NICT-2 Translation System at WAT-2021: Applying a Pretrained Multilingual Encoder-Decoder Model to Low-resource Language Pairs

Kenji Imamura and **Eiichiro Sumita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{kenji.imamura,eiichiro.sumita}@nict.go.jp

Abstract

In this paper, we present the NICT system (NICT-2) submitted to the NICT-SAP shared task at the 8th Workshop on Asian Translation (WAT-2021). A feature of our system is that we used a pretrained multilingual BART (Bidirectional and Auto-Regressive Transformer; mBART) model. Because publicly available models do not support some languages in the NICT-SAP task, we added these languages to the mBART model and then trained it using monolingual corpora extracted from Wikipedia. We fine-tuned the expanded mBART model using the parallel corpora specified by the NICT-SAP task. The BLEU scores greatly improved in comparison with those of systems without the pretrained model, including the additional languages.

1 Introduction

In this paper, we present the NICT system (NICT-2) that we submitted to the NICT-SAP shared task at the 8th Workshop on Asian Translation (WAT-2021) (Nakazawa et al., 2021). Because the NICT-SAP task expects to perform translations with little parallel data, we developed a system to improve translation quality by applying the following models and techniques.

Pretrained model: An encoder-decoder model pretrained using huge monolingual corpora was used. We used a multilingual bidirectional auto-regressive Transformer (mBART) (i.e., multilingual sequence-to-sequence denoising auto-encoder (Liu et al., 2020)) model, which supports 25 languages. Because it includes English and Hindi, but does not include Indonesian, Malay, and Thai, we expanded it to include the unsupported languages and additionally pretrained it on these five languages.[1]

Multilingual models: We tested multilingual models trained using multiple parallel corpora to increase resources for training.

Domain adaptation: We tested two domain adaptation techniques. The first technique is training multi-domain models. Similar to multilingual models, this technique trains a model using the parallel corpora of multiple domains. The domains are identified by domain tags in input sentences. The second technique is adaptation based on fine-tuning. This method fine-tunes each domain model (using its domain corpus) from a model trained by a mixture of multi-domain corpora.

Our experimental results showed that the pretrained encoder-decoder model was effective for translating low-resource language pairs. However, the effects of multilingual models and domain adaptation became low when we applied the pretrained model.

The following sections are organized as follows. We first summarize the NICT-SAP shared task in Section 2, and briefly review the pretrained mBART model in Section 3. Details of our system is explained in Section 4. In Section 5, we present experimental results. Finally, we conclude our paper in Section 6.

2 NICT-SAP Shared Task

The NICT-SAP shared task was to translate text between English and four languages, that is, Hindi (Hi), Indonesian (Id), Malay (Ms), and Thai (Th), for which the amount of data in parallel corpora is relatively low. The task contained two domains.

The data in the Asian Language Translation (ALT) domain (Thu et al., 2016) consisted of translations obtained from WikiNews. The ALT

[1] The mBART-50 model (Tang et al., 2020) supports 50 languages including Indonesian and Thai. However, Malay is not supported by either the mBART model or mBART-50 models. Therefore, we applied additional pretraining to the mBART model.

Proceedings of the 8th Workshop on Asian Translation, pages 90–95
Bangkok, Thailand (online), August 5-6, 2021.

Domain	Set	En-Hi	En-Id	En-Ms	En-Th
ALT	Train	18,088	18,087	18,088	18,088
	Dev		1,000		
	Test		1,018		
IT	Train	252,715	158,200	504,856	73,829
	Dev	2,016	2,023	2,050	2,049
	Test	2,073	2,037	2,050	2,050

Table 1: Data sizes for the NICT-SAP task after filtering.

Language	#Sentences	#Tokens
English (En)	7,000,000 (*1)	174M
Hindi (Hi)	1,968,984	51M
Indonesian (Id)	6,997,907	151M
Malay (Ms)	2,723,230	57M
Thai (Th)	2,233,566 (*2)	60M

Table 2: Statistics of the training data for the pretrained model. The number of tokens indicates the number of subwords. (*1) The English data were sampled from 150M sentences to fit the number of sentences into the maximum number of the other languages. (*2) Sentences in Thai were detected using an in-house sentence splitter.

data is a multilingual parallel corpus, that is, it contains the same sentences in all languages. The training, development, and test sets were provided from the WAT organizers.

The data in the IT domain consisted of translations of software documents. The WAT organizers provided the development and test sets (Buschbeck and Exel, 2020). For the training set, we obtained GNOME, KDE, and Ubuntu sub-corpora from the OPUS corpus (Tiedemann, 2012). Therefore, the domains for the training and dev/test sets were not identical.

The data sizes are shown in Table 1. There were fewer than 20K training sentences in the ALT domain. Between 73K and 504K training sentences were in the IT domain. Note that there were inadequate sentences in the training sets. We filtered out translations that were longer than 512 tokens, or where source/target sentences were three times longer than the target/source sentences if they had over 20 tokens.

3 mBART Model

In this section, we briefly review the pretrained mBART model (Liu et al., 2020).

The mBART model is a multilingual model of bidirectional and auto-regressive Transformers (BART; (Lewis et al., 2020)). The model is based on the encoder-decoder Transformer (Vaswani et al., 2017), in which the decoder uses an auto-regressive method (Figure 1).

Two tasks of BART are trained in the mBART model. One is the token masking task, which restores masked tokens in input sentences. The other is the sentence permutation task, which predicts the original order of permuted sentences. Both tasks learn using monolingual corpora.

To build multilingual models based on BART, mBART supplies language tags (as special tokens) at the tail of the encoder input and head of the decoder input. Using these language tags, a mBART model can learn multiple languages.

The published pretrained mBART model[2] consists of a 12-layer encoder and decoder with a model dimension of 1,024 on 16 heads. This model was trained on 25 languages in the Common Crawl corpus (Wenzek et al., 2019). Of the languages for the NICT-SAP task, English and Hindi are supported by the published mBART model, but Indonesian, Malay, and Thai are not supported.

The tokenizer for the mBART model uses byte-pair encoding (Sennrich et al., 2016) of the SentencePiece model (Kudo and Richardson, 2018)[3]. The vocabulary size is 250K subwords.

4 Our System

4.1 Language Expansion/Additional Pretraining of mBART

As described above, the published mBART model does not support Indonesian, Malay, and Thai. We expanded the mBART model to support these three languages, and additionally pretrained the model on the five languages in the NICT-SAP task.

The corpus for additional pretraining was extracted from Wikipedia dump files as follows. Unlike the XLM models (Lample and Conneau, 2019), which were also pretrained using Wikipedia corpora, we divided each article into sentences in our corpus, to train the sentence permutation task. Additionally, we applied sentence filtering to clean each language.

1. First, Wikipedia articles were extracted from

[2] https://dl.fbaipublicfiles.com/fairseq/models/mbart/mbart.cc25.v2.tar.gz

[3] https://github.com/google/sentencepiece

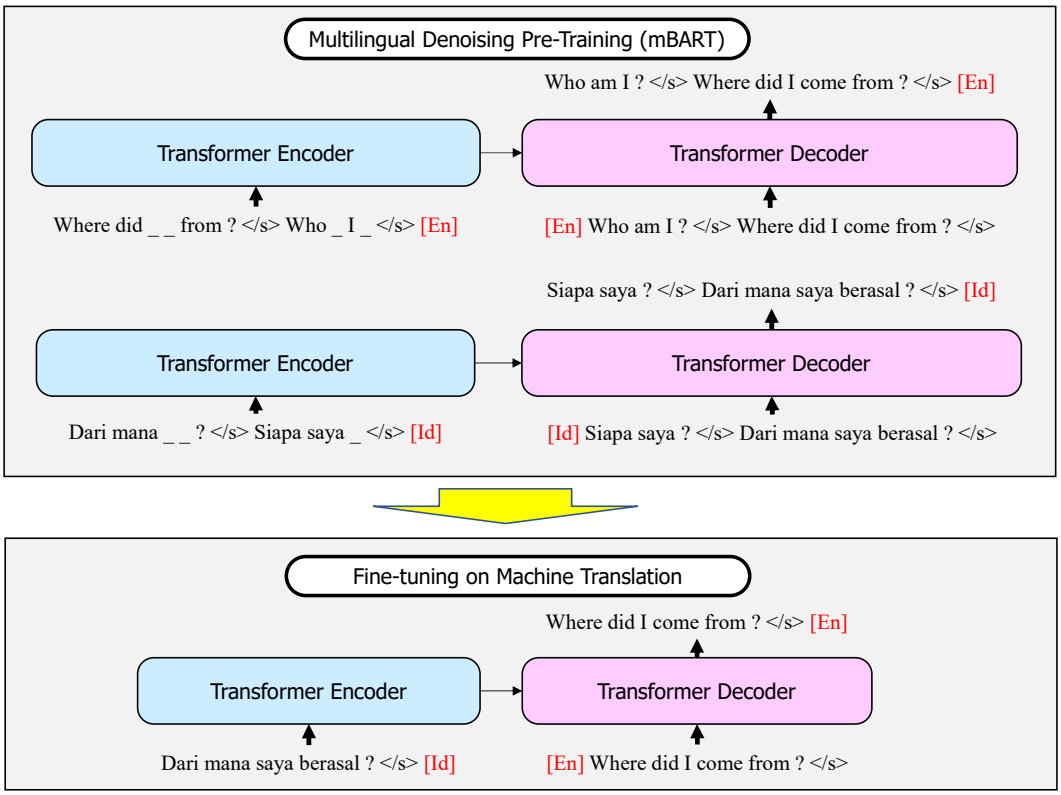

Figure 1: Example of mBART pretraining and fine-tuning for the machine translation task from Indonesian to English (arranged from (Liu et al., 2020)).

the dump files using `WikiExtractor`[4] while applying the NFKC normalization of Unicode.

2. Sentence splitting was performed based on sentence end marks, such as periods and question marks. However, because Thai does not have explicit sentence end marks, we applied a neural network-based sentence splitter (Wang et al., 2019), which was trained using in-house data.

3. We selected valid sentences, which we regarded as sentences that consisted of five to 1,024 letters, where and 80% of the letters were included in the character set of the target language. In the case of Hindi, for example, we regarded a sentence as valid if 80% of the letters were in the set of Devanagari code points, digits, and spaces.

The number of sentences for the mBART additional pretraining is shown in Table 2. We sampled 7M English sentences to balance the sizes of the other languages because the number of English sentences was disproportionately large (about 150M sentences).

We first expanded the word embeddings of the published mBART large model using random initialization and trained it. This is similar to the training procedure for mBART-50 (Tang et al., 2020), except for the corpora and hyperparameters. The settings for the additional pretraining are shown in Table 3.

We conducted the additional pretraining using the Fairseq translator (Ott et al., 2019)[5] on eight NVIDIA V100 GPUs. It took about 15 days.

For the tokenizer, we used the SentencePiece model in the published mBART large model. This model does not support Indonesian, Malay, and Thai. Indonesian and Malay use Latin characters, hence we divert the model to tokenize these languages. Thai uses a special character set. However, we diverted the SentencePiece model because almost all characters in the Thai corpus were included in the vocabulary of the model.

[4] https://github.com/attardi/wikiextractor

[5] https://github.com/pytorch/fairseq

Attribute	Value
LR	0.0003
Warm-up	Linear warm-up in 10K updates
Decay	Linear decay
Tokens per sample	512
Batch size	640K tokens
# updates	500K (around 77 epochs)
Dropout schedule	0.10 until 250K updates, 0.05 until 400K updates, and 0.0 until 500K updates
Loss function	Cross entropy
Token masking	mask=0.3, mask-random=0.1, mask-length=span-poisson, poisson-lambda=3.5, replace-length=1
Sentence permutation	permute-sentence=1.0

Table 3: Hyperparameters for the mBART additional pretraining

Phase	Attribute: Value
Fine-tuning	LR: 0.00008, Dropout: 0.3, Batch size: 16K tokens, Loss function: label smoothed cross entropy, Warm-up: linear warm-up in five epochs, Decay: invert square-root, Stopping criterion: early stopping on the dev. set.
Translation	Beam width: 10, Length penalty: 1.0.

Table 4: Hyper-parameters for fine-tuning and translation.

4.2 Other Options

We fine-tuned the pretrained model using the NICT-SAP parallel corpora shown in Table 1. We also used Transformer base models (six layers, the model dimension of 512 on 8 heads) for comparison without the pretrained model. In addition to the effect of the pretrained models, we investigated the effects of multilingual models and domain adaptation.

4.2.1 Multilingual Models

Similar to the multilingual training of mBART, the multilingual model translated all the language pairs using one model by supplying source and target language tags to parallel sentences.

By contrast, bilingual models were trained using the corpora of each language pair. When we use the mBART model, we supplied source and target language tags to parallel sentences, even for the bilingual models.

4.2.2 Domain Adaptation

We tested two domain adaptation methods; multi-domain models and fine-tuning-based methods. Both methods utilize parallel data of the other domains.

Similar to the multilingual models, we trained the multi-domain models by supplying domain tags (this time, we used <__WN__> for the ALT domain and <__IT__> for the IT domain) at the head of sentences in the source language.

The fine-tuning method did not use domain tags. First, a mixture model was trained using a mixture of multiple domain data. Next, domain models were fine-tuned from the mixture model using each set of domain data. Therefore, we created as many domain models as the number of domains.

5 Experiments

The models and methods described above were fine-tuned and tested using the hyperparameters in Table 4.

Tables 5 and 6 show the official BLEU scores (Papineni et al., 2002) for the test set in the ALT and IT domains, respectively. Similar results were obtained on the development sets, but they were omitted in this paper. We submitted the results using the pretrained mBART model, which were good on the development sets, on average.

The results are summarized as follows;

- For all language pairs in both domains, the BLEU scores with our extended mBART model were better than those under the same conditions without the pretrained models.

  When we focus on Indonesian, Malay, and Thai, which were not supported in the original mBART model, the BLEU scores of the submitted results were increased over 8 points from the baseline results in the ALT domain. We conclude that language expansion and additional pretraining were effective for translating new languages.

  For verification, we checked sentences in the test sets and the corpus for the pretrained model (c.f., Table 2). There were no identical sentences in the two corpora in the ALT domain. (Between 0% and 10% of the test sentences were included in the IT domain.) Therefore, these improvements were not caused by the memorization of the test sentences in the pretrained model.

Setting			Translation Direction								
PT	ML	FT/MD	**En→Hi**	**Hi→En**	**En→Id**	**Id→En**	**En→Ms**	**Ms→En**	**En→Th**	**Th→En**	Remark
		FT	12.26	8.23	24.71	23.65	31.02	27.52	13.73	2.04	
		MD	9.74	6.97	24.17	21.91	28.62	25.48	10.48	1.45	
	✓	FT	22.31	14.41	31.77	21.65	36.40	21.61	46.11	14.37	Baseline
	✓	MD	15.25	12.18	29.48	22.76	29.76	23.49	43.46	14.65	
✓		FT	**34.97**	**35.21**	41.15	**43.90**	**45.17**	44.53	55.69	28.96	Submitted
✓		MD	33.31	32.71	**41.80**	42.35	44.09	44.03	54.21	28.92	
✓	✓	FT	33.43	31.37	42.16	40.80	45.06	42.21	**55.80**	27.74	
✓	✓	MD	28.03	33.14	41.69	43.56	43.28	**45.24**	55.65	**29.77**	

Table 5: Official BLEU scores in the ALT domain. PT, ML, FT, and MD in the setting columns represent pretraining, multilingual models, fine-tuning, and multi-domain models, respectively.

Setting			Translation Direction								
PT	ML	FT/MD	**En→Hi**	**Hi→En**	**En→Id**	**Id→En**	**En→Ms**	**Ms→En**	**En→Th**	**Th→En**	Remark
		FT	7.97	4.92	23.33	22.64	29.62	26.53	10.24	0.99	
		MD	7.01	4.37	23.63	21.40	28.43	25.01	5.60	0.59	
	✓	FT	19.77	18.09	33.47	26.15	34.75	26.48	45.66	12.73	Baseline
	✓	MD	14.60	15.70	30.83	26.28	30.99	26.24	42.25	12.97	
✓		FT	**29.05**	35.32	43.25	40.69	**40.76**	38.42	50.91	21.89	Submitted
✓		MD	28.24	34.60	43.73	40.36	40.42	**39.29**	50.26	23.10	
✓	✓	FT	26.54	34.82	44.19	40.45	40.56	37.79	51.34	22.22	
✓	✓	MD	25.96	**35.55**	**44.40**	**42.44**	39.25	39.28	**52.00**	**24.24**	

Table 6: Official BLEU scores in the IT domain. PT, ML, FT, and MD in the setting columns represent pretraining, multilingual models, fine-tuning, and multi-domain models, respectively.

- The multilingual models were effective only without pretrained models. For example, for English to Hindi translation in the ALT domain, the BLEU scores improved from 12.26 to 22.31 when we used multilingual models without the pretrained model. However, they degraded from 34.97 to 33.43 when we used multilingual models with the pretrained model.

  The multilingual models were effective under low-resource conditions because the size of parallel data increased during training. However, they were ineffective if the models had learned sufficiently in advance, like pretrained models.

- Regarding domain adaptation, the fine-tuning method was better than the multi-domain models in many cases without the pretrained model.

6 Conclusions

In this paper, we presented the NICT-2 system submitted to the NICT-SAP task at WAT-2021.

A feature of our system is that it uses the mBART pretrained model. Because the published pretrained model does not support Indonesian, Malay, and Thai, we expanded it to support the above languages using additional training on the Wikipedia corpus. Consequently, the expanded mBART model improved the BLEU scores, regardless of whether multilingual models or domain adaptation methods were applied.

Acknowledgments

Part of this work was conducted under the commissioned research program 'Research and Development of Advanced Multilingual Translation Technology' in the 'R&D Project for Information and Communications Technology (JPMI00316)' of the Ministry of Internal Affairs and Communications (MIC), Japan.

References

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation. arXiv 2008.04550.

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.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. arXiv 1901.07291.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

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. arXiv 2001.08210.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand.

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.

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 of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA.

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.

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. arXiv 2008.00401.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. 2019. Online sentence segmentation for simultaneous interpretation using multi-shifted recurrent neural network. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 1–11, Dublin, Ireland.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CCNet: Extracting high quality monolingual datasets from web crawl data. arXiv 1911.00359.