

Overview of the 9th Workshop on Asian Translation

Toshiaki Nakazawa

The University of Tokyo

nakazawa@nlab.ci.i.u-tokyo.ac.jp

Hideya Mino and Isao Goto

NHK

{mino.h-gq, goto.i-es}@nhk.or.jp

Raj Dabre and Shohei Higashiyama

National Institute of

Information and Communications Technology

{raj.dabre, shohei.higashiyama}@nict.go.jp

Shantipriya Parida

Silo AI

shantipriya.parida@silo.ai

Anoop Kunchukuttan

Microsoft AI and Research

anoop.kunchukuttan@microsoft.com

Makoto Morishita

NTT Communication Science Laboratories

makoto.morishita.gr@hco.ntt.co.jp

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@tohoku.ac.jp

Yusuke Oda

Inspired Cognition, Tohoku University

odashi@inspiredco.ai

Sadao Kurohashi

Kyoto University

kuro@i.kyoto-u.ac.jp

Abstract

This paper presents the results of the shared tasks from the 9th workshop on Asian translation (WAT2022). For the WAT2022, 8 teams submitted their translation results for the human evaluation. We also accepted 4 research papers. About 300 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-WAT2021 (Nakazawa et al., 2021c), WAT2022 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 9th WAT, we included the following new tasks/languages:

- Structured Document Translation Task: English \leftrightarrow Japanese, Chinese and Korean translation.

- Video Guided Ambiguous Subtitling Task: Japanese \rightarrow English video guided translation for ambiguous subtitles.
- Khmer Speech Translation Task: Low-resource Khmer \rightarrow English/French speech translation.
- Two new translation tasks to the Restricted Translation task: Chinese \leftrightarrow Japanese.
- Parallel Corpus Filtering: Japanese \leftrightarrow English parallel corpus filtering.
- Bengali Visual Genome Task: English \rightarrow Bengali multi-modal translation has been added, similar to the recurring Hindi and Malayalam multi-modal translation tasks.
- 5 new languages to the Multilingual Indic Machine Translation Task (MultiIndicMT): Assamese, Sindhi, Sinhala, Nepali and Urdu.
- 1 new language to the Wikinews and Software Documentation Translation Task (NICT-SAP): Vietnamese.

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.

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 ParaNatCom¹ (Parallel English-Japanese abstract corpus made from Nature Communications articles) for the development and test sets of the Document-level Scientific Paper Translation sub-task. We cannot provide document-level training corpus, but you can use ASPEC and any other extra resources.

2.2 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

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

Lang	Train	Dev	DevTest	Test-2022
zh-ja	1,000,000	2,000	2,000	10,204
ko-ja	1,000,000	2,000	2,000	7,230
en-ja	1,000,000	2,000	2,000	10,668

Lang	Test-N1	Test-N2	Test-N3	Test-N4
zh-ja	2,000	3,000	204	5,000
ko-ja	2,000	–	230	5,000
en-ja	2,000	3,000	668	5,000

Table 1: Statistics for JPC

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 Corpus² (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 collaboration with NICT. The corpus consists of Chinese-Japanese, Korean-Japanese, and English-Japanese parallel sentences of patent descriptions. Most sentences were extracted from documents with one of four International Patent Classification (IPC) sections: chemistry, electricity, mechanical engineering, and physics. As shown in Table 1, each parallel corpus consists of training, development, development-test, and three or four test datasets, including two test datasets introduced at WAT2022: test-2022 and test-N4. The test datasets have the following characteristics:

- test-2022: the 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;
- test-N3: patent documents published between 2016 and 2017 with manually translated target sentences; and
- test-N4: patent documents from patent families published between 2019 and 2020.

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

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

2.4 Newswire (JIJI) Task

The Japanese \leftrightarrow English newswire task uses JIJI Corpus which was constructed by Jiji Press Ltd. in collaboration with NICT and NHK. The corpus consists of news text that comes from Jiji Press news of various categories including politics, economy, nation, business, markets, sports and so on. The corpus is partitioned into training, development, development-test and test data, which consists of Japanese-English sentence pairs. In addition to the test set (test set I) that has been provided from WAT 2017, 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 OPUS³, other WAT tasks (past and

³<http://opus.nlpl.eu/>

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-way parallel.

present) and WMT⁴. 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) corpora⁵ (Buschbeck and Exel, 2020) and encouraged participants to consider GNOME, UBUNTU and KDE corpora from OPUS. We 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 list⁶ all available corpora here and participants were not restricted from using any corpora as long as they are freely available.

2.7 Structured Document Translation Task

For the first time we introduce a structured document translation task for English \leftrightarrow Japanese, Chinese and Korean translation. The goal is to translate sentences with XML annotations in them. The key challenge is to accurately transfer the XML annotations from the marked source language words/phrases to their translations in the target language. The evaluation dataset for this task was created by SAP and is an extension of the software documentation dataset, which is used for the **NICT-SAP task**. It consists of 2,011 and 2,002 segments in the development and test sets respectively. Note that the dataset also comes with its XML stripped equivalent and can be used to evaluate English \leftrightarrow Japanese, Chinese and Korean translation for the software documentation domain. Given that there is no training data available for this task, it becomes more challenging.

2.8 Indic Multilingual Task (MultiIndicMT)

Owing to the increasing interest in Indian language translation and the success of the multilingual Indian languages tasks in 2018 (Nakazawa et al., 2018), 2020 (Nakazawa et al., 2020) and 2021 (Nakazawa et al., 2021b), we decided to enlarge the scope of the 2021 task by adding 5 new languages to the MultiIndicMT task, namely, Assamese (As), Urdu (Ur), Sindhi (Si), Sinhala (Sd) and Nepali (Ne). In addition to the original 10 Indic languages, alongside English (En), namely, Hindi (Hi), Marathi (Mr), Kannada (Kn), Tamil

⁴<http://www.statmt.org/wmt20/>

⁵Software Domain Evaluation Splits

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

(Ta), Telugu (Te), Gujarati (Gu), Malayalam (MI), Bengali (Bn), Oriya (Or) and Punjabi (Pa), we have a total of 15 Indic languages being evaluated this year. We used the FLORES-101 dataset’s⁷ dev and devtest sets for development and testing both containing roughly 1000 sentences each per language. FLORES-101 is N-way parallel which ensures Indic to Indic translation evaluation although we did not consider it this year.

The objective of this task, like the Indic languages tasks in 2018, 2020, and 2021, is to evaluate the performance of multilingual NMT models for English to Indic and Indic to English translation. The desired solution could be one-to-many, many-to-one or many-to-many NMT models. In general, we encouraged participants to focus on multilingual NMT (Dabre et al., 2020) solutions. For training, we encouraged the use of the Samanantar corpus (Ramesh et al., 2022) which covers 11 of the 15 Indic languages. For other languages, we asked users to use the corpora from Opus, specifically the Paracrawl datasets⁸ for Nepali and Sinhala. We also listed additional sources of monolingual corpora for participants to use.

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, 2021a).

For English→Hindi multi-modal translation task, we asked the participants to use Hindi Visual Genome 1.1 corpus (HVG, Parida et al., 2019a,b).⁹

The statistics of HVG 1.1 are given in Table 6. 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

⁷<https://github.com/facebookresearch/flores>

⁸<https://opus.nlpl.eu/ParaCrawl.php>

⁹<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 6: 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.

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.

2.10 English→Malayalam Multi-Modal Task

This task was introduced in WAT2021 using the first multi-modal 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).¹⁰

The statistics of MVG are given in Table 7. As in Hindi Visual Genome (see Section 2.9), 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.

¹⁰<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533>



	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 कारें	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 2022 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.

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 7: 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.

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 En-

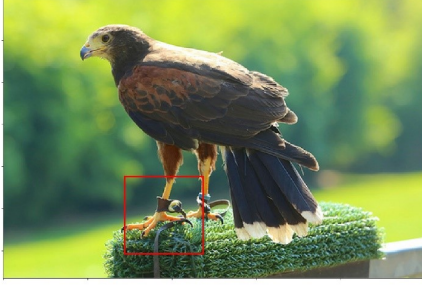
glish text into Malayalam. Both textual and visual information can be used.

2.11 English→Bengali Multi-Modal Task

This new task, introduced in WAT2022, uses a multimodal machine translation dataset in *Bengali* language. The task mimics the structure of English→Hindi (Section 2.9) and English→Malayalam (Section 2.10) multi-modal tasks. For English→Bengali multi-modal translation task we asked the participants to use the Bengali Visual Genome corpus (BVG for short, [Sen et al., 2022](#)).¹¹

The statistics of BVG are given in Table 8. One “item” in BVG again consists of an image with a rectangular region highlighting a part of the image, the original English caption of this region and the

¹¹<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3722>



English Text: The sharp bird talon.

Bengali Text: ধারালো পাখি টাল

Figure 3: Sample item from Bengali Visual Genome (BVG), Image with specific region and its description.

Bengali reference translation as shown in Figure 3. Depending on the track (see Section 2.11.1 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.

2.11.1 English→Bengali 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 Bengali. 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. Bengali Captioning (labeled “BN”): The participants are asked to generate captions in Bengali 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 Bengali. Both textual and visual information can be used.

2.12 Ambiguous MS COCO Japanese↔English Multimodal Task

This is the 2nd year that we have organized this task. We provide the Japanese–English Ambiguous MS COCO dataset (Merritt et al., 2020) for validation and testing, which contains ambiguous

Dataset	Items	Tokens	
		English	Bengali
Training Set	28,930	143,115	113,978
D-Test	998	4,922	3,936
E-Test (EV)	1,595	7,853	6,408
C-Test (CH)	1,400	8,186	6,657

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

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.¹²

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) dataset¹³ can be used as training data. In the unconstrained setting, the MS COCO English data¹⁴ and STAIR Japanese image captions¹⁵ 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 Japanese→English Video Guided MT Task for Ambiguous Subtitles

This is a new Japanese→English multimodal task. We provide VISA (Li et al., 2022), an ambiguous subtitles dataset, including 35, 880, 2, 000, and 2, 000 samples for training, validation, and testing, respectively. The dataset contains parallel subtitles in which the Japanese source subtitles are ambiguous and may require visual information in corresponding video clips for disambiguation. Furthermore, according to the cause of ambiguity, the dataset is divided into Polysemy and Omission.

Participants can use the constrained and unconstrained training data to train their multimodal machine translation system. In the constrained setting,

¹²<http://www.statmt.org/wmt17/multimodal-task.html>

¹³<https://github.com/nlab-mpg/Flickr30kEnt-JP>

¹⁴<https://cocodataset.org/#captions-2015>

¹⁵<https://stair-lab-cit.github.io/STAIR-captions-web/>

only the VISA dataset¹⁶ can be used as training data. In the unconstrained setting, pre-trained models, additional data from other sources can be used as additional training sources.

We prepare a baseline using the spatial hierarchical attention network following (Gu et al., 2021) with both motion and spatial features.

2.14 Low-Resource Khmer→English/French Speech Translation Task

This is the first time that WAT has hosted a speech translation task. The purpose of this task is to identify effective techniques for speech translation of Khmer into English and French. We expect that the low-resource nature of Khmer will pose a reasonable challenge. To this end, we have curated a dataset from the ECCC corpus (Soky et al., 2021), which is an international court dataset consisting of text and speech in Khmer, English, and French. The dataset used for WAT 2022 contains 11,563, 624, and 626 utterances for training, validation, and testing, respectively. This dataset has a wide range of speakers: witnesses, defendants, judges, clerks or officers, co-prosecutors, experts, defense counsels, civil parties, and interpreters.

Participants can use the constrained and unconstrained training data to train their speech machine translation system. In the constrained setting, only the provided ECCC dataset¹⁷ can be used as training data. Additionally, participants may use pre-trained models such as BART, mBART, mT5, and wav2vec 2.0 as applicable. In the unconstrained setting, additional data from other sources can also be used.

We prepare a baseline using the transformer-based model presented in (Soky et al., 2021) for both Khmer→English and Khmer→French directions.

2.15 Restricted Translation Task

The Restricted Translation task was first introduced in WAT2021 (Nakazawa et al., 2021c). In this task, participants are required to submit a system that translates source texts under given constraints about the target vocabulary. At inference time, vocabulary constraints are provided as a list of target words and phrases, consisting of scientific technical terms in the target language. The system

outputs must contain all these target words. We introduced English↔Japanese tasks in the previous campaign, and we also added Chinese↔Japanese tasks this year. We employ the ASPEC corpus for all the translation tasks and allow participants to use any other external data sources.

Restricted Vocabulary List Creation We built a new vocabulary constraints for the Chinese↔Japanese tasks by extracting phrase pairs from the evaluation data (“dev/devtest/test”) by the following steps: (1) extracting term candidates for each language, and (2) making the alignments between the extracted terms in both languages to make phrase-level translation pairs. More concretely, we automatically extracted the term candidates from the ASPEC corpus using `termextractor`¹⁸. We then obtained term lists for each sentence pair in the ASPEC corpus according to the extracted term candidates. To this end, we asked one Japanese-Chinese bilingual speaker to make alignments between the term lists for each sentence pair, and obtained the phrase pair lists. We conducted the source-based direct assessment (Cettolo et al., 2017; Federmann, 2018) on the dictionaries created by the process above. We employed another two bilingual annotators to give translation scores ranging [0, 100] for Chinese→Japanese and Japanese→Chinese directions respectively. We then filtered out the translation pairs with average scores less than 50. Thus, we publicized the restricted vocabulary lists for each language direction, along with the corresponding source-side terms and annotation scores¹⁹. Table 9 reports the statistics of the vocabulary constraints in the evaluation data for English↔Japanese and Chinese↔Japanese tasks.

Evaluation Metrics We evaluate submitted 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 match-

¹⁶<https://github.com/ku-nlp/VISA>

¹⁷https://github.com/ksoky/ECCC_DATASET

¹⁸We used `termex_janome.py` and `termex_nlpir.py` for Japanese and Chinese texts, respectively. <http://genssen.dl.itc.u-tokyo.ac.jp/pytermextract/>

¹⁹All scores are publicly available at the task page: <https://sites.google.com/view/restricted-translation-task/2022>.

	En-Ja (# phrase, # char)	Ja-En (# phrase, # word)	Zh-Ja (# phrase, # char)	Ja-Zh (# phrase, # char)
Dev.	(2.8, 16.4)	(2.8, 6.6)	(1.2, 4.7)	(1.2, 3.8)
Devtest	(3.2, 18.2)	(3.2, 7.3)	(1.5, 5.5)	(1.5, 4.5)
Test	(3.3, 18.1)	(3.2, 7.4)	(1.4, 5.2)	(1.4, 4.2)

Table 9: Statistics of the restricted vocabulary in the evaluation data. We report average number of phrases and characters/words per source sentence.

ing (including whitespaces) for each constraint. In Chinese and Japanese, we judge character-level sequence matching (including whitespaces) for each constraint without any preprocessing. For the final ranking, we also calculate the combined score of both: calculating BLEU with only the exactly matched sentences. We note that, in this scenario, the brevity score in BLEU does not carry its usual meaning, but the n -gram scores will maintain their consistency.

2.16 Parallel Corpus Filtering Task

Machine translation systems are trained from usually large corpora obtained from noisy data sources. Noisy examples in the training corpora are known as the main cause of reducing the translation accuracy of the resulting models (Khayrallah and Koehn, 2018), and this problem can be mitigated by corpus filtering (Koehn et al., 2020), which removes problematic examples from the training corpus, so that the model is eventually trained by cleaner dataset than the data source.

The motivation for this task is inspired by the Parallel Corpus Filtering Tasks held in 2018, 2019, and 2020 Workshop on Machine Translation (Koehn et al., 2020), in which the participants are asked to filter the web crawled corpora, train the NMT model on the cleaner subsets, and evaluate its quality on a multi-domain test set. Unlike the tasks in the WMT, the Parallel Corpus Filtering Task in this workshop focuses on both filtering and domain adaptation.

Specifically, this task lets the participants train machine translation models under the following restrictions:

- The model architecture is fixed. The training program is provided as a fixed Docker image by the organizer, and participants can only run a specific training command to build their own model. The same image is used in the final evaluation.

Dataset	# sentences
JParaCrawl v3.0	25.7M
ASPEC Train	3M
ASPEC Dev	1.8K
ASPEC Devtest	1.8K
ASPEC Test	1.8K

Table 10: Number of sentence pairs in the corpora used in the parallel corpus filtering task.

- Training corpus is fixed. The whole corpus is provided by the organizer, and participants are requested to find a subset of the corpus that is more effective in achieving higher translation accuracy on the given model architecture.
- The test set is from a single domain (scientific paper domain) and its in-domain data is provided.

We adopted the Transformer model as the shared architecture for this task.²⁰

We asked the participants to select a subset from JParaCrawl (Morishita et al., 2020), the noisy English-Japanese web-crawled parallel corpus, based on its cleanliness and domain-similarity. The baseline model is obtained by training the model on the whole set of this dataset. We also provide the in-domain clean English-Japanese corpus, the ASPEC (Nakazawa et al., 2016) dataset except for the ‘test’ sub-set, which is used in the evaluation.

We trained the model with the submitted data for both English-Japanese and English-Japanese. We evaluated the submission on both BLEU score (Papineni et al., 2002) and JPO adequacy as described in Section 6.1 on the ASPEC test set.

The corpus statistics are summarized in Table 10.

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NICT-5	NICT	Japan
sakura	Rakuten Institute of Technology Singapore, Rakuten Asia.	Singapore
CNLP-NITS-PP	NIT Silchar	India
NITR	NIT Rourkela	India
HwTscSU	Huawei Translation Services Center, 2012 Lab, Huawei co. LTD; School of Computer Science and Technology, Soochow University	China
SILO_NLP	Silo AI	Finland
nlp_novices	SCTR's Pune Institute of Computer Technology	India

Table 11: List of participants who submitted translations for the human evaluation in WAT2022

	ASPEC		NICT-SAP				Parallel
	Restricted		Unstructured		Structured		Corpus
Team ID	En-Ja	Ja-En	En-Ms (IT)	Ms-En (IT)	En-Ja/Ko/Zh	Ja/Ko/Zh-En	Filtering
TMU	✓	✓			✓	✓	
NICT-5							
sakura							✓
HwTscSU			✓	✓			

	Multimodal			Indic										
	En-X (TX)			En-X					X-En					
Team ID	Hi	MI	Bn	As	Bn	Sd	Si	Ur	Ne	As	Bn	Sd	Si	Ur
CNLP-NITS-PP	✓		✓	✓	✓				✓	✓	✓			
NITR				✓		✓	✓	✓	✓			✓	✓	✓
SILO_NLP	✓	✓	✓											
nlp_novices	✓	✓	✓											

Table 12: Submissions for each task by each team.

3 Participants

Table 11 shows the participants in WAT2022. The table lists 8 organizations from various countries, including Japan, China, India, Singapore and Finland.

300 translation results by 8 teams were submitted for automatic evaluation. Table 12 summarizes the participation of teams across WAT2022 tasks and indicates which tasks included manual evaluation.

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 2022, 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. We also have SMT baseline systems for the tasks that started at WAT 2017 or before 2017. 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 7.0²¹ for Japanese segmentation.

²⁰The Dockerfile for constructing the training pipeline can be obtained from <https://github.com/MorinoseMorizo/wat2022-filtering>

²¹<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN>

- Stanford Word Segmenter version 2014-01-04²² (Chinese Penn Treebank (CTB) model) for Chinese segmentation.
- The Moses toolkit for English and Indonesian tokenization.
- Mecab-ko²³ for Korean segmentation.
- Indic NLP Library²⁴ (Kunchukuttan, 2020) for Indic language segmentation.
- The tools included in the ALT corpus for Myanmar and Khmer segmentation.
- subword-nmt²⁵ 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 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.

4.1.3 For Structured Document Translation Task

- No tokenization was explicitly performed.

4.1.4 For English→Hindi, English→Malayalam, and English→Bengali Multi-Modal Tasks

- Hindi Visual Genome 1.1, Malayalam Visual Genome, and Bengali Visual Genome come 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.

²²<http://nlp.stanford.edu/software/segmenter.shtml>

²³<https://bitbucket.org/eunjeon/mecab-ko/>

²⁴https://github.com/anoopkunchukuttan/indic_nlp_library

²⁵<https://github.com/rsennrich/subword-nmt>

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, JIJ, and ALT tasks, tensor2tensor²⁶ for the NICT-SAP task, HuggingFace transformers²⁷ for the Structured Document Translation task and OpenNMT-py²⁸ for other tasks.

4.2.1 NMT with Attention (OpenNMT)

For ASPEC, JPC, JIJ, 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.

²⁶<https://github.com/tensorflow/tensor2tensor>

²⁷<https://github.com/huggingface/transformers>

²⁸<https://github.com/OpenNMT/OpenNMT-py>

4.2.2 Transformer (Tensor2Tensor)

For the News Commentary task, we used tensor2tensor’s²⁹ 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 trained models for all languages except Vietnamese. 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.

4.2.3 Transformer (HuggingFace)

For the Structured Document Translation task, we used the official mbart-50 model fine-tuned³⁰ for machine translation to directly translate the test sets. We used the HuggingFace transformers implementation to decode sentences using a beam of size 4 and length penalty of 1.0. The tokenization was handled by the mbart-50 tokenizer. Surprisingly, this naive approach actually yielded good results.

4.2.4 Transformer (OpenNMT-py)

For the English→Hindi, English→Malayalam, and English→Bengali Multimodal tasks we used the Transformer model (Vaswani et al., 2018) as implemented in OpenNMT-py (Klein et al., 2017)

²⁹<https://github.com/tensorflow/tensor2tensor>

³⁰<https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt>

and used the “base” model with default parameters for the multi-modal task baseline. We have generated the vocabulary of 32k sub-word types jointly for both the source and target languages. The vocabulary is shared between the encoder and decoder.

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 SacreBLEU (Post, 2018). RIBES scores were calculated using RIBES.py version 1.02.4.³¹ AMFM scores were calculated using scripts created by the technical collaborators listed in the WAT2022 web page.³² Note that AMFM scores were not produced for all tasks. For the Structured Document Translation task, we used only the XML-BLEU metric (Hashimoto et al., 2019), which takes into account the accuracy of XML annotation transfer. All scores for each task were calculated using the corresponding reference translations.

Except for XML-BLEU, which uses [this implementation](#) for evaluation, the following preprocessing is done prior to computing scores. Before the calculation of the automatic evaluation scores, the translation results were tokenized or segmented with tokenization/segmentation tools for each language. For Japanese segmentation, we used three different tools: Juman version 7.0 (Kurohashi et al., 1994), KyTea 0.4.6 (Neubig et al., 2011) with full SVM model³³ and MeCab 0.996 (Kudo, 2005) with IPA dictionary 2.7.0.³⁴ For Chinese segmentation, we used two different tools: KyTea 0.4.6 with full SVM Model in MSR model and Stanford Word Segmenter (Tseng, 2005) version 2014-06-16 with Chinese Penn Treebank (CTB) and Peking University (PKU) model.³⁵ For Korean segmentation, we used mecab-ko.³⁶ For Myanmar and Khmer segmentations, we used myseg.py³⁷ and

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

³²<http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2022/>

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

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

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

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

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

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.
 - JIJEn-ja and JIJja-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 4: The interface for translation results submission

kmseg.py.³⁸ For English, French and Russian tokenizations, we used tokenizer.perl³⁹ in the Moses toolkit. For Indonesian, Malay, and Vietnamese tokenizations, we used tokenizer.perl

my-en-data/wat2020.my-en.zip

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

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

actually sticking to the English tokenization settings. For Thai tokenization, we segmented the text at each individual character. For Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Odia, Punjabi, Sindhi, Sinhala, Tamil, Telugu, and Urdu tokenizations, we used Indic NLP Library⁴⁰ (Kunchukuttan, 2020). The de-

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

tailed procedures for the automatic evaluation are shown on the WAT evaluation web page.⁴¹

5.2 Automatic Evaluation System

The automatic evaluation system receives translation results by participants and automatically gives evaluation scores to the uploaded results. As shown in Figure 4, 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 WAT2022 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 WAT2022 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 WAT2022. Anybody can register an account for the system by the procedures described in the application site.⁴²

5.3 A Note on AMFM Scores

Unlike previous years we do not compute AMFM scores on all tasks due to low participation this year. 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 WAT2022, we conducted *JPO adequacy evaluation* (Section 6.1).

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 13: 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.⁴³ 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 13 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 in evaluating. The percentages in each grade are rough indications for the transmission degree of the source sentence meanings. For Structured Document Translation, we instructed the evaluators to consider the XML structure accuracy between the source, the translation and the reference. The detailed criterion is described in the JPO document (in Japanese).⁴⁴

⁴¹<http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html>

⁴²<http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2022/application/index.html>

⁴³The number of systems varies depending on the subtasks.

⁴⁴http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

7 Evaluation Results

In this section, the evaluation results for WAT2022 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2022 website.⁴⁵

7.1 Official Evaluation Results

Figures 5 and 6 show those of ASPEC-RT subtasks, Figures 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17 show those of Indic Multilingual subtasks, Figures 18, 19 and 20 show those of Multimodal subtasks, Figures 21 and 22 show those of Parallel Corpus Filtering subtasks, and Figures 23, 24, 25, 26, 27, 28, 29 and 30 show those of NICT-SAP subtasks. Each figure contains the JPO adequacy evaluation result and evaluation summary of top systems. The detailed automatic evaluation results are shown in Appendix A.

8 Findings

8.1 NICT-SAP Task

This year we only had 1 submission from team “HwTscSU” who outperformed all previous years submissions. They claim to have fine-tuned on the development set, which has a high degree of similarity to the test set. This action ended up giving large improvements in translation quality over non fine-tuned baselines. The human evaluation showed that around 80% of translations had a score of 4 or 5 indicating the high translation quality.

8.2 Structured Document Translation Task

We only had 1 submission this year from team “NICT-5” who used similar ideas as our organizer baseline where they used the M2M-100 model for directly translating test sets. They also used the detag-and-project approach where they translated the sentences without XML and then inserted the XML content using word alignment. They got better scores in 3 out of 6 directions, but were not too far behind in others. The human evaluation showed that over 60% of the translations were scored 4 or 5 for English to Japanese/Korean/Chinese whereas this number increased to 80% for the reverse direction.

8.3 Indic Multilingual Task

In contrast to WAT 2021, this year we had two teams who participated in the task, namely, “NITR” and “CNLP-NITS-PP”. They did not submit results for all pairs. Human evaluation was done only for Nepali to English translation, where the human ratings were mostly low. Due to poor and inconsistent participation, it is difficult to make any further observations.

8.4 English→Hindi Multi-Modal Task

This year three teams participated in the different sub-tasks (TEXT, MM) of the English→Hindi Multi-Modal task. The WAT2022 automatic evaluation scores for the participating teams are shown in Tables 43 to 46. The team “nlp_novices” 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 *Transformer* using OPUS Corpus as an additional resource. For the multimodal sub-task (MM), we received two submissions from the teams “CNLP-NITS-PP”, and “Silo_NLP”, respectively. The team “Silo_NLP” obtained the highest BLEU score for the multimodal translation (MM) for the evaluation (E-Test) by extracting object tags from images and using fine-tuned *mBART*. They used Flickr8 as an additional resource. The team “CNLP-NITS-PP” obtained the highest BLEU score for the challenge (C-Test) test set following transliteration-based phrase pairs augmentation and visual features in training using BRNN encoder and doubly-attentive-rnn decoder.

Human evaluation was done for the challenge test set text-only translation (TEXT) as shown in Figure 19. We observe that both BLEU and RIBES can correctly predict the quality of translation as measured by the manually annotated Adequacy. “nlp_novices” is the best submission with almost 35% of sentences reaching the highest rank of “Almost all information is transmitted correctly”, see the JPO adequacy scale in Table 13 which was used in the evaluation.

8.5 English→Malayalam Multi-Modal Task

This year two teams “Silo_NLP”, and “nlp_novices” participated in the different sub-tasks (TEXT, MM) of the English→ Malayalam Multi-Modal task. The WAT2022 automatic evaluation scores are shown in the Table 47, 48,

⁴⁵<http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/>

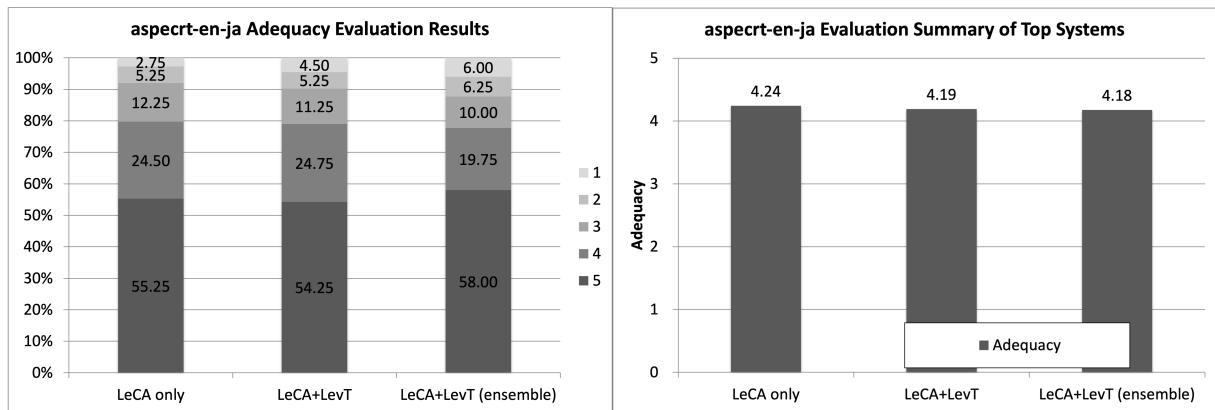


Figure 5: Official evaluation results of aspectr-en-ja.

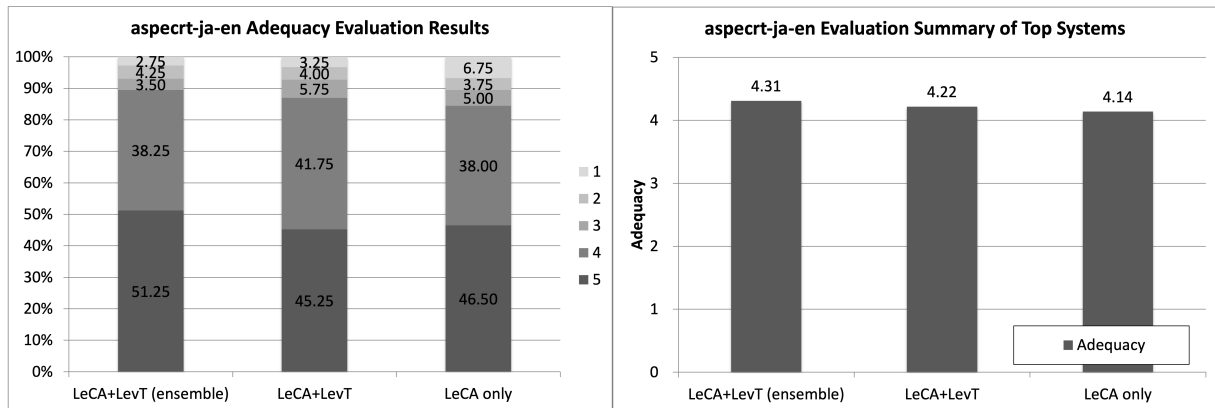


Figure 6: Official evaluation results of aspectr-ja-en.

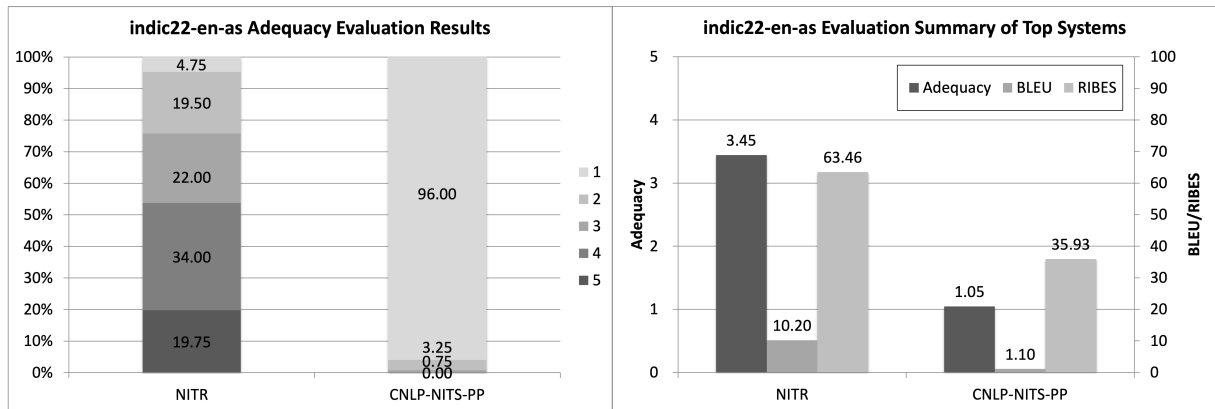


Figure 7: Official evaluation results of indic22-en-as.

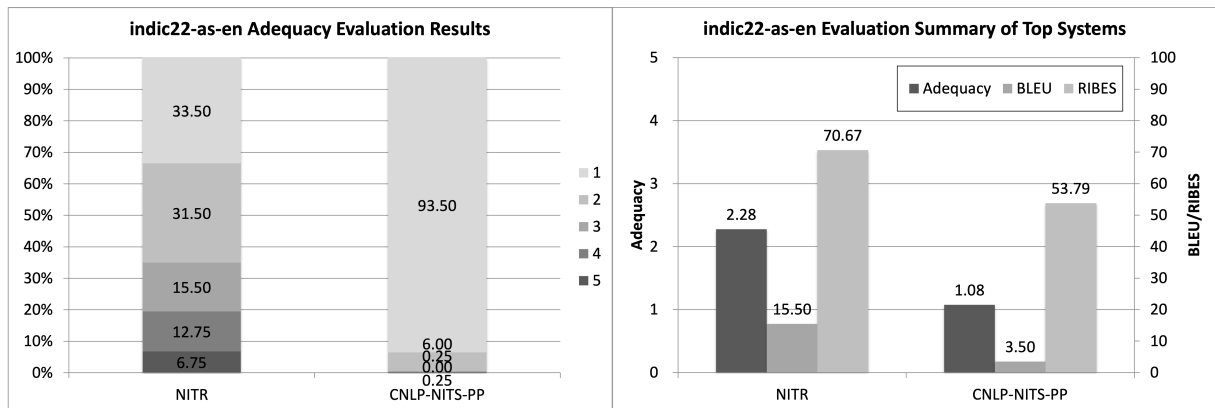


Figure 8: Official evaluation results of indic22-as-en.

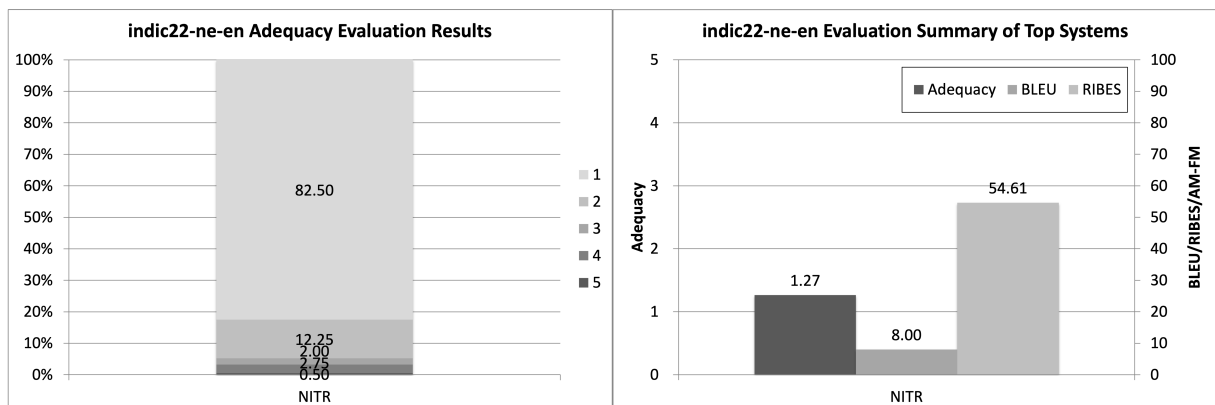


Figure 9: Official evaluation results of indic22-ne-en.

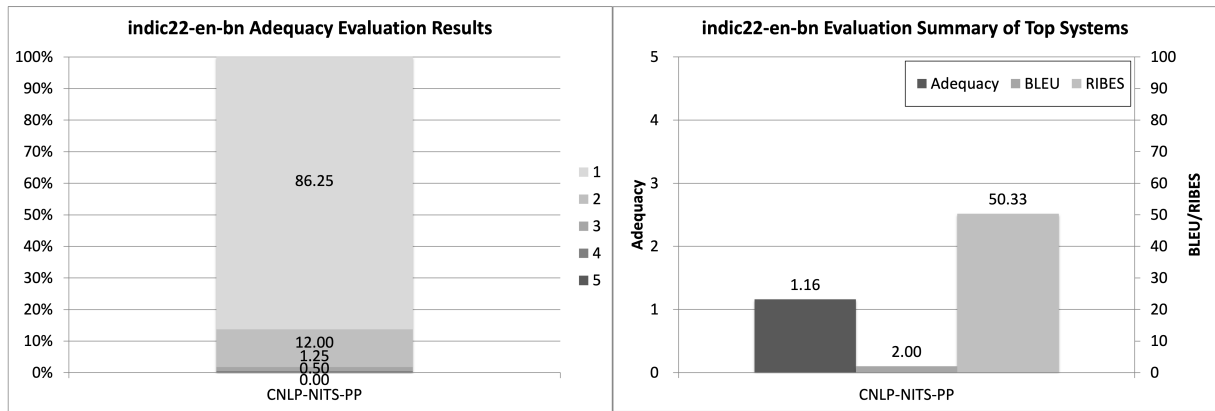


Figure 10: Official evaluation results of indic22-en-bn.

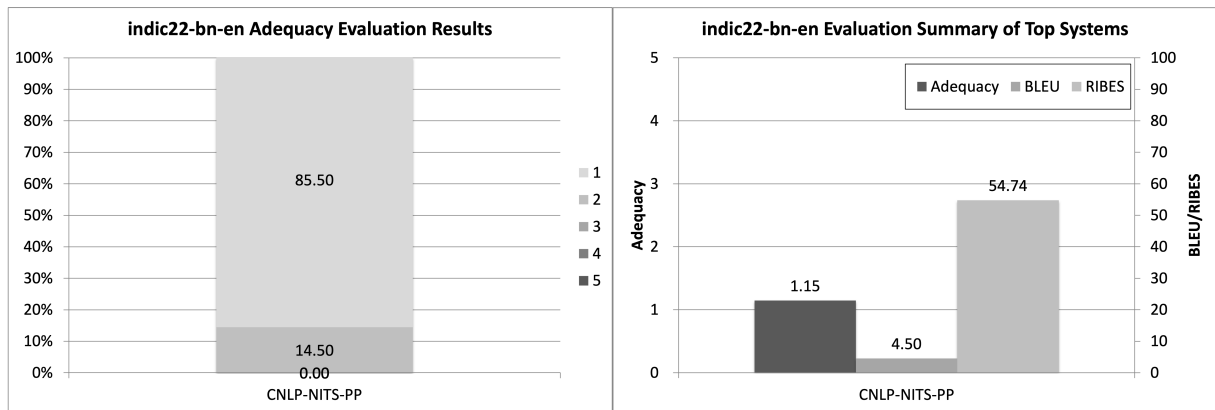


Figure 11: Official evaluation results of indic22-bn-en.

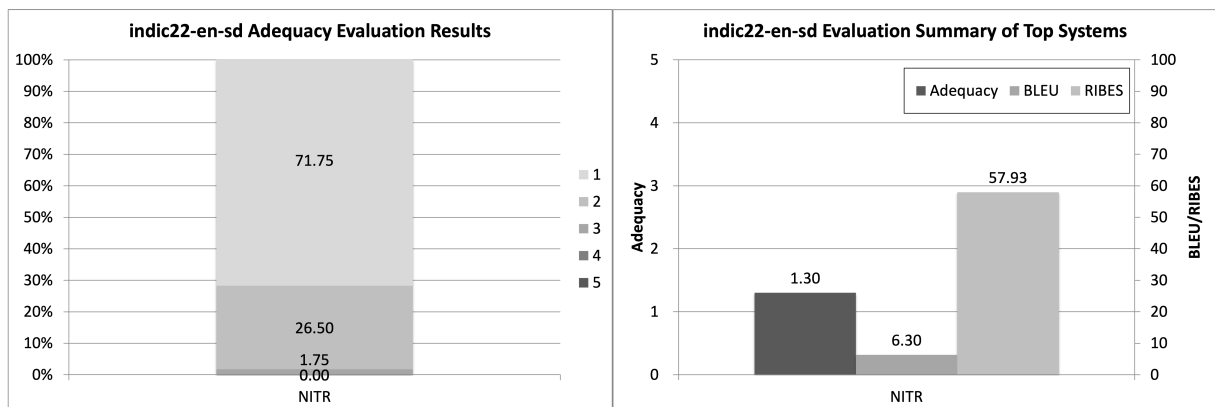


Figure 12: Official evaluation results of indic22-en-sd.

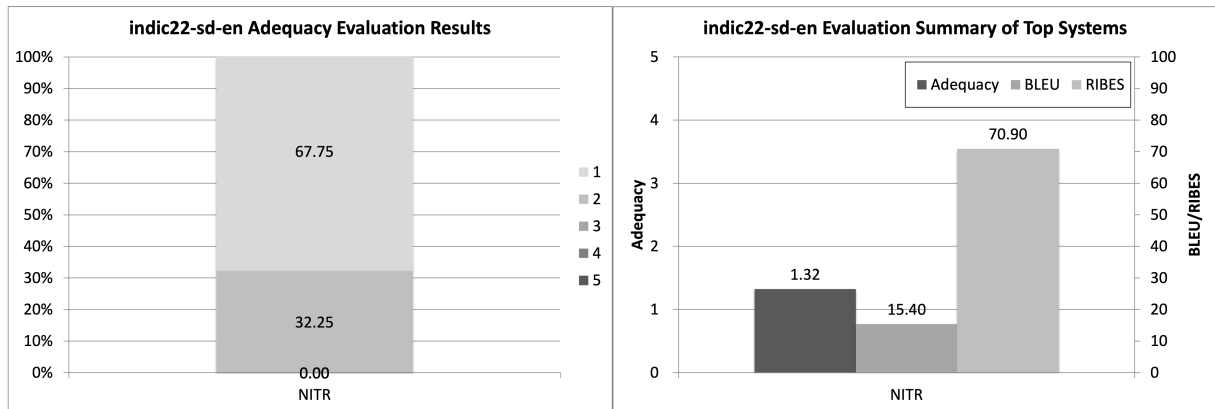


Figure 13: Official evaluation results of indic22-sd-en.

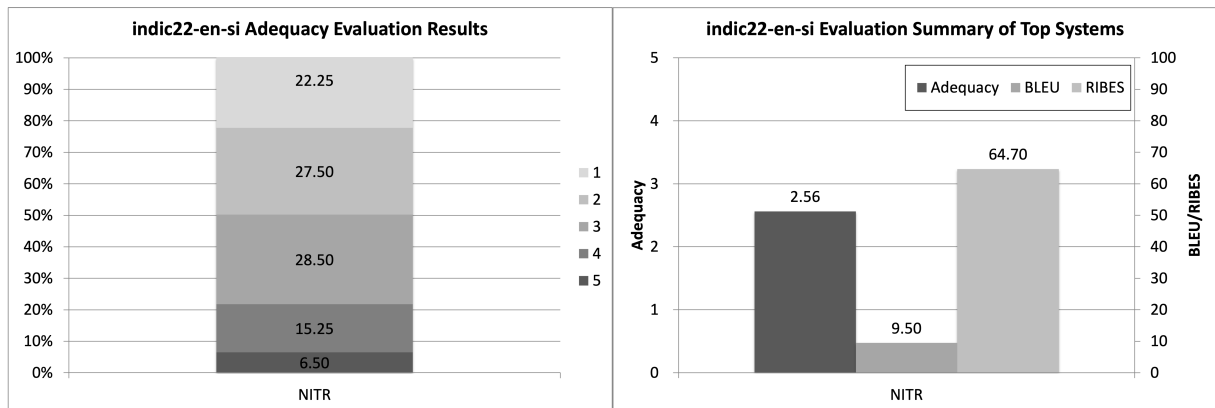


Figure 14: Official evaluation results of indic22-en-si.

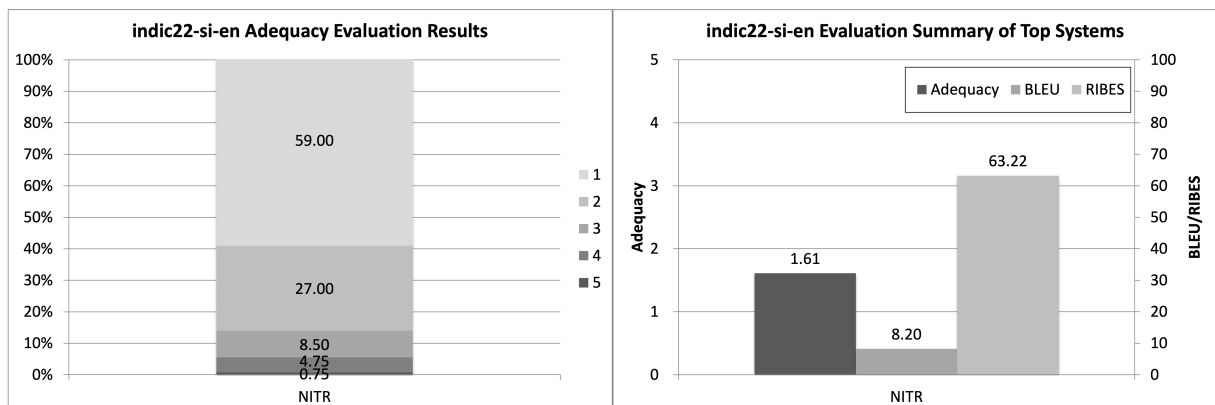


Figure 15: Official evaluation results of indic22-si-en.

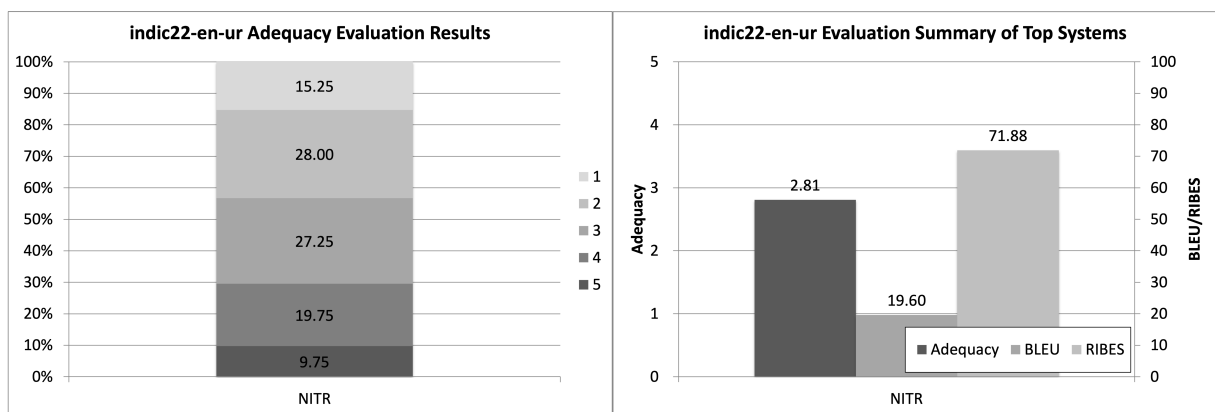


Figure 16: Official evaluation results of indic22-en-ur.

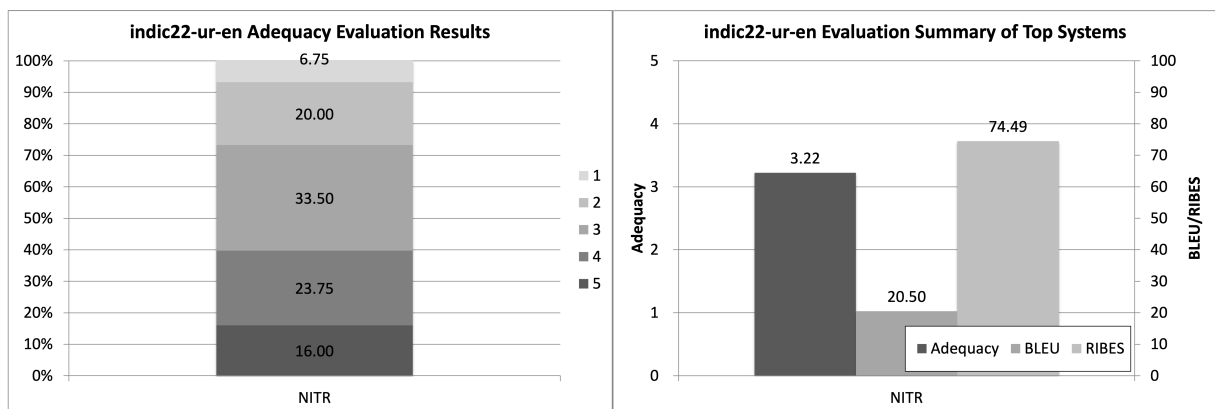


Figure 17: Official evaluation results of indic22-ur-en.

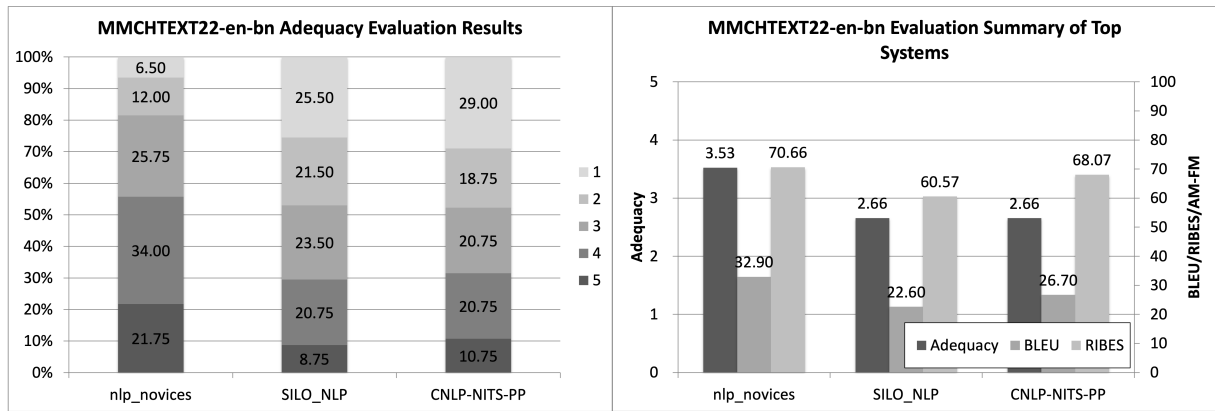


Figure 18: Official evaluation results of mmchtext22-en-bn.

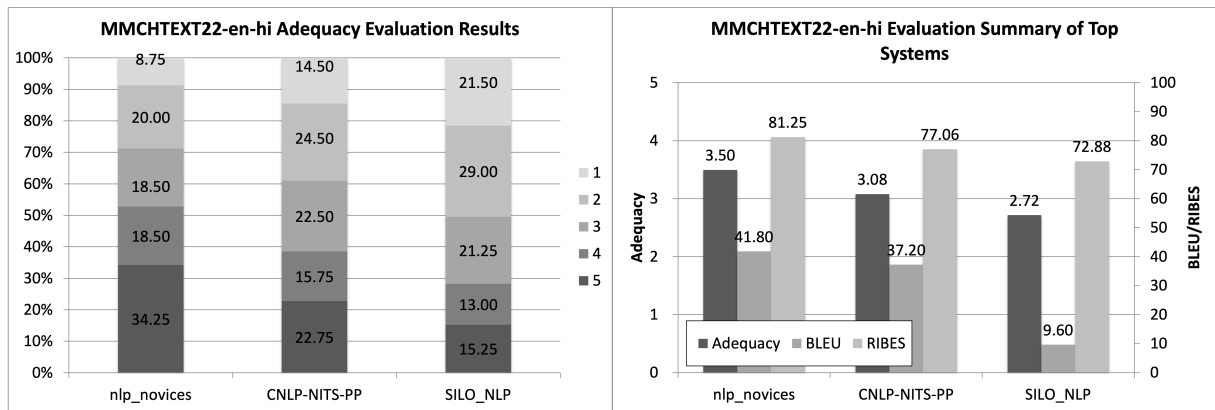


Figure 19: Official evaluation results of mmchtext22-en-hi.

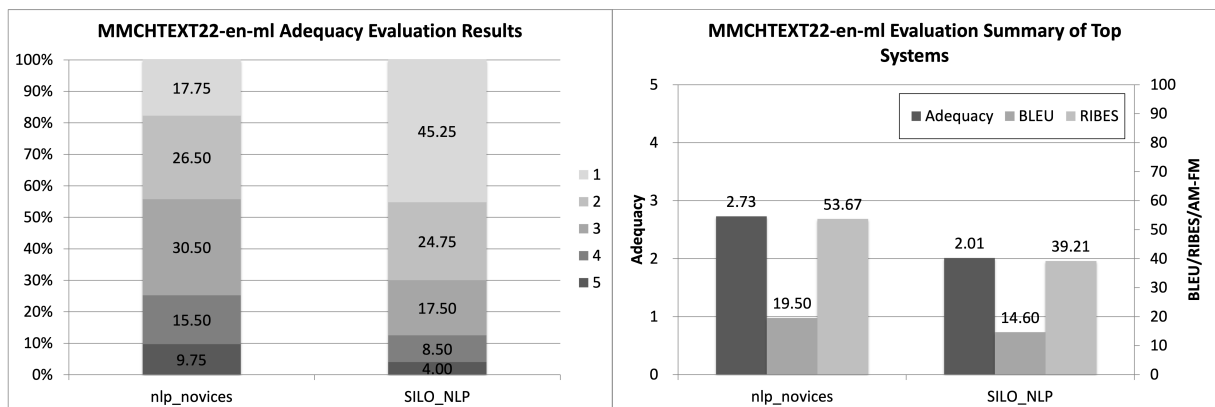


Figure 20: Official evaluation results of mmchtext22-en-ml.

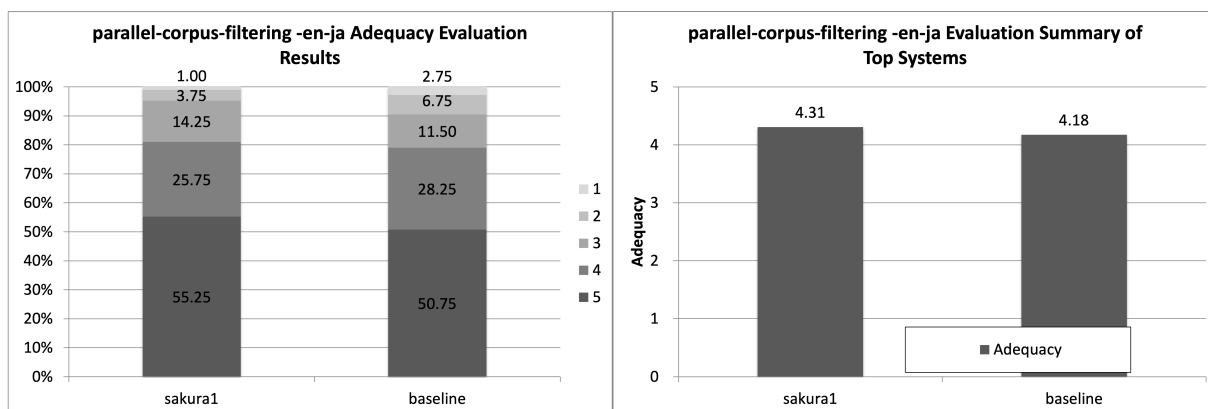


Figure 21: Official evaluation results of parallel-corpus-filtering-en-ja.

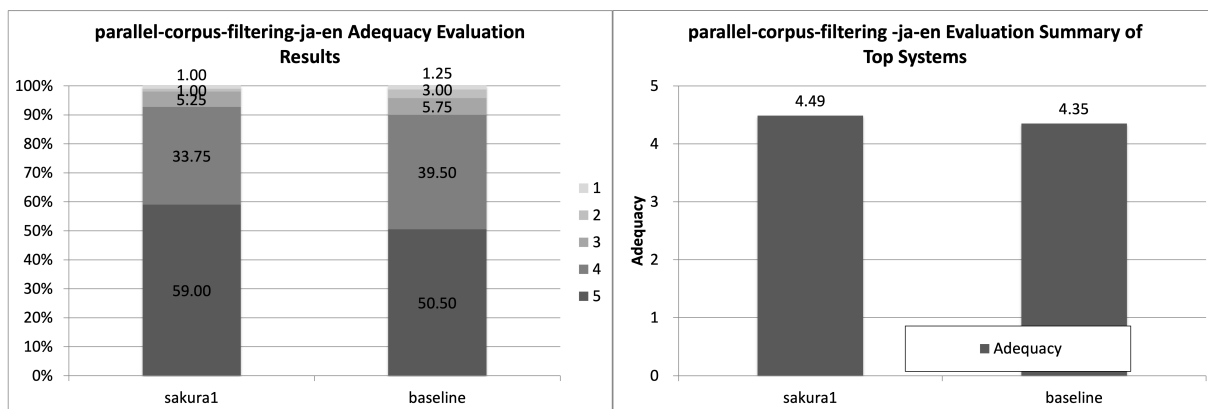


Figure 22: Official evaluation results of parallel-corpus-filtering-ja-en.

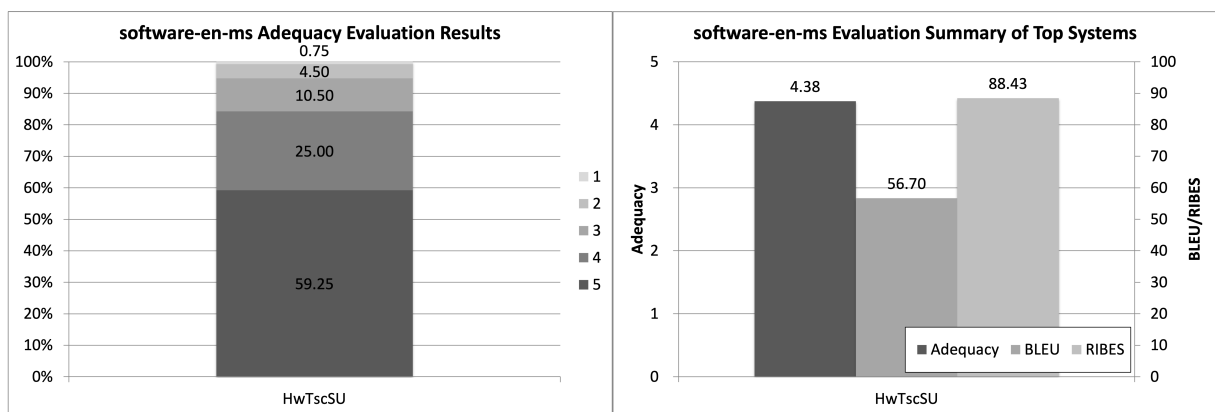


Figure 23: Official evaluation results of software-en-ms.

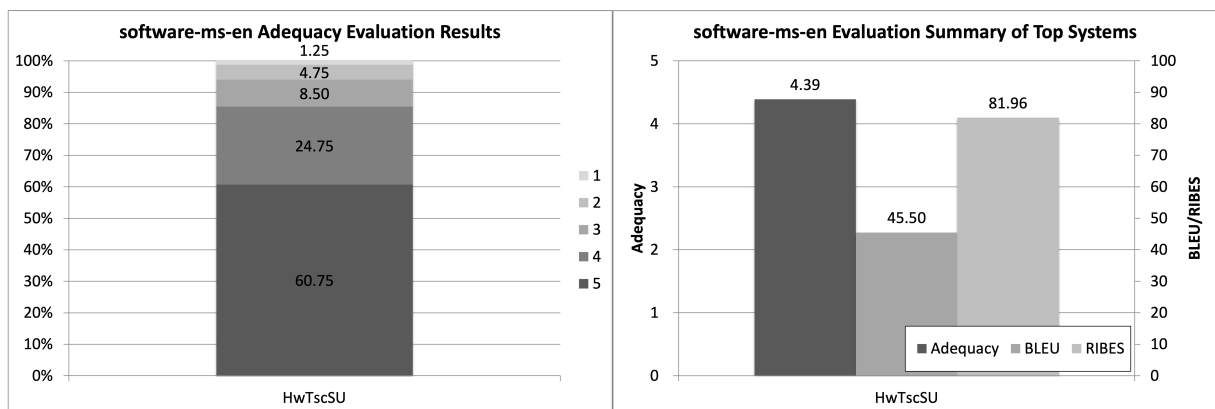


Figure 24: Official evaluation results of software-ms-en.

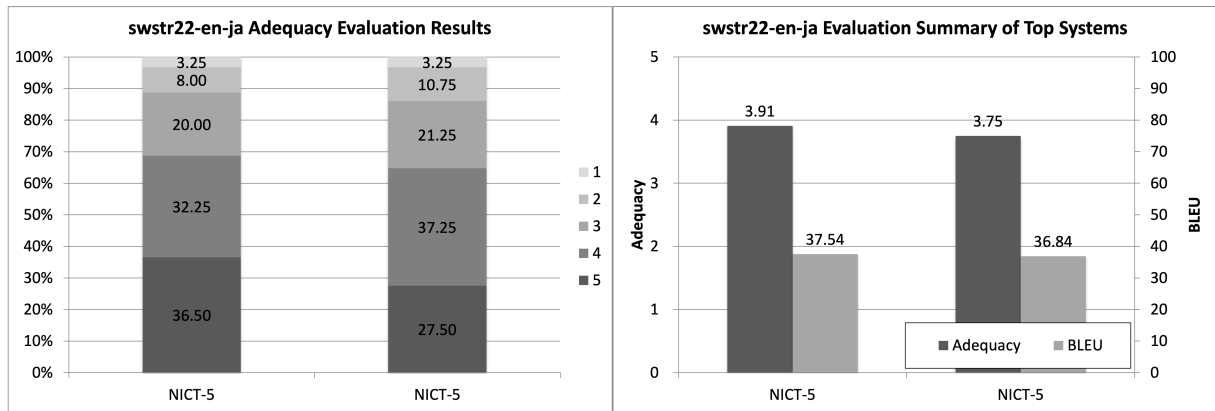


Figure 25: Official evaluation results of swstr22-en-ja.

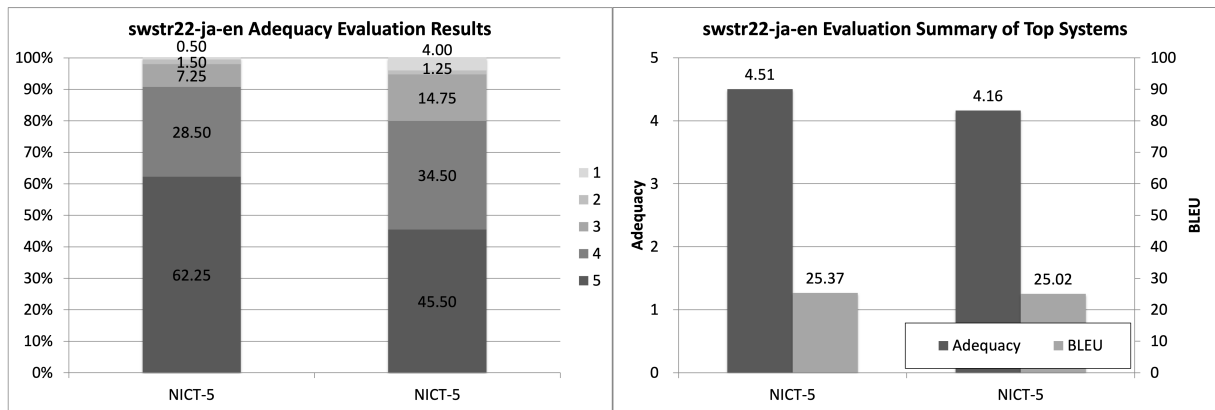


Figure 26: Official evaluation results of swstr22-ja-en.

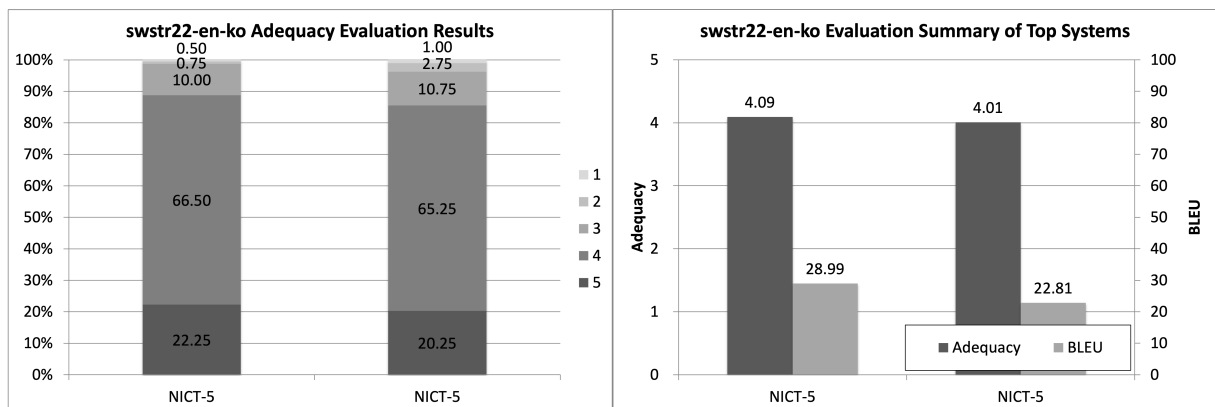


Figure 27: Official evaluation results of swstr22-en-ko.

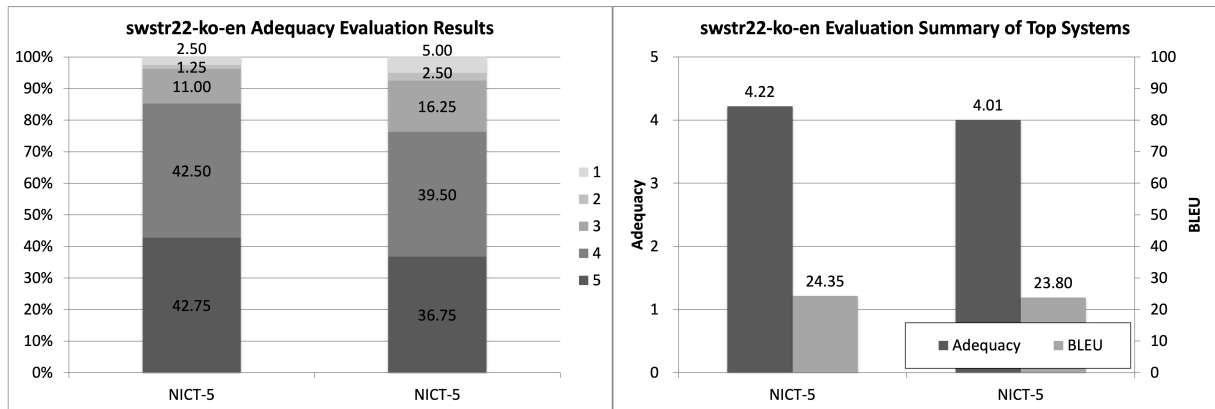


Figure 28: Official evaluation results of swstr22-ko-en.

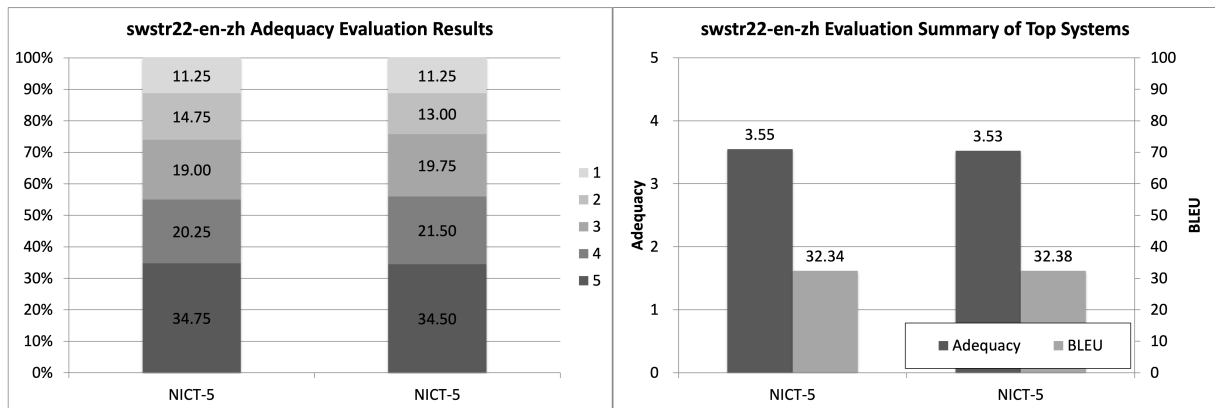


Figure 29: Official evaluation results of swstr22-en-zh.

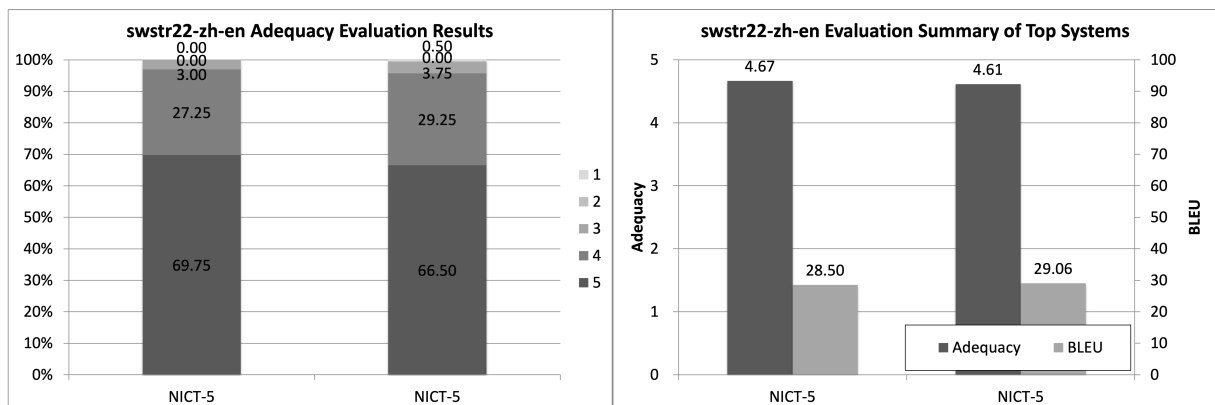


Figure 30: Official evaluation results of swstr22-zh-en.

49, 50.

For English to Malayalam text-only translation the team “Silo_NLP” obtained a BLEU score of 30.80 fine-tuning with pre-trained mBART-50 model for the evaluation test set and team “nlp_novices” obtained a BLEU score of 19.60 using the *Simple Transformer* model. For multimodal, the team “Silo_NLP” obtained a BLEU score of 41.00 for the evaluation test set and a BLEU score of 20.40 for the challenge test set. They extracted the object tags from the images with fine-tuning *mBART* for the multimodal task.

Human evaluation was done for the challenge test set text-only translation (TEXT) as shown in Figure 20. Automatic (both BLEU and RIBES) scores agree with the manual judgement on the JPO adequacy scale (see Table 13), but it is important to mention that even for the better system (“nlp_novices”) only a little less than 10% of sentences have all information transmitted correctly. If we consider Adequacy ranks 3–5 together (i.e. about a half or more of information transmitted correctly), “nlp_novices” can produce about 55% of sentences like that while “Silo_NLP” has only 30% of sentences in these levels.

8.6 English→Bengali Multi-Modal Task

This year three teams participated in the different sub-tasks (TEXT, MM) of the English→Bengali Multi-Modal task. The WAT2022 automatic evaluation scores are shown in the Table 51, 52, 53, 54.

The team “Silo_NLP” obtained the highest BLEU score for the text-only translation (TEXT) for the evaluation (E-Test) set by using *Transformer* model and utilizing BNLIT Corpus as an additional resource. The team “nlp_novices” obtained the highest BLEU score on the challenge (C-Test) test set by fine-tuning the *Transformer* model. For the multimodal sub-task (MM), we received two submissions from the teams “CNLP-NITS-PP”, and “Silo_NLP”, respectively. The team “CNLP-NITS-PP” obtained the highest BLEU score for the evaluation (E-Test) test set following transliteration-based phrase pairs augmentation and visual features in training using BRNN encoder and doubly-attentive-rnn decoder. The team “Silo_NLP” and “nlp_novices” obtained the same BLEU score for the challenge (C-Test) test set. The team “nlp_novices” followed the same approach as that for E-Test while team “Silo_NLP” extracted the object tags from images and fine-

tuned *mBART*.

Human evaluation was done for the challenge test set text-only translation (TEXT) as shown in Figure 18. Automatic scores (both BLEU and RIBES) agree on the best system (“nlp_novices”) with the manual judgement on the JPO adequacy scale (see Table 13) but they diverge for “Silo_NLP” and “CNLP-NITS-PP” where both receive Adequacy of 2.66 and differ in BLEU and RIBES. “CNLP-NITS-PP” scores higher in automatic metrics and actually very close to the winning “nlp_novices”, see RIBES of 70.66 (“nlp_novices”) vs. 68.07 (“CNLP-NITS-PP”). This suggests a problem with RIBES because the difference in Adequacy is important: 3.53 vs 2.66. One can also see a striking difference in the distribution of Adequacy levels between the winning “CNLP-NITS-PP” (55% of sentences reach Adequacy of 4 or 5) and its competitors (only 30% of sentences reach 4 or 5), see the left part of Figure 18.

8.7 Restricted Translation Task

In this year, we received system submissions from the team “TMU” for the English→Japanese and Japanese→English tasks, and no systems submitted to the Chinese↔Japanese tasks. The TMU team employed a soft-constrained system that combined two methods, namely the Lexical Constraint Aware NMT (LeCA; Chen et al., 2020) and the Multi-Source Levenshtein Transformer (MSLevT Susanto et al., 2020). In case the soft constrained method of LeCA does not satisfy the target-side term requirements, the authors applied one of the automatic post-editing methods to compensate for those terms in the system outputs, such as MSLevT, and achieved 100% performance on the output of the constrained phrase pairs.

Table 14 reports the final score and two distinct human evaluation results⁴⁶. Regarding the final automatic evaluation score, we used SacreBLEU⁴⁷ to calculate BLEU scores. More details are described in Section 2.15. Moreover, we asked human bilingual speakers to assess three systems on

⁴⁶In the final automatic score for En-Ja, we received an inquiry from TMU that the specification of the submission form included backslashes before quotations, and they were detrimental to the evaluation of some constrained terms. The final score without the backslash is as follows; LeCA+LevT (*ensemble*): 42.1, LeCA+LevT: 39.3, LeCA only: 23.8.

⁴⁷Detail settings: case=mixed, numrefs=1, smooth.exp, version.1.5.1, (en-ja) lang=en-ja, tok=ja-mecab-0.996-IPA, (ja-en) lang=ja-en.

the English↔Japanese translation tasks.⁴⁸ Two annotators were asked to assess the systems’ translation accuracy, and we also conducted another system assessment by the source-based direct assessment (src-based DA) (Cettolo et al., 2017; Federmann, 2018), with two bilingual annotators.

In the English→Japanese direction (En-Ja), we do not observe any consistent tendency among three results. LeCA only is the most preferred system by annotators in terms of translation accuracy. However, the other two systems also achieve higher evaluation scores as well as src-based DA scores. These systems can not be statistically distinguished from the human reference. On the other hand, the LeCA+LevT ensemble model achieved the top performance in all metrics in the Japanese→English direction (Ja-En), while LeCA+LevT is less preferred in the src-based DA.

According to the HE (accuracy) results, we observe that the LeCA+LevT (ensemble) system achieves both the highest number of outputs with score=5 (58%) and score=1 (6%) in the human evaluation. For the outputs with score=1 in LeCA+LevT (ensemble), texts other than the constrained terms were often omitted. This indicates that the lack of the effects from the brevity penalty in our final score can not capture under-generation problems on the ensemble model. Therefore, we eventually need to consider an alternative scoring to address this issue in future work. Another observation is that annotators do not necessarily have specific domain knowledge that would be required to provide more accurate assessment. To address this issue, we need to allow annotators to look up the generated dictionaries during the assessment. In conclusion, the trade-off between completing vocabulary constraints and achieving high translation performance remains challenging, even in the soft-constrained model.

8.8 Parallel Corpus Filtering Task

We received a single submission from team ‘sakura’, Rakuten Institute of Technology. They submitted two systems, one leverages feature decay algorithms(FDA) and the other one uses probability scores of the NMT model trained on the AS-

⁴⁸Two systems, that is “LeCA only” and “LeCA+LevT”, were originally designated by the team “TMU” for human evaluation, however, none of those systems are top-ranked on our metrics. Therefore, we decided to additionally include each top-ranked system (LeCA+LevT (*ensemble*)) to the human evaluation.

PEC corpus. They submitted the top 5M-scored sentence pairs as a clean dataset.

Table 15 summarized the evaluation results. We carried out human evaluation only for the baseline and the FDA-based method since the NMT probability-based model was inferior to the baseline in terms of the BLEU scores.

The results show that the submission based on the FDA surpasses the baseline in both language directions on both BLEU and human evaluation while reducing the data size to 20%.

9 Conclusion and Future Perspective

This paper summarizes the shared tasks of WAT2022. We had 8 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.

This year we had smaller number of participants compared to the last year. For the next WAT workshop, we want attract much more people to join our shared tasks.

Acknowledgement

The English→Hindi English→Malayalam, and English→Bengali Multi-Modal shared tasks were supported by the following grants at Silo AI 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 not contain any personal data. Personal data of annotators needed when the datasets were prepared and when the outputs were evaluated were processed in compliance with the GDPR and national law.

- At Silo AI, the work was supported by the NLP Innovation.
- 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).

The Restricted Translation is supported by Microsoft.

En-Ja Outputs	final	HE (Accuracy)	HE (src-based DA)
LeCA+LevT (<i>ensemble</i>)	52.7	4.18	76.4*
LeCA+LevT	50.5	4.19	76.6*
LeCA only	37.6	4.24	74.9
Human Reference	—	—	76.6

Ja-En Outputs	final	HE (Accuracy)	HE (src-based DA)
LeCA+LevT (<i>ensemble</i>)	40.8	4.31	74.1*
LeCA+LevT	38.1	4.22	72.0
LeCA only	23.0	4.14	73.3
Human Reference	—	—	74.7

Table 14: Human evaluation results of translation accuracy run by WAT and source-based direct assessments, ranging [0, 5] and [0, 100], respectively. The “final” column reports the final score of the automatic evaluation metric described in the Section 2.15. * indicates that the systems and Human Reference are not statistically distinguishable to the annotators.

En-Ja Team	BLEU	Human Eval.
sakura (FDA)	28.8	4.31
baseline	27.4	4.18
sakura (NMT Prob.)	26.7	—

Ja-En Team	final	Human Eval.
sakura (FDA)	21.8	4.49
sakura (NMT Prob.)	19.9	—
baseline	19.4	4.35

Table 15: Results of the parallel corpus filtering task evaluated on the ASPEC test set.

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.
- Guanhua Chen, Yun Chen, Yong Wang, and Victor O.K. Li. 2020. [Lexical-constraint-aware neural machine translation via data augmentation](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3587–3593. International Joint Conferences on Artificial Intelligence Organization. Main track.
- 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.
- Weiqi Gu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2021. Video-guided machine translation with spatial hierarchical attention network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 87–92.
- Kazuma Hashimoto, Raffaella Buschiazzi, James Bradbury, Teresa Marshall, Richard Socher, and Caiming Xiong. 2019. [A high-quality multilingual dataset for structured documentation translation](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 116–127,

- Florence, Italy. 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: 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.
- Huda Khayrallah and Philipp Koehn. 2018. [On the impact of various types of noise on neural machine translation](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.
- 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, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. [Findings of the WMT 2020 shared task on parallel corpus filtering and alignment](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.
- 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.
- 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.
- Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. 2022. Visa: An ambiguous subtitles dataset for visual scene-aware machine translation. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 6735–6743.
- 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](#).
- Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. [JParaCrawl: A large scale web-based English-Japanese parallel corpus](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.
- 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, et al. 2021a. Overview of the 8th workshop on asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45.
- Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi.

- 2021b. [Overview of the 8th workshop on Asian translation](#). In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.
- 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, Hideki Nakayama, Isao Goto, Hideya Mino, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Shohei Higashiyama, Hiroshi Manabe, Win Pa Pa, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, Katsuhito Sudoh, Sadao Kurohashi, and Pushpak Bhattacharyya, editors. 2021c. [Proceedings of the 8th Workshop on Asian Translation \(WAT2021\)](#). Association for Computational Linguistics, Online.
- 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.
- Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. [ASPEC: Asian scientific paper excerpt corpus](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).
- 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.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravindh Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Didee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. [Samanantar: The largest publicly available parallel corpora collection for 11 indic languages](#). *Transactions of the Association for Computational Linguistics*, 10:145–162.
- 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.
- Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. [Bengali 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.
- Kak Soky, Masato Mimura, Tatsuya Kawahara, Sheng Li, Chenchen Ding, Chenhui Chu, and Sethserey Sam. 2021. Khmer speech translation corpus of the extraordinary chambers in the courts of cambodia (eccc). In *2021 24th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 122–127.
- Raymond Hendy Susanto, Shamil Chollampatt, and Lil-ing Tan. 2020. [Lexically constrained neural machine translation with Levenshtein transformer](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. 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 Kajiwar, and Chenhui Chu. 2020. [Double attention-based multimodal neural machine translation with semantic image regions](#). In *EAMT*, pages 105–114.

Appendix A Submissions

Tables 16 to 63 summarize translation results submitted to WAT2022. Type and RSRC columns indicate type of method and use of other resources.

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
TMU	6947	NMT	NO	49.80	49.80	50.00	0.864394	0.865434	0.869480	0.788640
TMU	6948	NMT	NO	50.80	51.60	51.30	0.867732	0.869859	0.873110	0.800450

Table 16: ASPECRT en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
TMU	6942	NMT	NO	39.60	0.799376	0.640000
TMU	6949	NMT	NO	39.30	0.795787	0.653280

Table 17: ASPECRT ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6932	NMT	NO	1.70	0.222952	—

Table 18: ECCC km-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6933	SMT	NO	5.70	0.202578	—
ORGANIZER	6934	NMT	NO	5.70	0.202578	—

Table 19: ECCC km-fr submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6963	NMT	YES	3.50	0.537859	—
NITR	6998	NMT	NO	15.50	0.706743	—

Table 20: INDIC22 as-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6966	NMT	YES	4.50	0.547407	—

Table 21: INDIC22 bn-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6965	NMT	YES	1.10	0.359265	—
NITR	7016	NMT	NO	10.20	0.634631	—

Table 22: INDIC22 en-as submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6969	NMT	YES	2.00	0.503286	—

Table 23: INDIC22 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NITR	7009	NMT	NO	6.30	0.579323	—

Table 24: INDIC22 en-sd submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NITR	7012	NMT	NO	9.50	0.647028	—

Table 25: INDIC22 en-si submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NITR	7014	NMT	NO	19.60	0.718763	—

Table 26: INDIC22 en-ur submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NITR	7007	NMT	NO	8.00	0.546125	—

Table 27: INDIC22 ne-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NITR	7005	NMT	NO	15.40	0.709039	—

Table 28: INDIC22 sd-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NITR	7003	NMT	NO	8.20	0.632228	—

Table 29: INDIC22 si-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
NITR	7000	NMT	NO	20.50	0.744934	—

Table 30: INDIC22 ur-en submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	6543	NMT	NO	47.04	48.86	46.96	0.870867	0.870604	0.869950	—

Table 31: JPC22 en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6544	NMT	NO	44.51	0.857963	—

Table 32: JPC22 ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6542	NMT	NO	72.79	0.952385	—

Table 33: JPC22 ja-ko submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				kytea	ctb	pku	kytea	ctb	pku	
ORGANIZER	6540	NMT	NO	44.73	45.77	45.48	0.871424	0.877354	0.875780	—

Table 34: JPC22 ja-zh submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	6541	NMT	NO	73.55	74.58	73.89	0.956442	0.956203	0.956269	—

Table 35: JPC22 ko-ja submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	6539	NMT	NO	51.03	51.64	51.14	0.887901	0.885180	0.887404	—

Table 36: JPC22 zh-ja submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	6536	NMT	NO	58.87	60.50	58.83	0.905725	0.907156	0.904626	—

Table 37: JPCN4 en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6537	NMT	NO	54.86	0.880671	—

Table 38: JPCN4 ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6535	NMT	NO	74.73	0.958438	—

Table 39: JPCN4 ja-ko submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				kytea	ctb	pku	kytea	ctb	pku	
ORGANIZER	6538	NMT	NO	57.51	57.92	57.99	0.898847	0.906742	0.904318	—

Table 40: JPCN4 ja-zh submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	6534	NMT	NO	76.69	78.17	77.09	0.963465	0.963548	0.963376	—

Table 41: JPCN4 ko-ja submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	6532	NMT	NO	64.31	65.00	64.62	0.924617	0.922020	0.924463	—

Table 42: JPCN4 zh-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6739	NMT	NO	37.00	0.795302	—
SILO_NLP	6836	NMT	NO	36.20	0.785673	—
nlp novices	6733	NMT	YES	43.10	0.816860	—

Table 43: MMEVTEXT22 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6740	NMT	NO	39.40	0.802635	—
SILO_NLP	6958	NMT	YES	42.00	0.796441	—

Table 44: MMEVMM22 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6742	NMT	NO	37.20	0.770640	—
SILO_NLP	6838	NMT	NO	29.60	0.728801	—
nlp novices	6725	NMT	YES	41.80	0.812500	—

Table 45: MMCHTEXT22 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6741	NMT	NO	39.30	0.791468	—
SILO_NLP	6959	NMT	YES	39.10	0.784169	—

Table 46: MMCHMM22 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
SILO_NLP	6848	NMT	NO	30.80	0.589471	—
nlp novices	6719	NMT	YES	30.60	0.643987	—

Table 47: MMEVTEXT22 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
SILO_NLP	6849	NMT	NO	14.60	0.392158	—
nlp novices	6720	NMT	YES	19.60	0.535043	—

Table 48: MMCHTEXT22 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
SILO_NLP	6936	NMT	NO	41.00	0.705349	—

Table 49: MMEVMM22 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
SILO_NLP	6937	NMT	NO	20.40	0.533737	—

Table 50: MMCHMM22 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6703	NMT	NO	40.90	0.758246	—
CNLP-NITS-PP	6746	NMT	NO	40.90	0.752543	—
SILO_NLP	6954	NMT	NO	41.00	0.767212	—

Table 51: MMEVTEXT22 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6704	NMT	NO	22.50	0.614267	—
CNLP-NITS-PP	6745	NMT	NO	26.70	0.680655	—
SILO_NLP	6843	NMT	NO	22.60	0.605676	—
nlp novices	6970	NMT	YES	32.90	0.706596	—

Table 52: MMCHTEXT22 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6743	NMT	NO	43.90	0.780669	—
SILO_NLP	6939	NMT	NO	42.10	0.754291	—

Table 53: MMEVMM22 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
CNLP-NITS-PP	6744	NMT	NO	28.70	0.688931	—
SILO_NLP	6940	NMT	NO	28.70	0.666817	—

Table 54: MMCHMM22 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
HwTscSU	6751	NMT	NO	56.70	0.884286	—

Table 55: SOFTWARE en-ms submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
HwTscSU	6752	NMT	NO	45.50	0.819582	—

Table 56: SOFTWARE ms-en submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				juman	kytea	mecab	juman	kytea	mecab	
ORGANIZER	6806	NMT	NO	—	—	40.27	—	—	—	—
NICT-5	6821	NMT	NO	—	—	36.84	—	—	—	—
NICT-5	6974	NMT	NO	—	—	37.54	—	—	—	—

Table 57: SWSTR22 en-ja submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6809	NMT	NO	21.87	—	—
NICT-5	6823	NMT	NO	22.81	—	—
NICT-5	6976	NMT	NO	28.99	—	—

Table 58: SWSTR22 en-ko submissions

System	ID	Type	RSRC	BLEU			RIBES			AMFM
				kytea	ctb	pku	kytea	ctb	pku	
ORGANIZER	6811	NMT	NO	28.03	—	—	—	—	—	—
NICT-5	6827	NMT	NO	32.34	—	—	—	—	—	—
NICT-5	6978	NMT	NO	32.38	—	—	—	—	—	—

Table 59: SWSTR22 en-zh submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6807	NMT	NO	28.20	—	—
NICT-5	6822	NMT	NO	25.02	—	—
NICT-5	6975	NMT	NO	25.37	—	—

Table 60: SWSTR22 ja-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6810	NMT	NO	10.80	—	—
NICT-5	6824	NMT	NO	23.80	—	—
NICT-5	6977	NMT	NO	24.35	—	—

Table 61: SWSTR22 ko-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6812	NMT	NO	29.14	—	—
NICT-5	6826	NMT	NO	28.50	—	—
NICT-5	6979	NMT	NO	29.06	—	—

Table 62: SWSTR22 zh-en submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ORGANIZER	6706	OTHER	NO	14.50	0.465183	—

Table 63: VIDEOGAS ja-en submissions