SRCB Neural Machine Translation Systems in WAT 2018

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

Abstract

This is the first time SRCB participates in WAT. This paper describes the Neural Machine Translation systems for the shared translation tasks of WAT 2018. We participated in ASPEC tasks, and submitted results on English-Japanese, Japanese-English and Japanese-Chinese three language pairs. We employed Transformer as baseline model, and experimented subword segmentation, relative position representation and model ensembling. 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 2018 (Nakazawa et al., 2018). We participated in ASPEC tasks, and submitted results on three language pairs, including English-Japanese, Japanese-English and Japanese-Chinese.

In the ASPEC tasks, we employed Transformer (Vaswani et al., 2017) as our baseline model and built our translation system based on OpenNMT (Klein et al., 2017) open source toolkit [1]. To enhance the performance of the model, we made the following changes: 1) To deal with out of vocabulary (OOV) and rare words problem in translation, we used subword unit, that is Joint Byte Pair Encoding (BPE) (Sennrich et al., 2016c) scheme, to encoder vocabulary for both source and target language. 2) We proposed a synthetic data augmentation method and it was observed to be useful in Japanese-English corpus. 3) We incorporated relative position representation (Shaw et al., 2018) into Transformer model. 4) We used two ensemble techniques to further improve translation quality.

The remainder of this paper is organized as follows: Section 2 describes our NMT system. 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 NMT system is built upon Transformer (Vaswani et al., 2018) model. We used open source OpenNMT[1] and imported some changes and new features such as relative position embedding (Shaw et al., 2018) and ensembling.

The Transformer (Vaswani et al., 2017) also adopts sequence to sequence architecture and it consists of an encoder layer and a decoder layer. Different from traditional Seq2Seq model (Bahdanau et al., 2014), the encoder layer consists of two sublayers: a multi-head self-attention layer and a position-wise fully connected feed-forward layer. Instead of employing a single attention function mechanism (Luong et al., 2015), the multi-head self-attention adopts several different learnt linear projections to queries, keys and values respectively. This mechanism allows model to jointly attend to information from different representation subspaces at different positions. The decoder layer consists of three sublayers: a masked multi-head self-attention, followed by encoder-

[1] http://opennmt.net/

decoder attention and a position-wise feed-forward layer. Residual connections (He et al., 2016) followed by layer normalization (Ba et al., 2016) are used between each sublayer, which can prevent gradient vanishing (Hochreiter et al., 1998) and propagate information to higher layers. The masked multi-head self-attention uses masking in its self-attention to prevent a given output position from incorporating information about future output positions during training.

To make use of the order of the sequence, the Transformer models add positional encodings to capture information about the absolute position of the tokens in the sequence. The positional encodings are based on sinusoids of varying frequency and are added to the input embeddings at the bottoms of the encoder and decoder stacks. The absolute position representations hypothesized that sinusoidal position encodings would help the model to generalize to sequence lengths unseen during training.

2.2 Relative Position Representation

Instead of using absolute position encodings, Shaw et al. (2018) presented an alternative approach, extending the self-attention mechanism to efficiently consider representations of the relative positions, or distance between sequence elements. The approach can be cast as a special case of considering arbitrary relations between any two elements of the inputs.

The approach learns two relative position representations. The first representation is to propagate edge information to the sublayer output when computing weighted sum of a linearly transformed input elements. Similarly, the second representation is to consider edges when computing a compatibility function that compares two input elements.

In their experiments, they observed significant improvement over absolute position representations. Furthermore, they observed that combing relative and absolute position representations yields no further improvement in translation quality. Thus, in our translation system, we incorporated relative position representation and removed absolute position encodings from encoder layers.

2.3 Ensembling

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.

2.4 Data Augmentation

For ASPEC dataset (Nakazawa et al., 2014), the quality of first 1M data is better than other 2M data. Thus, first 1M data are included in our training data. Furthermore, in order to use more data, we proposed an approach to select pairs from the second 1M data. Specifically, we first train a translation model using first 1M data, and then for each source sentence in the second 1M data, we generate a predicted sentence based on trained model. A BLEU score is calculated by comparing predicted sentence with target sentence. If the BLEU score is zero, then the source sentence and corresponding target sentence is added to training dataset. Finally, we train a new model based on augmented dataset. The intuition is that predicted sentences with zero BLEU scores are supposed to be not fully understood by trained model, so these sentences are added to training dataset and translation model is expected to learn more pattern from these data.

3 Experiments

We experimented our NMT system on Japanese-English, English-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-Chinese subtask, all the sentences in ASPEC corpora are used as training data. For Japanese-

English and English-Japanese subtasks, we used only the first 1 million sentences sorted by sentence-alignment similarity. Furthermore, for Japanese-English subtask, we augmented training data to nearly 2 million by our proposed method. And, for all corpora, Japanese sentences were segmented by the morphological analyzer Juman[2] and English sentences were tokenized by tokenizer.perl of Moses (Koehn et al., 2007). Sentences with more than 60 words were excluded. We used subword unit, that is Joint Byte Pair Encoding (BPE) (Sennrich et al., 2016c) scheme, to encoder vocabulary for both source and target sentences. Table 1 shows the numbers of the sentences in each parallel corpus.

	Ja-En	En-Ja	Ja-Ch
Train	1,770,818	1,000,000	672,315
Dev	1,790	1,790	2,741
test	1,812	1,812	2,300

Table 1: Number of parallel sentences

3.2 Results

Table 2 lists our final results for three languages subtasks. According to evaluation systems, for all these subtask, we got best performance in BLEU measure. We described experimental results in detail in the following subsections.

	Ja-En	En-Ja	Ja-Zh
BLEU	30.59	43.43	37.60

Table 2: Results of Subtasks

Japanese-English subtask:

The Japanese-English subtask is the main subtask that we participated in and all the technical points mentioned in section 2 have been used. All the useful ones with corresponding contributions are listed in Table 3.

System	BLEU
Baseline	26.27
BPE Subword 1M	28.34
Data Augmentation	29.41
Relative Position	29.77
Average checkpoint	30.18
Step-wise	30.59

Table 3: Technical point contributions

The baseline model is the transformer model with 3M parallel sentences. There can be more than 2 BLEU scores improvement only by using the first 1M parallel sentences with BPE subword. In relation to data augmentation, if the second 1M data are all added to the first 1M training data, the BLEU score decreases to 28.28. On the other hand, if only the sentences with zero BLEU scores are added, the BLEU score for the model increases by more than 1 point. According to the BLEU measurement, the zero score indicates that there is no matched 4-gram between predictions and references, which means these sentences have not been trained well. Relative position can also yield more than 1 score improvement based on the baseline system. As for the ensemble part, both average checkpoint and step-wise can increase almost 0.5 BLEU score. And the ensemble turn is using average checkpoint first followed by step-wise.

The results show each technical points has an obvious contribution, however, they are not precisely superimposed, adding the contributions together is an important part in our real work.

English-Japanese subtask:

	Baseline	3 models	4 models
BLEU	41.98	42.49	43.43

Table 4: Results of step-wise Ensemble

The English-Japanese subtask is the opposite direction of the Japanese-English subtask, so the language characteristics are almost the same. Thus we used the same technical points as Japanese-English subtask except that turn over the training dataset. In addition, due to the limited time, we don’t use data augmentation in this subtask. The results are shown in table 4. As we can figure out the BLEU score of 4 models is higher nearly 1 score than 3 models, so there is potential to get higher scores with more models.

Japanese-Chinese subtask:

	2 models	4 models	10 models
BLEU	36.31	37.03	37.53

Table 5: Results of step-wise ensembling

The results of step-wise ensembling for Japanese-Chinese are showed in table 5. To introduce variation, models are trained with different hyper-

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

parameters. Due to the resource limitation, only 10 models with BLEU score between 35.93 and 36.33 are selected to perform the final ensemble. It can be observed that the BLEU score goes up as the number of models increases. Based on the translation candidates generated by the model with BLEU score 37.53, we apply rerank method to achieve the final result shown in table 2.

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 and ensembling. We evaluated our Transformer system on Japanese-English, English-Japanese and Japanese-Chinese scientific paper translation subtasks at WAT 2018. The experimental results show that the implementation of relative position representation and ensembling decoding 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 the weighted transformer (Ahmed et al., 2017), which replaces the multi-head attention by multiple self-attention branches that the model learns to combine during training process. We also plan to investigate the impact of parameters, such as batch size and learning rate, on translation quality in future.

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.

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

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

Shaw P, Uszkoreit J, Vaswani A. 2018. Self-Attention with Relative Position Representations. arXiv preprint arXiv:1803.02155.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450 .

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.

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

Ahmed K, Keskar N S, Socher R. 2017. Weighted Transformer Network for Machine Translation. arXiv preprint arXiv:1711.02132.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

He K, Zhang X, Ren S, et al. 2016 Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770-778.

Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. 1998. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02): 107-116.

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.

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

Philipp Koehn, Hieu Hoang, and et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the ACL-2007.