# Rakuten's Participation in WAT 2021: Examining the Effectiveness of Pre-trained Models for Multilingual and Multimodal Machine Translation

**Raymond Hendy Susanto, Dongzhe Wang, Sunil Kumar Yadav, Mausam Jain, Ohnmar Htun**
Rakuten Institute of Technology
Rakuten Group, Inc.
{raymond.susanto,dongzhe.wang,sunilkumar.yadav,
mausam.jain,ohnmar.htun}@rakuten.com

## Abstract

This paper introduces our neural machine translation systems' participation in the WAT 2021 shared translation tasks (team ID: *sakura*). We participated in the (i) NICT-SAP, (ii) Japanese-English multimodal translation, (iii) Multilingual Indic, and (iv) Myanmar-English translation tasks. Multilingual approaches such as mBART (Liu et al., 2020) are capable of pre-training a complete, multilingual sequence-to-sequence model through denoising objectives, making it a great starting point for building multilingual translation systems. Our main focus in this work is to investigate the effectiveness of multilingual finetuning on such a multilingual language model on various translation tasks, including low-resource, multimodal, and mixed-domain translation. We further explore a multimodal approach based on universal visual representation (Zhang et al., 2019) and compare its performance against a unimodal approach based on mBART alone.

## 1 Introduction

This paper introduces our neural machine translation (NMT) systems' participation in the 8th Workshop on Asian Translation (WAT-2021) shared translation tasks (Nakazawa et al., 2021). We participated in the (i) NICT-SAP's IT and Wikinews, (ii) Japanese-English multimodal translation, (iii) Multilingual Indic, and (iv) Myanmar-English translation tasks.

Recent advances in language model pre-training have been successful in advancing the state-of-the-art in various natural language processing tasks. Multilingual approaches such as mBART (Liu et al., 2020) are capable of pre-training a full sequence-to-sequence model through multilingual denoising objectives, which leads to significant gains in downstream tasks, such as machine translation. Building upon our success with utilizing mBART25 in the 2020 edition of WAT (Wang and Htun, 2020), we put more focus on multilingual and multimodal translation this year. In particular, instead of performing *bilingual finetuning* on mBART for each language pair, we train a single, multilingual NMT model that is capable of translating multiple languages at once. As first proposed by Tang et al. (2020), we apply *multilingual finetuning* to mBART50 for the NICT-SAP task (involving 4 Asian languages) and Multilingual Indic task (involving 10 Indic languages). Our findings show the remarkable effectiveness of mBART pre-training on these tasks. On the Japanese-English multimodal translation task, we compare a unimodal text-based model, which is initialized based on mBART, with a multimodal approach based on universal visual representation (UVR) (Zhang et al., 2019). Last, we continue our work on Myanmar-English translation by experimenting with more extensive data augmentation approaches. Our main findings for each task are summarized in the following:

- **NICT-SAP task:** We exploited mBART50 to improve low-resource machine translation on news and IT domains by finetuning them to create a mixed-domain, multilingual NMT system.
- **Multimodal translation:** We investigated multimodal NMT based on UVR in the constrained setting, as well as a unimodal text-based approach with the pre-trained mBART model in the unconstrained setting.
- **Multilingual Indic task:** We used the pre-trained mBART50 models, extended them for various Indic languages, and finetuned them on the entire training corpus followed by finetuning on the PMI dataset.



*Proceedings of the 8th Workshop on Asian Translation*, pages 96–105
Bangkok, Thailand (online), August 5-6, 2021. ©2021 Association for Computational Linguistics

| Split | Domain | Language | | | |
|---|---|---|---|---|---|
| | | 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 1: Statistics of the NICT-SAP datasets. Each language is paired with English.

- **Myanmar-English translation:** We designed contrastive experiments with different data combinations for Myanmar↔English translation and validated the effectiveness of data augmentation for low-resource translation tasks.

## 2 NICT-SAP Task

### 2.1 Task Description

This year, we participated in the NICT-SAP translation task, which involves two different domains: IT domain (Software Documentation) and Wikinews domain (ALT). These are considered low-resource domains for Machine Translation, combined with the fact that it involves four low-resource Asian languages: Hindi (hi), Indonesian (id), Malay (ms), and Thai (th).

For training, we use parallel corpora from the Asian Language Treebank (ALT) (Thu et al., 2016) for the Wikinews domain and OPUS[1] (GNOME, KDE4, and Ubuntu) for the IT domain. For development and evaluation, we use the datasets provided by the organizer: SAP software documentation (Buschbeck and Exel, 2020)[2] and ALT corpus.[3] Table 1 shows the statistics of the datasets.

### 2.2 Data Processing

We tokenized our data using the 250,000 SentencePiece model (Kudo and Richardson, 2018) from mBART (Liu et al., 2020), which was a joint vocabulary trained on monolingual data for 100 languages from XLMR (Conneau et al., 2020). Moreover, we prepended each source sentence with a domain indicator token to distinguish the ALT (<2alt>) and IT domain (<2it>).

We collect parallel corpora from all the language pairs involved in this task, namely {hi,id,ms,th}↔en. Following mBART, we prepend source and target language tokens to each source and target sentences, respectively. The size of each dataset varies across language pairs. For instance, the size of the Malay training corpus for the IT domain is roughly 5× larger than that of Thai. To address this data imbalance, we train our model with a temperature-based sampling function following Arivazhagan et al. (2019):

$$p_{i,j} \propto \left( \frac{|B_{i,j}|}{\sum_{i,j} |B_{i,j}|} \right)^{1/T}$$

where $B_{i,j}$ corresponds to the parallel corpora for a language pair $(i, j)$ and $T$ the temperature for sampling.

### 2.3 Model

We use the pre-trained mBART50 model (Tang et al., 2020) as our starting point for finetuning our translation systems. Unlike the original mBART work that performed bilingual finetuning (Liu et al., 2020), Tang et al. (2020) proposed *multilingual finetuning* where the mBART model is finetuned on many directions at the same time, resulting in a single model capable of translating many languages to many other languages. In addition to having more efficient and storage maintenance benefits, such an approach greatly helps low-resource language pairs where little to no parallel corpora are available.

While the mBART50 has great coverage of 50 languages, we found that it does not include all languages involved in this task, particularly Malay. Following Tang et al. (2020), who extended mBART25 to create mBART50, we extended mBART50's embedding layers with one additional randomly initialized vector for the Malay language token.[4] We use the same model architecture as mBART50, which is based on Transformer (Vaswani et al., 2017). The model was finetuned for 40,000 steps with Adam (Kingma and Ba, 2015) using $\beta 1 = 0.9$, $\beta 2 = 0.98$, and $\epsilon = 1e^{-6}$. We use a maximum batch size of 512 tokens and gradients were accumulated every 4 mini-batches on each GPU. We ran our experiments on 4 NVIDIA

[1] https://opus.nlpl.eu/

[2] https://github.com/SAP/software-documentation-data-set-for-machine-translation

[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/NICT-SAP-Task/altsplits-sap-nict.zip

[4] Our modifications to the original mBART code are accessible at https://github.com/raymondhs/fairseq-extensible-mbart.

| Domain | System | Translation Direction | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | en→hi | hi→en | en→id | id→en | en→ms | ms→en | en→th | th→en |
| ALT | Dabre and Chakrabarty (2020) | 24.23 | 12.37 | 32.88 | 17.39 | 36.77 | 18.03 | 42.13 | 10.78 |
| | mBART50 - pre-trained | 29.79 | 32.27 | 39.07 | 42.62 | 41.74 | 43.36 | 54.15 | 28.02 |
| | mBART50 - ft.nn | 34.00 | 35.75 | 41.47 | 44.09 | 43.92 | 45.14 | 55.87 | 29.70 |
| | +ensemble of 3* | 34.25 | 36.17 | 41.57 | 44.72 | 44.01 | 45.70 | 55.98 | 30.10 |
| IT | Dabre and Chakrabarty (2020) | 14.03 | 16.89 | 32.52 | 25.95 | 34.62 | 26.33 | 28.24 | 10.00 |
| | mBART50 - pre-trained | 26.03 | 36.38 | 43.97 | 43.17 | 40.15 | 39.37 | 52.67 | 25.06 |
| | mBART50 - ft.nn | 28.43 | 40.30 | 45.01 | 44.41 | 41.92 | 40.92 | 55.60 | 26.05 |
| | +ensemble of 3* | 28.50 | 40.17 | 45.39 | 44.70 | 42.26 | 40.97 | 55.64 | 26.30 |

Table 3: BLEU results on the NICT-SAP task. Our final submission is marked by an asterisk.

| Domain | Translation Direction | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | en→hi | hi→en | en→id | id→en | en→ms | ms→en | en→th | th→en |
| ALT | 84.92 | 83.29 | 86.80 | 85.10 | 87.19 | 85.15 | 83.71 | 82.26 |
| IT | 82.68 | 86.13 | 86.30 | 86.30 | 87.33 | 84.94 | 82.99 | 80.91 |

Table 4: AMFM results on the NICT-SAP task

| | |
|---|---|
| Vocab size | 250k |
| Embed. dim. | 1024 |
| Tied embed. | Yes |
| FFN dim. | 4096 |
| Attention heads | 16 |
| En/Decoder layers | 12 |
| Label smoothing | 0.2 |
| Dropout | 0.3 |
| Attention dropout | 0.1 |
| FFN dropout | 0.1 |
| Learning rate | $3e^{-5}$ |

Table 2: Models settings for both NICT-SAP and Multilingual Indic tasks

Quadro RTX 6000 GPUs. Table 2 shows the details of our experimental settings.

### 2.4 Results

Table 3 and Table 4 show our experimental results in terms of BLEU (Papineni et al., 2002) and AMFM (Banchs et al., 2015) scores, respectively. We first show our multilingual finetuning results on the released mBART50 model (Tang et al., 2020),[5] which was pre-trained as a denoising autoencoder on the monolingual data from XLMR (Conneau et al., 2020) (*mBART50 - pre-trained*). Compared to one submission from previous year's WAT from Dabre and Chakrabarty (2020), which is a multilingual many-to-many model without any pre-training, we observe a significant improvement from multilingual finetuning across all language pairs for both domains. For instance, we obtain the largest improvement of 25.23 BLEU points for id→en on the ALT domain. These findings clearly show that multilingual models greatly benefit from pre-training as compared to being trained from scratch, and more so for low resource languages.

Second, Tang et al. (2020) released a many-to-many multilingual translation that was finetuned from mBART on publicly available parallel data for 50 languages, including all language pairs in this task, except Malay. We adapt this model by performing a further finetuning on the NICT-SAP dataset (*mBART50 - ft.nn*). On average, this model further improves BLEU by 2.37 points on ALT and 1.98 points on IT.

Finally, we trained three independent models with different random seeds to perform ensemble decoding. This is our final submission, which achieves the first place in AMFM scores on this year's leaderboard for 7 translation directions for ALT (all except en→ms) and 6 directions for IT (all except for en→hi and en→id).

For the human evaluation on the IT task, our systems obtained 4.24 adequacy score for en→id and 4.05 for en→ms, which were the highest among all participants this year. We refer readers to the overview paper (Nakazawa et al., 2021) for the complete evaluation results.

## 3 Japanese↔English Multimodal Task

### 3.1 Task Description

Multimodal neural machine translation (MNMT) has recently received increasing attention in the NLP research fields with the advent of visually-grounded parallel corpora. The motivation of Japanese↔English multimodal task is to improve

[5] https://github.com/pytorch/fairseq/tree/master/examples/multilingual

translation performance with the aid of heterogeneous information (Nakazawa et al., 2020). In particular, we performed the experiments based on the benchmark Flickr30kEnt-JP dataset (Nakayama et al., 2020), where manual Japanese translations are newly provided to the Flickr30k Entities image captioning dataset (Plummer et al., 2015) that consists of 29,783 images for training and 1,000 images for validation, respectively. For each image, the original Flickr30k has five sentences, while the extended Flickr30kEnt-JP has corresponding Japanese translation in parallel[6].

In terms of input sources, this multimodal task has been divided into four sub-tasks: **constrained** and **unconstrained** Japanese↔English translation tasks. In the constrained setting, we investigated the MNMT models with universal visual representation (UVR) (Zhang et al., 2019), which is obtained from the pre-trained bottom-up attention model (Anderson et al., 2018). In contrast, we also explored the capability of unimodal translation (i.e., text modality only) under the unconstrained setting, where the pre-trained mBART25 model (Liu et al., 2020) was employed as the external resource.

### 3.2 Data Processing

**Text preparation** For the constrained setting, we firstly exploited Juman analyzer[7] for Japanese and Moses tokenizer for English. Then, we set the vocabulary size to 40,000 to train the byte-pair encoding (BPE)-based subword-nmt[8] (Sennrich et al., 2016) model. Moreover, we merged the source and target sentences and trained a joint vocabulary for the NMT systems. Under the unconstrained setting, we used the same 250,000 vocabulary as in the pre-trained mBART model for the text input to mBART finetuning, which was automatically tokenized with a SentencePiece model (Kudo and Richardson, 2018) based on BPE method.

**Universal visual retrieval** For the constrained setting particularly, we propose to extract the pre-computed global image features from the raw Flickr30k images using the bottom-up attention Faster-RCNN object detector that is pre-trained on the Visual Genome dataset (Krishna et al., 2017).

[6]During training, we dismissed the 32 out of 29,783 training images having blank Japanese sentences, which ended up with 148,756 lines of Japanese↔English bitext.

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

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

| Models | MNMT | mBART |
|---|---|---|
| Vocabulary size | 40k | 250k |
| Embedding dim. | 1024 | 1024 |
| Image dim. | 2048 | - |
| Tied embeddings | Yes | Yes |
| FFN dim. | 4096 | 4096 |
| Attention heads | 16 | 16 |
| En/Decoder layers | 12 | 12 |
| Label smoothing | 0.1 | 0.2 |
| Dropout | 0.3 | 0.3 |
| Attention dropout | 0.1 | 0.1 |
| FFN dropout | 0.1 | 0.1 |
| Learning rate | $5e^{-4}$ | $3e^{-5}$ |

Table 4: Multimodal model parameter settings

Specifically, we adopted the pre-trained model[9] to extract the spatial image features corresponding to 36 bounding boxes regions per image, which were then encoded into a global image feature vector by taking the global average pooling of them. In practice, we followed (Zhang et al., 2019) and presented the UVR relying on image-monolingual annotations (i.e., source sentences). To retrieve the universal visual information from the source sentences, the sentence-image pairs have been transformed into two topic-image lookup tables from the Flickr30kEnt-JP dataset for Japanese→English and English→Japanese tasks, respectively. Note that no image information has been learned in our unconstrained models due to the text-only property.

### 3.3 Model

In this section, we will elaborate on our proposed model architectures for the constrained and unconstrained tasks, respectively.

**Multimodal model with UVR** Following (Zhang et al., 2019), we built the multimodal models based on the standard Transformer (Vaswani et al., 2017) with an additional cross-attention layer in the encoder, followed by a gating mechanism that fused the visual modality and text modality information. In particular, visual representation retrieved from the topic-image lookup table has been encoded by a self-attention network that is in parallel with the source sentence encoder. Then, a cross attention mechanism has been applied to append

[9]Download from https://storage.googleapis.com/up-down-attention/resnet101_faster_rcnn_final.caffemodel

the image representation to the text representation. Using a learnable weighting gate $\lambda \in [0, 1)$, we obtained the aggregated multimodal representation corresponding to the significance distribution of either modality, which would be used as input to the decoder for predicting target translations. The hyper-parameter setting is shown in Table 4.

**mBART25 finetuning** Regardless of the image representation, we also finetuned on the Flickr30kEnt-JP corpus using the mBART25 pretrained model under the unconstrained task setting. Following (Liu et al., 2020), we used the same mBART25-large model[10] and finetuned for 40,000 steps with early stopping control if the validation loss has not been improved for 3 iterations. We used the learning rate schedule of 0.001 and maximum of 4000 tokens in a batch, where the parameters were updated after every 2 epochs. More details of model hyper-parameters setting can be found in Table 4.

We trained the MNMT models and finetuned the mBART25 models using the Fairseq toolkit (Ott et al., 2019) on 4 V100 GPUs. Finally, the best performing models on the validation sets were selected and applied for decoding the test sets. Furthermore, we trained three independent models with different random seeds to perform ensemble decoding.

### 3.4 Results

In Table 5, we show the evaluation scores that the multimodal NMT with universal visual representation and mBART25 finetuning models achieve. In the constrained setting (a.k.a, task (a)), we observed that the MNMT single model (MNMT$_{sin.}$) decoding results unexceptionally lagged behind that of the ensemble decoding (MNMT$_{ens.}$) in both directions. Without any other resources except pretrained image features, our best submissions of NNMT with UVR win the first place in BLEU as well as human adequacy scores on the WAT leaderboard for the Japanese→English task (a). Moreover, the MNMT$_{ens.}$ model can outperform the mBART25 finetuning model (mBART$_{sin.}$) using external models/embeddings by 0.17 BLEU score in the English→Japanese task (a), which validates the effectiveness of exploring visual information for machine translation.

Under the unconstrained setting, the text-only mBART$_{sin.}$ models achieved significant improvement over the MNMT (UVR) single models by 0.83 and 3.47 BLEU scores in the English→Japanese and Japanese→English tasks, respectively. Compared with other submissions, our mBART$_{sin.}$ model decoding achieve the first place in both BLEU scores and AMFM scores on the WAT leaderboard for the Japanese→English (b). It indicates that the advantages of pre-training are substantial in the Flickr30kEnt-JP translation tasks, in spite of the help of another modality (i.e., images) associated to the input sentences.

| Task | Model | BLEU | AMFM | Human |
|---|---|---|---|---|
| en-ja (a) | MNMT$_{sin.}$ | 42.09 | - | - |
| en-ja (a) | MNMT$_{ens.}$ | **43.09** | - | 4.67 |
| en-ja (b) | mBART$_{sin.}$ | 42.92 | 64.83 | - |
| ja-en (a) | MNMT$_{sin.}$ | 51.53 | - | - |
| ja-en (a) | MNMT$_{ens.}$ | 52.20 | - | 4.54 |
| ja-en (b) | mBART$_{sin.}$ | **55.00** | 58.00 | - |

Table 5: Comparisons of MNMT with UVR and mBART25 finetuning best models results in the Japanese↔English multimodal task: (a) constrained setting, (b) unconstrained setting. Note that the human evaluation scores shown in the table are referred to be the adequacy scores.

## 4 Multilingual Indic Task

### 4.1 Task Description

The Multilingual Indic task covers English (en) and 10 Indic (in) Languages: Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Oriya (or), Punjabi (pa), Tamil (ta) and Telugu (te). Multilingual solutions spanning 20 translation directions, en↔in were encouraged in form of many2many, one2many and many2one models. We train one2many for en→in and many2one for in→en directions.

We use the parallel corpora provided by the organizer for training, validation, and evaluation. Table 6 shows the statistics of the entire training data and PMI dataset specific statistics (Haddow and Kirefu, 2020).

### 4.2 Data Processing

We normalize entire Indic language data using Indic NLP Library[11] version 0.71. After that, we use the 250,000-token SentencePiece model from mBART and prepend source and target tokens to

[10] https://github.com/pytorch/fairseq/blob/master/examples/mbart/

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

| | Language | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **bn** | **gu** | **hi** | **kn** | **ml** | **mr** | **or** | **pa** | **ta** | **te** |
| Train | 1,756,197 | 518,015 | 3,534,387 | 396,865 | 1,204,503 | 781,872 | 252,160 | 518,508 | 1,499,441 | 686,626 |
| - PMI | 23,306 | 41,578 | 50,349 | 28,901 | 26,916 | 28,974 | 31,966 | 28,294 | 32,638 | 33,380 |
| Dev | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |
| Test | 2,390 | 2,390 | 2,390 | 2,390 | 2,390 | 2,390 | 2,390 | 2,390 | 2,390 | 2,390 |

Table 6: Statistics of the Multilingual Indic datasets. Each language is paired with English. The PMI dataset is used for adaptation.

| **Direction** | **System** | **Indic Language** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **bn** | **gu** | **hi** | **kn** | **ml** | **mr** | **or** | **pa** | **ta** | **te** |
| en2in | ORGANIZER | 5.58 | 16.38 | 23.31 | 10.11 | 3.34 | 8.82 | 9.08 | 21.77 | 6.38 | 2.80 |
| | mBART50 - ft.1n | 11.09 | 23.25 | 35.57 | 13.57 | 10.94 | 15.99 | 17.81 | 29.37 | 12.58 | 11.86 |
| | +adaptation on PMI | 13.83 | 25.27 | 36.92 | 18.83 | 8.13 | 17.87 | 17.88 | 30.93 | 13.25 | 15.48 |
| in2en | ORGANIZER | 11.27 | 26.21 | 28.21 | 20.33 | 13.64 | 15.10 | 16.35 | 23.66 | 16.07 | 14.70 |
| | mBART50 - ft.nn | 26.69 | 38.73 | 41.58 | 34.11 | 32.23 | 31.76 | 32.67 | 40.38 | 31.09 | 33.87 |
| | +adaptation on PMI | 27.92 | 39.27 | 42.61 | 35.46 | 33.21 | 32.06 | 32.82 | 41.18 | 31.94 | 35.44 |

Table 7: BLEU results on the Multilingual Indic task

each source and target sentence, respectively. We then binarize the data using Fairseq (Ott et al., 2019) framework. Following Section 2.2, we also train with temperature-based sampling to address dataset imbalance.

### 4.3 Model

Similar to our use of the pre-trained mBART50 model from Section 2.3, we use multilingual finetuning and model extension for Oriya, Punjabi, and Kannada using randomly initialized vectors. We use the same model architecture as mBART50 and run Adam optimization using $\beta 1 = 0.9$, $\beta 2 = 0.98$, and $\epsilon = 1e^{-6}$. We use a maximum batch size of 512 tokens and gradients were accumulated every 4 mini-batches on each GPU. We ran our experiments on 8 NVIDIA V100 GPUs. Table 2 shows the details of our experimental settings.

We finetune one2many pre-trained mBART50 (*mBART50 - ft.1n*) for en→in on entire training set for six epochs. We further adapt this model on PMI dataset given as part of the training set for nine epochs. Similarly, we finetune many2many pre-trained mBART50 (*mBART50 - ft.nn*) for in→en on entire training set for six epochs and adaptation on PMI dataset for one epoch.

### 4.4 Results

Table 7 shows our experimental results in terms of BLEU scores. As a baseline, we compare our models with the organizer's bilingual base Transformer model trained on the PMI dataset (*ORGANIZER*). We observe an average improvement of 7.4 BLEU points over this baseline across all en→in pairs by finetuning the *mBART50 - ft.1n* model for 6 epochs. Further adaptation on the PMI dataset for 12 epochs results in an average improvement of 1.6 BLEU points. For en→ml, we observe a drop from 10.94 to 8.13 on adaptation. Similarly, we observe an average improvement of 15.76 BLEU points over baseline across all in→en pairs by finetuning the *mBART50 - ft.nn* model for 4 epochs. Further adaptation on the PMI dataset for a single epoch results in an average improvement of 0.88 BLEU points. Table 8 and 9 show official AMFM and human evaluation results (top three systems for ten translation directions) respectively. Our systems ranked second 6 times out of the 10 directions for which human evaluation results are available, while SRPOL has consistently outperformed all systems. This demonstrates the efficacy of using mBART models for multilingual models. Complete evaluation results are available in the overview paper (Nakazawa et al., 2021).[12]

## 5 Myanmar-English Translation Task

### 5.1 Task Description

In the ALT+ tasks, we conducted experiments on the Myanmar-English parallel data which was provided by the organizers and consist of two corpora, the ALT corpus (Ding et al., 2019, 2020) and UCSY corpus (Yi Mon Shwe Sin and Khin Mar Soe, 2018). The ALT corpus consists of 18,088 training sentences, 1,000 validation sentences, and 1,018 test sentences. The UCSY dataset contains 204,539 training sentences. The quality of the UCSY corpus used in WAT2021 was improved by correcting

[12]Our training scripts are available at https://github.com/sukuya/indic-mnmt-wat2021-sakura.

| Direction | System | Indic Language | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | bn | gu | hi | kn | ml | mr | or | pa | ta | te |
| en2in | ORGANIZER | 70.15 | 75.71 | 75.97 | 74.19 | 70.68 | 73.07 | 71.45 | 76.24 | 72.32 | 70.81 |
| | mBART50 - ft.1n | 73.77 | 81.02 | 81.09 | 80.19 | 79.45 | 79.09 | 76.74 | 80.14 | 79.11 | 77.21 |
| | +adaptation on PMI | 76.47 | 81.34 | 81.70 | 81.78 | 80.19 | 80.36 | 76.99 | 80.22 | 79.57 | 78.51 |
| in2en | ORGANIZER | 61.31 | 72.66 | 73.61 | 69.20 | 64.66 | 65.81 | 73.08 | 70.15 | 67.60 | 63.60 |
| | mBART50 - ft.nn | 77.24 | 82.07 | 83.42 | 80.51 | 80.55 | 79.58 | 80.82 | 82.35 | 79.61 | 80.20 |
| | +adaptation on PMI | 77.29 | 81.86 | 83.45 | 80.97 | 80.68 | 79.55 | 80.60 | 82.34 | 79.04 | 80.40 |

Table 8: AMFM results on the Multilingual Indic task

| Direction | Rank | | |
|---|---|---|---|
| | I | II | III |
| en→bn | 4.65 (SRPOL) | **4.39** (sakura) | 3.94 (IIITH) |
| bn→en | 4.80 (SRPOL) | 3.82 (IIITH) | 3.59 (mcairt) |
| en→kn | 4.72 (SRPOL) | **4.57** (sakura) | 4.00 (IIITH) |
| kn→en | 4.72 (SRPOL) | **4.49** (sakura) | 3.94 (IIITH) |
| en→ml | 4.41 (SRPOL) | 3.54 (CFILT) | 2.72 (IIITH) |
| ml→en | 4.03 (SRPOL) | **3.99** (sakura) | 3.71 (IITP-MT) |
| en→mr | 4.34 (SRPOL) | 4.14 (CFILT) | 3.84 (IIITH) |
| mr→en | 4.57 (SRPOL) | **4.35** (sakura) | 4.01 (IIITH) |
| en→or | 4.26 (SRPOL) | 3.82 (IIITH) | 3.76 (CFILT) |
| or→en | 4.37 (SRPOL) | **4.25** (sakura) | 3.42 (IIITH) |

Table 9: Human evaluation results for the top three systems on the Multilingual Indic task. Bold values represent our system.

| Dataset | English | Myanmar |
|---|---|---|
| $P_1$ | original | original |
| $P_2$ | clean + tokenize | original |
| $P_3$ | clean | clean |
| $P_4$ | clean | clean + word tokenize |
| $P_5$ | clean | clean + syllable tokenize |
| $P_6$ | clean + tokenize | clean + word tokenize |
| $P_7$ | clean + tokenize | clean + syllable tokenize |

Table 10: Preprocessing variations for the Myanmar-English dataset

translation mistakes, spelling errors, and typographical errors.[13] The model was trained and evaluated by using the dataset provided by the organizer, mainly for research around simple hyperparameter tuning of Marian NMT (Junczys-Dowmunt et al., 2018) without any additional data.

### 5.2 Data Processing

For the ALT+ tasks, the ALT and UCSY training datasets were merged first. For cleaning, we removed redundant whitespaces and double quotation marks. We tokenized English sentences using Moses (Koehn et al., 2007) and Myanmar sentences using Pyidaungsu Myanmar Tokenizer[14] with syllable and word level segments, which were then fed into a SentencePiece model to produce subword units. Slightly different from previous approach (Wang and Htun, 2020), we generated three English datasets with different types: (i) original, (ii) clean, and (iii) clean and tokenized versions. For Myanmar, we have four types: (i) original, (ii) clean, (iii) word-level tokenized, and (iv) syllable-level tokenized. Table 10 describes the resulting datasets with different preprocessing steps.

### 5.3 Model

For training, we generated multiple training datasets by using different combinations of the datasets in Table 10:

- $D_1 = \{P_1\}$
- $D_2 = \{P_1, P_2, P_6, P_7\}$
- $D_3 = \{P_1, P_3, P_4, P_6, P_7\}$
- $D_4 = \{P_3, P_4, P_6, P_7\}$

For both directions on each dataset, we trained individual Transformer models using the Marian[15] toolkit. We created two different parameter configurations as shown in Table 11. We used the first configuration (*Config. 1*) on $D_1$ and the second configuration (*Config. 2*) on the rest ($D_2$, $D_3$, and $D_4$). Note that our second configuration has a larger vocabulary size and increased regularization (dropout, label smoothing). All experimental models in this task were trained on 3 GP104 machines with 4 GeForce GTX 1080 GPUs in each, and the experimental results will be shown and analyzed in the following section.

### 5.4 Results

Table 12 presents the results of our experiments on the given ALT test dataset evaluation for two directions. As our baseline, we trained on the original training set ($D_1$) without further preprocessing and using the first model configuration. After using data augmentation, we observed consistent

[13] http://lotus.kuee.kyoto-u.ac.jp/WAT/my-en-data/

[14] https://github.com/kaunghtetsan275/pyidaungsu

[15] https://marian-nmt.github.io

| Models | Config. 1 | Config. 2 |
|---|---|---|
| Vocabulary size | 160k | 380k |
| Embedding dim. | 1024 | 1024 |
| Tied embeddings | Yes | Yes |
| Transformer FFN dim. | 4096 | 4096 |
| Attention heads | 8 | 8 |
| En/Decoder layers | 4 | 4 |
| Label smoothing | 0.1 | 0.2 |
| Dropout | 0.1 | 0.2 |
| Batch size | 12 | 12 |
| Attention weight dropout | 0.1 | 0.2 |
| Transformer FFN dropout | 0.1 | 0.2 |
| Learning rate | $1e^{-3}$ | $1e^{-4}$ |
| Learning rate warmup | 8000 | 16000 |
| Trained positional embeddings | No | Yes |

Table 11: Myanmar-English model parameter settings

improvements in BLEU scores in any combination. This indicates that proper preprocessing steps such as cleaning and tokenization are crucial for this task. On en-my, we obtained the highest BLEU of 29.62 when training on $D_4$, which does not include the original segments $P_1$. On my-en, however, the highest BLEU is achieved on $D_2$, i.e., 19.75. It includes the cleaning and tokenization steps, particularly on the English side. Any forms of tokenization, be it word-level or syllable-level, appear to be helpful for Myanmar. Our best submission obtained the 6$^{\text{th}}$ place on the en-my leaderboard and the 5$^{\text{th}}$ place on my-en.

| Task | Dataset | Config. | BLEU |
|---|---|---|---|
| ALT+ en-my | $D_1$ | 1 | 21.70 |
| ALT+ en-my | $D_2$ | 2 | 29.25 |
| ALT+ en-my | $D_3$ | 2 | 29.07 |
| ALT+ en-my | $D_4$ | 2 | **29.62** |
| ALT+ my-en | $D_1$ | 1 | 14.80 |
| ALT+ my-en | $D_2$ | 2 | **19.75** |
| ALT+ my-en | $D_3$ | 2 | 18.70 |
| ALT+ my-en | $D_4$ | 2 | 18.50 |

Table 12: Results on the Myanmar-English translation task

## 6 Conclusion

We presented our submissions (team ID: *sakura*) to the WAT 2021 shared translation tasks in this paper. We showed the remarkable effectiveness of pre-trained models in improving multilingual and multimodal neural machine translation. On multilingual translation, models initialized with mBART50 achieved substantial performance gains on both NICT-SAP and Multilingual Indic tasks. On multimodal translation, a text-only model with mBART25 pre-training improves upon an MNMT model based on UVR. Finally, we extended our data augmentation approaches on the Myanmar-English translation tasks and obtained further improvements.

## References


Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges.

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

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 160–169, Suzhou, China. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre and Abhisek Chakrabarty. 2020. NICT's submission to WAT 2020: How effective are simple many-to-many neural machine translation models? In *Proceedings of the 7th Workshop on Asian Translation*, pages 98–102, Suzhou, China. Association for Computational Linguistics.

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, Sann Su Su Yee, Win Pa Pa, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. A Burmese (Myanmar) treebank: Guildline

and analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3):40.

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India. *arXiv e-prints*, page arXiv:2001.09907.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4204–4210, Marseille, France. European Language Resources Association.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, OndÅ™ej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. 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, et al. 2020. Overview of the 7th workshop on asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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 *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

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).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dongzhe Wang and Ohnmar Htun. 2020. Goku's participation in WAT 2020. In *Proceedings of the 7th Workshop on Asian Translation*, pages 135–141, Suzhou, China. Association for Computational Linguistics.

Yi Mon Shwe Sin and Khin Mar Soe. 2018. Syllable-based myanmar-english neural machine translation. In *Proc. of ICCA*, pages 228–233.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.