Overview of the 10th Workshop on Asian Translation

Toshiaki Nakazawa nakazawa@nlab.ci.i.u-tokyo.ac.jp
The University of Tokyo

Kazutaka Kinugawa kinugawa.k-jg@nhk.or.jp
Hideya Mino mino.h-gq@nhk.or.jp
Isao Goto goto.i-es@nhk.or.jp
NHK

Raj Dabre raj.dabre@nict.go.jp
Shohei Higashiyama shohei.higashiyama@nict.go.jp
National Institute of Information and Communications Technology

Shantipriya Parida shantipriya.parida@silo.ai
Silo AI

Makoto Morishita makoto.morishita@ntt.com
NTT Communication Science Laboratories

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

Akiko Eriguchi akikoe@microsoft.com
Microsoft

Yusuke Oda yusuke.oda.c1@tohoku.ac.jp
Tohoku University

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

Abstract

This paper presents the results of the shared tasks from the 10th workshop on Asian translation (WAT2023). For the WAT2023, 2 teams submitted their translation results for the human evaluation. We also accepted 1 research paper. About 40 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-WAT2022 Nakazawa

Proceedings of the 10th Workshop on Asian Translation, pages 1–28
September 4, 2023, Macau SAR, China

et al. (2022), WAT2023 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 10th WAT, we included the following new tasks/languages:

- Non-Repetitive Translation Task: Japanese → English style-controlled translation in the news domain.
- 4 new languages to the Multilingual Indic Machine Translation Task (MultiIndicMT): Sindhi, Santali, Kashmiri, Maithili.

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 ParaNatCom[1] (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 tasks are sentence-level and the translation quality of them seem to be saturated. To move to a highly topical setting of translation of dialogues evaluated at the level of documents, WAT uses BSD Corpus[2] (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.

[1] `http://www2.nict.go.jp/astrec-att/member/mutiyama/paranatcom/`

[2] `https://github.com/tsuruoka-lab/BSD`

Lang	Train	Dev	DevTest	Test-2022	Test-N1	Test-N2	Test-N3	Test-N4
zh-ja	1,000,000	2,000	2,000	10,204	2,000	3,000	204	5,000
ko-ja	1,000,000	2,000	2,000	7,230	2,000	–	230	5,000
en-ja	1,000,000	2,000	2,000	10,668	2,000	3,000	668	5,000

Table 1: Statistics for JPC

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. 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.

2.4 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 2. Notice that both of the corpora have been modified from the data used in WAT2018.

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

Table 2: Statistics for the data used in Myanmar-English translation tasks

		Language Pair			
Split	**Domain**	**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 3: The NICT-SAP task corpora splits. The corpora belong to two domains: wikinews (ALT) and software documentation (IT). The Wikinews corpora are N-way parallel.

2.5 NICT-SAP Task

In WAT2021, we decided to continue the WAT2020 task for joint multi-domain multilingual neural machine translation involving 4 low-resource Asian languages: Thai (Th), Hindi (Hi), Malay (Ms), Indonesian (Id). English (En) is the source or the target language for the translation directions being evaluated. The purpose of this task was to test the feasibility of multi-domain multilingual solutions for extremely low-resource language pairs and domains. Naturally the solutions could be one-to-many, many-to-one or many-to-many NMT models. The domains in question are Wikinews and IT (specifically, Software Documentation). The total number of evaluation directions are 16 (8 for each domain). There is very little clean and publicly available data for these domains and language pairs and thus we encouraged participants to not only utilize the small Asian Language Treebank (ALT) parallel corpora Thu et al. (2016) but also the parallel corpora from OPUS[3], other WAT tasks (past and present) and WMT[4]. The ALT dataset contains 18,088, 1,000 and 1,018 training, development and testing sentences. As for corpora for the IT domain we only provided evaluation (dev and test sets) corpora[5] 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 3 we give statistics of the aforementioned corpora which we used for the organizer's baselines. Note that the evaluation corpora for both domains are created from documents and thus contain document level meta-data. Participants were encouraged to use document level approaches. Note that we do not exhaustively list[6] all available corpora here and participants were not restricted from using any corpora as long as they are freely available.

[3] `http://opus.nlpl.eu/`

[4] `http://www.statmt.org/wmt20/`

[5] Software Domain Evaluation Splits

[6] `http://lotus.kuee.kyoto-u.ac.jp/WAT/NICT-SAP-Task`

2.6 Structured Document Translation Task

For the first time we introduce a structured document translation task for English ↔ 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 ↔ 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.

2.7 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. (2020a), 2021 Nakazawa et al. (2021b) and 2022 Nakazawa et al. (2022), we decided to enlarge the scope of the 2022 task by adding 4 new languages to the MultiIndicMT task, namely, Santali, Sindhi, Kashmiri and Maithili. In addition to the original 15 Indic languages, alongside English (En), namely, Hindi (Hi), Marathi (Mr), Kannada (Kn), Tamil (Ta), Telugu (Te), Gujarati (Gu), Malayalam (Ml), Bengali (Bn), Oriya (Or), Punjabi (Pa), Assamese (As), Urdu (Ur), Sindhi (Si), Sinhala (Sd) and Nepali (Ne), we have a total of 19 Indic languages being evaluated this year. We used the FLORES-200 dataset's[7] dev and devtest sets for development and testing both containing roughly 1000 sentences each per language. FLORES-200 is N-way parallel which ensures Indic to Indic translation evaluation.

The objective of this task, like the Indic languages tasks in 2018, 2020-2022, 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 as well as exploiting pre-trained models like IndicBART Dabre et al. (2022) or IndicTrans2 AI4Bharat et al. (2023). For training, we encouraged the use of the Samanantar corpus Ramesh et al. (2022) and its extension, the BPCC corpus AI4Bharat et al. (2023) which covers 18 of the 19 Indic languages. For Sinhala which is not covered by BPCC, we asked users to use the corpora from Opus, specifically the Paracrawl datasets[8]. We also listed additional sources of monolingual corpora for participants to use, namely IndicCorp v2 Doddapaneni et al. (2023).

2.8 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, 2020a, 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 4. 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 2.8.1 below), some of these item components are available as the source and some serve as the reference or play the role of a competing candidate solution.

[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

		Tokens	
Dataset	Items	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 4: 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.

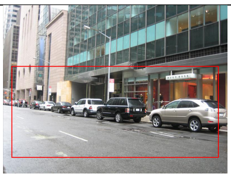
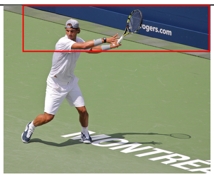
	Text-Only MT	**Hindi Captioning**	**Multi-Modal MT**
Image	–		
Source Text	The woman is waiting to cross the street	–	A blue wall beside tennis court
System Output	महिला सड़क पार करने का इंतजार कर रही है	सड़क पर कार	टेनिस कोर्ट के बगल में एक नीली दीवार
Gloss	Woman waiting to cross the street	Car on the 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 2023 English→Hindi Multi-Modal Task.

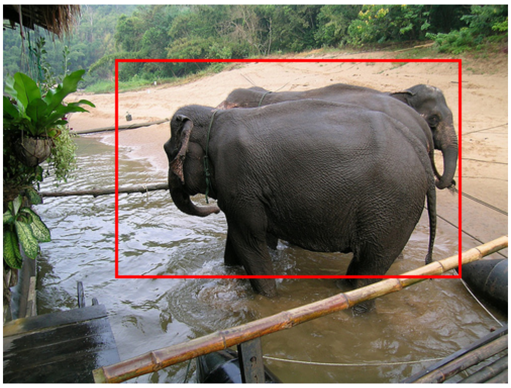
2.8.1 English→Hindi Multi-Modal Task Tracks

1. Text-Only Translation (labeled "TEXT" in WAT official tables): The participants are asked to translate short English captions (text) into Hindi. No visual information can be used. On the other hand, additional text resources are permitted (but they need to be specified in the corresponding system description paper).

2. Hindi Captioning (labeled "HI"): The participants are asked to generate captions in Hindi for the given rectangular region in an input image.

3. Multi-Modal Translation (labeled "MM"): Given an image, a rectangular region in it and an English caption for the rectangular region, the participants are asked to translate the English text into Hindi. Both textual and visual information can be used.

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



English Text: Two elephants standing in the water.

Malayalam Text: വെള്ളത്തിൽ നിൽക്കുന്ന രണ്ട് ആനകൾ

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

2.9 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 5. As in Hindi Visual Genome (see Section 2.8), 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 2.9.1 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→Malayalam Multi-Modal Task Tracks

1. Text-Only Translation (labeled "TEXT" in WAT official tables): The participants are asked to translate short English captions (text) into Malayalam. No visual information can be used. On the other hand, additional text resources are permitted (but they need to be specified in the corresponding system description paper).

2. Malayalam Captioning (labeled "ML"): The participants are asked to generate captions in Malayalam for the given rectangular region in an input image.

3. Multi-Modal Translation (labeled "MM"): Given an image, a rectangular region in it and an English caption for the rectangular region, the participants are asked to translate the English text into Malayalam. Both textual and visual information can be used.

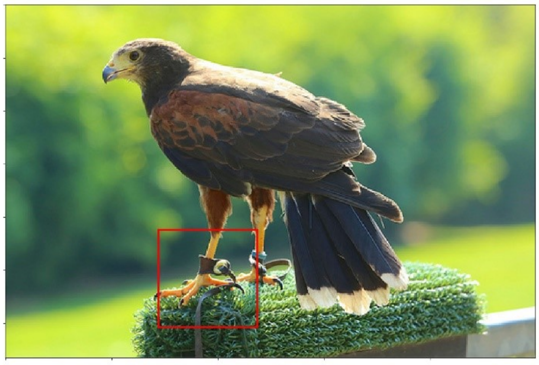
2.10 English→Bengali Multi-Modal Task

This new task, introduced in WAT2022, uses a multimodal machine translation dataset in *Bengali* language. The task mimics the structure of English→Hindi (Section 2.8) and English→Malayalam (Section 2.9) multi-modal tasks. For English→Bengali multi-modal trans-

[10] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533`

		Tokens	
Dataset	Items	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 5: 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.



English Text: The sharp bird talon.

Bengali Text: ধারালো পাখি টাল

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

lation 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 6. 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 Bengali reference translation as shown in Figure 3. Depending on the track (see Section 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.

2.10.1 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 Ben-

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

		Tokens	
Dataset	Items	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 6: 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.

gali 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.

2.11 Ambiguous MS COCO Japanese↔English Multimodal Task

This is the 3rd 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 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) dataset[13] can be used as training data. In the unconstrained setting, the MS COCO English data[14] and STAIR Japanese image captions[15] 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.

2.12 Japanese→English Video Guided MT Task for Ambiguous Subtitles

This is the 2nd year that we have organized this 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, only the VISA dataset[16] 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.

[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/

[16] https://github.com/ku-nlp/VISA

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

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

This is the 2nd year that we have organized this 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 2023 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 dataset[17] 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.

2.14 Restricted Translation Task

Despite recent success of NMT, the MT systems still struggle to generate translation with a consistent terminology. Consistency is the key to clear and accurate translation, especially when translating documents in a specific field, for instance, science or business and marketing contexts, requiring technical terms and proper nouns to get translated into the corresponding unique expressions continuously in the entire documents. To tackle this inconsistent translation issue, we have introduced *Restricted Translation task* since WAT 2021 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. There exist English↔Japanese tasks and Chinese↔Japanese tasks. We employ the ASPEC corpus for all the translation tasks and allow participants to use any other external data sources.

2.15 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.

This task lets the participants train machine translation models under the following restrictions:

[17] `https://github.com/ksoky/ECCC_DATASET`

Dataset	**# sentences**
JParaCrawl v3.0	25.7M
WMT22 generaltest2022.en-ja	2,037
WMT22 generaltest2022.ja-en	2,008

Table 7: Number of sentence pairs in the corpora used in the parallel corpus filtering task.

- 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.

- Training corpus is fixed. The organizer provides the whole corpus, and participants are requested to rank sentences in the corpus by their quality.

- The model will be trained with high-scored sentences (top 100k, 1M, and 10M sentences), and evaluate their translation performance.

- For evaluation, we used WMT22 General Translation Task test-set Kocmi et al. (2022), which includes various domains. Thus domain adaptation by selecting training data is not our scope.

We adopted the Transformer model as the shared architecture for this task.[18] 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. The baseline model is obtained by training the model on the whole set of this dataset.

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. (2002a) and JPO adequacy as described in Section 6.1 on the WMT22 General Translation Task test-set. The corpus statistics are summarized in Table 7.

The ultimate goal of this shared task is to create a cleaner JParaCrawl corpus. After this shared task ends, we plan to ensemble all participant scores and make a cleaner corpus.

2.16 Non-Repetitive Translation Task

We introduce a novel non-repetitive translation task for Japanese → English sentence-level translation. The underlying motivation is to guide a machine translation (MT) system to follow the writing style of the English news domain. To realize high-quality text, English news has many rules, such as using the active rather than the passive voice, using the affirmative rather than the negative, and avoiding redundant phrases (Block, 1994; Cappon, 2019; Papper, 2021). Our goal is to produce high-quality translations that follow a set of writing rules used by professional news translators. For the first year, we focus on the repetition of words or phrases. Here is an example:

(Ja) 入学$_{(1)}$予定者７人が教育方針や私立小への入学$_{(2)}$などを理由に入学$_{(3)}$を辞退した。

(En) ..., seven children dropped plans to enter the school$_{(3)}$, with parents citing disagreements with its education policy, decisions to join$_{(2)}$ private schools or other reasons, ...

[18] The Dockerfile for constructing the training pipeline can be obtained from `https://github.com/MorinoseiMorizo/wat2022-filtering`

In this sentence pair, “入学$_{(1)}$” has been intentionally removed in the translation, probably because it is contextually obvious (*reduction*).[19] In addition, “入学$_{(2)}$” and “入学$_{(3)}$” are translated differently as “join$_{(2)}$” and “enter$_{(3)}$,” respectively (*substitution*). Unlike technical terms, common words and phrases that are repeated in a sentence can create a monotonous or awkward impression, and should be avoided where appropriate. In this task, participants are required to control an MT system in applying reduction or substitution so that it does not output the same words/phrases for certain repeated words/phrases in the source sentence. We refer to such translations as *non-repetitive translations*. The key point of this task is to control lexical redundancy and diversity while maintaining an accurate translation.

We provide development and test sets containing 70 and 173 examples, respectively. In each set, about one-third of the examples are reductions and the remaining two-thirds are substitutions. This evaluation dataset was constructed from Jiji news articles. In each example, the Japanese source sentence contains one type of repeated word/phrase that is translated with reduction or substitution into the English reference sentence. No training set has been prepared specifically for this task. Although we also provide the dataset including 200K training sentence pairs from the WAT2020 Newswire tasks (Nakazawa et al., 2020b),[20] which was also constructed from Jiji news articles, participants can use any data for training as long as it does not contain the test set in this task. Note that the evaluation dataset for this task partially overlaps with that of the WAT2020 Newswire tasks (Nakazawa et al., 2020b).

To verify that the reductions and substitutions are appropriate, a two-step manual inspection is used instead of automatic metrics. First, three human annotators check the output for mistranslations, undertranslations, or overtranslations, and assign a 0/1 acceptability score to each output. Here, we stress to the annotators that they should be aware of the difference between reduction (removing contextually obvious words/phrases) and undertranslation (failing to output necessary words/phrases). Unacceptable outputs in this stage do not affect the final result. Next, the annotators check whether the target words/phrases have been successfully substituted or reduced, and judge whether the outputs are written in either non-repetitive style or repetitive style. Although an MT system must choose either substitution or reduction to produce a non-repetitive translation style, the choice does not have to be consistent with the reference translation. In addition, the MT system does not necessarily have to choose the same word/phrase as that used in the reference. The final decisions on acceptability and translation style are made by a majority vote of the three annotators at each stage. The reference translation is not shown to the annotators in either evaluation step. The final result is determined by the number of translations that are both acceptable and non-repetitive.

As a baseline, we use the vanilla “big” Ja→En Transformer model (Vaswani et al., 2017) pre-trained on JparaCrawl v3.0 (Morishita et al., 2022), which was downloaded from the authors’ website.[21]

3 Participants

Table 8 shows the participants in WAT2023. Both teams participated the Indic Multimodal Tasks. About 40 translation results by 2 teams were submitted for automatic evaluation.

[19] *Reduction* includes sharing a noun head, e.g., “the reopened school and provisional school” → “the reopened and provisional schools.”

[20] `https://lotus.kuee.kyoto-u.ac.jp/WAT/jiji-corpus/2020/TaskDescription.html`

[21] `https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/`

Team ID	Organization	Country
ODIAGEN	Odia Generative AI	India
BITS-P	Birla Institute of Technology and Science, Pilani	India

Table 8: List of participants who submitted translations for the human evaluation in WAT2023

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 2023, 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, and ALT+UCSY

- Juman version 7.0[22] for Japanese segmentation.
- Stanford Word Segmenter version 2014-01-04[23] (Chinese Penn Treebank (CTB) model) for Chinese segmentation.
- The Moses toolkit for English and Indonesian tokenization.
- Mecab-ko[24] for Korean segmentation.
- Indic NLP Library[25] Kunchukuttan (2020) for Indic language segmentation.
- The tools included in the ALT corpus for Myanmar and Khmer segmentation.
- subword-nmt[26] 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.

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

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

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

[25] https://github.com/anoopkunchukuttan/indic_nlp_library

[26] https://github.com/rsennrich/subword-nmt

- 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.

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, and ALT tasks, tensor2tensor[27] for the NICT-SAP task, HuggingFace transformers[28] for the Structured Document Translation task and OpenNMT-py[29] for other tasks.

4.2.1 NMT with Attention (OpenNMT)

For ASPEC, JPC, 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.

[27] https://github.com/tensorflow/tensor2tensor

[28] https://github.com/huggingface/transformers

[29] https://github.com/OpenNMT/OpenNMT-py

4.2.2 Transformer (Tensor2Tensor)

For the News Commentary task, we used tensor2tensor's[30] 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-tuned[31] 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) 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. (2002a), 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.[32] AMFM scores were calculated using scripts created by the technical collaborators listed in the WAT2023 web page.[33] 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.

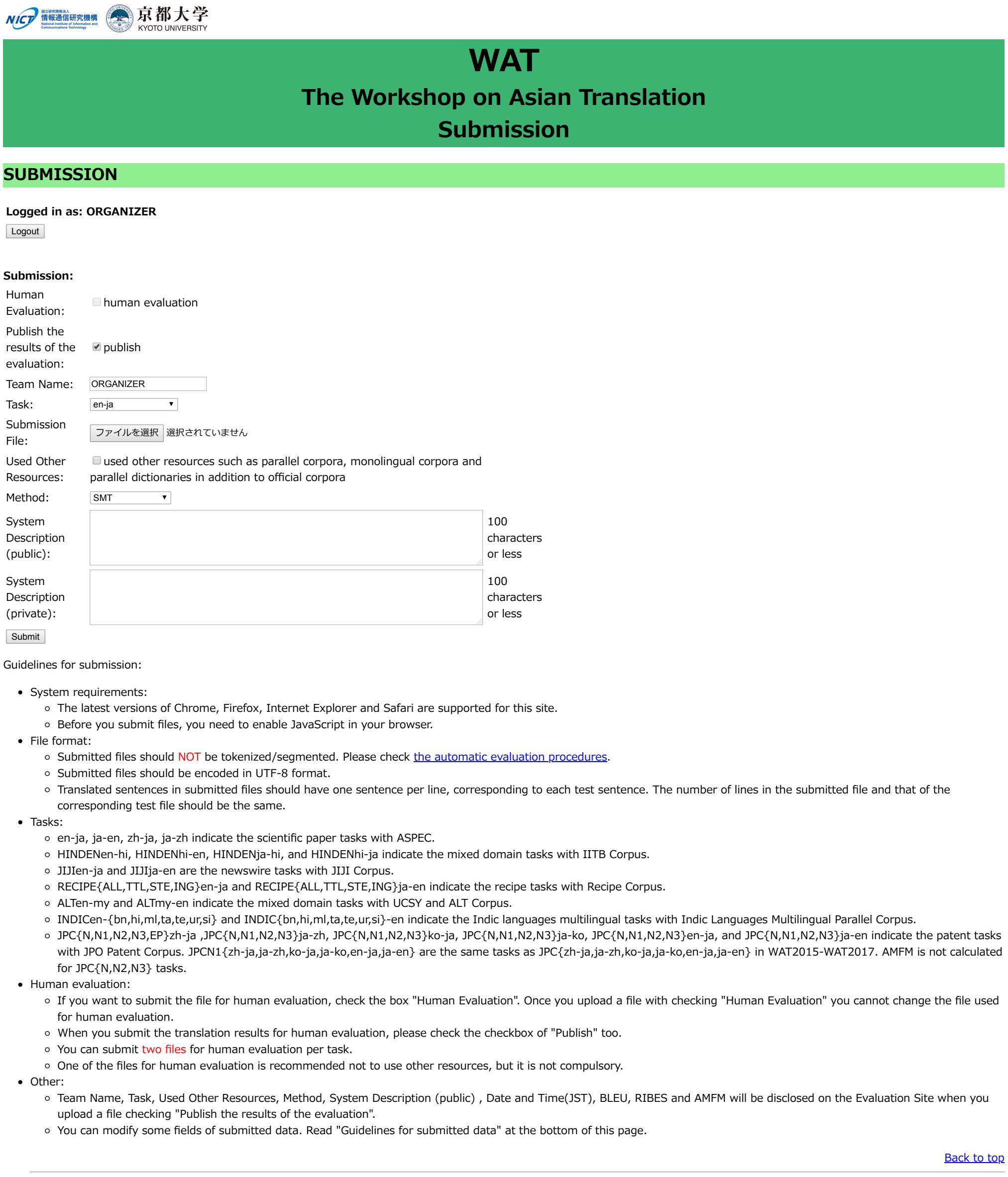
Except for XML-BLEU, which uses this implementation for evaluation, the following preprocessing is done prior to computing scores. Before the calculation of the automatic evaluation scores, the translation results were tokenized or segmented with tokenization/segmentation tools for each language. For Japanese segmentation, we used three different tools: Juman version 7.0

[30] https://github.com/tensorflow/tensor2tensor

[31] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[32] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html

[33] lotus.kuee.kyoto-u.ac.jp/WAT/WAT2023/



WAT

The Workshop on Asian Translation

Submission

SUBMISSION

Logged in as: ORGANIZER

Submission:

Human Evaluation:	□ human evaluation
Publish the results of the evaluation:	☑ publish
Team Name:	ORGANIZER
Task:	en-ja
Submission File:	ファイルを選択 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:	SMT
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.
  - JIJIen-ja and JIJIja-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.

Figure 4: The interface for translation results submission

Kurohashi et al. (1994), KyTea 0.4.6 Neubig et al. (2011) with full SVM model[34] and MeCab 0.996 Kudo (2005) with IPA dictionary 2.7.0.[35] 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)

[34] http://www.phontron.com/kytea/model.html

[35] http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz

model.[36] For Korean segmentation, we used mecab-ko.[37] For Myanmar and Khmer segmentations, we used `myseg.py`[38] and `kmseg.py`.[39] For English, French and Russian tokenizations, we used `tokenizer.perl`[40] in the Moses toolkit. For Indonesian, Malay, and Vietnamese tokenizations, we used `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 Library[41] Kunchukuttan (2020). The detailed procedures for the automatic evaluation are shown on the WAT evaluation web page.[42]

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 WAT2023 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 WAT2023 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 WAT2023. Anybody can register an account for the system by the procedures described in the application site.[43]

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 WAT2023, we conducted *JPO adequacy evaluation* (Section 6.1).

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

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

[38] `http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/wat2020.my-en.zip`

[39] `http://lotus.kuee.kyoto-u.ac.jp/WAT/km-en-data/km-en.zip`

[40] `https://github.com/moses-smt/mosesdecoder/tree/RELEASE-2.1.1/scripts/tokenizer/tokenizer.perl`

[41] `https://github.com/anoopkunchukuttan/indic_nlp_library`

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

[43] `http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2023/application/index.html`

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

Table 9: 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.[44] 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 9 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).[45]

7 Evaluation Results

In this section, the evaluation results for WAT2023 are reported from several perspectives. Some of the results for both automatic and human evaluations are also accessible at the WAT2023 website.[46]

7.1 Official Evaluation Results

Figures 5, 6 and 7 show the evaluation results of Multimodal 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.

[44] The number of systems varies depending on the subtasks.

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

[46] `http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/`

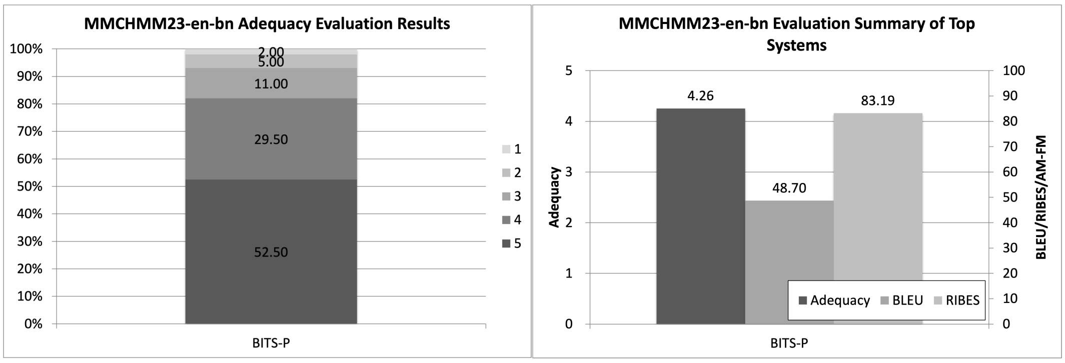


Figure 5: Official evaluation results of mmchmm23-en-bn.

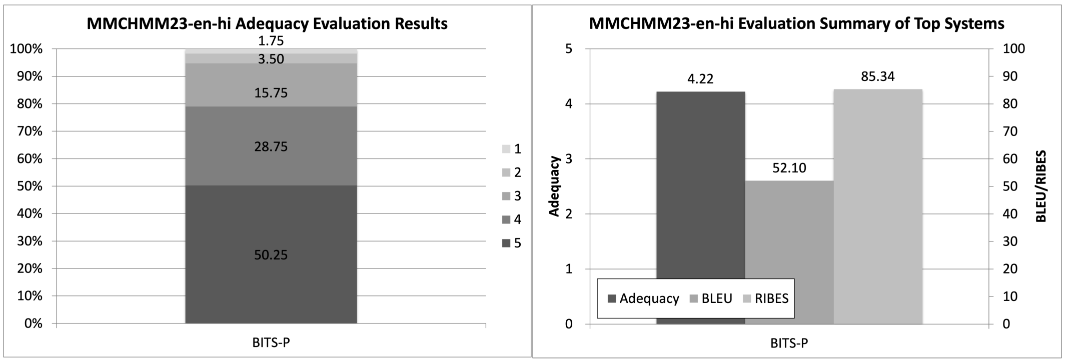


Figure 6: Official evaluation results of mmchmm23-en-hi.

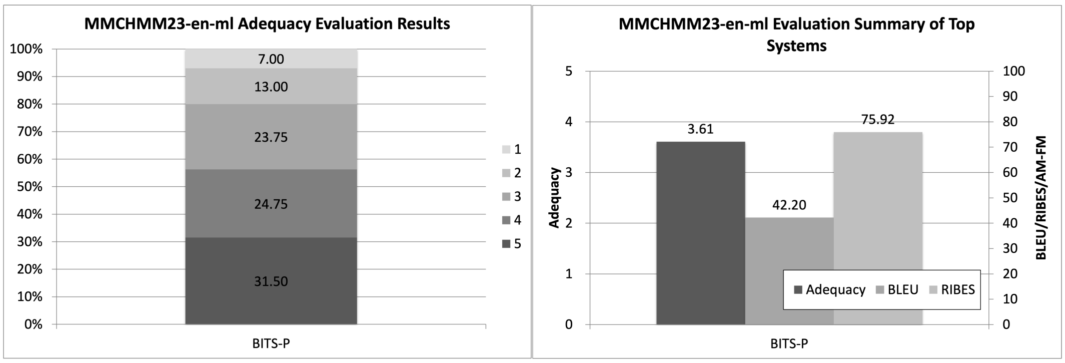


Figure 7: Official evaluation results of mmchmm23-en-ml.

8 Findings

8.1 English→Hindi Multi-Modal Task

This year two teams participated in the different sub-tasks (TEXT, MM) of the English→Hindi Multi-Modal task. The WAT2023 automatic evaluation scores for the participating teams are shown in Tables 11, 14, 17 and 20.

For the text-only sub-task (TEXT), one team "ODIAGEN" participated in the evaluation (E-Test) and challenge (C-Test) set by fine-tuning the *Transformer* model using NLLB-200 from Facebook. Their scores were outperformed in comparison to all previous years' submissions. It is worth mentioning, they did not use any additional resources.

For the multimodal sub-task (MM), we received two submissions from the teams "ODIAGEN", and "BITS-P", respectively. The team "BITS-P" obtained a BLEU score of *45.00* for the evaluation (E-Test) by NLLB model finetuning on captions with object tags of original and synthetic images using DETR model. The team "BITS-P" used additional resources for their model building. The team "ODIAGEN" obtained BLEU score of *41.60* by using image features (extracting object tags) appended with text and MBART finetuning. For the challenge (C-Test) set, both teams ("BITS-P", and "ODIAGEN") obtained BLEU scores of *52.10* and *42.80* respectively following the same approaches as in E-Test.

Human evaluation was done for the challenge test set multimodal translation (MM) as shown in Figure 6.

8.2 English→Malayalam Multi-Modal Task

This year two teams "ODIAGEN", and "BITS-P" participated in the different sub-tasks (TEXT, MM) of the English→ Malayalam Multi-Modal task. The WAT2023 automatic evaluation scores are shown in the Table 21, 15, 18, 12.

For the English to Malayalam text-only translation, team "ODIAGEN" obtained a BLEU score of *46.60*, and *39.70* for the evaluation (E-Test) and challenge (C-Test) respectively. They used fine-tuning Transformer using NLLB-200 from Facebook. For multimodal, the team "BITS-P" obtained a BLEU score of *51.90* for the evaluation test set and a BLEU score of *42.20* for the challenge test set. They used NLLB model finetuned on captions along with object tags of original and synthetic images using the DETR model.

Human evaluation was done for the challenge test set multimodal translation (MM) as shown in Figure 7.

8.3 English→Bengali Multi-Modal Task

This year two teams participated in the different sub-tasks (TEXT, MM) of the English→Bengali Multi-Modal task. The WAT2023 automatic evaluation scores are shown in the Table 22, 16, 19, 13.

For the text-only sub-task (TEXT), one team "ODIAGEN" participated in the evaluation (E-Test) and challenge (C-Test) set by fine-tuning the *Transformer* model using NLLB-200 from Facebook. Their scores were outperformed in comparison to all previous years' submissions.

For the multimodal sub-task (MM), we received two submissions from the teams "ODIAGEN", and "BITS-P", respectively. The team "BITS-P" obtained a BLEU score of *50.60* for the evaluation (E-Test) test set using the NLLB model finetuned on captions along with object tags of original and synthetic images using the DETR model. They used additional resources. The team "ODIAGEN" obtained a BLEU score of *43.90* by using transliteration-based phrase pairs augmentation and visual features in training using a BRNN encoder and doubly-attentive-rnn decoder. For the challenge (C-Test) test, for the same configuration, both teams obtained a BLEU score of *48.70* and 30.50 respectively.

Human evaluation was done for the challenge test set multimodal translation (MM) as

Model	**Test (WAT 2023)**				**Test (WAT 2020)**
		# Non-repetitive	**# Repetitive**	**# Error**	**BLEU (%)**
baseline	**# Acceptable**	**19 (11.0%)**	71 (41.0%)	0 (0.0%)	15.6
	# Unacceptable	29 (16.8%)	49 (28.3%)	5 (2.9%)	

Table 10: Evaluation results of the non-repetitive translation task. # Error indicates the number of translations where the target words/phrases themselves are mistranslated, undertranslated or overtranslated. As a reference, we also computed a BLEU score (Papineni et al., 2002b) on the test set II of the WAT2020 Newswire tasks (Nakazawa et al., 2020b) using SacreBLEU (Post, 2018).[47]

shown in Figure 5.

8.4 Non-Repetitive Translation Task

Although we did not receive any submissions in the non-repetitive translation task, we report the evaluation results of the baseline model in this new task. The results are presented in Table 10. The number of acceptable translations was about half of the test set. In addition, 79% $(71/(19+71))$ of the acceptable translations were written in a repetitive style. For the acceptable and non-repetitive outputs, the numbers of reductions and substitutions were 7 and 12, respectively. (For the unacceptable and non-repetitive outputs, the numbers of reductions and substitutions were 10 and 19, respectively.) This indicates that there is a lot of room for improvement in this task.

9 Conclusion and Future Perspective

This paper summarizes the shared tasks of WAT2023. This year, we had 2 participants who submitted their translation results. Both teams participated to the Indic multimodal translation tasks. This year we had smaller number of participants compared to the previous years. 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 SiloGen.
- 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. The Non-Repetitive Translation task was supported by the commissioned research (No. 225) by National Institute of Information and Communications Technology (NICT), Japan.

[47] The signature is `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`.

References

AI4Bharat, Gala, J., Chitale, P. A., AK, R., Doddapaneni, S., Gumma, V., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M. M., Dabre, R., and Kunchukuttan, A. (2023). Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv: 2305.16307*.

Banchs, R. E., D'Haro, L. F., and Li, H. (2015a). Adequacy-fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(3):472–482.

Banchs, R. E., D' Haro, L. F., and Li, H. (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.

Block, M. (1994). *Broadcast Newswriting: The RTNDA Reference Guide*. Bonus Books.

Buschbeck, B. and Exel, M. (2020). A parallel evaluation data set of software documentation with document structure annotation.

Cappon, R. J. (2019). *The Associated Press Guide to News Writing*. Peterson's, fourth edition.

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Dabre, R., Shrotriya, H., Kunchukuttan, A., Puduppully, R., Khapra, M., and Kumar, P. (2022). IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

D'Haro, L. F., Banchs, R. E., Hori, C., and Li, H. (2019). Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. *Computer Speech and Language*, 55:200–215.

Ding, C., Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Utiyama, M., and Sumita, E. (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.

Ding, C., Utiyama, M., and Sumita, E. (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.

Doddapaneni, S., Aralikatte, R., Ramesh, G., Goyal, S., Khapra, M. M., Kunchukuttan, A., and Kumar, P. (2023). Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Gu, W., Song, H., Chu, C., and Kurohashi, S. (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.

Hashimoto, K., Buschiazzo, R., Bradbury, J., Marshall, T., Socher, R., and Xiong, C. (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.

Imankulova, A., Dabre, R., Fujita, A., and Imamura, K. (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.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (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.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Khayrallah, H. and Koehn, P. (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.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., and Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., and Guzmán, F. (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.

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

Kunchukuttan, A. (2020). The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Kurohashi, S., Nakamura, T., Matsumoto, Y., and Nagao, M. (1994). Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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.

Li, Y., Shimizu, S., Gu, W., Chu, C., and Kurohashi, S. (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.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Merritt, A., Chu, C., and Arase, Y. (2020). A corpus for english-japanese multimodal neural machine translation with comparable sentences.

Morishita, M., Chousa, K., Suzuki, J., and Nagata, M. (2022). JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Morishita, M., Suzuki, J., and Nagata, M. (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.

Nakazawa, T., Doi, N., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Oda, Y., Parida, S., Bojar, O., and Kurohashi, S. (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.

Nakazawa, T., Higashiyama, S., Ding, C., Mino, H., Goto, I., Kazawa, H., Oda, Y., Neubig, G., and Kurohashi, S. (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.

Nakazawa, T., Mino, H., Goto, I., Dabre, R., Higashiyama, S., Parida, S., Kunchukuttan, A., Morishita, M., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., and Kurohashi, S. (2022). Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 1–36, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., et al. (2021a). Overview of the 8th workshop on asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa Pa, W., Kunchukuttan, A., Parida, S., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., and Kurohashi, S. (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.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa Pa, W., Kunchukuttan, A., Parida, S., Bojar, O., and Kurohashi, S. (2020a). 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.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Kunchukuttan, A., Pa, W. P., Bojar, O., Parida, S., Goto, I., Mino, H., Manabe, H., Sudoh, K., Kurohashi, S., and Bhattacharyya, P., editors (2020b). *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. Association for Computational Linguistics.

Nakazawa, T., Nakayama, H., Goto, I., Mino, H., Ding, C., Dabre, R., Kunchukuttan, A., Higashiyama, S., Manabe, H., Pa, W. P., Parida, S., Bojar, O., Chu, C., Eriguchi, A., Abe, K., Oda, Y., Sudoh, K., Kurohashi, S., and Bhattacharyya, P., editors (2021c). *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, Online. Association for Computational Linguistics.

Nakazawa, T., Sudoh, K., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., and Kurohashi, S. (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.

Neubig, G., Nakata, Y., and Mori, S. (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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002a). Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002b). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Papper, R. A. (2021). *Broadcast News and Writing Stylebook*. Routledge, seventh edition.

Parida, S. and Bojar, O. (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.

Parida, S., Bojar, O., and Dash, S. R. (2019a). Hindi visual genome: A dataset for multimodal english-to-hindi machine translation. *arXiv preprint arXiv:1907.08948*.

Parida, S., Bojar, O., and Dash, S. R. (2019b). Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *Computación y Sistemas*. In print. Presented at CICLing 2019, La Rochelle, France.

Post, M. (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.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., and Khapra, M. S. (2022). Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Riza, H., Purwoadi, M., Uliniansyah, T., Ti, A. A., Aljunied, S. M., Mai, L. C., Thang, V. T., Thai, N. P., Chea, V., Sam, S., Seng, S., Khin Mar Soe, Khin Thandar Nwet, Utiyama, M., and Ding, C. (2016). Introduction of the asian language treebank. In *In Proc. of O-COCOSDA*, pages 1–6.

Sen, A., Parida, S., Kotwal, K., Panda, S., Bojar, O., and Dash, S. R. (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.

Soky, K., Mimura, M., Kawahara, T., Li, S., Ding, C., Chu, C., and Sam, S. (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 Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 122–127.

Thu, Y. K., Pa, W. P., Utiyama, M., Finch, A., and Sumita, E. (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).

Tseng, H. (2005). A conditional random field word segmenter. In *In Fourth SIGHAN Workshop on Chinese Language Processing*.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., and Uszkoreit, J. (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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (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.

Zhang, C., D'Haro, L. F., Banchs, R. E., Friedrichs, T., and Li, H. (2021a). *Deep AM-FM: Toolkit for Automatic Dialogue Evaluation*, pages 53–69. Springer Singapore, Singapore.

Zhang, C., Lee, G., D'Haro, L. F., and Li, H. (2021b). D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1.

Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (2020). Double attention-based multimodal neural machine translation with semantic image regions. In *EAMT*, pages 105–114.

Appendix A Submissions

Tables 11 to 22 summarize translation results submitted to WAT2023. Type and RSRC columns indicate type of method and use of other resources.

System	ID	Type	RSRC	BLEU	RIBES	AMFM
BITS-P	7124	NMT	YES	52.10	0.853388	–
ODIAGEN	7106	NMT	NO	42.80	0.815156	–

Table 11: MMCHMM23 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
BITS-P	7126	NMT	YES	42.20	0.759248	–

Table 12: MMCHMM23 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
BITS-P	7122	NMT	YES	48.70	0.831946	–
ODIAGEN	7108	NMT	NO	30.50	0.690706	–

Table 13: MMCHMM23 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIAGEN	7088	NMT	NO	53.60	0.858033	–
ODIAGEN	7110	NMT	NO	53.10	0.854334	–

Table 14: MMCHTEXT23 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIAGEN	7112	NMT	NO	39.70	0.752401	–

Table 15: MMCHTEXT23 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIAGEN	7090	NMT	NO	47.80	0.821982	–

Table 16: MMCHTEXT23 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
BITS-P	7125	NMT	YES	45.00	0.829320	–
ODIAGEN	7105	NMT	NO	41.60	0.811420	–

Table 17: MMEVMM23 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
BITS-P	7127	NMT	YES	51.90	0.799683	–

Table 18: MMEVMM23 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
BITS-P	7123	NMT	YES	50.60	0.814207	–
ODIAGEN	7107	NMT	NO	42.40	0.763497	–

Table 19: MMEVMM23 en-bn submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIAGEN	7087	NMT	NO	44.60	0.829217	–
ODIAGEN	7109	NMT	NO	44.60	0.829213	–

Table 20: MMEVTEXT23 en-hi submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIAGEN	7091	NMT	NO	46.60	0.746474	–
ODIAGEN	7111	NMT	NO	46.20	0.737472	–

Table 21: MMEVTEXT23 en-ml submissions

System	ID	Type	RSRC	BLEU	RIBES	AMFM
ODIAGEN	7089	NMT	NO	49.20	0.797703	–

Table 22: MMEVTEXT23 en-bn submissions