NICT's Machine Translation Systems for CCMT-2019 Translation Task

Kehai Chen, Rui Wang(✉), Masao Utiyama, and Eiichiro Sumita

National Institute of Information and Communications Technology, Kyoto, Japan
{khchen,wangrui,mutiyama,eiichiro.sumita}@nict.go.jp

Abstract. This paper describes the NICT's neural machine translation systems for Chinese↔English directions in the CCMT-2019 shared news translation task. We used the provided parallel data augmented with a large quantity of back-translated monolingual data to train state-of-the-art NMT systems. We then employed techniques that have been proven to be most effective, such as fine-tuning, and model ensembling, to generate the primary submissions of Chinese↔English translation tasks.

Keywords: Neural machine translation · CCMT-2019 · NICT

1 Introduction

This paper presents the neural machine translation (NMT) systems built for National Institute of Information and Communications Technology (NICT)'s participation in the CCMT-19 shared News Translation Task for Chinese↔English directions. Specifically, we used the Transformer architecture to build our translation systems. We then employed techniques that have been proven to be most effective, such as back-translation, fine-tuning, and model ensembling, to generate the primary submissions of Chinese↔English translation tasks. All of our systems are constrained, i.e., we used only the parallel and monolingual data provided by the organizers to train and tune our systems. This system is also a part of our system for WMT19 [1][1].

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

2 Datasets

2.1 Data

As parallel data to train our systems, we used all the provided parallel data for all our targeted translation directions. The training data for the Chinese↔English

[1] The Chinese-English task is jointly held by CCMT-2019 and WMT19. Therefore, part of these two system description papers are overlapped.

S. Huang and K. Knight (Eds.): CCMT 2019, CCIS 1104, pp. 77–81, 2019.
https://doi.org/10.1007/978-981-15-1721-1_8

(ZH↔EN) translation tasks consists of two parts: (1) we selected the first 10 million lines of the News Crawl 2018 English corpus according to the finding of [6,11], (2) the corresponding synthetic data was generated through back-translation [5,8].

2.2 Pre-processsing

We applied tokenizer and truecaser of Moses [4] to the English sentences. For Chinese, we used `Jieba`[2] for tokenization but did not perform truecasing. For cleaning, we filtered out sentences longer than 80 tokens in the training data by using `Moses` script clean-n-corpus.perl, and replaced characters forbidden by `Moses`. Tables 1 and 2 present the statistics of the parallel and monolingual data, respectively, after pre-processing.

Table 1. Statistics of our pre-processed parallel data

Language pair	#Sentence pairs	#Tokens	
		Chinese	English
Chinese↔English	24.8M	509.9M	576.2M

Table 2. Statistics of our pre-processed monolingual data

Language	#Sentences	#Tokens
English	338.7M	7.5B
Chinese	130.5M	2.3B

3 MT Systems

3.1 NMT

We used Marian toolkit [2][3] to build competitive NMT systems based on the Transformer [10] architecture. We used the byte pair encoding (BPE) algorithm [9] for obtaining the sub-word vocabulary whose size was set to 50,000. The number of dimensions of all input and output layers was set to 512, and that of the inner feed-forward neural network layer was set to 2048. The number of attention heads in each encoder and decoder layer was set to eight. During training, the value of label smoothing was set to 0.1, and the attention dropout and residual dropout were set to 0.1. The Adam optimizer [3] was used to tune the parameters of the model. The learning rate was varied under a warm-up strategy with warm-up steps of 16,000. We validated the model with an interval

[2] https://github.com/fxsjy/jieba.
[3] https://marian-nmt.github.io.

of 5,000 batches on the development set and selected the best model according to BLEU [7] score on the development set. All our NMT systems were consistently trained on 4 GPUs,[4] with the following parameters for `Marian` (Table 3):

Table 3. Parameters for training `Marian`.

```
--type transformer  --max-length
100  --transformer-dim-ffn  4096
--dim-vocabs  50000  50000  -w  12000
--mini-batch-fit  --valid-freq 5000
--save-freq 5000 --disp-freq 500
--valid-metrics  ce-mean-words  perplexity
translation  --quiet-translation
--sync-sgd --beam-size  12
--normalize=1  --valid-mini-batch
16  --keep-best  --early-stopping
20 --cost-type=ce-mean-words
--enc-depth 6 --dec-depth 6
--tied-embeddings  --transformer-dropout
0.1 --label-smoothing  0.1
--learn-rate  0.0003 --lr-warmup  16000
--lr-decay-inv-sqrt  16000 --lr-report
--optimizer-params  0.9  0.98  1e-09
--clip-norm  5  --exponential-smoothing
```

3.2 Back-Translation of Monolingual Data

The so-called "back-translation" of monolingual has been shown to be one of the most efficient ways to exploit monolingual data for NMT [8]. It is simply to translate target monolingual data into the source language, using a pre-trained target-to-source NMT models, in order to produce a new synthetic parallel data that can be used to train NMT models. We concatenated the resulting synthetic parallel data to the original parallel data to train better NMT models. For En→Zh, we back-translated the entire XMU Chinese monolingual corpus containing 5.4M sentences as the source to produce synthetic English data. For Zh→En, we empirically compared the impact of back-translating different sizes of English monolingual data, using the first 10M lines of the concatenation of News Crawl-2016 and News Crawl-2017 English corpora to produce synthetic Chinese data.

3.3 Fine-Tuning and Ensemble of NMT Models

After the back-translation, we performed the training run independently for five times on the mixture of the original parallel data and the pseudo-parallel

[4] NVIDIA® Tesla® P100 16 Gb.

data, and thus obtain the translation models. The new model was further fine-tuned on the ccmt2018_newstest set for 20 epochs. Finally, we decoded the ccmt2019_newstest set with an ensemble of the five fine-tuned models to generate the primary submissions for the ZH↔EN tasks.

4 Results

Our systems are evaluated on the `WMT2019NewsTest` test set[5] for ZH↔EN tasks and the results are shown in Table 4. For EN→ZH, BLEU scores were computed on the basis of character-based segmentation. “w/backtr” and “w/o backtr” indicate with and without back-translation, respectively. “w/ft” indicates that this single model was fine-tuned on the ccmt2018_newstest sets. “ensemble” indicates that five fine-tuned single models were ensembled at decoding time.

Table 4. Results (BLEU-cased) of our MT systems on the ccmt2018_newstest test set.

System	ZH→EN	EN→ZH
Single model (w/o backtr)	23.3	30.3
Single model (w/backtr)	25.3	31.8
Single model (w/ft)	27.5	33.1
Five fine-tuned single models (ensemble)	31.0	34.5

Our observations from Table 4 are as follows: It is obvious that the back-translation, fine-tuning, and ensemble methods are greatly effective for the ZH↔EN tasks. In particular, the ensemble gave more improvements on the ZH→EN task over the “Single model+back-translation+fine-tuning” model than the EN→ZH task.

5 Conclusion

We presented in this paper the NICT’s participation in the CCMT-2019 shared Chinese↔English news translation task. Our primary submissions to the tasks were the results of a simple combination of back-translation, fine-tuning, and ensemble methods. Our results confirmed that these three methods can incrementally improve translation performance of the Transformer NMT.

Acknowledgments. We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions. Rui Wang was partially supported by JSPS grant-in-aid for early-career scientists (19K20354): “Unsupervised Neural Machine Translation in Universal Scenarios” and NICT tenure-track researcher startup fund “Toward Intelligent Machine Translation”.

[5] http://www.statmt.org/wmt19/translation-task.html.

References

1. Dabre, R., et al.: NICT's supervised neural machine translation systems for the WMT19 news translation task. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Association for Computational Linguistics, Florence, Italy, pp. 168–174, August 2019. https://www.aclweb.org/anthology/W19-5313
2. Junczys-Dowmunt, M., et al.: Marian: fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia, pp. 116–121 (2018). http://aclweb.org/anthology/P18-4020
3. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014). http://arxiv.org/abs/1412.6980
4. Koehn, P., et al.: 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, Prague, Czech Republic, pp. 177–180 (2007). http://aclweb.org/anthology/P07-2045
5. Marie, B., et al.: NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Association for Computational Linguistics, Florence, Italy, pp. 294–301, August 2019. https://www.aclweb.org/anthology/W19-5330
6. Marie, B., Wang, R., Fujita, A., Utiyama, M., Sumita, E.: NICT's neural and statistical machine translation systems for the WMT18 news translation task. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, pp. 449–455, October 2018. https://www.aclweb.org/anthology/W18-6419
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318, July 2002. https://doi.org/10.3115/1073083.1073135, http://www.aclweb.org/anthology/P02-1040
8. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp. 86–96 (2016). http://aclweb.org/anthology/P16-1009
9. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic, Berlin, Germany, pp. 1715–1725 (2016). http://aclweb.org/anthology/P16-1162
10. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30 (2017). https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
11. Wang, R., Marie, B., Utiyama, M., Sumita, E.: NICT's corpus filtering systems for the WMT18 parallel corpus filtering task. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Association for Computational Linguistics, Belgium, Brussels, pp. 963–967, October 2018. https://doi.org/10.18653/v1/W18-6489, https://www.aclweb.org/anthology/W18-6489