Supervised neural machine translation based on data augmentation and improved training & inference process

Yixuan Tong Liang Liang Boyan Liu Shanshan Jiang Bin Dong
Ricoh Software Research Center Beijing Co., Ltd.
{ yixuan.tong, liang.liang, boyan.liu, shanshan.jiang, bin.dong}@srcb.ricoh.com

Abstract

This is the second time for SRCB to participate in WAT. This paper describes the neural machine translation systems for the shared translation tasks of WAT 2019. We participated in ASPEC tasks and submitted results on English-Japanese, Japanese-English, Chinese-Japanese, and Japanese-Chinese four language pairs. We employed the Transformer model as the baseline and experimented relative position representation, data augmentation, deep layer model, ensemble. Experiments show that all these methods can yield substantial improvements.

1 Introduction

The advent of neural networks in machine translation has brought great improvement on translation quality over traditional statistical machine translation (SMT) in recent years (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2014). A lot of research efforts have been attracted to investigate neural networks in machine translation. This paper describes the Neural Machine Translation systems of Ricoh Software Research Center Beijing (SRCB) for the shared translation tasks of WAT 2019 (Nakazawa et al., 2019). We participated in ASPEC tasks, and submitted results on four language pairs, including English-Japanese, Japanese-English, Japanese-Chinese and Chinese-Japanese.

In the ASPEC tasks, we employed Transformer (Vaswani et al., 2018) as our baseline model and built our translation system based on OpenNMT (Klein et al., 2017) open source toolkit. To enhance the performance of the model, we made the following changes: 1) We proposed data augmentation method (Yihan et al., 2018) and back translation algorithm (Sennrich et al., 2015), which was observed to be useful in Japanese-English and Japanese-Chinese corpus. 2) We incorporated weighted loss function and Sentence-wise regularization method (Gong et al., 2019) into Transformer model. 3) We used deep layer (Wang et al., 2019) technique to further improve translation quality. 4) We used ensemble techniques and model stabilization to further improve translation quality.

The remainder of this paper is organized as follows: Section 2 describes our NMT system and algorithms. Section 3 describes the processing of the data and all experimental results and analysis. Finally, we conclude in section 4.

2 Systems

2.1 Base Model

Our system is based on the Transformer model. Transformer model is a paradigm model for neural machine translation which can achieve start-of-the-art translation quality.

2.2 Data augmentation and back translation

We use data augmentation algorithm (Yihan et al., 2018) to select the parallel sentences that original model cannot train well, then retrain the model using the new dataset to improve the translation quality. However, data augmentation algorithm is mutual exclusion with back translation algorithm (Sennrich et al., 2015). Back translation algorithm translates monolingual sentences to corresponding predictions to generate parallel sentences, which

Proceedings of the 6th Workshop on Asian Translation, pages 147–151

can augment the training set. In the real translation model, we use both data augmentation and back translation to train different models, then combine the models by ensemble learning.

2.3 Weighted loss function

The loss function of the neural network is a standard to judge whether it is convergent. When calculating the cross-entropy between predicted words with the original references, there is no consideration about words' length. So we add the length influence weight to the loss function which can represent the real loss score more accurately.

2.4 Sentence-wise smooth Regularization

Sentence-wise regularization method (Gong et al., 2019) is used in our system, which aims to output smooth prediction probabilities for all tokens in the target sequence. Compared with maximum-likelihood estimation, this method could adjust the weights and gradients in the target sequence automatically to ensure the predictions in a sequence uniformly. We implement grid search to find the best parameters for smooth regularization in different subtasks.

2.5 Deep layer model

Wang et al. (2019) showed that the location of layer normalization played a vital role when training deep Transformer. They also proved that pre-norm Transformer is more efficient for training than post-norm (vanilla Transformer) when the model goes deeper. Dynamic linear combination of previous layers was introduced which improves the translation quality as well. Note that we built our deep layer model in pre-norm way as default. In the state of practice, we find that more layers in decoders could enhance the ability of our real model. We use grid search to find proper parameters to achieve a balance between efficiency and performance

2.6 Ensemble

It has been investigated that ensembling different model can yield significant improvement in translation quality (Denkowski and Neubig, 2017). In our systems, we adopted two ensembling schemes. For one configured translation model, once the model finishes training, the last 8 checkpoints of the model are averaged to get one trained model. Then, we make different configurations and train several models independently. After averaging checkpoints for each model, we do step-wise ensembling.

Specifically, these models are run at each time step and an arithmetic mean of predicted probability is obtained, which is used to determine the next word.

2.7 Model Stabilization

We observed unneglectable level of instability in the Transformer models (up to 0.4 BLEU diverse for models with the same settings).

The first remedy to fight against instability is by introducing noise (Devlin et al., 2018). We randomly deletes tokens from the source side in the training dataset. It turns out this method would bring marginal improvement. We believe that by introducing noise, models would turn from over confident, thus result in better stability and generalization.

The other strategy is batch filtering. In our experiments, there are special batches of training which lead to considerable up going of training loss. We believe the outliers are to be blame. Thus batch filtering mechanism (Chen et al., 2018) is hired which eliminate bathes with gradient norm exceeding certain threshold.

3 Experiments

We experimented our NMT system on Japanese-English, English-Japanese, Chinese-Japanese, and Japanese-Chinese scientific paper translation subtasks.

3.1 Datasets

We used Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2014) as parallel corpora for all language pairs. For Japanese-English subtask and English-Japanese subtask, we used the first 1M sentences with augmented the second 1M sentences. Furthermore, for Japanese-English subtask, we augmented training data to nearly 2M by data augmentation. And, we also trained back translation models using the second 1M and the third 1M sentences as difference training datasets. For Chinese-Japanese subtask, all the sentences in ASPEC corpora are used as training data. And, we also trained back translation models using the first 1M Japanese sentences in Japanese-English subtask as difference training datasets.

For all corpora, Japanese sentences were segmented by the morphological analyzer Juman[1] and English sentences were tokenized by tokenizer.perl of Moses[2], while Chinese sentences were segmented by KyTea[3]. Sentences with more than 100 words were excluded. We used the subword unit, that is Joint Byte Pair Encoding (BPE) (Sennrich et al., 2016c) scheme, to encoder vocabulary for both source and target sentences.

3.2 Results

As shown in Table 1, we rank 1st in the direction of Japanese-English, Japanese-Chinese and Chinese-Japanese, and 2nd in one English-Japanese.

	Ja-En	En-Ja	Ja-Zh	Zh-Ja
Rank	1st	2nd	1st	1st
BLEU	30.92	45.71	38.63	52.37

Table 1: Results of subtasks

Japanese-English subtask:

The baseline model is a vanilla Transformer model with the first 1M data. Using the second 1M sentences to do data augmentation, the BLEU score has increased 1.34 to 30.20 which is the biggest improvement in this direction. What's more, the relative position representation has improved more than 1 BLEU score in WAT 2018 system. However, there is only more than a 0.2 increase in 2019's model. Changing the loss function weight with length, the new BLEU score become 30.78. Besides, the re-ranking algorithm using max function has 0.16 improvement.

System	BLEU
Baseline	28.86
Data augmentation	30.20
Relative position	30.42
Weighted loss function	30.78
Re-ranking	30.92

Table 2: Technical point contributions

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[2] http://www.statmt.org/moses/
[3] http://www.phontron.com/kytea/index.html

English-Japanese subtask:

As for this subtask, this is our first time to participate in that we tried many other algorithms. As for the training data, we tried four kinds of combinations shown in Table 3. The baseline model is the big Transformer model with the first 1M parallel sentences. For data augmentation and back translation is mutual exclusion, we trained different models and did ensemble to combine all the features. The BLEU score of first 1M data with the second 1M data using data augmentation is 43.33. The first 1M data with the second 1M data or the remaining 2M data using back translation is 43.32 and 43.66, respectively. The best combination rate for original data and back translation data is 1:4. So the results meet the exception. The last category is the 1M data with the second 1M data using back translation and the third 1M data using data augmentation which BLEU score is 43.57. It's lower than the model with back translation only. In practice, we choose different kinds of combination for model ensemble.

System	BLEU
Baseline	42.57
1M + Da 1M	43.33
1M + Bt 1M	43.32
1M + Bt 2M	43.66
1M + Bt 1M + Da 1M	43.57

Table 3: Results of different data combination. 'Da' is the abbreviation of data augmentation, 'Bt' is the abbreviation of back-translation.

The baseline model is a vanilla Transformer model with the first 1M data. Using the relative position representation has improved 0.56 BLEU score. Using the last 2M sentences to do data augmentation, the BLEU score has increased 0.53 to 43.66. Besides, the sentence-wise smooth has improved 0.12 BLEU score in WAT 2019 system. There is a 0.14 increase in 2019's model when introducing deep layer model. Finally, the model ensemble algorithm has 1.79 improvement.

System	BLEU
Baseline	42.57
Relative position	43.13
Data augmentation	43.66
Sentence-wise smooth	43.78

Deep layer model	43.92
Ensemble	45.71

Table 4: Technical point contributions

Japanese-Chinese subtask:

For this subtask, we utilized only the data in ASPEC, no data augmentation was used. We implemented the system based on OpenNMT 1.22.0, and adapted the beam search bug fix in the afterwards versions. We hired sentence-wise smooth, encoder side token deletion and batch filtering. The hyper parameters were searched with respect to the devtest.txt dataset. After generating 8 models (with BLEU above 37.0), we ensembled those models with step-wise ensemble system. The results of applying thus technologies was in Table 5 (results after averaging of last 8 checkpoints, best in two models).

System	BLEU
Baseline	35.92
Relative position	36.71
Sentence-wise smooth	36.98
Encoder side token deletion & batch filtering	37.21
Ensemble of 8 models	38.63

Table 5: Technical point contributions

Chinese-Japanese subtask:

The baseline model is a vanilla Transformer model with all ASPEC data. We using the first 1M sentences in English-Japanese subtask to do data augmentation, the BLEU score has increased 0.81 to 50.46. What's more, the relative position representation has improved 0.59 BLEU score in WAT 2019 system. Sentence-wise smooth increases 0.19. Besides, deep layer model algorithm has 0.57 improvement. Finally, the model ensemble algorithm has 0.56 improvement.

System	BLEU
Baseline	49.65
Data augmentation	50.46
Relative position	51.05
Sentence-wise smooth	51.24
Deep layer model	51.81
Ensemble	52.37

Table 6: Technical point contributions

4 Conclusion

In this paper, we described our NMT system, which is based on Transformer model. We made several changes to original Transformer model, including relative position representation, deep layer model, ensembling and other technical points. We evaluated our Transformer system on Japanese-English, English-Japanese, Japanese-Chinese and Chinese-Japanese scientific paper translation subtasks at WAT 2019.The results show that the implementation of these points can effectively improve the translation quality.

In our future work, we plan to explore more vocabulary encoding schemes and compare with byte pair encoding (BPE) (Sennrich et al., 2016). In addition, we will attempt to implement other transformer structures, which combine other advanced technologies.

References

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In Proceeding of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), 1700-1709.

Ilya Sutskever, Oriol Vinyals,and Quoc Le. 2014. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NIPS 2014), December.

Kyunghyun Cho, Bart Van and et al. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014), October.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Nakazawa, Toshiaki and Ding, Chenchen and Dabre, Raj and Mino, Hideya and Goto, Isao and Pa, Win Pa and Doi, Nobushige and Oda, Yusuke and Kunchukuttan, Anoop and Parida, Shantipriya and Bojar, Ondřej and Kurohashi, Sadao. 2019. Overview of the 6th Workshop on Asian Translation. In Proceedings of the 6th Workshop on Asian Translation (WAT2019).

Vaswani A, Shazeer N, Parmar N, et al. 2017 Attention is all you need. Advances in Neural Information Processing Systems, 5998-6008.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. arXiv preprint arXiv:1701.02810.

Li Y, Liu B, Tong Y, et al. SRCB Neural Machine Translation Systems in WAT 2018[C]//Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation. 2018.

Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data[J]. arXiv preprint arXiv:1511.06709, 2015.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation (WNMT), pages 18–27.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2014. ASPEC : Asian Scientific Paper Excerpt Corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), pages 2204–2208.

Gong C, Tan X, He D, et al. Sentence-wise smooth regularization for sequence to sequence learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 6449-6456.

Qiang Wang, Bei Li, Tong Xiao, and Jingbo Zhu. 2019. Learning deep transformer models for machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Italy, Florence. Association for Computational Linguistics

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

Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

Chen M X, Firat O, Bapna A, et al. The best of both worlds: Combining recent advances in neural machine translation[J]. arXiv preprint arXiv:1804.09849, 2018.