# English-Myanmar Supervised and Unsupervised NMT: NICT's Machine Translation Systems at WAT-2019

**Rui Wang**[1*], **Haipeng Sun**[2,1*], **Kehai Chen**[1],
**Chenchen Ding**[1], **Masao Utiyama**[1], **and Eiichiro Sumita**[1]
[1] National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0289, Japan
[2] Harbin Institute of Technology, Harbin, China
{wangrui, sun.haipeng, khchen, chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

## Abstract

This paper presents the NICT's participation (team ID: NICT) in the 6th Workshop on Asian Translation (WAT-2019) shared translation task, specifically Myanmar (Burmese) - English task in both translation directions. We built neural machine translation (NMT) systems for these tasks. Our NMT systems were trained with language model pretraining. Back-translation technology is adopted to NMT. Our NMT systems rank the third in English-to-Myanmar and the second in Myanmar-to-English according to BLEU score.

## 1 Introduction

This paper describes the neural machine translation (NMT) systems[1] built for National Institute of Information and Communications Technology (NICT)'s participation in the the 6th Workshop on Asian Translation (WAT-2019) translation task(Nakazawa et al., 2019), specifically Myanmar (My) - English (En) for both translation directions.

The remainder of this paper is organized as follows. In Section 2, we present the data preprocessing. In Section 3, we introduce the details of our NMT systems. Empirical results obtained with our systems are analyzed in Section 4 and we conclude this paper in Section 5.

## 2 Data Preprocessing

As parallel data to train our systems, we used all the provided parallel data for all our targeted translation directions, including the training corpus "ALT" and "UCSY", and the "ALT" dev/test data. The statistics of our preprocessed parallel data are illustrated in Table 1.

| Corpus | #lines | #tokens (My/En) |
|---|---|---|
| train(ALT) | 17.9K | 1.0M / 410.2K |
| train(UCSY) | 208.6K | 5.8M / 2.6M |
| dev(ALT) | 0.9K | 57.4K / 22.1K |
| test(ALT) | 1.0K | 58.3K / 22.7K |

Table 1: Statistics of our preprocessed parallel data.

In WAT2019, two Myanmar monolingual corpora consist of Myanmar Wikipedia and Myanmar Common Crawl. For English monolingual corpus, we randomly extracted 10 million sentences from WMT monolingual News Crawl datasets.[2] The statistics of our preprocessed monolingual data are illustrated in Table 2.

| Corpus | #lines | #tokens |
|---|---|---|
| My | 6.7M | 125.6M |
| En | 10.0M | 229.8M |

Table 2: Statistics of our preprocessed monolingual data.

We used `Moses` tokenizer and truecaser for English. The truecaser was trained on the English data, after tokenization. For Myanmar, we used the original tokens. For cleaning, we only applied the `Moses` script `clean-n-corpus.perl` to remove lines in the parallel data containing more than 80 tokens and replaced characters forbidden by `Moses`.

[*] Rui and Haipeng have equal contribution to this paper. This work was conductd when Haipeng visited NICT as an internship student.

[1] This system is based on our WMT-2019 system (Marie et al., 2019).

[2] http://data.statmt.org/news-crawl/



*Proceedings of the 6th Workshop on Asian Translation*, pages 90–93
Hong Kong, China, November 4, 2019. 

## 3 MT Systems

To build competitive NMT systems, we chose to rely on the Transformer architecture (Vaswani et al., 2017) since it has been shown to outperform, in quality and efficiency, the two other mainstream architectures for NMT known as deep recurrent neural network (deep-RNN) and convolutional neural network (CNN). We chose to rely on the Transformer-based NMT initialized by a pretrained cross-lingual language model (Lample and Conneau, 2019) to train our NMT systems since it had been shown to be efficient in the low-resource language pairs. In order to limit the size of the vocabulary of the NMT model, we segmented tokens in the training data into sub-word units via byte pair encoding (BPE) (Sennrich et al., 2016b). We determined 60k BPE operations jointly on the training data for English and Myanmar, and used a shared vocabulary for both languages with 60k tokens based on BPE.

### 3.1 TLM

Before training NMT, we used all training corpora including parallel data and monolingual data to train a translation language model (TLM) using `XLM`[3] in order to pretrain the NMT model on 8 GPUs[4]. The parameters for training the language model were set as listed in Table 3.

```
--lgs 'en-my' --mlm_steps
'en,my,en-my,my-en'
--emb_dim 1024 --n_layers
6 --n_heads 8 --dropout
0.1 --attention_dropout
0.1 --gelu_activation true
--batch_size 32 --bptt 256
--optimizer adam,lr=0.0001
```

Table 3: Parameters for training TLM.

### 3.2 NMT

We trained a Transformer-based NMT model with the pre-trained TLM using `XLM` toolkit. Our NMT system was consistently trained on 8 GPUs, with the following parameters listed in Table 4.

We performed NMT decoding with a single model according to the best BLEU (Papineni et al., 2002) and the perplexity scores.

[3] https://github.com/facebookresearch/XLM

[4] NVIDIA @ Tesla @ V100 32Gb.

```
--lgs 'en-my' --encoder_only
false --emb_dim 1024 --n_layers
6 --n_heads 8 --dropout
0.1 --attention_dropout
0.1 --gelu_activation
true --tokens_per_batch
2000 --batch_size 32
--bptt 256 --optimizer
adam_inverse_sqrt,beta1=0.9,
beta2=0.98,lr=0.0001
--eval_bleu true
```

Table 4: Parameters for training NMT.

### 3.3 Back-translation

We also tried back-translation method (Sennrich et al., 2016a) to make use of monolingual corpora for English-to-Myanmar translation task. Parallel data for training NMT can be augmented with synthetic parallel data, generated through back-translation, to significantly improve translation quality. For back-translation generation, we used an NMT system, trained on the parallel data provided by the organizers, to translate target monolingual sentences into the source language to generate pseudo parallel corpora. Then, the pseudo parallel corpora were simply mixed with the original parallel data to train from scratch a new source-to-target NMT system.

### 3.4 UNMT

To the best of our knowledge, unsupervised NMT (UNMT) (Artetxe et al., 2018; Lample et al., 2018a; Yang et al., 2018; Lample et al., 2018b; Sun et al., 2019; Lample and Conneau, 2019) has achieved remarkable results on some similar language pairs. To obtain a better picture of the feasibility of UNMT, we also set up a UNMT system for one truly low-resource and distant language pair: En-My. We tried to train a Transformer-based UNMT model that relies solely on monolingual corpora, with the pre-trained cross-lingual language model using `XLM` toolkit. Note that this cross-lingual language model was trained solely on monolingual corpora shown in Section 2.

We used these monolingual corpora to train the UNMT model for 50000 iterations. The En-My UNMT system was trained on 8 GPUs, with the parameters listed in Table 6.

| Systems | ALT | UCSY | MONO | My-En | En-My |
|---|---|---|---|---|---|
| UNMT | | | ✓ | 0.81 | 0.31 |
| NMT | ✓ | | | 8.06 | 10.50 |
| NMT | ✓ | ✓ | | 14.97 | 14.15 |
| NMT+TLM | ✓ | ✓ | | 18.42 | 16.12 |
| NMT+TLM | ✓ | ✓ | ✓ | 21.33 | **19.73** |
| NMT+TLM+back-translation | ✓ | ✓ | ✓ | **29.89** | 19.01 |

Table 5: Results (BLEU-cased) of our MT systems on the test set. ALT denotes that ALT training data was used in this system; UCSY denotes that UCSY training data was used in this system; MONO denotes monolingual training data was used in this system. +TLM denotes that language model pretraining was used in this system; +back-translation denotes that back-translation was used in this system.

```
--lgs 'en-my' --ae_steps
'en,my' --bt_steps
'en-my-en,my-en-my'
--word_shuffle 3
--word_dropout 0.1
--word_blank 0.1 --lambda_ae
'0:1,100000:0.1,300000:0'
--encoder_only false
--emb_dim 1024 --n_layers
6 --n_heads 8 --dropout
0.1 --attention_dropout
0.1 --gelu_activation
true --tokens_per_batch
2000 --batch_size 32
--bptt 256 --optimizer
adam_inverse_sqrt,beta1=0.9,
beta2=0.98,lr=0.0001
--eval_bleu true
```

Table 6: Parameters for training UNMT.

## 4 Results

Our systems are evaluated on the ALT test set and the results[5] are shown in Table 5. Our observations from are as follows:

1) The results of UNMT are very low, highlighting that UNMT is still very far from exploitable for low-resource distant language pairs.

2) Language model pretraining showed significant improvement in the NMT systems for both translation directions. This demonstrates that language model pretraining is effective for low-resource machine translation.

3) For My-En translation direction, back-translation could further improve translation performance, achieving 8 BLEU scores improvement. However, back-translation for En-My translation direction was unable to improve or even harm the NMT performance since the My monolingual data was noisy.

[5]The results of BLEU are based on our own evaluation. For the official results, please refer to http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html.

## 5 Conclusion

We presented in this paper the NICT's participation in the WAT-2019 shared translation task. Our primary NMT submissions to the task performed the third in English-to-Myanmar and the second in Myanmar-to-English according to BLEU score. Our results also confirmed the positive impact of language model pretraining in NMT. Moreover, our results for UNMT highlighted that unsupervised machine translation is still very far from exploitable for low-resource distant language pairs.

## References


Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *ICLR*, Vancouver, Canada.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *ICLR*, Vancouver, Canada.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *EMNLP*, pages 5039–5049, Brussels, Belgium.

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *WMT*, pages 294–301, Florence, Italy.

Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725, Berlin, Germany.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *ACL*, pages 1235–1245, Florence, Italy.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008, Long Beach, CA, USA.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *ACL*, pages 46–55, Melbourne, Australia.