

# Overview of the 8th 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

**Chenhui Chu**

Kyoto University  
chu@i.kyoto-u.ac.jp

**Akiko Eriguchi**

Microsoft  
akikoe@microsoft.com

**Kaori Abe**

Tohoku University  
abe-k@ecei.tohoku.ac.jp

**Yusuke Oda**

Tohoku University, LegalForce  
yusuke.oda.c1@tohoku.ac.jp

**Sadao Kurohashi**

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

## Abstract

This paper presents the results of the shared tasks from the 8th workshop on Asian translation (WAT2021). For the WAT2021, 28 teams participated in the shared tasks and 24 teams submitted their translation results for the human evaluation. We also accepted 5 research papers. About 2,100 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-WAT2020 (Nakazawa et al., 2020), WAT2021 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 8th WAT, we included the following new tasks:

- Malayalam Visual Genome Task: English → Malayalam multi-modal translation

- MultiIndicMT: Bengali / Gujarati / Hindi / Kannada / Malayalam / Marathi / Odia / Punjabi / Tamil / Telugu ↔ English translation
- Restricted Translation Task: Japanese ↔ English translation
- Ambiguous MSCOCO Task: Japanese ↔ English multi-modal translation

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.

- 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.

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 1: Statistics for JPC

## 2 Tasks

### 2.1 ASPEC+ParaNatCom Task

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 ParaNatCom<sup>1</sup> (Parallel English-Japanese abstract corpus made from Nature Communications articles) for the development and test sets of the Document-level Scientific Paper Translation subtask. We cannot provide document-level training corpus, but you can use ASPEC and any other extra resources.

### 2.2 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. To move to a highly topical setting of translation of dialogues evaluated at the level of documents, WAT uses BSD Corpus<sup>2</sup> (The Business Scene Dialogue corpus) for the dataset including training, development and test data for the first time this year. Participants of this task must get a copy of BSD corpus by themselves.

### 2.3 JPC Task

JPO Patent Corpus (JPC) for the patent tasks was constructed by the Japan Patent Office (JPO) in

<sup>1</sup><http://www2.nict.go.jp/astrec-att/member/mutiyama/paranacom/>

<sup>2</sup><https://github.com/tsuruoka-lab/BSDBSD>

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 WAT2021, the patent task 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 datasets 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 1. Note that training, development, development-test and test-N1 data are the same as those used in WAT2017.

### 2.4 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 con-

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 2: Statistics for JIJI Corpus

sists of Japanese-English sentence pairs. In addition to the test set (test set I) that has been provided from WAT 2017, a test set (test set II) with document-level context has also been provided from 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** : 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 JIJI Corpus are shown in Table 2.

The definition of data use is shown in Table 3.

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).

## 2.5 ALT and UCSY Corpus

The parallel data for Myanmar-English translation tasks at WAT2021 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 4. Notice that both of the corpora have been modified from the data used in WAT2018.

## 2.6 NICT-SAP Task

In WAT2021, we decided to continue the WAT2020 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 OPUS<sup>3</sup>, other WAT tasks (past and present) and WMT<sup>4</sup>. 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) corpora<sup>5</sup> (Buschbeck and Exel, 2020) and encouraged participants to consider GNOME, UBUNTU and KDE corpora from OPUS. We

<sup>3</sup><http://opus.nlpl.eu/>

<sup>4</sup><http://www.statmt.org/wmt20/>

<sup>5</sup>Software Domain Evaluation Splits

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

Table 3: Definition of data use in the Japanese  $\leftrightarrow$  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 4: Statistics for the data used in Myanmar-English translation tasks

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

also encouraged the use of monolingual corpora expecting that it would be for pre-trained NMT models such as BART/MBART (Lewis et al., 2020; Liu et al., 2020). In Table 5 we give statistics of the aforementioned corpora which we used for the organizer’s baselines. Note that the evaluation corpora for both domains are created from documents and thus contain document level meta-data. Participants were encouraged to use document level approaches. Note that we do not exhaustively list<sup>6</sup> all available corpora here and participants were not restricted from using any corpora as long as they are freely available.

## 2.7 News Commentary Task

For the Russian $\leftrightarrow$ Japanese task we asked participants to use the JaRuNC corpus<sup>7</sup> (Imankulova

<sup>6</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/NICT-SAP-Task>

<sup>7</sup><https://github.com/aizhanti/JaRuNC>

Lang.pair	Partition	#sent.	#tokens	#types
Ja $\leftrightarrow$ 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 $\leftrightarrow$ 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 $\leftrightarrow$ 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 6: In-Domain data for the Russian–Japanese task.

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 $\leftrightarrow$ English translation quality as well but this is beyond the scope of this years sub-task. Refer to Table 6 for the statistics of the in-domain parallel corpora. In addition, we encouraged the participants to use out-of-domain parallel corpora from various sources such as KFTT,<sup>8</sup> JESC,<sup>9</sup> TED,<sup>10</sup> ASPEC,<sup>11</sup> UN,<sup>12</sup> Yandex<sup>13</sup> and Russian $\leftrightarrow$ English news-commentary corpus.<sup>14</sup> This year we also encouraged participants to use any corpora from WMT 2020<sup>15</sup> and WMT 2021<sup>16</sup> 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 unintentionally seen during training.

<sup>8</sup><http://www.phontron.com/kftt/>

<sup>9</sup><https://datarepository.wolframcloud.com/resources/Japanese-English-Subtitle-Corpus>

<sup>10</sup><https://wit3.fbk.eu/>

<sup>11</sup><http://lotus.kuee.kyoto-u.ac.jp/ASPEC/>

<sup>12</sup><https://cms.unov.org/UNCorpus/>

<sup>13</sup><https://translate.yandex.ru/corpus?lang=en>

<sup>14</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/News-Commentary/news-commentary-v14.en-ru.filtered.tar.gz>

<sup>15</sup><http://www.statmt.org/wmt20/translation-task.html>

<sup>16</sup><http://www.statmt.org/wmt21/translation-task.html>

source	bn	gu	hi	kn	ml	mr	or	pa	ta	te	Grand Total
alt	20,106		20,106								40,212
bibleudin		15,609	62,073	61,707	61,300	60,876				62,191	323,756
cvit-pib	91,985	58,264	266,545		43,087	114,220	94,494	101,092	115,968	44,720	930,375
iitb			1,603,080								1,603,080
jw	278,307	310,094	509,594	303,991	362,816	270,346		388,364	673,232	192,904	3,289,648
mtenglish2odia							34,846				34,846
nlpc									31,373		31,373
odiencorp							90,854				90,854
opensubtitles	411,097		92,319		383,313				32,140	27,063	945,932
pmi	23,306	41,578	50,349	28,901	26,916	28,974	31,966	28,294	32,638	33,380	326,302
tanzil	187,052		187,080		187,081				93,540		654,753
ted2020	10,318	15,691	46,759	2,253	5,990	22,608		749	11,105	5,236	120,709
ufal									166,866		166,866
urst		65,000									65,000
wikimatrix	280,566		231,459		71,508	124,304			95,159	91,908	894,904
wikititles		11,665							102,131		113,796
<b>Grand Total</b>	<b>1,302,737</b>	<b>517,901</b>	<b>3,069,364</b>	<b>396,852</b>	<b>1,142,011</b>	<b>621,328</b>	<b>252,160</b>	<b>518,499</b>	<b>1,354,152</b>	<b>457,402</b>	<b>9,632,406</b>

Table 7: Statistics of the filtered parallel corpora provided by the organizers. The target language is English.

Language	#Lines
as	1.39M
bn	39.9M
en	54.3M
gu	41.1M
hi	63.1M
kn	53.3M
ml	50.2M
mr	34.0M
or	6.94M
pa	29.2M
ta	31.5M
te	47.9M

Table 8: Monolingual corpora statistics.

## 2.8 Indic Multilingual Task

Owing to the increasing interest in Indian language translation and the success of the multilingual Indian languages tasks in 2018 (Nakazawa et al., 2018) and 2020 (Nakazawa et al., 2020), we decided to enlarge the scope of the 2020 task by adding new languages, scouring new data and creating an N-way parallel evaluation set. In 2020, the evaluation data came from the CVIT-PIB dataset<sup>17</sup> but it did not contain sufficient N-way parallel sentences to evaluate on additional languages. To this end, we decided to obtain evaluation corpora from the PMI dataset<sup>18</sup> which contains sufficient N-way parallel corpora spanning 10 Indian languages and English and is similar (domain wise) to the CVIT-PIB dataset.

The evaluation data consists of various articles

<sup>17</sup>[http://preon.iiit.ac.in/~jerin/resources/datasets/pib\\_v1.3.tar](http://preon.iiit.ac.in/~jerin/resources/datasets/pib_v1.3.tar)

<sup>18</sup><http://data.statmt.org/pmindia>

composed by the Prime Minister of India. The languages involved are Hindi (Hi), Marathi (Mr), Kan-  
nada (Kn), Tamil (Ta), Telugu (Te), Gujarati (Gu),  
Malayalam (Ml), Bengali (Bn), Oriya (Or), Pun-  
jabi (Pa) and English (En). Compared to 2020, we  
have 3 additional languages leading to a total of 10  
Indian languages, 4 of which are Dravidian and the  
rest are Indo-Aryan. English is either the source or  
the target language during evaluation leading to a  
total of 20 translation directions. Due to the N-way  
nature of the evaluation corpus we can also evalu-  
ate 90 Indian language to Indian language transla-  
tion pairs but this may be the focus in future work-  
shops.

The objective of this task, like the Indic lan-  
guages tasks in 2018 and 2020, 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 parallel corpus collection spanning all  
languages<sup>19</sup> which was split into training, devel-  
opment and test sets. This dataset was created by  
first creating an evaluation set of 3,390 11-way  
sentences (1,000 for development and 2,390 for  
testing) and then filtering them out from all parallel  
corpora we could obtain at the time. Furthermore,  
we made sure to filter out sentences from the 2020  
evaluation set. This way the provided parallel  
corpus can be safely used for benchmarking the  
2020 evaluation set as well. The filtered training  
parallel corpora came from a variety of sources  
such as: CVIT-PIB, PMIndia, IITB 3.0,<sup>20</sup> JW,<sup>21</sup>

<sup>19</sup>[http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/indic\\_wat\\_2021.tar.gz](http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/indic_wat_2021.tar.gz)

<sup>20</sup>[http://www.cfilt.iitb.ac.in/iitb\\_parallel/](http://www.cfilt.iitb.ac.in/iitb_parallel/)

<sup>21</sup><http://opus.nlpl.eu/JW300.php>

NLPC,<sup>22</sup>UFAL EnTam,<sup>23</sup>Uka Tarsadia,<sup>24</sup>Wiki Titles (ta,<sup>25</sup>gu,<sup>26</sup>)ALT,<sup>27</sup>OpenSubtitles,<sup>28</sup> Bibleuedin,<sup>29</sup> MTEnglish2Odia,<sup>30</sup>OdiEnCorp 2.0,<sup>31</sup> TED,<sup>32</sup> and WikiMatrix<sup>33</sup>. Additionally we listed the CCAIaligned corpus<sup>34</sup> to be used despite its poor quality which applies to WikiMatrix as well. We also provided filtered monolingual corpora<sup>35</sup> sourced from PMI and we also encouraged the use of monolingual corpora from the IndicCorp.<sup>36</sup>The statistics of this corpus are given in table 8. We expected that this year, the novel way of using the monolingual corpora would be to pre-train NMT models such as BART/MBART (Lewis et al., 2020; Liu et al., 2020). In general we encouraged participants to focus on multilingual NMT (Dabre et al., 2020) solutions.

Detailed statistics for the aforementioned corpora can be found in Table 7. We also listed additional sources of corpora for participants to use. Our organizer’s baselines used the PMI corpora for training as it is the in-domain corpus.

## 2.9 English→Hindi Multi-Modal Task

This task is running successfully in WAT since 2019 and attracted many teams working on multimodal machine translation and image captioning in Indian languages (Nakazawa et al., 2019, 2020).

For English→Hindi multi-modal translation task, we asked the participants to use Hindi Visual Genome 1.1 corpus (HVG, Parida et al.,

<sup>22</sup><https://github.com/nlpc-uom/English-Tamil-Parallel-Corpus>

<sup>23</sup><http://ufal.mff.cuni.cz/~ramasamy/parallel/html/>

<sup>24</sup>[https://github.com/shahparth123/eng\\_guj\\_parallel\\_corpus](https://github.com/shahparth123/eng_guj_parallel_corpus)

<sup>25</sup><http://data.statmt.org/wikititles/v2/wikititles-v2.ta-en.tsv.gz>

<sup>26</sup><http://data.statmt.org/wikititles/v1/wikititles-v1.gu-en.tsv.gz>

<sup>27</sup><http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT>

<sup>28</sup><http://opus.nlpl.eu/OpenSubtitles-v2018.php>

<sup>29</sup><http://opus.nlpl.eu/bible-uedin.php>

<sup>30</sup><https://github.com/soumendrak/MTEnglish2Odia>

<sup>31</sup><https://ufal.mff.cuni.cz/odiencorp>

<sup>32</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/>

<sup>33</sup><https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix>

<sup>34</sup><http://www.statmt.org/cc-aligned/>

<sup>35</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/filteredmono.tar.gz>

<sup>36</sup><https://indicnlp.ai4bharat.org/corpora>

<sup>37</sup><https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3267>

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 9: 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.

2019a,b).<sup>37</sup>

The statistics of HVG 1.1 are given in Table 9. One “item” in HVG consists of an image with a rectangular region highlighting a part of the image, the original English caption of this region and the Hindi reference translation. Depending on the track (see 2.9.1 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.

### 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.

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

<sup>37</sup><https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3267>



	Text-Only MT	Hindi Captioning	Multi-Modal MT
Image	—		
Source Text	The woman is waiting to cross the street	—	A blue wall beside tennis court
System Output	महिला सड़क पार करने का इंतजार कर रही है	सड़क पर कार	टेनिस कोर्ट के बगल में एक नीली दीवार
Gloss	Woman waiting to cross the street	Car on the road	a blue wall next to the tennis court
Reference Solution	एक महिला सड़क पार करने के लिए इंतजार कर रही है	सड़क के किनारे खड़ी करें	टेनिस कोर्ट के बगल में एक नीली दीवार
Gloss	the woman is waiting to cross the street	Cars parked along the side of the road	A blue wall beside the tennis court

Figure 1: An illustration of the three tracks of WAT 2021 English→Hindi Multi-Modal Task.



English Text: Two elephants standing in the water.

Malayalam Text: വെള്ളത്തിൽ നിൽക്കുന്ന രണ്ട് ആനകൾ

Figure 2: Sample item from Malayalam Visual Genome (MVG), Image with specific region and its description.

## 2.10 English→Malayalam Multi-Modal Task

This task is introduced this year using the first multimodal machine translation dataset in *Malayalam* language. For English→Malayalam multi-modal translation task we asked the participants to use the Malayalam Visual Genome corpus (MVG for short Parida and Bojar, 2021)<sup>38</sup>.

The statistics of MVG are given in Table 10. One “item” in MVG consists of an image with a rectangular region highlighting a part of the image, the original English caption of this region and the Malayalam reference translation as shown in Figure 2. Depending on the track (see 2.10.1 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.

<sup>38</sup><https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533>

### 2.10.1 English→Malayalam 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 Malayalam. 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. Malayalam Captioning (labeled “ML”): The participants are asked to generate captions in Malayalam 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 Malayalam. Both textual and visual information can be used.

### 2.11 Flickr30kEnt-JP Japanese↔English Multi-Modal Tasks

The goal of Flickr30kEnt-JP Japanese↔English multi-modal task<sup>39</sup> 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

<sup>39</sup><https://nlab-mpg.github.io/wat2021-mmt-jp/>

Dataset	Items	Tokens	
		English	Malayalam
Training Set	28,930	143,112	107,126
D-Test	998	4,922	3,619
E-Test (EV)	1,595	7,853	6,689
C-Test (CH)	1,400	8,186	6,044

Table 10: Statistics of Malayalam Visual Genome used for the English→Malayalam 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 Malayalam 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	148,915/1.99M	148,910*/2.50M
Dev	1,000	5,000/67,288	5,000/84,017
Test	1,000	1,000/10,876	1,000/16,113

Table 11: Statistics of the dataset used for Japanese↔English multi-modal tasks. Here we use the MeCab tokenizer to count Japanese tokens. \*Some of the original English sentences are actually broken so we did not provide their translations.

extended dataset of the Flickr30k<sup>40</sup> and Flickr30k Entities<sup>41</sup> 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.

This year, from the same shared tasks in WAT 2020, we increased the number of parallel sentences for training and validation. We summarize the statistics of the dataset for this year in Table 11. 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 2.0 which is publicly available.<sup>42</sup> The original Flickr30k has five English sentences for each image. While the Japanese set for WAT 2020 had the translations of only the first two sentences, this year we have all five translations for each image. Therefore, we can use five parallel sentences for each image to train and validate the systems. The test data remain exactly the same as in WAT 2020, where phrase-to-region annotation is not included.

There are two settings of submission: with and

<sup>40</sup><http://shannon.cs.illinois.edu/DenotationGraph/>

<sup>41</sup><http://bryanplummer.com/Flickr30kEntities/>

<sup>42</sup><https://github.com/nlab-mpg/Flickr30kEnt-JP>

without resource constraints. In the constrained setting, external resources such as additional data and pre-trained models (with external data) are not allowed, except for pre-trained convolutional neural networks (for visual analysis) and basic linguistic tools such as taggers, parsers, and morphological analyzers.

## 2.12 Ambiguous MS COCO Japanese↔English Multimodal Task

This is another Japanese–English multimodal machine translation task. We provide the Japanese–English Ambiguous MS COCO dataset (Merritt et al., 2020) for validation and testing, which contains ambiguous verbs that may require visual information in images for disambiguation. The validation and testing sets contain 230 and 231 Japanese–English sentence pairs, respectively. The Japanese sentences are translated from the English sentences in the original Ambiguous MS COCO dataset.<sup>43</sup>

Participants can use the constrained and unconstrained training data to train their multimodal machine translation system. In the constrained setting, only the Flickr30kEntities Japanese (F30kEnt-Jp) dataset<sup>44</sup> can be used as training data. In the unconstrained setting, the MS COCO English data<sup>45</sup> and STAIR Japanese image captions<sup>46</sup> can be used as additional training data.

We prepare a baseline using the double attention on image region method following (Zhao et al., 2020) for both Japanese→English and English→Japanese directions.

## 2.13 Restricted Translation Task

Despite recent success of NMT, the MT systems still struggle to generate translation with a consistent terminology. Consistency is the key to clear and accurate translation, especially when translating documents in a specific field, for instance, science or business and marketing contexts, requiring technical terms and proper nouns to get translated into the corresponding unique expressions continuously in the entire documents. To tackle this inconsistent translation issue, we have designed *Restricted Translation task* at WAT 2021.

In the restricted translation task, participants are required to submit a system that translates source

<sup>43</sup><http://www.statmt.org/wmt17/multimodal-task.html>

<sup>44</sup><https://github.com/nlab-mpg/Flickr30kEnt-JP>

<sup>45</sup><https://cocodataset.org/#captions-2015>

<sup>46</sup><https://stair-lab-cit.github.io/STAIR-captions-web/>



	En-Ja (# phrase, # char)	Ja-En (# phrase, # word)
Dev.	(2.8, 164)	(2.8, 6.6)
Devtest	(3.2, 18.2)	(3.2, 7.3)
Test	(3.3, 18.1)	(3.2, 7.4)

Table 12: Statistics of the restricted vocabulary in the evaluation data. We report average number of phrases and characters/words per source sentence.

texts under target vocabulary constraints. At inference time, such a restricted vocabulary is provided as a list of target words, consisting of scientific technical terms in the target language, and the system outputs must contain all these target words. For the English↔Japanese translation tasks, we employ the ASPEC corpus and allow to use other external data source. We built the restricted vocabulary lists by asking 10 bilingual speakers to manually extract the scientific technical terms from the evaluation data sets (“*dev/devtest/test*”). Table 12 reports the data statistics of the restricted vocabulary in the evaluation data.

We evaluate systems with two distinct metrics: 1) BLEU score as a conventional translation accuracy and 2) a consistency score: the ratio of the number of sentences satisfying exact match of given constraints over the whole test corpus. For the “exact match” evaluation, we conduct the following process. In English, we simply lowercase hypotheses and constraints, then judge character-level sequence matching (including whitespaces) for each constraint. In Japanese, we judge character-level sequence matching (including whitespaces) for each constraint without preprocessing. For the final ranking, we also calculate the combined score of both: calculating BLEU with only the exact match sentences. We note that, in this scenario, the brevity score in BLEU does not carry its usual meaning, but the n-gram scores maintain their consistency.

### 3 Participants

Table 13 shows the participants in WAT2021. The table lists 24 organizations from various countries, including Japan, India, USA, Singapore, Myanmar, Thailand, Korea, Poland, Denmark and Switzerland.

2,100 translation results by 28 teams were submitted for automatic evaluation and about 360 translation results by 24 teams were submitted for the human evaluation. Table 14 summarizes the

participation of teams across WAT2021 tasks and indicates which tasks included manual evaluation. The human evaluation was conducted only for the tasks with the check marks in “human eval” line.

There were no participants in the Newswire (JIJI) task, BSD task and JaRuNC task.

## 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 served as the standard for human evaluation. At WAT 2021, we adopted some of neural machine translation (NMT) as baseline systems. The details of the NMT baseline systems are described in this section.

The NMT baseline systems consisted of publicly available software, and the procedures for building the systems and for translating using the systems were published on the WAT web page.<sup>47</sup> We also have SMT baseline systems for the tasks that started at WAT 2017 or before 2017. The baseline systems are shown in Tables 15, 16, and 17. 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 their systems. 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, JIJI, and ALT+UCSY

- Juman version 7.0<sup>48</sup> for Japanese segmentation.
- Stanford Word Segmenter version 2014-01-04<sup>49</sup> (Chinese Penn Treebank (CTB) model) for Chinese segmentation.

<sup>47</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2021/baseline/baselineSystems.html>

<sup>48</sup><http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

<sup>49</sup><http://nlp.stanford.edu/software/segmenter.shtml>

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NTT	NTT Corporation	Japan
NICT-2	NICT	Japan
NICT-5	NICT	Japan
NLPHut	Idiap Research Institute Switzerland, IIT BHU, BITS Pilani India, KIIT University India, Silicon Techlab pvt. Ltd India, University of Chicago	Switzerland, India, USA
TMEKU	Tokyo Metropolitan University, Ehime University, Kyoto University	Japan
*goodjob	Dalian University of Technology	China
YCC-MT1	University of Technology (Yatanarpon Cyber City)	Myanmar
YCC-MT2	University of Technology (Yatanarpon Cyber City)	Myanmar
NECTEC	National Electronics and Computer Technology Center (NECTEC)	Thailand
mcairt	CAIR	India
nictrb	NICT	Japan
sakura	Rakuten Institute of Technology Singapore, Rakuten Asia.	Singapore
IIT-H	International Institute of Information Technology	India
*gauvar	Amazon	Singapore
*JBKJB	Individual participant	Korea
SRPOL	Samsung R&D Poland	Poland
NHK	NHK	Japan
CFILT	Computing for Indian Language Technology	India
iitp	IIT Patna	India
Volta	International Institute of Information Technology Hyderabad	India
coastal	University of Copenhagen	Denmark
CFILT-IITB	Indian Institute of Technology Bombay	India
CNLP-NITS-PP	NIT Silchar	India
Bering Lab	Bering Lab	South Korea
tpt_wat	Transperfect Translations	USA

Table 13: List of participants who submitted translations for the human evaluation in WAT2021 (Note: teams with ‘\*’ marks did not submit their system description papers, therefore the evaluation results are UNOFFICIAL according to our policy)

Team ID	ASPEC +	ASPEC		ALT +		NICT-SAP			
	ParaNatCom	Restricted		UCSY		En-Hi/Id/Ms/Th		Hi/Id/Ms/Th-En	
	EJ	EJ	JE	En-My	My-En	IT	Wikinews	IT	Wikinews
TMU		✓	✓						
NTT									
NICT-2						✓	✓	✓	✓
goodjob	✓								
YCC-MT1				✓					
YCC-MT2				✓					
NECTEC					✓				
nictrb		✓	✓						
sakura				✓	✓	✓	✓	✓	✓
NHK		✓	✓						
human eval	✓	✓	✓			✓		✓	

Team ID	JPC						En-Hi			Multimodal				MS COCO	
	EJ	JE	CJ	JC	KJ	JK	TX	HI	MM	TX	HI	EJ	JE		EJ
TMU	✓	✓			✓	✓									
NLP Hut							✓	✓		✓	✓				
TMEKU												✓	✓		✓
sakura												✓	✓		
iitp									✓						
Volta							✓		✓						
CNLP-NITS-PP							✓		✓						
Bering Lab	✓	✓	✓	✓	✓	✓									
tpt_wat	✓	✓	✓	✓	✓	✓									
human eval	✓	✓					✓	✓	✓	✓	✓	✓	✓		✓

Team ID	Indic21										X-En										
	En-X					X-En					En-X				X-En						
	Bn	Kn	Ml	Mr	Or	Hi	Gu	Pa	Ta	Te	Bn	Kn	Ml	Mr	Or	Hi	Gu	Pa	Ta	Te	
NICT-5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NLP Hut	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
mcairt	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
sakura	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IIT-H	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
gauvar	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
JBKJB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SRPOL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CFILT	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
coastal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CFILT-IITB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
human eval	✓	✓	✓	✓	✓						✓	✓	✓	✓	✓						

Table 14: Submissions for each task by each team. E and J denote English and Japanese 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 $\leftrightarrow$ ko
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 15: Baseline Systems I

System ID	System	Type	JJI		ALT+UCSY	
			ja-en	en-ja	my $\leftrightarrow$ en	km $\leftrightarrow$ en
NMT	System	NMT	✓	✓	✓	✓
SMT Phrase	OpenNMT's NMT with attention	SMT	✓	✓	✓	✓
SMT Hiero	Moses' Phrase-based SMT	SMT	✓	✓	✓	✓
SMT S2T	Moses' Hierarchical Phrase-based SMT	SMT	✓	✓	✓	✓
SMT T2S	Moses' String-to-Tree Syntax-based SMT and Berkeley parser	SMT	✓	✓	✓	✓
RBMT X	Moses' Tree-to-String Syntax-based SMT and Berkeley parser	RBMT	✓	✓	✓	✓
RBMT X	The Honyaku V15 (Commercial system)	RBMT	✓	✓	✓	✓
Online X	PC-Transer V13 (Commercial system)	Other	✓	✓	✓	✓
Online X	Google translate	Other	✓	✓	✓	✓
Online X	Bing translator	Other	✓	✓	✓	✓

Table 16: Baseline Systems II

System ID	System	Type	NewsCommentary	NICT+SAP IT&Wikinews {hi,i,d,ms,th} ↔en	Indic {bn,gu,hi,kn,ml,mr,or,pa,ta,te} ↔en	en-hi	en-ml	Multimodal Flickr ja ↔en	MS COCO ja ↔en
NMT	OpenNMT's NMT with attention	NMT	ru ↔ ja						
NMT T2T	Tensor2Tensor's Transformer	NMT	✓	✓	✓	✓	✓	✓	
NMT OT	OpenNMT-py's Transformer	NMT							
MNMT	Multimodal NMT	NMT							
MNMT2	Double-attention based Multimodal NMT	NMT							✓

Table 17: Baseline Systems III

- The Moses toolkit for English and Indonesian tokenization.
- Mecab-ko<sup>50</sup> for Korean segmentation.
- Indic NLP Library<sup>51</sup> (Kunchukuttan, 2020) for Indic language segmentation.
- The tools included in the ALT corpus for Myanmar and Khmer segmentation.
- subword-nmt<sup>52</sup> for all languages.

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

#### 4.1.2 For News Commentary

- The Moses toolkit for English and Russian only for the News Commentary data.
- Mecab<sup>53</sup> for Japanese segmentation.
- Corpora are further processed by tensor2tensor’s internal pre/post-processing which includes sub-word segmentation.

#### 4.1.3 For 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 English→Malayalam Tasks

- Hindi Visual Genome 1.1 and Malayalam Visual Genome comes untokenized and we did not use or recommend any specific external tokenizer.
- The standard OpenNMT-py sub-word segmentation was used for pre/post-processing for the baseline system and each participant used what they wanted.

<sup>50</sup><https://bitbucket.org/eunjeon/mecab-ko/>

<sup>51</sup>[https://github.com/anoopkunchukuttan/indic\\_nlp\\_library](https://github.com/anoopkunchukuttan/indic_nlp_library)

<sup>52</sup><https://github.com/rsenrich/subword-nmt>

<sup>53</sup><https://taku910.github.io/mecab/>

#### 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 NMT models for all tasks. Unless mentioned otherwise we use the Transformer model (Vaswani et al., 2017). We used OpenNMT (Klein et al., 2017) (RNN-model) for ASPEC, JPC, JIJI, and ALT tasks, tensor2tensor<sup>54</sup> for the News Commentary (JaRuNC), NICT-SAP and MultiIndicMT tasks and OpenNMT-py<sup>55</sup> for other tasks.

#### 4.2.1 NMT with Attention (OpenNMT)

For ASPEC, JPC, JIJI, and ALT tasks, we used OpenNMT (Klein et al., 2017) as the implementation of the baseline NMT systems of NMT with attention (System ID: NMT). We used the following OpenNMT configuration.

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

The default values were used for the other system parameters.

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.

<sup>54</sup><https://github.com/tensorflow/tensor2tensor>

<sup>55</sup><https://github.com/OpenNMT/OpenNMT-py>

### 4.2.2 Transformer (Tensor2Tensor)

For the News Commentary task, we used tensor2tensor’s<sup>56</sup> 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.

For the NICT-SAP task, 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 (separated by 1000 batches) and performed decoding with a beam of size 4 and a length penalty of 0.6.

For the MultiIndicMT task we trained unidirectional models using only the PMI corpus instead of the entire training data. We intentionally used the PMI corpus because its domain is the same as that of the evaluation set. Due to lack of time and resources we did not train multilingual models nor did we use additional data. We trained “transformer\_base” models with shared vocabularies of 8,000 subwords. We trained our models on 1 GPU till convergence on the development set BLEU scores, chose the model with the best development set BLEU 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 English→Malayalam Multimodal 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

<sup>56</sup><https://github.com/tensorflow/tensor2tensor>

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.

## 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., 2015a). BLEU scores were calculated using `multi-bleu.perl` in the Moses toolkit (Koehn et al., 2007). RIBES scores were calculated using `RIBES.py` version 1.02.4.<sup>57</sup> AMFM scores were calculated using scripts created by the technical collaborators listed in the WAT2021 web page.<sup>58</sup> All scores for each task were calculated using the corresponding reference translations.

Before the calculation of the automatic evaluation scores, the translation results were tokenized or segmented with tokenization/segmentation tools for each language. For Japanese segmentation, we used three different tools: Juman version 7.0 (Kurohashi et al., 1994), KyTea 0.4.6 (Neubig et al., 2011) with full SVM model<sup>59</sup> and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0.<sup>60</sup> For Chinese segmentation, we used two different tools: KyTea 0.4.6 with full SVM Model in MSR model and Stanford Word Segmenter (Tseng, 2005) version 2014-06-16 with Chinese Penn Treebank (CTB) and Peking University (PKU) model.<sup>61</sup> For Korean segmentation, we used `mecab-ko`.<sup>62</sup> For Myanmar and Khmer segmentations, we used `myseg.py`<sup>63</sup> and `kmseg.py`.<sup>64</sup> For English and Russian tokenizations, we used `tokenizer.perl`<sup>65</sup> in the Moses toolkit. For

<sup>57</sup><http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html>

<sup>58</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2021/>

<sup>59</sup><http://www.phontron.com/kytea/model.html>

<sup>60</sup><http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz>

<sup>61</sup><http://nlp.stanford.edu/software/segmenter.shtml>

<sup>62</sup><https://bitbucket.org/eunjeon/mecab-ko/>

<sup>63</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/wat2020.my-en.zip>

<sup>64</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/km-en-data/km-en.zip>

<sup>65</sup><https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl>

# WAT

## The Workshop on Asian Translation Submission

### SUBMISSION

Logged in as: ORGANIZER

[Logout](#)

**Submission:**

Human Evaluation:  human evaluation

Publish the results of the evaluation:  publish

Team Name:

Task:

Submission File:  選択されていません

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

Method:

System Description (public):  100 characters or less

System Description (private):  100 characters or less

Guidelines for submission:

- System requirements:
  - The latest versions of Chrome, Firefox, Internet Explorer and Safari are supported for this site.
  - Before you submit files, you need to enable JavaScript in your browser.
- File format:
  - Submitted files should **NOT** be tokenized/segmented. Please check [the automatic evaluation procedures](#).
  - Submitted files should be encoded in UTF-8 format.
  - Translated sentences in submitted files should have one sentence per line, corresponding to each test sentence. The number of lines in the submitted file and that of the corresponding test file should be the same.
- Tasks:
  - en-ja, ja-en, zh-ja, ja-zh indicate the scientific paper tasks with ASPEC.
  - HINDENen-hi, HINDENhi-en, HINDENja-hi, and HINDENhi-ja indicate the mixed domain tasks with IITB Corpus.
  - JJIen-ja and JJIja-en are the newswire tasks with JIJI Corpus.
  - RECIPE{ALL,TTL,STE,ING}en-ja and RECIPE{ALL,TTL,STE,ING}ja-en indicate the recipe tasks with Recipe Corpus.
  - ALTen-my and ALTmy-en indicate the mixed domain tasks with UCSY and ALT Corpus.
  - INDICen-{bn,hi,ml,ta,te,ur,si} and INDIC{bn,hi,ml,ta,te,ur,si}-en indicate the Indic languages multilingual tasks with Indic Languages Multilingual Parallel Corpus.
  - JPC{N,N1,N2,N3,EP}zh-ja ,JPC{N,N1,N2,N3}ja-zh, JPC{N,N1,N2,N3}ko-ja, JPC{N,N1,N2,N3}ja-ko, JPC{N,N1,N2,N3}en-ja, and JPC{N,N1,N2,N3}ja-en indicate the patent tasks with JPO Patent Corpus. JPCN1{zh-ja,ja-zh,ko-ja,ja-ko,en-ja,ja-en} are the same tasks as JPC{zh-ja,ja-zh,ko-ja,ja-ko,en-ja,ja-en} in WAT2015-WAT2017. AMFM is not calculated for JPC{N,N2,N3} tasks.
- Human evaluation:
  - If you want to submit the file for human evaluation, check the box "Human Evaluation". Once you upload a file with checking "Human Evaluation" you cannot change the file used for human evaluation.
  - When you submit the translation results for human evaluation, please check the checkbox of "Publish" too.
  - You can submit **two files** for human evaluation per task.
  - One of the files for human evaluation is recommended not to use other resources, but it is not compulsory.
- Other:
  - Team Name, Task, Used Other Resources, Method, System Description (public) , Date and Time(JST), BLEU, RIBES and AMFM will be disclosed on the Evaluation Site when you upload a file checking "Publish the results of the evaluation".
  - You can modify some fields of submitted data. Read "Guidelines for submitted data" at the bottom of this page.

[Back to top](#)

Figure 3: The interface for translation results submission

Indonesian and Malay tokenizations, we used `tokenizer.perl` actually sticking to the English tokenization settings. For Thai tokenization, we segmented the text at each individual character. For Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, and Telugu tokenizations, we used Indic NLP Library<sup>66</sup>

<sup>66</sup>[https://github.com/anoopkunchukuttan/indic\\_nlp\\_library](https://github.com/anoopkunchukuttan/indic_nlp_library)

(Kunchukuttan, 2020). The detailed procedures for the automatic evaluation are shown on the WAT evaluation web page.<sup>67</sup>

## 5.2 Automatic Evaluation System

The automatic evaluation system receives translation results by participants and automatically gives

<sup>67</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html>



evaluation scores to the uploaded results. As shown in Figure 3, the system requires participants to provide the following information for each submission:

- Human Evaluation: whether or not they submit the results for human evaluation;
- Publish the results of the evaluation: whether or not they permit to publish automatic evaluation scores on the WAT2021 web page;
- Task: the task you submit the results for;
- Used Other Resources: whether or not they used additional resources; and
- Method: the type of the method including SMT, RBMT, SMT and RBMT, EBMT, NMT and Other.

Evaluation scores of translation results that participants permit to be published are disclosed via the WAT2021 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 WAT2021. Anybody can register an account for the system by the procedures described in the application site.<sup>68</sup>

### 5.3 A Note on AMFM Scores

Up until WAT 2020, we used an older generation AMFM evaluation approach which did not use deep neural networks. Given the advances in multilingual pre-trained models, this year, our collaborators provided us with deep AMFM models. With the exception of ASPEC and restricted translation tasks we used the provided deep AMFM models to compute AMFM scores. Given that these deep models need GPUs to run quickly, we have not yet integrated it into our evaluation server as it is not equipped with GPUs. Instead, we compute the AMFM scores offline and add them to the evaluation scoreboard. For readers interested in AMFM and recent advances we refer readers to the following literature: Zhang et al. (2021b,a); D’Haro et al. (2019); Banchs et al. (2015b).

## 6 Human Evaluation

In WAT2021, we conducted *JPO adequacy evaluation* (other than En-Hi and En-MI multi-modal task, Section 6.1).

<sup>68</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2021/application/index.html>

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 18: The JPO adequacy criterion

## 6.1 JPO Adequacy Evaluation

We conducted JPO adequacy evaluation for the top two or three participants’ systems of pairwise evaluation for each subtask.<sup>69</sup> The evaluation was carried out by translation experts based on the JPO adequacy evaluation criterion, which is originally defined by JPO to assess the quality of translated patent documents.

### 6.1.1 Sentence Selection and Evaluation

For the JPO adequacy evaluation, the 200 test sentences were randomly selected from the test sentences.

For each test sentence, input source sentence, translation by participants’ system, and reference translation were shown to the annotators. To guarantee the quality of the evaluation, each sentence was evaluated by two annotators. Note that the selected sentences are basically the same as those used in the previous workshop.

### 6.1.2 Evaluation Criterion

Table 18 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).<sup>70</sup>

## 7 Evaluation Results

In this section, the evaluation results for WAT2021 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2021 web-

<sup>69</sup>The number of systems varies depending on the subtasks.

<sup>70</sup>[http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku\\_hyouka.htm](http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm)

site.<sup>71</sup>

## 7.1 Official Evaluation Results

Figures 4 and 5 show those of JPC subtasks, Figures 6 and 7 show those of MMT subtasks, Figures 8, 9, 10, 11, 12, 13, 14, 15, 16, and 17 show those of Indic Multilingual subtasks and Figures 18 and 19 show those of . 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 19. The weights for the weighted  $\kappa$  (Cohen, 1968) is defined as  $|Evaluation1 - Evaluation2|/4$ .

## 8 Findings

### 8.1 JPC Task

Three teams participated in JPC task. Bering Lab and tpt\_wat submitted results for all language pairs and TMU submitted results for J $\leftrightarrow$ K and J $\leftrightarrow$ E pairs. Similarly to WAT 2020, participants’ systems were transformer-based or BART-based. Bering Lab trained Transformer models with additional corpora, which were crawled patent document pairs aligned by a sentence encoding method and contained more than 13M sentences for each language pair. Their system achieved the best BLEU, RIBES, and AMFM scores for J $\rightarrow$ C/K/E and the best BLEU and RIBES scores for K $\rightarrow$ J among the past and this year’s systems. tpt\_wat used Transformer and back-translation with a single setting for six language pairs. TMU used fine-tuned Japanese BART models and achieved the best AMFM score for K $\rightarrow$ J. As for human adequacy evaluation, the evaluated system TMU did not show superior performance to past years’ systems for J $\leftrightarrow$ E, while the results cannot be directly compared.

Among the top-performing systems, Bering Lab’s systems obtained large BLEU improvements around two points over the past years’ systems for J $\leftrightarrow$ K. The improvements were probably due to their additional corpora, because their model without the additional corpus ranked second for J $\rightarrow$ K. Another finding by TMU was that pretrained Japanese BART brought gains for all J $\leftrightarrow$ K/E directions.

<sup>71</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/>

## 8.2 NICT-SAP Task

In contrast to 2020 where we had only 1 submission, this year we received submissions from 5 teams, 4 of which submitted system description papers. The submitted models were trained using a variety of techniques such as domain adaptation, corpora selection and weighing, MBART pre-training and multilingual NMT training. All submissions significantly outperformed the organizers baselines as well as the best submission in 2020. The gains showed by this year’s submissions range from approximately 14 to 30 BLEU (depending on the task) compared to the baselines. The main reason was that this year’s submission rely on high quality data selection as well as on massively multilingual pre-trained models. Out of the 4 teams that submitted system description papers, only one relied on data selection and surprisingly obtained the best results for some language pairs. For other language pairs, this team obtained cometic results. Regardless, it is clear that models like MBART are extremely useful in extremely low-resource domains such as Wikinews and software documentation.

Regarding, human evaluation we did JPO adequacy evaluation for English to Indonesian and English to Malay for the Software Documentation domain. Kindly refer to Figure 18 and 19 for the results of human evaluation. For both translation directions, team “sakura” had the highest JPO as well as BLEU scores but the scores for team “NICT-2” were not that far behind. They were certainly significantly better than the organizer scores who only developed models using parallel corpora without any pre-training. We can certainly say that at high enough BLEU score levels (higher than 40), the large differences in BLEU do not necessarily correlate with large differences in human evaluation scores. To be specific, the gap between “sakura” and “NICT-2” in terms of BLEU for English to Indonesian is 2.14 and for English to Malay is 1.5 BLEU. However, the corresponding gaps in human evaluation are 0.08 and 0.15 which is not significant. Human evaluation on a larger scale might be needed but we were unable to do so due to budgetary limitations.

### 8.2.1 News Commentary (JaRuNC) task

Unfortunately we did not receive any submissions this year.

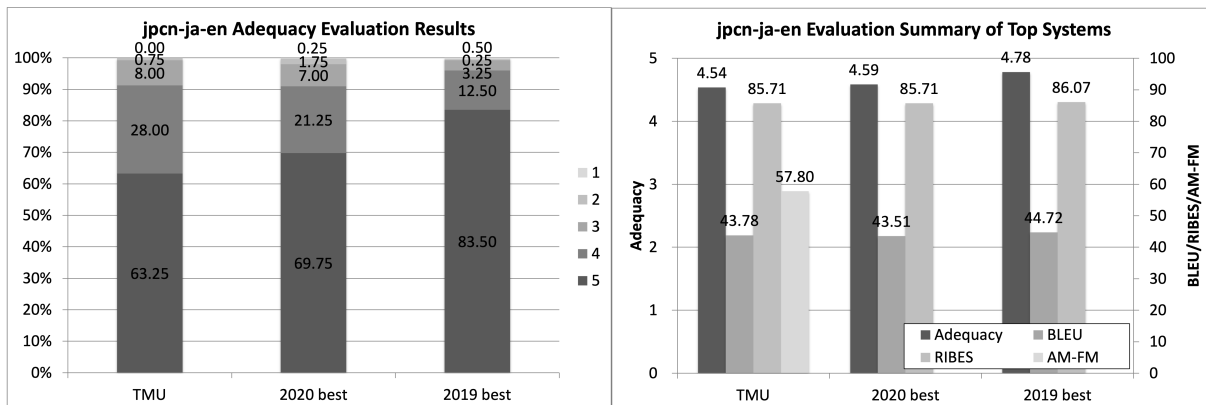


Figure 4: Official evaluation results of jpcn-ja-en.

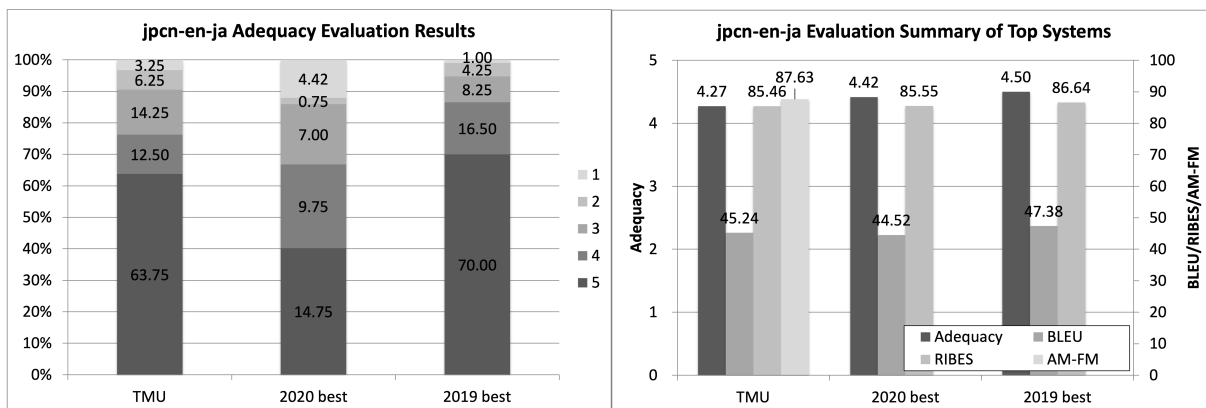


Figure 5: Official evaluation results of jpcn-en-ja.

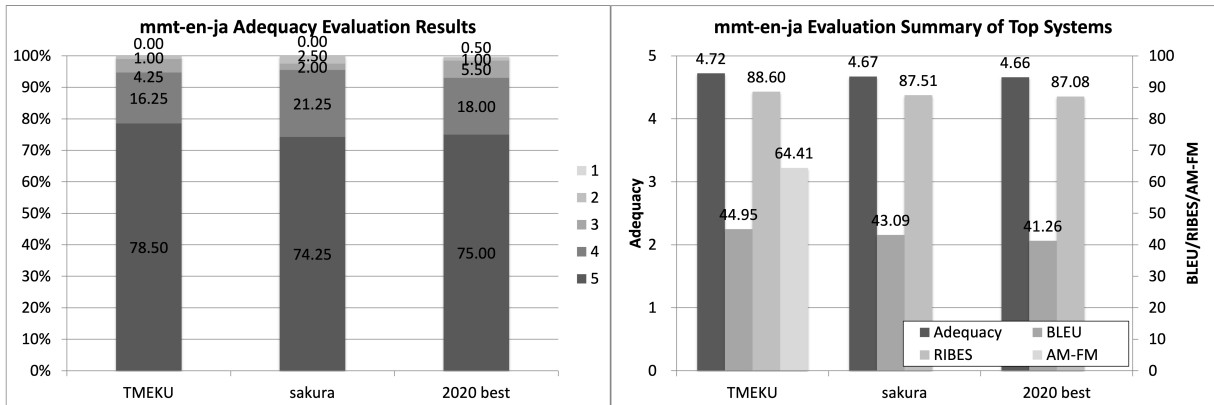


Figure 6: Official evaluation results of mmt-en-ja.

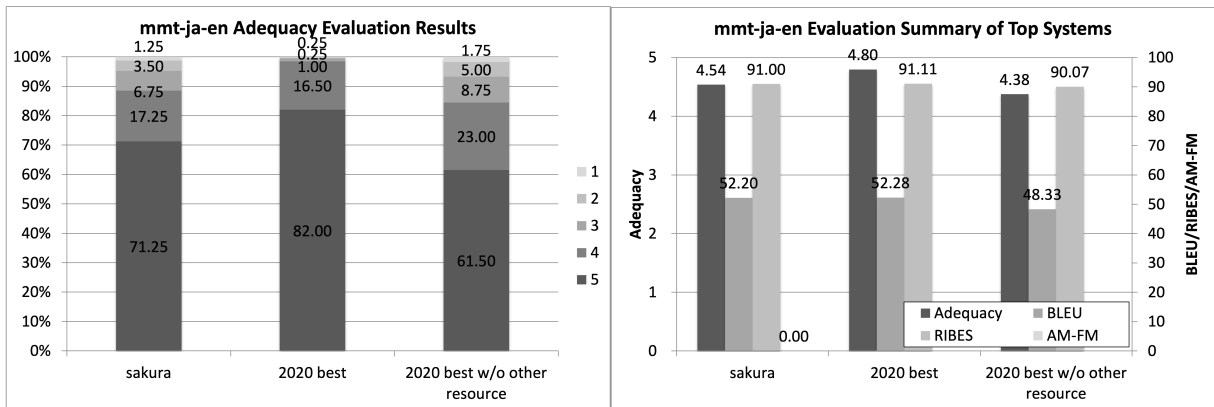


Figure 7: Official evaluation results of mmt-ja-en.

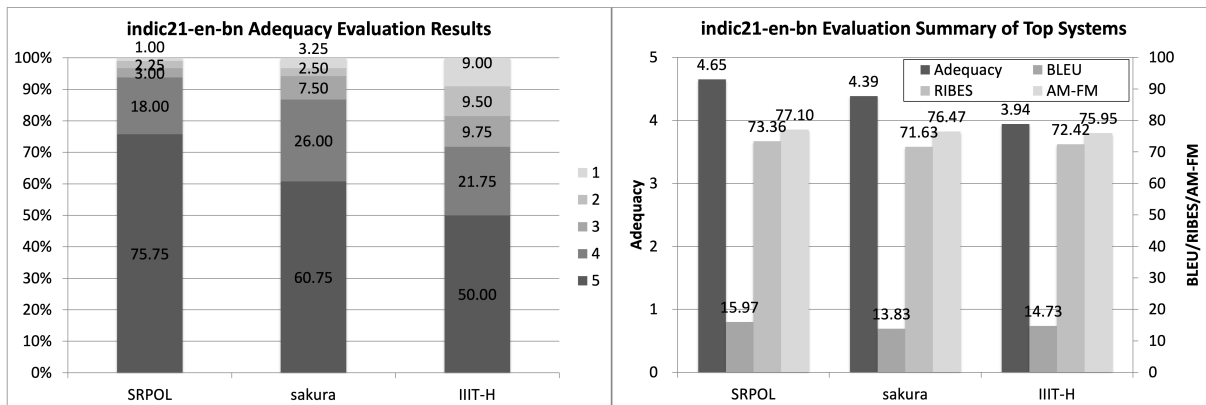


Figure 8: Official evaluation results of indic21-en-bn.

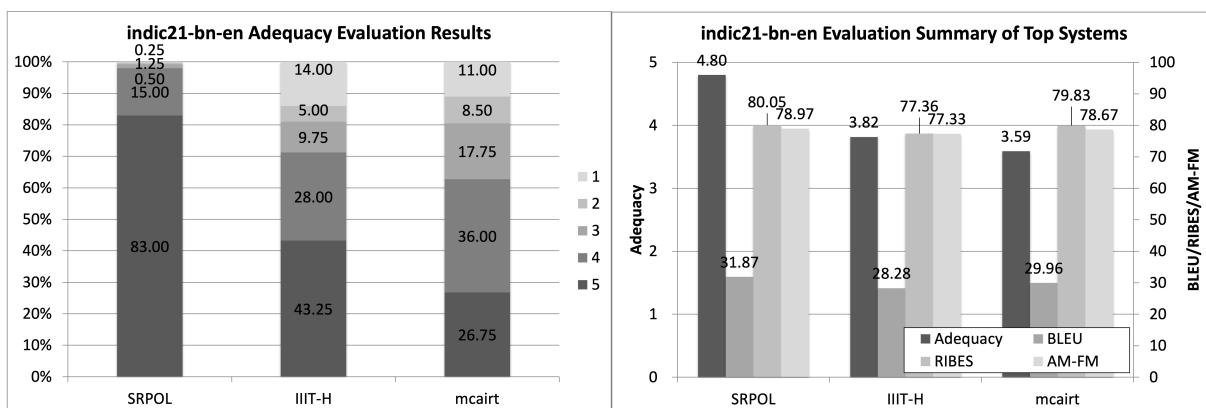


Figure 9: Official evaluation results of indic21-bn-en.

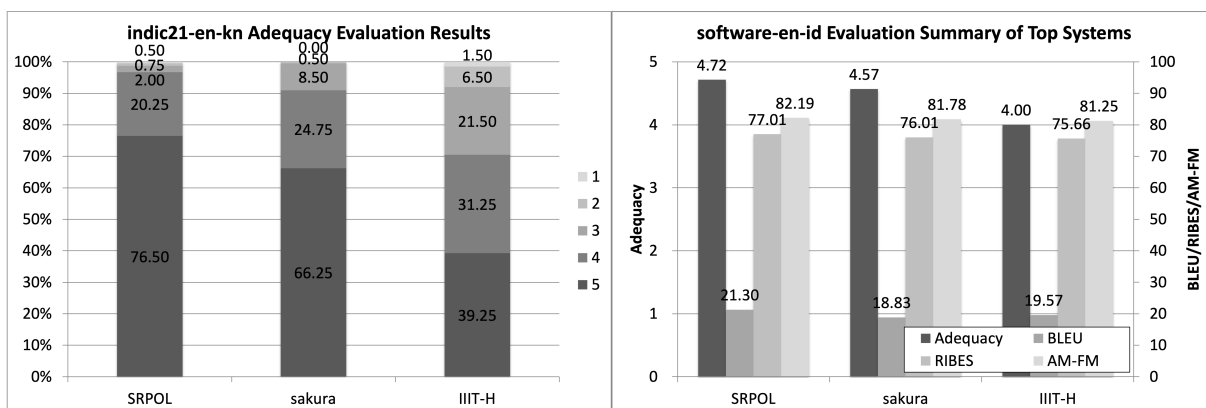


Figure 10: Official evaluation results of indic21-en-kn.

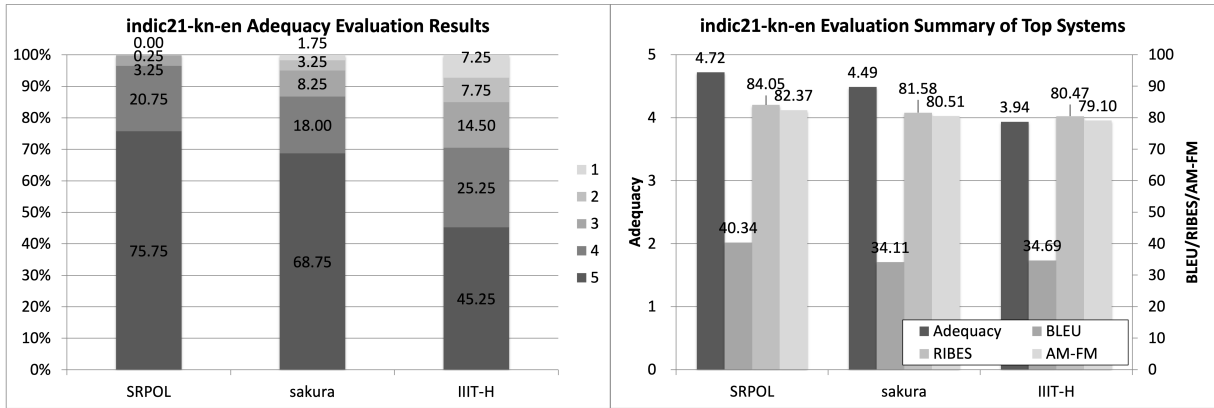


Figure 11: Official evaluation results of indic21-kn-en.

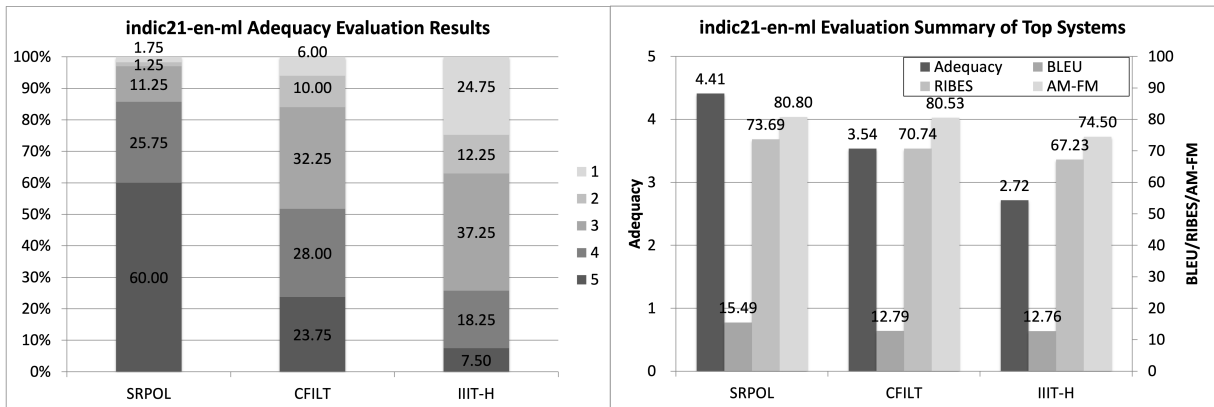


Figure 12: Official evaluation results of indic21-en-ml.

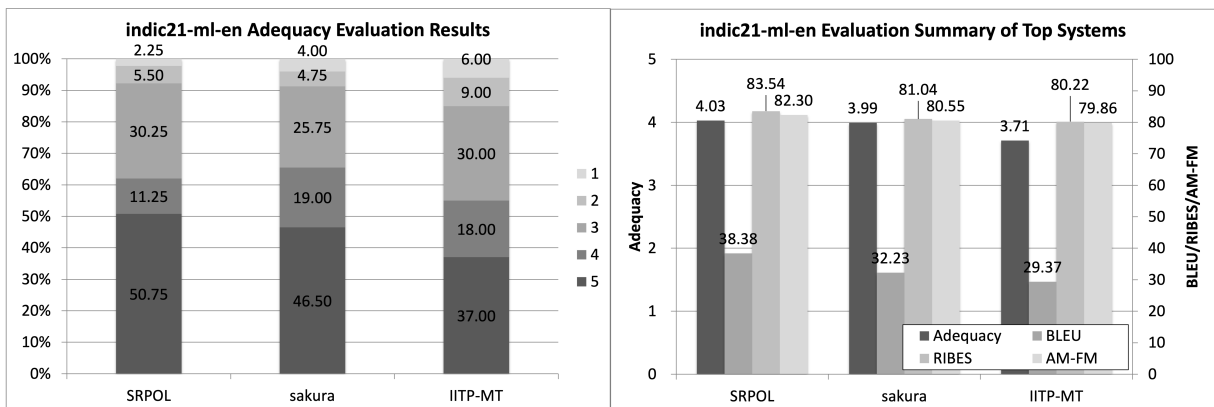


Figure 13: Official evaluation results of indic21-ml-en.

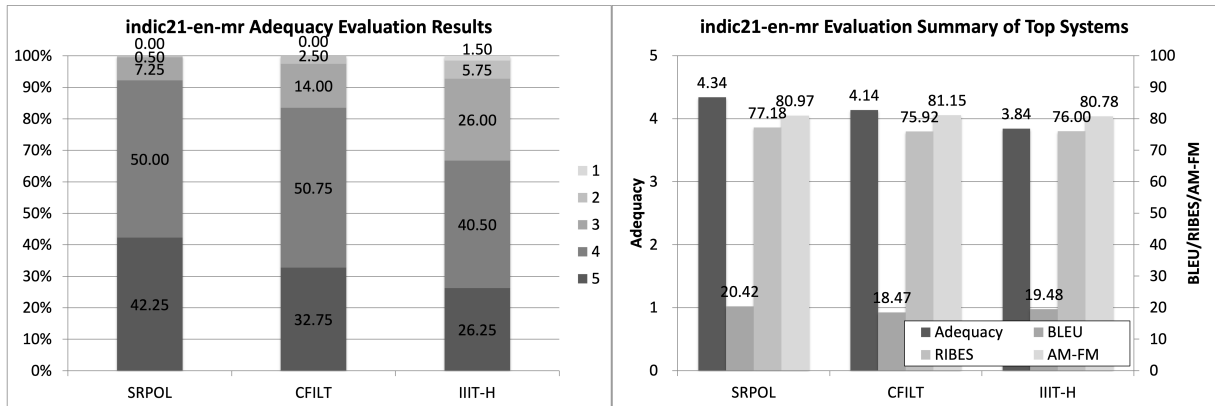


Figure 14: Official evaluation results of indic21-en-mr.

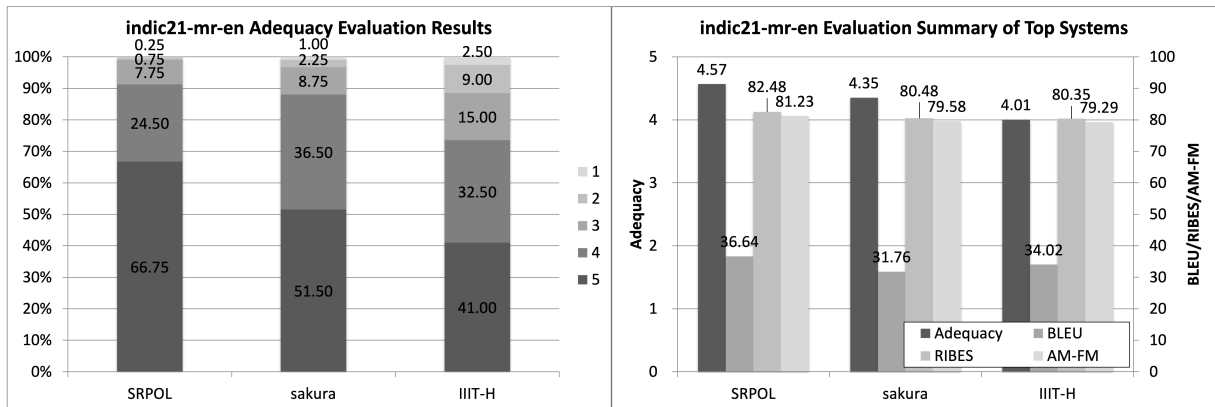


Figure 15: Official evaluation results of indic21-mr-en.

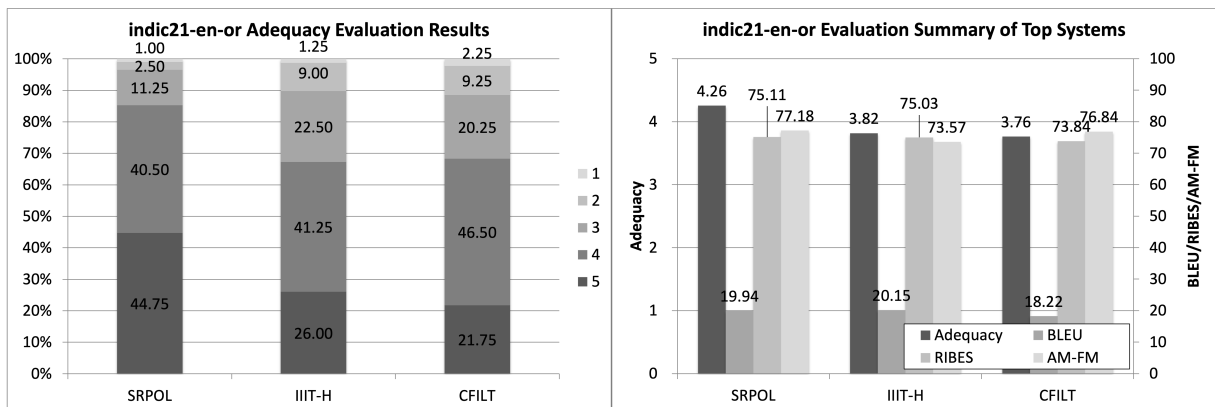


Figure 16: Official evaluation results of indic21-en-or.

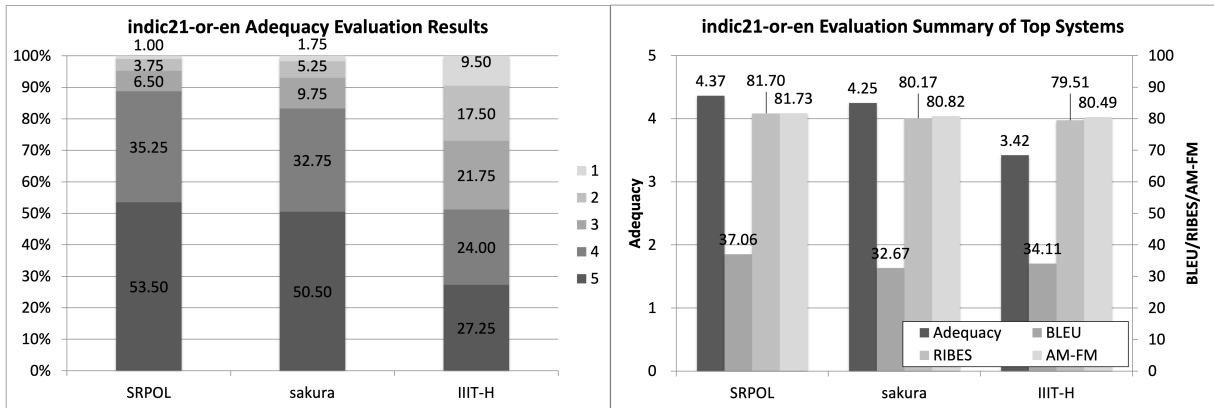


Figure 17: Official evaluation results of indic21-or-en.

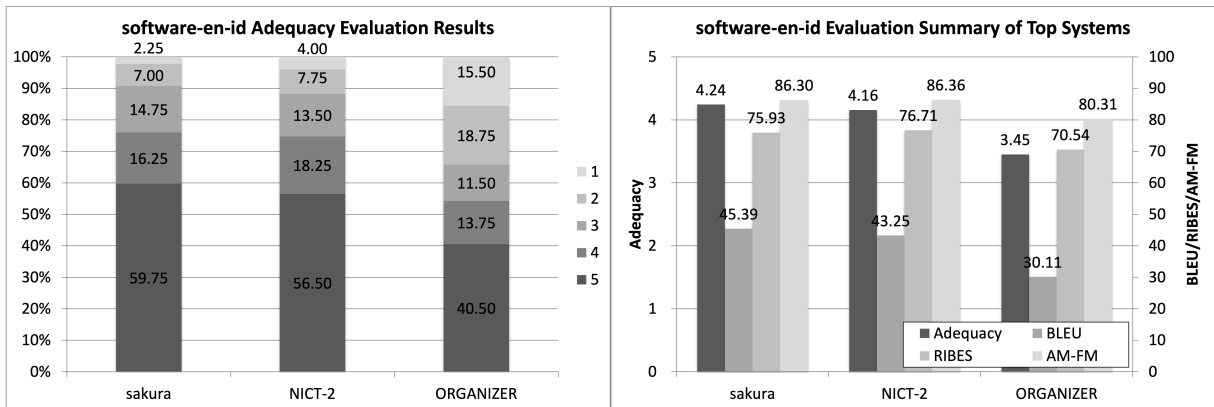


Figure 18: Official evaluation results of software-en-id.

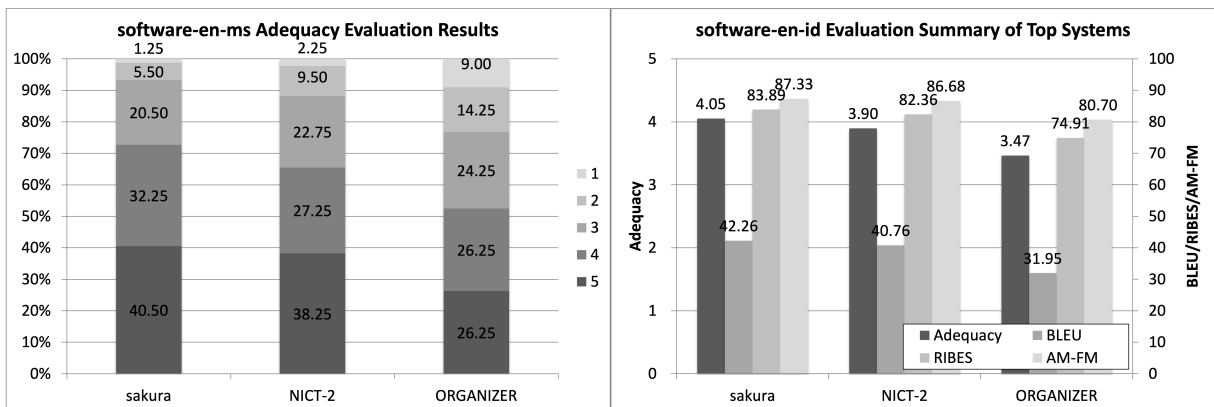


Figure 19: Official evaluation results of software-en-ms.



Subtask	SYSTEM DATA		Annotator A		Annotator B		all	weighted	
	ID	ID	average	variance	average	variance	average	$\kappa$	$\kappa$
jpcn-ja-en	TMU	5187	4.34	0.52	4.74	0.30	4.54	0.09	0.20
jpcn-en-ja	TMU	5347	4.21	1.34	4.34	1.14	4.27	0.33	0.53
indic21-en-bn	SRPOL	6232	4.57	0.71	4.74	0.36	4.65	0.30	0.36
	sakura	6150	4.32	1.25	4.46	0.60	4.38	0.20	0.34
	IIIT-H	6005	3.89	2.01	4.00	1.55	3.94	0.27	0.52
indic21-bn-en	SRPOL	6242	4.79	0.31	4.80	0.18	4.80	0.14	0.18
	IIIT-H	6015	3.67	2.38	3.96	1.49	3.81	0.31	0.54
	mcairt	6332	3.33	1.82	3.85	1.26	3.59	0.19	0.34
indic21-en-kn	SRPOL	6235	4.70	0.28	4.74	0.41	4.71	0.23	0.29
	sakura	6153	4.73	0.20	4.41	0.64	4.57	0.15	0.22
	IIIT-H	6008	4.11	0.63	3.90	1.35	4.00	0.33	0.48
indic21-kn-en	SRPOL	6245	4.63	0.29	4.81	0.25	4.72	0.25	0.30
	sakura	5873	4.62	0.38	4.36	1.23	4.49	0.21	0.32
	IIIT-H	6018	4.17	0.77	3.70	2.23	3.94	0.21	0.40
indic21-en-ml	SRPOL	6236	4.26	1.09	4.56	0.37	4.41	0.20	0.30
	CFILT	6046	3.46	1.30	3.60	1.26	3.54	0.16	0.30
	IIIT-H	6009	2.24	1.99	3.19	0.58	2.71	0.04	0.11
indic21-ml-en	SRPOL	6246	3.27	0.90	4.78	0.43	4.03	0.05	0.05
	sakura	5874	3.57	0.86	4.42	1.33	3.99	0.03	0.10
	IITP-MT	6289	3.31	1.23	4.12	1.41	3.71	0.11	0.21
indic21-en-mr	SRPOL	6237	4.26	0.34	4.42	0.44	4.34	0.05	0.04
	CFILT	6047	4.08	0.44	4.20	0.65	4.14	0.01	0.01
	IIIT-H	6010	3.63	0.71	4.05	0.93	3.84	0.09	0.18
indic21-mr-en	SRPOL	6247	4.34	0.55	4.79	0.31	4.57	0.07	0.11
	sakura	5875	4.14	0.70	4.56	0.53	4.35	0.12	0.18
	IIIT-H	6021	3.86	1.26	4.15	0.99	4.00	0.05	0.15
indic21-en-or	SRPOL	6238	4.12	0.65	4.38	0.69	4.25	0.31	0.49
	IIIT-H	6011	3.80	0.77	3.83	1.08	3.82	0.63	0.75
	CFILT	6048	3.75	0.90	3.77	0.97	3.76	0.77	0.85
indic21-or-en	SRPOL	6248	4.36	0.85	4.38	0.56	4.37	0.14	0.35
	sakura	5876	4.24	1.06	4.26	0.75	4.25	0.26	0.49
	IIIT-H	6022	3.34	2.08	3.50	1.32	3.42	0.32	0.63
software-en-id	sakura	5799	4.86	0.20	3.62	1.37	4.24	0.02	0.07
	NICT-2	5902	4.74	0.45	3.58	1.56	4.16	0.07	0.15
	organizer	3609	4.17	1.66	2.73	2.04	3.45	0.13	0.25
software-en-ms	sakura	5818	3.44	0.76	4.66	0.38	4.05	0.01	0.08
	NICT-2	5904	3.25	0.93	4.54	0.61	3.90	-0.03	0.09
	organizer	3610	2.88	1.18	4.05	1.34	3.46	0.06	0.27

Table 19: JPO adequacy evaluation results in detail.

### 8.3 Indic Multilingual Task

In WAT 2021, we received an overwhelming participation from 11 teams, 10 of which submitted system description papers. In contrast, in WAT 2020 there were only 4 system description papers. All participants trained multilingual NMT models. Some teams focused on leveraging monolingual corpora for pre-training MBART models or for backtranslation whereas other teams focused on script mapping to increase the similarity between the Indian languages and other teams focused on language family specific (Indo-Aryan vs Dravidian) models. Compared to the previous years, it is clear that backtranslation needs to be supplemented with pre-training as well as data selection for the best translation quality. The best performing team, “SRPOL”, used back-translation, pre-training, data selection and domain adapta-

tion. Following “SRPOL” teams such as “sakura”, “CFILT”, “IIIT-H”, “IITP-MT” and “mcairt” performed the best with ranks varying depending on the translation direction. One important observation we made was that “SRPOL” results for Indian to English translation were far higher than those of the other teams. In general their submission were 2 to 5 BLEU higher than the second best team. We suppose that this is due to their detailed experimentation with data selection and back-translation. On the other hand, for English to Indian language translation, although “SRPOL” had the highest BLEU for most directions, the gap between “SRPOL” and other participants was not that high. In a number of cases the differences were less than 0.5 BLEU which is not significant.

In general, we observed that translation into English had substantially high BLEU scores with

most participants obtaining higher than 25 BLEU for most directions. This makes sense because Indian languages are similar to each other and when the target language is the same, the increase in the target language data and transfer learning on the source side will lead to a large improvement in translation quality. In most cases, the scores for Indo-Aryan (Hindi, Marathi, Oriya, Punjabi, Gujarati and Bengali) to English translation were much higher than the scores for Dravidian (Tamil, Telugu, Kannada and Malayalam) to English translation.

On the other hand, for translation into Indian languages, BLEU scores were relatively lower. This is due to the morphological richness of Indian languages as well as the fact that multilingual English to Indian language translation does not benefit from the abundance of target language corpus like multilingual Indian language to English translation does. The BLEU scores for translation into Indo-Aryan languages such as Hindi and Punjabi showed the best translation quality exceeding 30 BLEU. This makes sense because Hindi and Punjabi are very similar and Hindi is the most resource rich among all Indian languages. It is certain that Punjabi benefits from the Hindi parallel data via transfer learning despite not sharing the same script. Script sharing, a technique used by some participants, could help enhance the amount of transfer learning taking place even further. For other Indo-Aryan languages the translation quality was a bit lower where English to Bengali exhibited the least translation quality compared to the other Indo-Aryan languages. This shows that linguistic similarity is not enough to lead to a high amount of transfer. In the case of translation into Dravidian languages we observed the lowest BLEU scores, usually around 15 BLEU or lower, with the exception of English to Kannada. Despite having larger corpora than some Indo-Aryan languages, translation into Dravidian languages is very hard as they are significantly morphologically richer than Indo-Aryan languages. Simply leveraging large monolingual corpora may not be enough and methods that take Dravidian linguistics into account may be necessary.

With regards to human evaluation, we observed that differences in BLEU scores do not always correspond to differences in human evaluation scores. For example, take the case of English to Malayalam translation where the gap between “SRPOL”

and “CFILT” in terms of BLEU is 2.7 and in terms of JPO scores is 0.87. For the same teams in case of English to Marathi, the gap in BLEU and JPO scores are 1.95 and 0.2 respectively. The difference between a gap of 2.7 and 1.95 is not very large as it is on a scale of 100<sup>72</sup> but the difference between 0.87 and 0.2 on a scale of 5<sup>73</sup> is quite large. In previous editions of this workshop we have always insisted that BLEU scores should not always be trusted in order to decide if translations truly are the best and this year’s human evaluation results show that this is still the case. Multi-metric evaluation helps us better understand different aspects of translation and we recommend readers to adopt the same even if automatic metrics are used. Although we are limited by budgetary constraints we hope to conduct larger scale human evaluation in the future.

#### 8.4 English→Hindi Multi-Modal Task

This year four teams participated in the different sub-tasks (TEXT, MM, and HI) of the English→Hindi Multi-Modal task. The WAT2021 automatic evaluation scores for the participating teams are shown in Tables 63, 60, 62, 58, 55, 57. The team “Volta” obtained the highest BLEU score for the text-only translation (TEXT) for both the evaluation (E-Test) and challenge (C-Test) test set. The best performance is obtained by fine-tuned *mBART* using IITB Corpus as an additional resource. For the captioning sub-task (HI) one team “NLPHut” participated and able to obtain better results compared to previous years’ best results based on region-specific image caption generation. For the multimodal sub-task (MM), we received three submissions from the teams “Volta”, “iitp” and “CNLP-NITS-PP”, respectively. The team “Volta” obtained the highest BLEU score for the multimodal translation (MM) for both the evaluation (E-Test) and challenge (C-Test) test set. They extracted object tags from images using visual information to enhance the textual input and achieve the BLEU score of 51.60 on the challenge test set, also the translation output able to resolve ambiguity as compared with text-only translation.

Due to constraints, no human evaluation was made this year for the English→Hindi Multi-Modal Task.

<sup>72</sup>BLEU scores go from 0 to 100.

<sup>73</sup>Human evaluation scores go from 1 to 5.

## 8.5 English→Malayalam Multi-Modal Task

This year one team “NLP Hut” participated in the different sub-tasks text-only translation (TEXT) and Malayalam captioning (ML) sub-tasks of the English→Multi-Modal task. The WAT2021 automatic evaluation scores are shown in the Table 64, 61, 59, 56.

For English to Malayalam text-only translation the team “NLP Hut” using the *Transformer* model obtained a BLEU score of 34.83 as compared to baseline of 30.49 on the evaluation test set and for the challenge test set obtained 12.15 compared to the baseline 12.98. For Malayalam image captioning, the team “NLP Hut” used the region-specific approach by extracting image features for the given specific region (bounding box) along with the whole image features and concatenating both to pass into an LSTM decoder to obtain the captions.

Due to constraints, no human evaluation was made this year for the English→Malayalam Multi-Modal Task.

## 8.6 Flickr30kEnt-JP Japanese↔English Multi-Modal Tasks

This year, two teams participated in the English→Japanese task, and one team participated in the Japanese→English task, respectively. It is notable that all submissions outperformed the best scores in WAT 2020, probably because of the increased size of the training dataset as well as the novel techniques introduced by the participants.

Overall, we observe the similar trend as in the last year. In the English→Japanese task, MMT systems constantly outperformed text-only NMT models including unconstrained ones, while in the Japanese→English task, unconstrained NMT model achieved the best performance. This is perhaps because the Flickr30kEnt-JP dataset itself is indeed constructed by English to Japanese human translation where images were actually referred to resolve ambiguity. One team developed an elegant method for soft alignment of word-region to realize better grounding of multimodal information, which is shown to achieve a favorable performance gain. This result again indicates the importance of text-image grounding in MMT, and we believe that we still have much room for improvements.

## 8.7 Ambiguous MS COCO Japanese↔English Multimodal Task

This year only one team participated in the English→Japanese task. Their system was based on a word-region alignment method to enhance the interaction between source tokens and image regions and then integrating aligned information to the visual features during decoding (Zhao et al., 2021). We observe that their system outperformed the organizer’s system, which is based on double attention to both source tokens and image regions. It verified that it is important to integrate visual information in a proper way for this task and multimodal MT in general that text is a strong clue for translation, but visual information can further improve translation if it is used properly.

Unfortunately, there is no team participating the Japanese→English task. We hope that we can have more participants next year for the tasks in both directions.

## 8.8 Restricted Translation Task

We received 3 systems for the English→Japanese translation task and 4 systems for the Japanese→English.<sup>74</sup> On the whole, all the submitted systems are basically lexical-constraint-aware NMT models with lexically constrained decoding method, where the restricted target vocabulary is concatenated into source sentences and, during the beam search at inference time, the models generate translation outputs containing the target vocabulary. We observed that these techniques boost the final translation performance of the NMT models in the restricted translation task.

For human evaluation, we conducted the source-based direct assessment (Cettolo et al., 2017; Federmann, 2018) and source-based contrastive assessment (Sakaguchi and Van Durme, 2018; Federmann, 2018), to have the top-ranked systems of each team appraised by bilingual human annotators. In the human evaluation campaign, we also include the human reference data. Table 20 reports the final automatic evaluation score and the human evaluation results. In both tasks, the systems from the team “NTT” are the most highly evaluated in all the submitted systems in the final score and the human evaluation, consistently. We also note that our designed automation metric is well correlated

<sup>74</sup>We discuss 3 submitted systems from the teams “NTT”, “NHK”, and “NICTRB” teams, as we do not have a system description paper from the team “TMU”.

En-Ja Team	final	Human Eval.	
		src-based DA	src-based CA
NTT	<b>57.2</b>	<b>77.5</b>	<b>79.7</b>
NHK	33.9	74.1	77.2
NICTRB	28.8	73.6	77.1
(human ref.)	—	73.4	76.4

Ja-En Team	final	Human Eval.	
		src-based DA	src-based CA
NTT	<b>44.1</b>	<b>75.6</b>	<b>74.4</b>
NHK	37.5	73.9	73.5
NICTRB	31.8	72.1	71.8
TMU	22.6	50.2	48.3
(human ref.)	—	74.1	72.9

Table 20: Human evaluation results of source-based direct assessment (src-based DA) and source-based contrastive assessment (src-based CA), ranging 0 to 100. The column of “final” reports the final score of the automatic evaluation metric described in Section 2.13

with the human evaluation results. Besides that, we found that the ASPEC human reference data might have a quality issue, consisting of low-quality examples that are annotated with a score of [0, 50], with the ratio of (En-Ja, Ja-En)=(13.30%, 12.43%). This is why a few systems are shown to surpass the original human reference data in the human evaluation.

## 9 Conclusion and Future Perspective

This paper summarizes the shared tasks of WAT2021. We had 24 participants worldwide who submitted their translation results for the human evaluation, and collected a large number of useful submissions for improving the current machine translation systems by analyzing the submissions and identifying the issues.

For the next WAT workshop, we will try to add more Indic languages to our MultiIndicMT task along with newer evaluation sets. Also, we will add a new English→Bengali Multi-Modal task into the Multimodal translation tasks.

## Acknowledgement

The English→Hindi and English→Malayalam Multi-Modal shared tasks were supported by the following grants at Idiap Research Institute and Charles University. The authors do not see any significant ethical or privacy concerns that would prevent the processing of the data used in the study. The datasets do contain personal data, and these are processed in compliance with the GDPR and national law.

- 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).

## References

- Rafael E. Banchs, Luis F. D’Haro, and Haizhou Li. 2015a. *Adequacy-fluency metrics: Evaluating mt in the continuous space model framework*. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(3):472–482.
- Rafael E. Banchs, Luis F. D’Haro, and Haizhou Li. 2015b. *Adequacy–fluency metrics: Evaluating mt in the continuous space model framework*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.
- Bianka Buschbeck and Miriam Exel. 2020. *A parallel evaluation data set of software documentation with document structure annotation*.
- M. Cettolo, Marcello Federico, L. Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsutho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213 – 220.
- Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. *A survey of multilingual neural machine translation*. *ACM Comput. Surv.*, 53(5).
- Luis Fernando D’Haro, Rafael E. Banchs, Chiori Hori, and Haizhou Li. 2019. *Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics*. *Computer Speech and Language*, 55:200–215.
- 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.
- Christian Federmann. 2018. [Appraise evaluation framework for machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.
- 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. European Association for Machine Translation.
- 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](https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf).
- 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.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Andrew Merritt, Chenhui Chu, and Yuki Arase. 2020. [A corpus for english-japanese multimodal neural machine translation with comparable sentences](#).
- 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, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. [Overview of the 7th workshop on Asian translation](#). In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Suzhou, China. Association for Computational Linguistics.
- Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. [Overview of the 5th workshop on Asian translation](#). In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

- Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. [Pointwise prediction for robust, adaptable japanese morphological analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 529–533, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.
- Shantipriya Parida and Ondřej Bojar. 2021. [Malayalam visual genome 1.0](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- 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.
- 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.
- Keisuke Sakaguchi and Benjamin Van Durme. 2018. [Efficient online scalar annotation with bounded support](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.
- 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).
- Huihsin Tseng. 2005. A conditional random field word segmenter. In *In Fourth SIGHAN Workshop on Chinese Language Processing*.
- Masao Utiyama and Hitoshi Isahara. 2007. A japanese-english patent parallel corpus. In *MT summit XI*, pages 475–482.
- Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Lukasz 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, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *CoRR*, abs/1706.03762.
- Yi Mon Shwe Sin and Khin Mar Soe. 2018. Syllable-based myanmar-english neural machine translation. In *In Proc. of ICCA*, pages 228–233.
- Chen Zhang, Luis Fernando D’Haro, Rafael E. Banchs, Thomas Friedrichs, and Haizhou Li. 2021a. [Deep AM-FM: Toolkit for Automatic Dialogue Evaluation](#), pages 53–69. Springer Singapore, Singapore.
- Chen Zhang, Grandee Lee, Luis Fernando D’Haro, and Haizhou Li. 2021b. [D-score: Holistic dialogue evaluation without reference](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1.
- Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. [Double attention-based multimodal neural machine translation with semantic image regions](#). In *EAMT*, pages 105–114.
- Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. [Neural machine translation with semantically relevant image regions](#). In *NLP*.

## Appendix A Submissions

Tables 21 to 76 summarize translation results submitted to WAT2021. Type and RSRC columns indicate type of method and use of other resources.

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5916	NMT	YES	34.970000	0.822350	0.839182
sakura	5791	NMT	NO	34.250000	0.820590	0.849202

Table 21: ALT20 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5918	NMT	YES	41.15	0.901974	0.867678
sakura	5798	NMT	NO	41.57	0.901977	0.868025

Table 22: ALT20 en-id submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5920	NMT	YES	45.17	0.912195	0.873476
sakura	5816	NMT	NO	44.01	0.908439	0.871875

Table 23: ALT20 en-ms submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5922	NMT	YES	55.690000	0.815863	0.832513
sakura	5843	NMT	NO	55.980000	0.818307	0.837062

Table 24: ALT20 en-th submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5917	NMT	YES	35.21	0.834649	0.814594
sakura	5793	NMT	NO	36.17	0.835220	0.832895

Table 25: ALT20 hi-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5919	NMT	YES	43.90	0.898700	0.844199
sakura	5800	NMT	NO	44.72	0.897314	0.850998

Table 26: ALT20 id-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5921	NMT	YES	44.53	0.904478	0.841632
sakura	5821	NMT	NO	45.70	0.901696	0.851471

Table 27: ALT20 ms-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5923	NMT	YES	28.96	0.829525	0.817972
sakura	5845	NMT	NO	30.10	0.832399	0.822585

Table 28: ALT20 th-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
YCC-MT1	6195	SMT	NO	20.880000	0.553319	0.655310
YCC-MT1	6201	SMT	NO	20.130000	0.545962	0.654820
YCC-MT2	6175	NMT	NO	14.820000	0.659582	0.663840
YCC-MT2	6178	NMT	NO	14.020000	0.639593	0.645470
sakura	6031	NMT	NO	29.620000	0.739320	0.752340

Table 29: ALT2 en-my submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NECTEC	6188	NMT	NO	6.24	0.620840	0.424640
NECTEC	6192	NMT	NO	4.62	0.587155	0.391710
sakura	5230	NMT	NO	19.75	0.742698	0.562680
sakura	5990	NMT	NO	18.70	0.736523	0.550430

Table 30: ALT2 my-en submissions



System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
NTT	5368	NMT	NO	56.39	56.87	56.57	0.882454	0.882322	0.887104	0.817290
NTT	5616	NMT	YES	56.20	56.67	56.47	0.885308	0.885612	0.889831	0.818190
nictrb	5591	NMT	YES	51.07	51.32	51.36	0.836874	0.839934	0.844141	0.799950
NHK	5502	NMT	NO	52.07	52.69	52.33	0.815612	0.823084	0.827300	0.801660

Table 31: ASPECRT en-ja submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
TMU	5994	NMT	NO	25.29	25.29	25.29	0.653597	0.612290	0.612290	0.612290
NTT	5209	NMT	NO	44.34	44.34	44.34	0.811700	0.672320	0.672320	0.672320
NTT	5615	NMT	YES	44.28	44.28	44.28	0.813155	0.676670	0.676670	0.676670
nictrb	5592	NMT	YES	37.01	37.01	37.01	0.753823	0.651570	0.651570	0.651570
NHK	5505	NMT	NO	42.94	42.94	42.94	0.801015	0.661560	0.661560	0.661560

Table 32: ASPECRT ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4789	NMT	NO	11.27	0.638781	0.613093
NICT-5	5274	NMT	NO	21.37	0.747435	0.744400
NICT-5	5349	NMT	NO	23.89	0.754772	0.758921
NLPHut	4583	NMT	NO	13.88	0.669588	0.657119
mcairt	6026	NMT	NO	25.22	0.773387	0.778620
mcairt	6332	NMT	NO	29.96	0.798326	0.786717
sakura	5870	NMT	NO	26.69	0.776808	0.772365
IIIT-H	6015	NMT	NO	28.28	0.773574	0.773292
gaurvar	5556	NMT	NO	11.33	0.634088	0.673457
gaurvar	5565	NMT	NO	11.83	0.629932	0.674034
IITP-MT	6280	NMT	NO	25.77	0.774004	0.777377
SRPOL	6242	NMT	NO	31.87	0.800501	0.789735
SRPOL	6268	NMT	NO	31.82	0.800145	0.792364
CFILT	6052	NMT	NO	25.98	0.760268	0.766461
coastal	6162	NMT	NO	24.39	0.772190	0.778356
CFILT-IITB	6112	NMT	NO	18.48	0.721176	0.730379
CFILT-IITB	6124	NMT	NO	20.18	0.732342	0.734491

Table 33: HINDEN21 bn-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4788	NMT	NO	5.580000	0.573377	0.701527
NICT-5	5273	NMT	NO	10.590000	0.677858	0.755363
NICT-5	5348	NMT	NO	12.840000	0.704620	0.767497
NLPHut	4582	NMT	NO	8.130000	0.645895	0.735005
mcairt	6000	NMT	NO	13.020000	0.715490	0.779592
sakura	6150	NMT	NO	13.830000	0.716347	0.764714
IIIT-H	6005	NMT	NO	14.730000	0.724245	0.759513
gaurvar	5588	NMT	NO	3.230000	0.452631	0.628707
gaurvar	5938	NMT	NO	2.950000	0.465755	0.641712
IITP-MT	6278	NMT	NO	11.040000	0.703372	0.731181
SRPOL	6232	NMT	NO	15.970000	0.733646	0.771033
SRPOL	6258	NMT	NO	15.580000	0.732792	0.772309
CFILT	6041	NMT	NO	13.240000	0.710664	0.777074
coastal	6074	NMT	NO	11.090000	0.694142	0.763665

Table 34: HINDEN21 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4790	NMT	NO	16.380000	0.748273	0.757069
NICT-5	5275	NMT	NO	23.040000	0.797371	0.801466
NICT-5	5350	NMT	NO	24.260000	0.806181	0.811717
NLPHut	4585	NMT	NO	17.760000	0.763222	0.768177
mcairt	6003	NMT	NO	23.210000	0.809389	0.816739
sakura	6151	NMT	NO	25.270000	0.814798	0.813350
IIIT-H	6006	NMT	NO	26.970000	0.820249	0.820127
gaurvar	5580	NMT	NO	6.810000	0.586360	0.628529
gaurvar	5927	NMT	NO	6.920000	0.599337	0.645669
IITP-MT	6281	NMT	NO	20.460000	0.750935	0.808824
SRPOL	6233	NMT	NO	27.800000	0.824866	0.821221
SRPOL	6259	NMT	NO	27.310000	0.822329	0.819923
CFILT	6042	NMT	NO	24.560000	0.806649	0.817681
coastal	6078	NMT	NO	20.420000	0.795314	0.809795

Table 35: HINDEN21 en-gu submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4792	NMT	NO	23.310000	0.778841	0.759679
NICT-5	5277	NMT	NO	29.590000	0.817892	0.800234
NICT-5	5352	NMT	NO	30.180000	0.820984	0.801680
NLPHut	5987	NMT	NO	25.370000	0.788001	0.747598
mcairt	6004	NMT	NO	35.850000	0.846656	0.822626
sakura	6152	NMT	NO	36.920000	0.848042	0.816999
IIIT-H	6007	NMT	NO	38.250000	0.854192	0.822836
gaurvar	5578	NMT	NO	17.020000	0.681760	0.676601
gaurvar	5928	NMT	NO	15.860000	0.647511	0.681511
IITP-MT	6283	NMT	NO	34.480000	0.844721	0.820543
SRPOL	6254	NMT	NO	38.650000	0.855879	0.824649
SRPOL	6260	NMT	NO	38.040000	0.852496	0.822371
CFILT	6043	NMT	NO	35.390000	0.843969	0.821713
coastal	6079	NMT	NO	31.750000	0.829731	0.801179

Table 36: HINDEN21 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4794	NMT	NO	10.110000	0.651048	0.741873
NICT-5	5279	NMT	NO	16.130000	0.732794	0.798654
NICT-5	5354	NMT	NO	18.220000	0.746230	0.813658
NLPHut	4591	NMT	NO	11.840000	0.689612	0.762931
mcairt	5998	NMT	NO	14.580000	0.726259	0.805963
sakura	6153	NMT	NO	18.830000	0.760100	0.817831
IIIT-H	6008	NMT	NO	19.570000	0.756613	0.812490
gaurvar	5581	NMT	NO	4.350000	0.477922	0.658271
gaurvar	5929	NMT	NO	3.900000	0.469815	0.657091
IITP-MT	6285	NMT	NO	13.220000	0.635288	0.791821
SRPOL	6235	NMT	NO	21.300000	0.770110	0.821941
SRPOL	6261	NMT	NO	20.910000	0.771246	0.821329
CFILT	6044	NMT	NO	17.980000	0.747233	0.816981
coastal	6113	NMT	NO	16.110000	0.736528	0.809687

Table 37: HINDEN21 en-kn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4796	NMT	NO	3.340000	0.475441	0.706782
NICT-5	5281	NMT	NO	5.980000	0.605053	0.764924
NICT-5	5356	NMT	NO	6.510000	0.623301	0.789337
NLPHut	4590	NMT	NO	4.570000	0.554478	0.740136
mcairt	6002	NMT	NO	6.170000	0.622598	0.793308
sakura	5886	NMT	NO	10.940000	0.686534	0.794481
IIIT-H	6009	NMT	NO	12.760000	0.672331	0.745043
gaurvar	5582	NMT	NO	1.790000	0.338533	0.666547
gaurvar	5930	NMT	NO	1.480000	0.306966	0.656847
IITP-MT	6287	NMT	NO	3.790000	0.437679	0.758960
SRPOL	6236	NMT	NO	15.490000	0.736915	0.807998
SRPOL	6262	NMT	NO	15.430000	0.734111	0.808089
CFILT	6046	NMT	NO	12.790000	0.707437	0.805291
coastal	6081	NMT	NO	6.270000	0.619774	0.784292

Table 38: HINDEN21 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4798	NMT	NO	8.820000	0.652134	0.730656
NICT-5	5283	NMT	NO	14.690000	0.720677	0.785952
NICT-5	5358	NMT	NO	16.380000	0.739171	0.800357
NLPHut	4594	NMT	NO	10.410000	0.684554	0.745915
mcairt	5999	NMT	NO	14.900000	0.740079	0.791850
sakura	6156	NMT	NO	17.870000	0.752439	0.803566
IIIT-H	6010	NMT	NO	19.480000	0.760009	0.807758
gaurvar	5583	NMT	NO	5.100000	0.482727	0.654698
gaurvar	5931	NMT	NO	4.490000	0.467281	0.658104
IITP-MT	6291	NMT	NO	13.950000	0.665934	0.798673
SRPOL	6237	NMT	NO	20.420000	0.771845	0.809721
SRPOL	6263	NMT	NO	19.930000	0.766897	0.810757
CFILT	6047	NMT	NO	18.470000	0.759182	0.811499
coastal	6082	NMT	NO	14.480000	0.727647	0.799538

Table 39: HINDEN21 en-mr submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4800	NMT	NO	9.080000	0.638520	0.714530
NICT-5	5285	NMT	NO	15.010000	0.716665	0.748319
NICT-5	5360	NMT	NO	16.690000	0.734028	0.757804
NLPHut	4596	NMT	NO	12.810000	0.693696	0.736638
mcairt	5996	NMT	NO	17.710000	0.743984	0.763064
sakura	6157	NMT	NO	17.880000	0.740263	0.769884
IIIT-H	6011	NMT	NO	20.150000	0.750260	0.735718
gaurvar	5584	NMT	NO	2.200000	0.380253	0.591864
gaurvar	5932	NMT	NO	2.600000	0.431373	0.611704
IITP-MT	6293	NMT	NO	12.570000	0.714731	0.737576
SRPOL	6238	NMT	NO	19.940000	0.751086	0.771831
SRPOL	6264	NMT	NO	19.150000	0.749740	0.771493
CFILT	6048	NMT	NO	18.220000	0.738397	0.768399
coastal	6084	NMT	NO	15.660000	0.727477	0.758199

Table 40: HINDEN21 en-or submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4802	NMT	NO	21.770000	0.765216	0.762364
NICT-5	5287	NMT	NO	26.940000	0.808173	0.794023
NICT-5	5362	NMT	NO	29.150000	0.820085	0.803326
NLPHut	4598	NMT	NO	22.600000	0.785047	0.778215
mcairt	6001	NMT	NO	30.560000	0.830405	0.810106
sakura	6158	NMT	NO	30.930000	0.829019	0.802223
IIIT-H	6012	NMT	NO	33.350000	0.837603	0.810972
gaurvar	5585	NMT	NO	9.350000	0.633937	0.620318
gaurvar	5933	NMT	NO	10.020000	0.632319	0.643473
IITP-MT	6298	NMT	NO	16.810000	0.785680	0.663206
SRPOL	6239	NMT	NO	33.430000	0.837542	0.814115
SRPOL	6265	NMT	NO	32.880000	0.835465	0.813158
CFILT	6049	NMT	NO	31.160000	0.826367	0.813658
coastal	6085	NMT	NO	27.250000	0.816792	0.803382

Table 41: HINDEN21 en-pa submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4804	NMT	NO	6.380000	0.588286	0.723160
NICT-5	5289	NMT	NO	10.330000	0.675039	0.776138
NICT-5	5364	NMT	NO	11.420000	0.701210	0.792622
NLPHut	4616	NMT	NO	7.680000	0.630830	0.739011
mcairt	5995	NMT	NO	11.980000	0.707054	0.801632
sakura	6159	NMT	NO	13.250000	0.721520	0.795712
IIIT-H	6013	NMT	NO	14.430000	0.711995	0.778991
gaurvar	5586	NMT	NO	4.090000	0.452271	0.694376
gaurvar	5934	NMT	NO	3.600000	0.431281	0.684232
IITP-MT	6303	NMT	NO	8.510000	0.578195	0.756693
SRPOL	6240	NMT	NO	14.150000	0.730705	0.798837
SRPOL	6266	NMT	NO	13.890000	0.728770	0.799382
CFILT	6050	NMT	NO	12.990000	0.715699	0.802920
coastal	6086	NMT	NO	9.990000	0.682220	0.788022

Table 42: HINDEN21 en-ta submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4806	NMT	NO	2.800000	0.479896	0.708086
NICT-5	5291	NMT	NO	4.590000	0.569735	0.754015
NICT-5	5366	NMT	NO	4.200000	0.576863	0.752068
NLPHut	5986	NMT	NO	4.880000	0.570112	0.713960
mcairt	5997	NMT	NO	11.170000	0.702337	0.783647
sakura	6160	NMT	NO	15.480000	0.725543	0.785055
IIIT-H	6014	NMT	NO	15.610000	0.728432	0.780218
gaurvar	5587	NMT	NO	2.310000	0.414016	0.634376
gaurvar	5935	NMT	NO	2.310000	0.389727	0.642502
IITP-MT	6305	NMT	NO	6.250000	0.530898	0.764977
SRPOL	6241	NMT	NO	16.850000	0.739835	0.791085
SRPOL	6267	NMT	NO	16.820000	0.734483	0.792970
CFILT	6051	NMT	NO	15.520000	0.725496	0.789820
coastal	6088	NMT	NO	12.860000	0.707817	0.778251

Table 43: HINDEN21 en-te submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4791	NMT	NO	26.21	0.764569	0.726576
NICT-5	5276	NMT	NO	33.65	0.810918	0.793874
NICT-5	5351	NMT	NO	33.53	0.811609	0.796604
NLPHut	4633	NMT	NO	23.10	0.755101	0.713984
mcairt	6334	NMT	NO	36.77	0.829389	0.819546
sakura	5871	NMT	NO	38.73	0.834934	0.820654
IIIT-H	6016	NMT	NO	39.39	0.830158	0.806061
gaurvar	5557	NMT	NO	16.79	0.715044	0.696879
gaurvar	5566	NMT	NO	17.50	0.712002	0.698257
IITP-MT	6282	NMT	NO	36.49	0.827301	0.814556
SRPOL	6243	NMT	NO	43.98	0.853263	0.835789
SRPOL	6269	NMT	NO	42.87	0.849734	0.833146
CFILT	6053	NMT	NO	35.31	0.807849	0.797069
coastal	6163	NMT	NO	34.60	0.824060	0.814168
CFILT-IITB	6114	NMT	NO	28.79	0.786408	0.765441
CFILT-IITB	6125	NMT	NO	31.02	0.795199	0.776935

Table 44: HINDEN21 gu-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4793	NMT	NO	28.21	0.782146	0.736131
NICT-5	5278	NMT	NO	35.80	0.828390	0.808180
NICT-5	5353	NMT	NO	36.20	0.832916	0.805716
NLPHut	5985	NMT	NO	24.55	0.785027	0.721805
mcairt	6333	NMT	NO	40.05	0.850322	0.832119
sakura	5872	NMT	NO	41.58	0.856469	0.834172
IIIT-H	6017	NMT	NO	43.23	0.853267	0.823007
gaurvar	5532	NMT	NO	20.90	0.729188	0.714649
gaurvar	5567	NMT	NO	21.33	0.759034	0.722822
IITP-MT	6284	NMT	NO	40.08	0.851601	0.831265
SRPOL	6244	NMT	NO	46.93	0.872874	0.847064
SRPOL	6270	NMT	NO	45.61	0.867712	0.843456
CFILT	6054	NMT	NO	39.71	0.837668	0.822034
coastal	6164	NMT	NO	36.47	0.840014	0.824040
CFILT-IITB	6115	NMT	NO	30.90	0.807304	0.775032
CFILT-IITB	6126	NMT	NO	33.70	0.820716	0.791408

Table 45: HINDEN21 hi-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4795	NMT	NO	20.33	0.717654	0.692019
NICT-5	5280	NMT	NO	29.29	0.793521	0.782087
NICT-5	5355	NMT	NO	30.87	0.796119	0.792622
NLPHut	4593	NMT	NO	17.72	0.710551	0.679617
mcairt	6374	NMT	NO	31.16	0.803525	0.799216
sakura	5873	NMT	NO	34.11	0.815837	0.805112
IIIT-H	6018	NMT	NO	34.69	0.804694	0.790977
gaurvar	5558	NMT	NO	13.45	0.683906	0.687726
gaurvar	5568	NMT	NO	13.86	0.674282	0.687810
IITP-MT	6286	NMT	NO	31.24	0.806170	0.798540
SRPOL	6245	NMT	NO	40.34	0.840458	0.823730
SRPOL	6271	NMT	NO	39.01	0.837287	0.820355
CFILT	6055	NMT	NO	30.23	0.772913	0.778602
coastal	6165	NMT	NO	31.04	0.811950	0.806951
CFILT-IITB	6121	NMT	NO	24.01	0.758489	0.751223
CFILT-IITB	6131	NMT	NO	24.18	0.759045	0.744802

Table 46: HINDEN21 kn-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4797	NMT	NO	13.64	0.673109	0.646559
NICT-5	5282	NMT	NO	26.55	0.780019	0.772691
NICT-5	5357	NMT	NO	28.23	0.786269	0.786909
NLPHut	4634	NMT	NO	15.47	0.700957	0.668778
mcairt	6344	NMT	NO	28.07	0.792884	0.794932
sakura	5874	NMT	NO	32.23	0.810429	0.805450
IIIT-H	6020	NMT	NO	29.19	0.780463	0.748518
gaurvar	5559	NMT	NO	12.99	0.678961	0.684370
gaurvar	5569	NMT	NO	13.64	0.657440	0.684483
IITP-MT	6289	NMT	NO	29.37	0.802153	0.798550
SRPOL	6246	NMT	NO	38.38	0.835444	0.823006
SRPOL	6272	NMT	NO	37.04	0.830449	0.820716
CFILT	6056	NMT	NO	29.28	0.784424	0.789095
coastal	6166	NMT	NO	28.55	0.803090	0.805091
CFILT-IITB	6117	NMT	NO	22.10	0.751437	0.744459
CFILT-IITB	6130	NMT	NO	22.84	0.763162	0.745908

Table 47: HINDEN21 ml-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4799	NMT	NO	15.10	0.676716	0.658130
NICT-5	5284	NMT	NO	25.45	0.771352	0.764852
NICT-5	5359	NMT	NO	27.88	0.783012	0.779746
NLPHut	5983	NMT	NO	17.07	0.706399	0.696839
mcairt	6335	NMT	NO	27.29	0.785579	0.780231
sakura	5875	NMT	NO	31.76	0.804834	0.795844
IIIT-H	6021	NMT	NO	34.02	0.803479	0.792878
gaurvar	5560	NMT	NO	13.38	0.679550	0.692897
gaurvar	5570	NMT	NO	13.96	0.669879	0.693109
IITP-MT	6292	NMT	NO	29.96	0.799383	0.797333
SRPOL	6247	NMT	NO	36.64	0.824831	0.812258
SRPOL	6273	NMT	NO	35.68	0.821164	0.810290
CFILT	6057	NMT	NO	29.71	0.786570	0.789075
coastal	6167	NMT	NO	27.71	0.795729	0.791157
CFILT-IITB	6118	NMT	NO	23.57	0.752476	0.751917
CFILT-IITB	6127	NMT	NO	25.40	0.765200	0.767347

Table 48: HINDEN21 mr-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4801	NMT	NO	16.35	0.679781	0.730819
NICT-5	5286	NMT	NO	25.81	0.762604	0.780431
NICT-5	5361	NMT	NO	27.93	0.769634	0.782917
NLPHut	4597	NMT	NO	18.92	0.720916	0.740606
mcairt	6338	NMT	NO	29.96	0.798326	0.795586
sakura	5876	NMT	NO	32.67	0.801734	0.808239
IIIT-H	6022	NMT	NO	34.11	0.795132	0.804930
gaurvar	5550	NMT	NO	13.71	0.634313	0.725121
gaurvar	5571	NMT	NO	13.69	0.662493	0.721531
IITP-MT	6294	NMT	NO	31.19	0.794791	0.803226
SRPOL	6248	NMT	NO	37.06	0.816956	0.817318
SRPOL	6274	NMT	NO	36.04	0.812816	0.814871
CFILT	6058	NMT	NO	30.46	0.772850	0.793769
coastal	6107	NMT	NO	19.61	0.737380	0.727657
CFILT-IITB	6119	NMT	NO	25.05	0.754313	0.770941
CFILT-IITB	6128	NMT	NO	26.34	0.761082	0.780009

Table 49: HINDEN21 or-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4803	NMT	NO	23.66	0.749459	0.701483
NICT-5	5288	NMT	NO	34.34	0.816975	0.792541
NICT-5	5363	NMT	NO	35.81	0.827528	0.800753
NLPHut	4615	NMT	NO	24.35	0.766047	0.717322
mcairt	6342	NMT	NO	38.42	0.840360	0.818332
sakura	5877	NMT	NO	40.38	0.844351	0.823464
IIIT-H	6023	NMT	NO	41.24	0.837608	0.811169
gaurvar	5551	NMT	NO	18.61	0.703876	0.693631
gaurvar	5572	NMT	NO	18.59	0.730487	0.694658
IITP-MT	6301	NMT	NO	38.41	0.839598	0.815989
SRPOL	6249	NMT	NO	46.39	0.865765	0.841641
SRPOL	6275	NMT	NO	44.87	0.861389	0.836440
CFILT	6059	NMT	NO	38.01	0.818396	0.804561
coastal	6168	NMT	NO	35.90	0.835327	0.814440
CFILT-IITB	6123	NMT	NO	29.87	0.795413	0.772655
CFILT-IITB	6129	NMT	NO	32.34	0.805722	0.782112

Table 50: HINDEN21 pa-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4805	NMT	NO	16.07	0.690144	0.675969
NICT-5	5290	NMT	NO	24.72	0.766631	0.758282
NICT-5	5365	NMT	NO	26.90	0.780120	0.772249
NLP Hut	5984	NMT	NO	15.40	0.702428	0.669984
mcairt	6346	NMT	NO	28.04	0.793839	0.790184
sakura	5878	NMT	NO	31.09	0.806993	0.796074
IIIT-H	6024	NMT	NO	29.61	0.785332	0.750297
gaurvar	5563	NMT	NO	13.36	0.677433	0.687892
gaurvar	5573	NMT	NO	13.77	0.660037	0.688325
IITP-MT	6304	NMT	NO	27.76	0.788181	0.786587
SRPOL	6250	NMT	NO	36.13	0.822312	0.806540
SRPOL	6276	NMT	NO	35.06	0.815951	0.803595
CFILT	6060	NMT	NO	29.34	0.784291	0.785098
coastal	6169	NMT	NO	26.69	0.794380	0.786098
CFILT-IITB	6122	NMT	NO	21.37	0.747748	0.742311
CFILT-IITB	6132	NMT	NO	22.75	0.756364	0.745090

Table 51: HINDEN21 ta-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4807	NMT	NO	14.70	0.665774	0.636031
NICT-5	5292	NMT	NO	27.76	0.777383	0.771109
NICT-5	5367	NMT	NO	28.77	0.782427	0.779053
NLP Hut	4619	NMT	NO	16.48	0.695348	0.674821
mcairt	6348	NMT	NO	29.26	0.790319	0.786396
sakura	5879	NMT	NO	33.87	0.810630	0.802030
IIIT-H	6025	NMT	NO	30.44	0.783709	0.754690
gaurvar	5564	NMT	NO	12.14	0.652408	0.668328
gaurvar	5574	NMT	NO	12.44	0.629617	0.666143
IITP-MT	6306	NMT	NO	28.13	0.784897	0.776964
SRPOL	6251	NMT	NO	39.80	0.836433	0.820889
SRPOL	6277	NMT	NO	38.57	0.831502	0.820360
CFILT	6061	NMT	NO	30.10	0.778981	0.783349
coastal	6170	NMT	NO	30.50	0.806646	0.799696
CFILT-IITB	6120	NMT	NO	22.37	0.746368	0.743435
CFILT-IITB	6133	NMT	NO	24.02	0.757702	0.745885

Table 52: HINDEN21 te-en submissions



System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
TMU	5347	NMT	NO	45.24	47.12	45.27	0.854558	0.853335	0.854298	0.876323

Table 53: JPCN en-ja submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				NMT	NO	NO	NMT	NO	NO	
TMU	5187	NMT	NO	43.78	43.78	43.78	0.857054	0.857054	0.857054	0.578009

Table 54: JPCN ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NLPHut	5231	NMT	NO	1.690000	0.095373	0.385495

Table 55: MMCHHI21 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NLPHut	5439	OTHER	NO	0.990000	0.024940	0.383880

Table 56: MMCHHI21 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
Volta	6430	NMT	YES	51.600000	0.859645	0.877000
CNLP-NITS-PP	5730	NMT	YES	39.280000	0.792097	0.817356
iitp	5942	NMT	NO	37.500000	0.790809	0.823429

Table 57: MMCHMM21 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
Volta	6429	NMT	YES	51.660000	0.855410	0.876300
NLPHut	4623	NMT	YES	43.290000	0.824521	0.841544
CNLP-NITS-PP	5732	NMT	YES	37.160000	0.770621	0.797409

Table 58: MMCHTEXT21 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6146	NMT	NO	12.980000	0.378045	0.603143
NLPHut	4621	NMT	NO	12.150000	0.373986	0.649550

Table 59: MMCHTEXT21 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NLPHut	5400	OTHER	NO	1.300000	0.093243	0.333490

Table 60: MMEVHI21 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NLPHut	5438	OTHER	NO	0.970000	0.047566	0.405275

Table 61: MMEVHI21 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
Volta	6428	NMT	YES	44.640000	0.823319	0.839100
iitp	5941	NMT	NO	42.470000	0.807123	0.629444
CNLP-NITS-PP	5731	NMT	YES	39.460000	0.802055	0.641430

Table 62: MMEVMM21 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
Volta	6427	NMT	YES	44.120000	0.821469	0.838180
NLPHut	4622	NMT	YES	42.110000	0.813837	0.634481
CNLP-NITS-PP	5733	NMT	YES	37.010000	0.795302	0.642785

Table 63: MMEVTEXT21 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6145	NMT	NO	30.490000	0.580807	0.726976
NLPHut	4620	NMT	NO	34.830000	0.636404	0.798859

Table 64: MMEVTEXT21 en-ml submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
TMEKU	4730	NMT	NO	44.95	53.50	48.57	0.886046	0.890507	0.886316	0.644124
TMEKU	5452	NMT	NO	43.40	51.81	47.02	0.874392	0.880350	0.874700	0.644113
sakura	6313	NMT	NO	43.09	51.17	46.32	0.875110	0.879799	0.875825	0.644507

Table 65: MMT en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
sakura	6349	NMT	NO	52.20	0.909991	0.577316

Table 66: MMT ja-en submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	4403	NMT	NO	22.65	28.16	24.75	0.781797	0.784997	0.778157	0.790050
ORGANIZER	4423	NMT	NO	25.73	31.46	27.77	0.804437	0.806973	0.801715	0.806910
TMEKU	4731	NMT	NO	28.79	34.33	31.04	0.809852	0.813066	0.810293	0.821745
TMEKU	5451	NMT	NO	28.23	33.71	30.23	0.806312	0.808009	0.800428	0.815016

Table 67: MSCOCO en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	4404	NMT	NO	30.04	0.800134	0.757189
ORGANIZER	4422	NMT	NO	30.70	0.798426	0.755753

Table 68: MSCOCO ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5900	NMT	YES	29.050000	0.651775	0.821077
sakura	5792	NMT	NO	28.500000	0.663932	0.826771

Table 69: SOFTWARE en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5902	NMT	YES	43.25	0.767124	0.863589
sakura	5799	NMT	NO	45.39	0.759304	0.863010

Table 70: SOFTWARE en-id submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5904	NMT	YES	40.76	0.823552	0.866766
sakura	5818	NMT	NO	42.26	0.838933	0.873296

Table 71: SOFTWARE en-ms submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5906	NMT	YES	50.910000	0.770522	0.809907
sakura	5844	NMT	NO	55.640000	0.813347	0.829860

Table 72: SOFTWARE en-th submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5901	NMT	YES	35.32	0.712675	0.843388
sakura	5795	NMT	NO	40.17	0.726708	0.861348

Table 73: SOFTWARE hi-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5903	NMT	YES	40.69	0.745225	0.852173
sakura	5810	NMT	NO	44.70	0.759751	0.862999

Table 74: SOFTWARE id-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5905	NMT	YES	38.42	0.818175	0.843418
sakura	5823	NMT	NO	40.97	0.819980	0.849354

Table 75: SOFTWARE ms-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NICT-2	5907	NMT	YES	21.89	0.673464	0.787909
sakura	5846	NMT	NO	26.30	0.694253	0.809105

Table 76: SOFTWARE th-en submissions