English-Myanmar NMT and SMT with Pre-ordering: NICT's Machine Translation Systems at WAT-2018

Rui Wang **Chenchen Ding** **Masao Utiyama** **Eiichiro Sumita**
National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0289, Japan
{wangrui, chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

This paper presents the NICT's participation (team ID: NICT) in the 5th Workshop on Asian Translation (WAT-2018) shared translation task, specifically Myanmar (Burmese) - English task in both translation directions. We built state-of-the-art neural machine translation (NMT) as well as phrase-based statistical machine translation (PBSMT) systems for these tasks. Our NMT systems were trained with the Transformer architecture on the provided parallel data. Pre-ordering technology is adopted to both NMT and PBSMT. Our NMT systems rank the first in English-to-Myanmar and the second in Myanmar-to-English according to the official human evaluation.

1 Introduction

This paper describes the machine translation systems[1] built for National Institute of Information and Communications Technology (NICT)'s participation in the the 5th Workshop on Asian Translation (WAT-2018) translation task (Nakazawa et al., 2018), specifically Myanmar (My) - English (En) for both translation directions. All of our systems are constrained, i.e., we used only the parallel adata provided by the organizers to train and tune our systems.

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 and SMT systems with pre-ordering technology. 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.

Table 1: Statistics of our preprocessed parallel data.

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

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`.

[1] This system is based on our WMT-2018 system (Marie et al., 2018).

Table 2: Results (BLEU-cased and official human evaluation) of our MT systems on the test set. We only submitted the top 2 systems on BLEU for human evaluation.

System	My→En (BLEU)	En→My (BLEU)	My→En (Human)	En→My (Human)
`Moses`	9.44	25.55	N/A	N/A
`Moses` Pre-order	N/A	25.75	N/A	N/A
`Marian`	16.32	26.02	7.250	42.50
`Marian` Pre-order	N/A	23.79	N/A	N/A
`Marian` ensemble	20.79	29.89	20.50	61.00

3 MT Systems

3.1 NMT

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 `Marian`[2] (Junczys-Dowmunt et al., 2018) to train our NMT systems since it supports many state-of-the-art features and is one of the fastest NMT frameworks publicly available.[3]

All our NMT systems were consistently trained on 4 GPUs,[4] with the following parameters for `Marian`:

```
--type transformer --max-length
80 --dec-depth 6 --normalize 1
--save-freq 5000 --workspace
8000 --disp-freq 500
--beam-size 12 --overwrite
--cost-type ce-mean-words
--keep-best --enc-depth 6
--transformer-dropout 0.1
--valid-mini-batch 16 --valid-freq
5000 --learn-rate 0.0003
--lr-decay-inv-sqrt 16000
--lr-warmup 16000 --lr-report
--sync-sgd --devices 0 1 2
3 --dim-vocabs 50000 50000
--exponential-smoothing
--optimizer-params 0.9 0.98 1e-09
--clip-norm 5 --tied-embeddings
--mini-batch-fit --early-stopping
5 --label-smoothing 0.1
--valid-metrics ce-mean-words
perplexity translation
```

We performed NMT decoding with an ensemble of a total of 4 models according to the best BLEU (Papineni et al., 2002) and the perplexity scores, produced by 4 independent training runs.

3.2 SMT

We also trained phrase-based SMT systems using `Moses`. Word alignments and phrase tables were trained on the tokenized parallel data using `mgiza`. Source-to-target and target-to-source word alignments were symmetrized with the `grow-diag -final-and` heuristic. We simply trained regular phrase-based models and used the default distortion limit of 6. We trained two 5-gram language models on the entire target side of the parallel data, with SRILM (Stolcke, 2002). To tune the SMT model weights, we used MERT (Och, 2003) and selected the weights giving the best BLEU score on the development data.

3.3 Pre-ordering

We also tried a classic pre-ordering method for English-to-Myanmar translation task. Specifically, the dependency-based head finalization in Ding et al. (2014) is exactly reproduced in our experiment. The source English part is pre-ordered before being input into NMT and SMT systems.

[2] `https://marian-nmt.github.io/`, version 1.4.0

[3] It is fully implemented in pure C++ and supports multi-GPU training.

[4] NVIDIA® Tesla® P100 16Gb.

4 Results

Our systems are evaluated on the ALT test set and the results[5] are shown in Table 2. Our observations from are as follows:

1) Our NMT (Marian) system performed much better than SMT (Moses) system in My-to-En. That is, nearly 7 BLEU scores. However, there is no significant difference in En-to-My.

2) Pre-ordering did not show significant improvement in the NMT systems, while it improved translation quality by 0.2 BLEU points in the SMT system.

3) Ensemble decoding significantly outperformed decoding with a single NMT model.

5 Conclusion

We presented in this paper the NICT's participation in the WAT-2018 shared translation task. Our primary NMT submissions to the task performed the best submitted system according to the official human evaluation. Our results also confirmed the slight positive impact of using pre-ordering in PBSMT.

References

Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2014. Empirical dependency-based head finalization for statistical chinese-, english-, and french-to-myanmar (burmese) machine translation. In *Proceedings of IWSLT*, pages 184–191.

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*, Melbourne, Australia.

Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2018. NICT's neural and statistical machine translation systems for the WMT18 news translation task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium, October. Association for Computational Linguistics.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Anoop Kunchukuttan, Win Pa Pa, Isao Goto, Hideya Mino, Katsuhito Sudoh, and Sadao Kurohashi. 2018. Overview of the 5th workshop on asian translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*, Hong Kong, China, December.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, Seattle.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*.

[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.