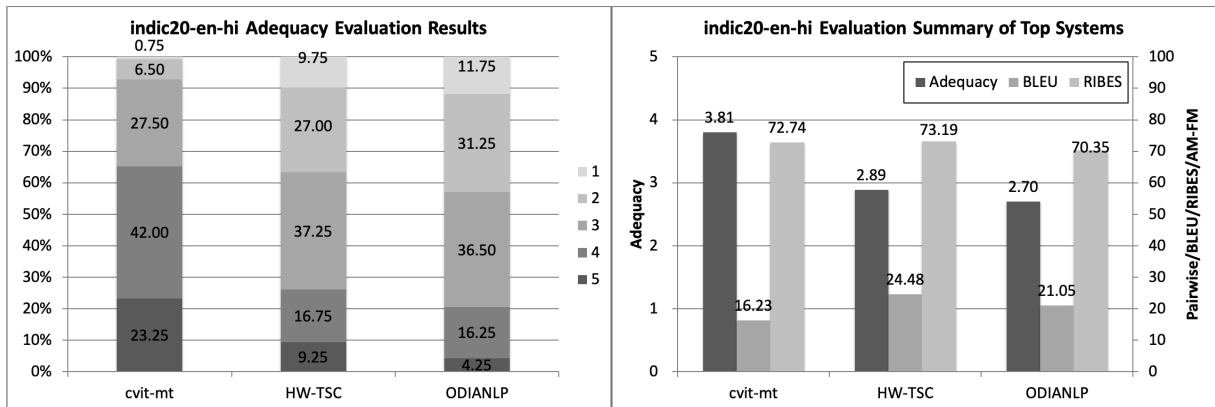


Figure 21: Official evaluation results of indic20-en-hi.

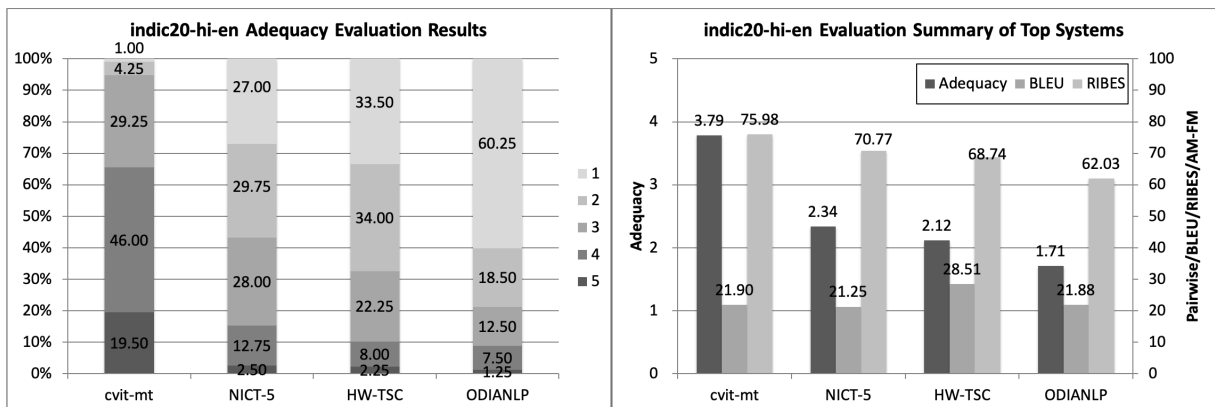


Figure 22: Official evaluation results of indic20-hi-en.

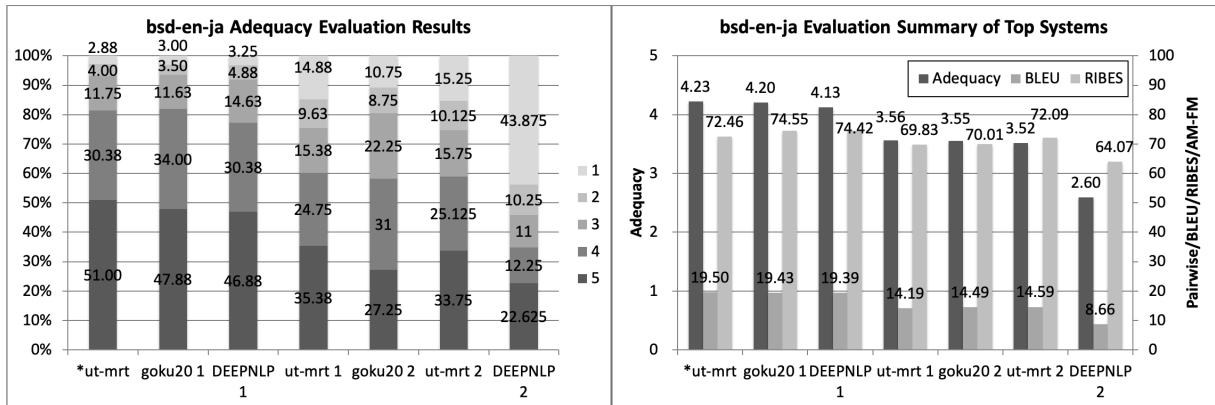


Figure 23: Official evaluation results of bsd-en-ja.

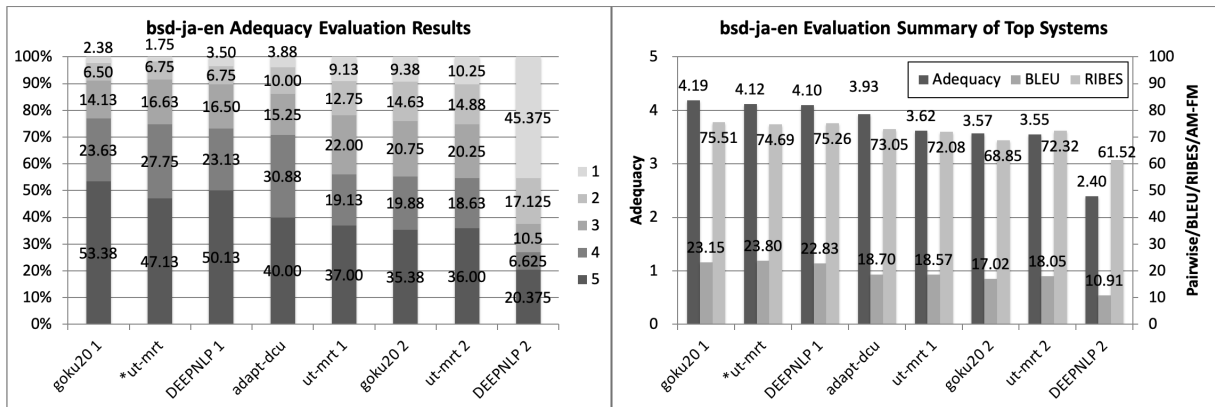


Figure 24: Official evaluation results of bsd-ja-en.

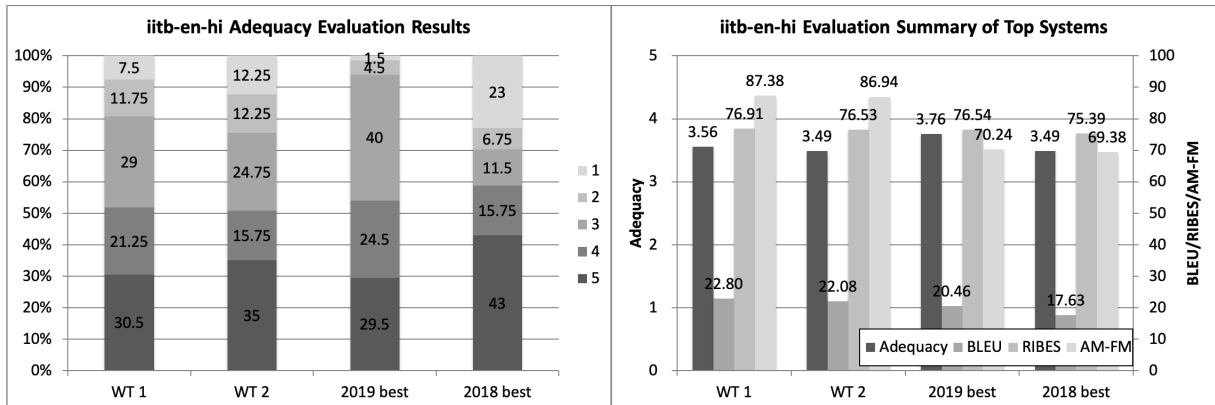


Figure 25: Official evaluation results of iitb-en-hi.

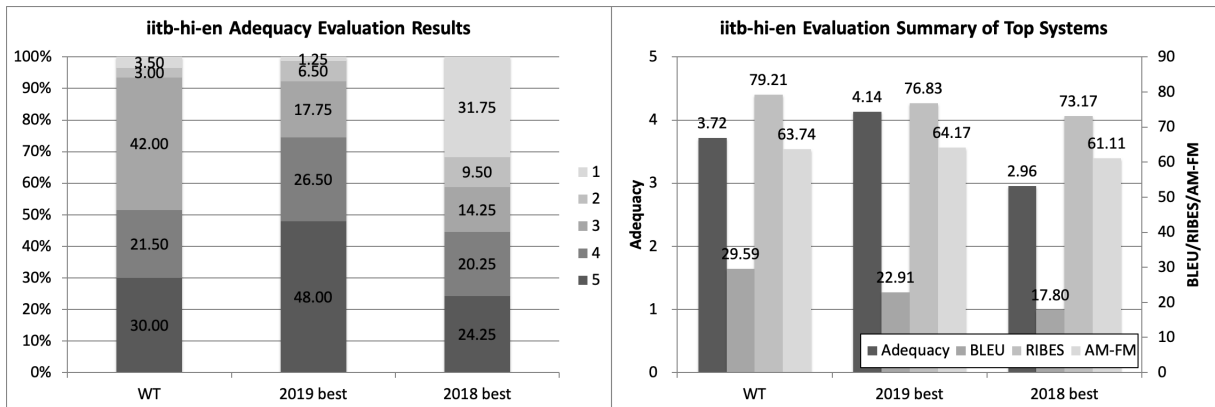


Figure 26: Official evaluation results of iitb-hi-en.

8.3 Newswire (JJI) Task

There were two submissions to the Japanese-to-English task from the NHK-NES team. The two submissions were translations without using context and translations using context. The team addressed the problem of translating zero subject sentences in Japanese into English by extracted subjects and topics from source context using deep analysis and added them to the input sentences as context. The automatic evaluation scores and the human evaluation scores of JPO Adequacy were improved by using context. Although official training data does not contain contextual information, the team used external training data that contained contextual information.

8.4 NICT-SAP Task

Despite the novelty of the task and the availability of clean evaluation data for Wikinews and Software Documentation domain, we had only 1 submission from the “NICT-5” team. They submitted a many-to-many model which in most cases, significantly outperformed the organizers baselines which were one-to-many and many-to-one models. Thai to English translation was significantly lower (around 10 BLEU) compared to all other translation directions. Human evaluation was not performed and at present it is difficult to draw any conclusions on the translation quality due to lack of participation.

8.5 Indic Multilingual Task

Of the several submissions we collected from 4 teams (excluding organizers), the best translations, as measured by BLEU, were submitted by “HW-TSC”. The BLEU scores for most translation directions for this team were significantly higher compared to the other participants. With regards to automatic evaluation scores, translation into English was observed to be substantially better than translation into the Indic languages. This is understandable because Indic languages are morphologically richer than English. Furthermore, translation quality to and from Dravidian languages such as Tamil, Telugu and Malayalam was observed to be the least when compared to the translation quality to and from the Indo-Aryan languages Bengali, Marathi, Hindi and Gujarati. Given that the Dravidian languages are morphologically richer than the Indo-Aryan ones, making them hard to translate or translate into. Marathi is a special language which ex-

hibits some properties of Dravidian languages such as agglutination making it morphologically richer than the other Indo-Aryan languages. This causes translation quality to and from Marathi to be lower than the translation quality to and from the other Indo-Aryan languages.

Human evaluation was done for Hindi–English and Bengali–English (both directions) which revealed that higher BLEU scores often did not correlate with what humans considered as higher quality translations. A deeper look showed that while “HW-TSC” had significantly higher BLEU scores, the percentage of sentences that were perfectly translated (a rating of 5 by human evaluators) were substantially lower than the percentage of sentences that were perfectly translated by the team “cvit-mt”. For all human evaluated directions, translations by “cvit-mt” was rated to be the best despite lagging behind in terms of BLEU when compared with “HW-TSC”. This indicates that human evaluation and automatic evaluation focus on different aspects of translation quality and thus both types of evaluation should be performed in order to better evaluate the quality of translations.

8.6 UFAL (EnOdia) Task

This year, four teams participated in this new task. For the English→Odia translation task, we received 10 submissions from four teams (excluding organizers) which includes five submissions from the team “cvit”, two submissions from the teams “ODIANLP” and “ADAPT” and a single submission from the team “NLPRL”. For the Odia→English translation task, we received 6 submissions from three teams (excluding organizers) which includes three submissions from the team “cvit”, two submissions from the team “ODIANLP”, and a single submission from the team “NLPRL”.

The team “ODIANLP” obtained the highest BLEU score for both English→Odia and Odia→English translation tasks.

Manual evaluation of EnOdia Task is provided in Table 24. Regardless the exact interpretation of the rankings (see the discussion in the following section on Hindi Multi-Modal task), manual translation is the best, followed by “cvit” in the translation into Odia and by “ODIANLP” in the translation into English.

8.7 English→Hindi Multi-Modal Task

This year three teams participated in the different sub-tasks (TEXT, MM, and HI) of the English→Multi-Modal task. The team “ODIANLP” obtained the highest BLEU score for the text-only translation (TEXT) for both the evaluation (E-Test) and challenge (C-Test) test set. For the captioning and multimodal sub-tasks (HI and MM), we received only one submission from the teams “ODIANLP” and “CNLP-NITS”, respectively.

In order to make the evaluation better grounded, we included also the outputs of the best system in each of the sub-tasks in the annotation. In manual ranking, the 2020 systems thus compete not only with the reference translation but also with the winner from 2019.

It should be noted that a revision of E-Test and C-Test files was carried out between the years but the source English did not really change.⁴⁹ The 2019 system outputs could be thus directly used in this year’s evaluation.

Table 23 in the appendix presents the results of the manual annotation. We compare the two interpretations, either ignoring items where the slider was not touched by the annotator, or interpreting it as the lowest value. The final ranking of the systems do not change (despite substantial changes in the actual average scores). Similarly, the standardization of the scores (“Ave” vs. “Ave Z”) do not change the overall ranking of the systems.

Across the tasks, 2020 systems perform better than the best system from 2019. The reference translation generally scores much better than the best system in each task, except for text-only translation of the E-Test. Same as last year, the best system comes out marginally better than the reference. Interestingly, IDIAP system ID 2956, which surpassed the reference in 2019 ended up fourth this year. This can be explained by the different random choice of evaluated sentences this year, but is still clearly illustrates that text-only competition is tight.

Across the various tasks, each of the three participating teams got its medal.

⁴⁹The revision affected 906 out of 1595 lines of E-Test but the changes were only in the Hindi (i.e. target) texts, with one exception, correcting the typo “mam” to “man” in one of the English source sentences. C-Test was corrected in 376 out of 1400 lines, again mainly in the target Hindi. The only two changes in the English side were a correction of a typo (“wres”→“wires”) and a rather unfortunate removal of quotation marks around a “press here” sticker text.

8.8 Japanese↔English Multi-Modal Tasks

This year two teams participated in both Japanese→English and English→Japanese tasks. Overall, we found that the positive effect of image information was not very significant. In English→Japanese task, MMT systems slightly outperform text-only NMT baselines, but in English→Japanese task, NMT achieved equal or better performance. This is perhaps because of the lack of sufficient training data which could make it difficult to properly optimized additional network parameters for visual inputs. We are planning to increase the size of the dataset so that we can hopefully mitigate this issue. Another observation is that no systems except for the baselines provided by organizers used phrase-to-phrase or phrase-to-region annotations in the dataset. We believe strong text–image grounding is the key to more successful MMT, and there is much room for research to utilize this information in future.

8.9 Document-level Translation Task

WAT2020 has 2 kinds of document-level translation tasks for the first time: English → Japanese Scientific Paper (ParaNatCom) and English ↔ Japanese Business Scene Dialogue (BSD). Unfortunately there was no participants for ParaNatCom (we think this is because there is no document-level training data available), and 4 participants for BSD.

The evaluation results of BSD tasks are in Figures 23 and 24. The ut-mrt team did not select the best-BLEU submissions for the human evaluation, we (organizers) additionally evaluate them and show the results in the figure with the ‘*’ mark with the team ID. From the results, the top 3 submissions (goku20 1, DEEPNLP 1 and *ut-mrt) are competitive while their systems are different. goku20 used mBART which considers previous 3 sentences together with the source sentence. The other two did not consider context in their system, but they used a lot of external resources with fine tuning on the provided training data.

Interestingly, goku20 reported that context-aware models fine-tuned only on the provided data did not improve the translation quality, but it has a good effect when it is fine-tuned on a larger data (they used JESC corpus). adapt-dcu also reported very similar results: mixed-fine tuning using JESC and OpenSubtitles along with BSD is much better than using only BSD. ut-mrt tried to use context-

aware model (ut-mrt 2), but it did not improve the translation quality compared to the model without context (ut-mrt 1). From these results, fine-tuning also requires a substantial amount of training data with context.

Another interesting point is that smaller (6,000) sub-word units leads to better translation quality, which is reported by adapt-dcu. We currently do not have any idea about the reason of this result. It is worth investigating the results deeper in the future.

8.10 IITB Hindi–English Task

This year we received two submissions for English to Hindi translation and one submission for Hindi to English translation, all from one team (“WT”). For English to Hindi translation, the submissions had about 1 to 2 BLEU points higher than the best submissions of 2018 and 2019. The AM-FM scores were substantially higher by approximately 16 to 17 points compared to previous years. However the human evaluation for adequacy showed that the translation quality was slightly worse than the best translation quality in 2019 and comparable to the best translation quality in 2018. For the reverse direction, Hindi to English, there was an explosive growth in translation quality as measured by BLEU. Where the best submission of 2019 had a BLEU score of 22.91, the best (and only) submission of 2020 had a BLEU score of 29.59. Human evaluation (adequacy), revealed that the overall adequacy score is less than the best adequacy score in 2019. Deeper investigation showed that this is due to most translations in 2020 being of average quality compared to the best translations in 2019 (42.00% in 2020 versus 17.17% in 2019). On the other hand the number of perfectly rated translations dropped substantially from 48.00% in 2019 to 30% in 2020. This explains why the overall human evaluation score for the best 2019 translations is higher than the one for the 2020 translations. This discrepancy between BLEU score and human evaluation shows the importance of manual investigations of translation instead of blindly relying on automatic scores.

9 Conclusion and Future Perspective

This paper summarizes the shared tasks of WAT2020. We had 14 participants worldwide who submitted their translation results for the human evaluation, and collected a large number of use-

ful submissions for improving the current machine translation systems by analyzing the submissions and identifying the issues.

For the next WAT workshop, we will include several new datasets and new languages (Arabic and additional Indic languages). We are also planning to provide document-level test sets for some translation tasks because sentence-level translation quality is almost saturated for some tasks. We also plan to have a new shared task named “Restricted Translation” task where we will investigate translation with restricted target vocabularies.

Acknowledgement

The English→Hindi Multi-Modal and UFAL En-Odia shared tasks were supported by the following grants at Idiap Research Institute and Charles University.

- At Idiap Research Institute, the work was 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 grant 19-26934X (NEUREM3) of the Czech Science Foundation 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

- Željko Agić and Ivan Vulić. 2019. *JW300: A wide-coverage parallel corpus for low-resource languages*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.
- 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.
- Ondrej Bojar, Vojtech Diatka, Pavel Rychlý, Pavel Stranák, Vít Suchomel, Ales Tamchyna, and Daniel Zeman. 2014. HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation. In *Language Resources and Evaluation Conference*.
- Bianka Buschbeck and Miriam Exel. 2020. [A parallel evaluation data set of software documentation with document structure annotation](#).
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement 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.
- 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.
- Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india. *arxiv 2001.09907*.
- 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, 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/>.
- Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.
- Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. [The IIT Bombay English-Hindi parallel corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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 *In Proceedings of EACL*, pages 241–248.
- Hideki Nakayama, Akihiro Tamura, and Takashi Nishimura. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of the 12th Language Resources and Evaluation Conference*.
- Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. [Overview of the 6th workshop on Asian translation](#). In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. Association for Computational Linguistics.
- 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, 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.
- Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of the 28th International Conference on 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. 2018. Odiencorp: Odia-english and odia-only corpus for machine translation. In *Proceedings of the Third International Conference on Smart Computing and Informatics*.
- 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 CILing 2019, La Rochelle, France.
- Shantipriya Parida, Satya Ranjan Dash, Ondřej Bojar, Petr Motlicek, Priyanka Pattnaik, and Debasish Kumar Mallick. 2020. Odiencorp 2.0: Odia-english parallel corpus for machine translation. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19.
- 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.
- Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. [Constructing parallel corpora for six Indian languages via crowdsourcing](#). In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. 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*.
- 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.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv 1907.05791*.
- Parth Shah and Vishvajit Bakrola. 2019. [Neural machine translation system of indic languages - an attention based approach](#). In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE.
- Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. [A multilingual parallel corpora collection effort for Indian languages](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.
- 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.
- Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. [Introducing the Asian language treebank \(ALT\)](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- 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, François 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 25 to 52 summarize translation results submitted for WAT2020 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 Tables 16, 17, and 18.

Subtask	SYSTEM ID	DATA ID	Annotator A		Annotator B		all average	weighted	
			average	variance	average	variance		κ	κ
aspec-ja-zh	Kyoto-U+ECNU	3814	4.00	0.75	4.35	0.60	4.17	0.07	0.31
	Kyoto-U+ECNU	4053	4.03	0.73	4.31	0.69	4.17	0.08	0.32
	2019 best	3170	4.44	0.47	4.29	0.85	4.36	0.15	0.15
aspec-zh-ja	Kyoto-U+ECNU	3813	3.81	0.83	4.61	0.47	4.21	0.16	0.24
	Kyoto-U+ECNU	3933	3.80	0.85	4.61	0.47	4.20	0.14	0.23
	2019 best	3210	4.80	0.24	4.46	0.61	4.63	0.27	0.31
jpcn-ja-en	goku20	3930	4.48	0.68	4.69	0.33	4.58	0.27	0.39
	2019 best	3188	4.73	0.40	4.83	0.22	4.78	0.36	0.46
jpcn-en-ja	goku20	3929	4.46	0.90	4.37	0.99	4.42	0.39	0.60
	2019 best	3192	4.43	0.81	4.57	0.77	4.50	0.36	0.49
jpcn-ja-zh	goku20	3926	4.61	0.49	4.66	0.47	4.63	0.35	0.51
	2019 best	3157	4.53	0.45	4.56	0.54	4.54	0.29	0.35
jpcn-zh-ja	goku20	3925	4.60	0.61	4.42	0.58	4.51	0.21	0.31
	2019 best	3152	4.72	0.26	4.57	0.55	4.65	0.26	0.35
jpcn-ja-ko	goku20	3928	4.78	0.37	4.68	0.60	4.73	0.48	0.65
	2019 best	2850	4.82	0.27	4.73	0.34	4.77	0.56	0.65
jpcn-ko-ja	TMU	3829	4.74	0.33	4.67	0.48	4.71	0.70	0.77
	TMU	3830	4.75	0.33	4.66	0.61	4.70	0.72	0.78
	goku20	3927	4.68	0.46	4.60	0.72	4.64	0.67	0.77
	2019 best	2924	4.72	0.39	4.58	0.68	4.65	0.59	0.69
jijic-ja-en	NHK-NES	3820	4.22	0.81	3.69	1.03	3.96	0.13	0.22
	organizer	3893	4.21	0.64	3.70	1.18	3.95	0.14	0.23
	NHK-NES	3818	4.18	0.88	3.60	1.10	3.89	0.16	0.24
	organizer	3895	4.09	0.73	3.48	1.05	3.78	0.13	0.25
indic20-en-bn	organizer	3642	1.76	0.83	1.90	1.01	1.83	0.22	0.40
	cvit-mt	3853	4.13	0.53	3.67	0.73	3.90	0.25	0.36
	HW-TSC	4032	3.91	0.85	3.61	0.86	3.76	0.20	0.44
	ODIANLP	3780	3.35	0.80	2.98	1.07	3.17	0.27	0.50
indic20-bn-en	cvit-mt	3837	3.94	0.66	4.03	0.85	3.98	0.15	0.35
	HW-TSC	4039	2.92	1.31	2.97	2.18	2.94	0.24	0.58
	NICT-5	3986	2.88	0.94	2.27	1.97	2.58	0.10	0.42
	ODIANLP	3822	1.79	1.41	1.76	1.95	1.78	0.41	0.72
indic20-en-hi	cvit-mt	3854	4.12	0.75	3.49	0.65	3.81	0.05	0.32
	HW-TSC	4033	3.29	1.09	2.48	0.96	2.89	-0.03	0.29
	ODIANLP	3786	2.94	0.92	2.46	1.02	2.70	0.14	0.42
indic20-hi-en	cvit-mt	3838	4.05	0.70	3.52	0.57	3.79	0.05	0.26
	NICT-5	4000	2.54	1.45	2.13	0.81	2.34	0.18	0.46
	HW-TSC	4040	2.41	1.18	1.82	0.78	2.12	0.11	0.41
	ODIANLP	3823	1.87	1.24	1.55	0.82	1.71	0.26	0.56
bsd-en-ja	goku20	3756	4.14	0.84	4.27	1.08	4.20	0.32	0.47
	DEEPNLP	4050	4.08	0.87	4.17	1.30	4.13	0.31	0.49
	ut-mrt	4074	3.57	1.74	3.55	2.33	3.56	0.33	0.59
	goku20	3753	3.67	1.18	3.44	2.02	3.55	0.35	0.53
	ut-mrt	4071	3.56	1.59	3.48	2.49	3.52	0.37	0.60
	DEEPNLP	4049	2.56	2.54	2.63	2.90	2.60	0.49	0.74
bsd-ja-en	goku20	3747	4.47	0.70	3.92	1.37	4.19	0.29	0.43
	DEEPNLP	4048	4.43	0.72	3.76	1.54	4.10	0.25	0.41
	adapt-dcu	3836	4.22	0.79	3.65	1.64	3.93	0.37	0.49
	ut-mrt	4073	3.90	1.37	3.34	2.03	3.62	0.31	0.53
	goku20	3748	3.92	1.30	3.23	2.08	3.57	0.25	0.47
	ut-mrt	4072	3.86	1.43	3.25	2.15	3.55	0.30	0.54
	DEEPNLP	4047	2.60	2.51	2.19	2.40	2.40	0.42	0.68
iitb-en-hi	WT	3640	3.71	1.89	3.40	1.15	3.56	0.34	0.60
	WT	3639	3.69	1.99	3.29	1.78	3.49	0.37	0.64
	2019 best	2680	3.94	1.02	3.58	0.82	3.76	0.53	0.58
iitb-hi-en	WT	3638	3.78	1.49	3.65	0.65	3.71	0.26	0.44
	2019 best	2681	4.53	0.53	3.74	1.18	4.13	0.05	0.13

Table 21: JPO adequacy evaluation results in detail.

		System	Run	BLEU	chrF3	nCDER	nCharacTER	nPER	nTER	nWER	BLEU _w
EV	TEXT	2019:IDIAP	2956	0.5142	0.5814	0.6201	0.5825	0.6926	0.5758	0.5548	40.21
		ODIANLP	3711	0.5024	0.5922	0.6141	0.6079	0.6851	0.5852	0.5633	40.85
		CNLP-NITS	3897	0.4848	0.5816	0.6051	0.6174	0.6796	0.5723	0.5530	38.84
		iiitsc	4030	0.4291	0.5188	0.5495	0.5186	0.6217	0.5126	0.4966	33.83
		ODIANLP	3713	0.4728	0.5563	0.5875	0.4815	0.6504	0.5672	0.5404	38.50
CH	TEXT	2019:IDIAP	3277	0.4037	0.4998	0.5255	0.4437	0.6010	0.4905	0.4614	30.72
		CNLP-NITS	3898	0.3691	0.4561	0.4984	0.4268	0.5667	0.4692	0.4475	27.75
		iiitsc	4031	0.2859	0.3730	0.4124	0.2629	0.4813	0.3903	0.3715	20.52
		CNLP-NITS	3896	0.5101	0.6098	0.6216	0.6485	0.7038	0.5867	0.5648	40.51
		2019:638	3271	0.4934	0.5576	0.6012	0.5063	0.6731	0.5545	0.5356	38.63
CH	MM	2019:NITSNLP	3288	0.3801	0.4442	0.4874	0.2652	0.5670	0.4398	0.4205	27.41
		CNLP-NITS	3894	0.4264	0.5128	0.5524	0.5013	0.6199	0.5226	0.4973	33.57
		2019:638	3270	0.2866	0.3784	0.4171	0.2036	0.4873	0.3851	0.3658	20.34
		ODIANLP	3779	0.0348	0.1048	0.0687	-0.3779	0.1636	-0.1298	-0.1349	0.78
		2019:NITSNLP	3297	0.0229	0.0889	0.0802	-0.5016	0.1297	0.0615	0.0566	0.00
ODIA	en-od	ODIANLP	3788	0.1160	0.1815	0.2525	-1.3391	0.2895	0.2171	0.2058	11.07
		cvit	3874	0.1038	0.1693	0.2336	-1.1605	0.2718	0.1919	0.1809	7.86
		cvit	3872	0.0323	0.2089	0.1308	-0.9498	0.1646	0.1173	0.1139	17.89
ODIA	od-en	ODIANLP	3772	0.0315	0.2072	0.1425	-1.1965	0.1746	0.1290	0.1253	18.31

Table 22: Multi-Modal and UFAL EnOdia Task automatic evaluation results. For each test set (EV, CH or Odia) and each track (TEXT, MM and HI; and Odia), 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_w denotes the WAT official BLEU scores.

		Team ID	Data ID	Ignoring Unscored		Unscored = Worst	
				Ave	Ave Z	Ave	Ave Z
EV	TEXT	ODIANLP	3711	83.38	0.34	78.25	0.53
		Reference	-	82.19	0.29	75.14	0.47
		CNLP-NITS	3897	80.01	0.23	70.46	0.37
		2019:IDIAP	2019:2956	76.94	0.15	67.64	0.30
		iiitsc	4030	74.27	0.07	58.60	0.07
		Reference	-	88.07	0.47	85.44	0.71
CH	TEXT	ODIANLP	3713	75.21	0.08	63.60	0.18
		2019:IDIAP	2019:3277	67.79	-0.10	56.29	0.03
		CNLP-NITS	3898	59.61	-0.38	40.40	-0.36
		iiitsc	4031	54.85	-0.53	36.78	-0.47
		Reference	-	86.82	0.45	83.82	0.68
		CNLP-NITS	3896	81.75	0.28	73.78	0.43
EV	MM	2019:638	2019:3271	74.82	0.07	63.47	0.18
		2019:NITSNLP	2019:3288	59.31	-0.39	42.88	-0.31
		Reference	-	90.66	0.53	88.53	0.78
		CNLP-NITS	3894	68.72	-0.11	55.80	0.01
CH	MM	2019:638	2019:3270	57.03	-0.45	41.79	-0.33
		Reference	-	90.26	0.53	80.45	0.58
		ODIANLP	3779	47.16	-0.73	10.69	-1.10
EV	HI	Reference	-	88.94	0.51	78.53	0.53
		2019:NITSNLP	2019:3297	58.56	-0.37	21.29	-0.84
		ODIANLP	3759	52.10	-0.57	10.47	-1.11
		Reference	-	88.94	0.51	78.53	0.53

Table 23: Manual evaluation result for WAT2020 English→Hindi Multi-Modal Tasks.

		Team ID	Data ID	Ignoring Unscored		Unscored = Worst	
				Ave	Ave Z	Ave	Ave Z
EN	OD	Reference	-	66.98	0.12	66.74	0.11
		cvit	3874	63.45	0.04	63.45	0.04
		ODIANLP	3788	63.35	0.01	63.24	0.01
		Reference	-	70.40	0.21	70.40	0.21
OD	EN	ODIANLP	3772	61.75	-0.05	61.55	-0.05
		cvit	3872	59.44	-0.09	59.25	-0.09

Table 24: Manual evaluation result for WAT2020 UFAL (EnOdia) Tasks.

System	ID	Type	RSRC			BLEU			RIBES			AMFM		
			kytea	stanford-ctb	stanford-pku	juman	kytea	mecab	juman	kytea	mecab	juman	kytea	mecab
baseline	1740	NMT	NO			46.87	47.30	47.00	0.880815	0.875511	0.880368	0.798110	0.798110	0.798110
Kyoto-U+ECNU	3813	NMT	YES			52.80	53.64	52.92	0.897053	0.894441	0.897199	0.823390	0.823390	0.823390
Kyoto-U+ECNU	3933	NMT	NO			52.65	53.48	52.80	0.896551	0.894073	0.896743	0.821660	0.821660	0.821660

Table 25: ASPEC zh-ja submissions

System	ID	Type	RSRC			BLEU			RIBES			AMFM		
			kytea	stanford-ctb	stanford-pku	juman	kytea	mecab	stanford-pku	kytea	mecab	stanford-ctb	stanford-pku	kytea
baseline	1738	NMT	NO			34.96	34.96	34.72	0.850199	0.850052	0.848394	0.787250	0.787250	0.787250
Kyoto-U+ECNU	3814	NMT	YES			38.56	38.56	38.43	0.858491	0.857645	0.858103	0.787730	0.787730	0.787730
Kyoto-U+ECNU	4053	NMT	NO			38.43	38.43	38.30	0.858229	0.857229	0.857722	0.786870	0.786870	0.786870

Table 26: ASPEC ja-zh submissions

System	ID	Type	RSRC			BLEU			RIBES				
			kytea	stanford-ctb	stanford-pku	juman	kytea	mecab	stanford-pku	kytea	mecab		
baseline	1995	NMT	NO			43.72	44.38	43.72	0.852623	0.849783	0.851797		
goku20	3925	NMT	NO			48.09	49.06	48.34	0.861858	0.859846	0.861534		

Table 27: JPCN zh-ja submissions

System	ID	Type	RSRC			BLEU			RIBES				
			kytea	stanford-ctb	stanford-pku	juman	kytea	mecab	stanford-pku	kytea	mecab		
baseline	1996	NMT	NO			38.07	39.18	38.65	0.845075	0.849118	0.848360		
goku20	3926	NMT	NO			41.52	42.59	42.27	0.853135	0.859118	0.857898		

Table 28: JPCN ja-zh submissions

System	ID	Type	RSRC	BLEU			RIBES		
				juman	kytea	mecab	juman	kytea	mecab
baseline	1997	NMT	NO	70.13	70.97	70.39	0.941950	0.941336	0.941905
TMU	3829	NMT	NO	73.13	74.03	73.45	0.950506	0.949816	0.950269
TMU	3830	NMT	NO	73.10	73.99	73.41	0.950791	0.950247	0.950528
goku20	3927	NMT	NO	69.37	70.41	69.68	0.947397	0.946878	0.947266

Table 29: JPCN ko-ja submissions

System	ID	Type	RSRC	BLEU			RIBES		
				juman	kytea	mecab	juman	kytea	mecab
baseline	2026	NMT	NO	71.62	71.62	71.62	0.944406	0.944406	0.944406
goku20	3928	NMT	NO	70.48	70.48	70.48	0.942321	0.942321	0.942321

Table 30: JPCN ja-ko submissions

System	ID	Type	RSRC	BLEU			RIBES		
				juman	kytea	mecab	juman	kytea	mecab
baseline	1999	NMT	NO	41.26	43.13	41.16	0.840117	0.838359	0.839361
goku20	3929	NMT	NO	44.52	46.63	44.57	0.855454	0.854173	0.855317

Table 31: JPCN en-ja submissions

System	ID	Type	RSRC	BLEU			RIBES		
				juman	kytea	mecab	juman	kytea	mecab
baseline	2000	NMT	NO	39.39	39.39	39.39	0.837932	0.837932	0.837932
goku20	3930	NMT	NO	43.51	43.51	43.51	0.857091	0.857091	0.857091

Table 32: JPCN ja-en submissions

System	ID	Type	RSRC	BLEU			RIBES		
				juman	kytea	mecab	juman	kytea	mecab
baseline	1943	NMT	NO	32.41	34.37	32.53	0.796011	0.796321	0.796411
goku20	3924	NMT	NO	40.60	42.34	40.66	0.813915	0.812992	0.813073
goku20	3932	NMT	NO	38.54	40.55	38.99	0.816579	0.817017	0.816112

Table 33: JPCEP zh-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
WT	3639	NMT	NO	22.08	0.765340	0.869400
WT	3640	NMT	YES	22.80	0.769138	0.873830

Table 34: HINDEN en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
WT	3638	NMT	NO	29.59	0.792065	0.637410

Table 35: HINDEN hi-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
baseline	3586	NMT	NO	39.78	0.776892	0.796570
ODIANLP	3711	NMT	YES	40.85	0.790908	0.808340
CNLP-NITS	3897	NMT	YES	38.84	0.793416	0.804250
iiitsc	4030	NMT	NO	33.83	0.753529	0.767900

Table 36: HINDENMMEVTEXT20 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIANLP	3779	OTHER	NO	0.78	0.064960	0.351940

Table 37: HINDENMMEVHI20 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS	3896	NMT	YES	40.51	0.803208	0.820980

Table 38: HINDENMMEVMM20en en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
baseline	3587	NMT	NO	28.35	0.691315	0.727010
ODIANLP	3713	NMT	YES	38.50	0.785252	0.804880
CNLP-NITS	3898	NMT	YES	27.75	0.714980	0.750320
iiitsc	4031	NMT	NO	20.52	0.623644	0.698600

Table 39: HINDENMMCHTEXT20 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIANLP	3759	OTHER	NO	0.00	0.040071	0.374450

Table 40: HINDENMMCHHI20 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS	3894	NMT	YES	33.57	0.754141	0.787320

Table 41: HINDENMMCHMM20en en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES
baseline	3584	NMT	NO	5.49	0.326116
ODIANLP	3592	NMT	NO	7.93	0.398153
ODIANLP	3788	NMT	YES	11.07	0.502642
cvit	3874	NMT	YES	7.86	0.425505
cvit	4022	NMT	YES	9.85	0.489535
cvit	4052	NMT	YES	9.48	0.487873
cvit	4062	NMT	NO	8.17	0.458266
cvit	4063	NMT	NO	8.17	0.458266
ADAPT	4118	NMT	YES	3.53	0.283107
ADAPT	4119	NMT	NO	4.94	0.319955

Table 42: ODI AEN en-od submissions

System	ID	Type	RSRC	BLEU	RIBES
baseline	3585	NMT	NO	8.92	0.349459
ODIANLP	3593	NMT	NO	12.54	0.464391
ODIANLP	3772	NMT	YES	18.31	0.558949
cvit	3872	NMT	YES	17.89	0.494172
cvit	3873	NMT	YES	15.06	0.487789
cvit	4064	NMT	YES	13.89	0.508536

Table 43: ODI AEN od-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
baseline	3642	NMT	NO	13.95	0.646680	0.470790
NHK-NES	3818	NMT	YES	28.02	0.736564	0.613710
NHK-NES	3820	NMT	YES	28.41	0.736924	0.624530

Table 44: JIJC ja-en submissions

System	ID	Type	RSRC	BLEU			RIBES			Pair
				juman	kytea	mecab	juman	kytea	mecab	
TMU	3651	NMT	NO	40.71	49.81	44.57	0.869433	0.874742	0.869886	-4.00
HW-TSC	3914	NMT	YES	41.26	49.79	44.76	0.870792	0.875362	0.870791	14.25
TMU	4046	NMT	NO	39.92	48.94	43.78	0.871343	0.876282	0.871542	-11.50

Table 45: MMT en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	Pair
TMU	3692	NMT	NO	48.38	0.899610	-1.50
TMU	3706	NMT	NO	48.33	0.900684	1.25
HW-TSC	3725	NMT	YES	52.28	0.911053	34.00

Table 46: MMT ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
baseline	3624	NMT	NO	15.03	0.627104	0.667009
ODIANLP	3780	NMT	NO	16.38	0.633248	0.672342
cvit	3853	NMT	YES	7.93	0.669877	0.679612
HW-TSC	4032	NMT	NO	19.64	0.680257	0.688318

Table 47: INDIC20 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
baseline	3631	NMT	NO	21.80	0.611506	0.677976
ODIANLP	3822	NMT	YES	19.71	0.588298	0.652696
cvit	3837	NMT	NO	16.50	0.729215	0.733984
NICT-5	3986	NMT	NO	16.14	0.654700	0.669477
HW-TSC	4039	NMT	NO	23.38	0.638013	0.717710

Table 48: INDIC20 bn-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
baseline	3625	NMT	NO	13.96	0.700555	0.636668
ODIANLP	3786	NMT	NO	21.05	0.703507	0.652905
cvit	3854	NMT	YES	16.23	0.709665	0.685356
HW-TSC	4033	NMT	NO	24.48	0.731919	0.691416

Table 49: INDIC20 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
baseline	3632	NMT	NO	25.68	0.668709	0.699228
ODIANLP	3823	NMT	YES	21.88	0.620309	0.670681
cvit	3838	NMT	NO	20.95	0.759791	0.756972
NICT-5	4000	NMT	NO	21.25	0.707683	0.711637
HW-TSC	4040	NMT	NO	28.51	0.687383	0.736056

Table 50: INDIC20 hi-en submissions

System	ID	Type	RSRC	BLEU			RIBES		
				juman	kytea	mecab	juman	kytea	mecab
goku20	3753	NMT	NO	14.19	20.89	15.75	0.700058	0.739468	0.715917
goku20	3756	NMT	YES	19.43	26.04	20.75	0.745493	0.771588	0.749985
DEEPNLP	4049	NMT	NO	8.66	15.10	10.10	0.640669	0.686233	0.658636
DEEPNLP	4050	NMT	YES	19.39	26.59	20.95	0.744209	0.770217	0.753233
ut-mrt	4071	NMT	YES	14.59	21.09	16.03	0.720883	0.744652	0.731101
ut-mrt	4074	NMT	YES	14.19	20.32	15.49	0.698327	0.737332	0.719519

Table 51: BSD en-ja submissions

System	ID	Type	RSRC	BLEU			RIBES
				BLEU	RSRC	RIBES	
goku20	3747	NMT	YES	23.15	YES	0.755099	
goku20	3748	NMT	NO	17.02	NO	0.688501	
adapt-dcu	3836	NMT	YES	18.70	YES	0.730460	
DEEPNLP	4047	NMT	NO	10.19	NO	0.615230	
DEEPNLP	4048	NMT	YES	22.83	YES	0.752619	
ut-mrt	4072	NMT	YES	18.05	YES	0.723202	
ut-mrt	4073	NMT	YES	18.57	YES	0.720809	

Table 52: BSD ja-en submissions

Source	Citation
CVIT-Mann ki Baat	(Siripragrada et al., 2020)
CVIT-PIB	(Siripragrada et al., 2020)
IITB en-hi v2.0	(Kunchukuttan et al., 2018)
MTurk Corpora	(Post et al., 2012)
JW300	(Agić and Vulić, 2019)
MTEnglish2Odia	
NLPC-Uom Corpus	
OdiEnCorp 1.0	(Parida et al., 2018)
OPUS	(Tiedemann, 2012)
PMIndia	(Haddow and Kirefu, 2020)
UFAL-en-ta-v2	(Ramasamy et al., 2012)
Urs Tarsadia Corpus	(Shah and Bakrola, 2019)
Wikimatrix	(Schwenk et al., 2019)
Wikittitles	

Table 53: Recommended Parallel Corpora for the Indic multilingual translation task. All download URLs can be obtained from https://github.com/AI4Bharat/indicnlp_catalog

Appendix B Parallel Corpora Sources for the Indic Multilingual Task

Table 53 lists the parallel corpora recommended for the Indic multilingual translation task.