ニューラル機械翻訳のためのデータ拡張

A Data Augmentation Method for Neural Machine Translation

劉　柏延[*]　　姜　珊珊[*]　　董　濱[*]
Boyan LIU　　Shanshan JIANG　　Bin DONG

要　旨

ニューラル機械翻訳（Neural Machine Translation: NMT）モデルは通常，対訳コーパスにより学習され，翻訳品質を改善するための新しい拡張コーパスを生成する．従来の研究では，NMTの品質の主要なテスト要素であるn-gramを考慮せずに対訳コーパスを生成ないし選択していた．本論文では，翻訳品質向上のための対訳コーパス選択標準としてn-gramを利用する，ニューラル機械翻訳のためのデータ拡張について説明する．実験により，この方法は，2018年に開催されたアジア諸言語翻訳に関するワークショップ（Workshop on Asian Translation: WAT）の基本的なNMTモデルよりも，英日と日英翻訳タスクにおけるBLEUスコアをそれぞれ最大0.86および1.34まで改善することが示された．

ABSTRACT

Neural machine translation (NMT) models typically operate with a parallel corpus and generate a new augmented corpus to improve translation quality. Previous studies created or selected the parallel corpus without considering the importance of the n-gram algorithm in BLEU. N-gram is a key test element in the NMT quality standard that should be taken into account. This paper describes a data augmentation method for neural machine translation. We regard n-gram as a select standard for a parallel corpus which can enhance translation quality. Experiments showed that compared with the basic NMT model, our method improved BLEU scores by 0.86 and 1.34, respectively, when performing English-Japanese and Japanese-English translation tasks from the Workshop on Asian Translation (WAT).

[*] リコーソフトウェア研究所（北京）有限公司
Ricoh Software Research Center (Beijing) Co., Ltd.

1. Introduction

Neural machine translation (NMT) has greatly improved translation quality in recent years, compared with statistical machine translation (SMT). The sequence to sequence model[1] is a milestone for NMT, which was the first model that outperformed the traditional phrased-based machine translation. The sequence to sequence model has an encoder and decoder scheme and can be built based on a recurrent neural network (RNN)[2], long short-term memory network (LSTM)[3], convolutional neural network (CNN)[4], etc. Compared with the original RNN-based NMT, the ConvS2S model[5] based on CNNs can be fully parallel trained, resulting in more accurate translations as well as faster training. Moreover, the transformer model[6] based on self-attention[7] and feed-forward connections can overcome the difficulties in parallel training and further improve both the convergence speed and the translation quality. Nowadays, the Transformer model has become the paradigm for neural machine translation that can produce state-of-the-art translation quality.

In addition to the model iteration, a high quality and abundant parallel corpus is also crucial for a translation system. The most widely used generation method of the augmented corpus is back translation[8]. By pairing monolingual training data with an automatic back-translation, we can treat it as additional parallel training data to improve translation quality. NTT used this method for their NMT system in the Workshop on Asian Translation (WAT) 2017, where they won first place[9]. However, this method does not take into account the n-gram algorithm, a key test element in the NMT quality standard. Our method adds an n-gram selection step to select the parallel sentences that the original model could not train well, thus improving the quality of the translation. Experiments on this algorithm showed significant improvements in BLEU scores.

2. System

2-1 Base Model

Our NMT system is based on the widely used Transformer model. As for the implementation framework, we used the open-source OpenNMT which is maintained by Harvard University. This framework provides several kinds of models from statistical machine translation to neural machine translation.

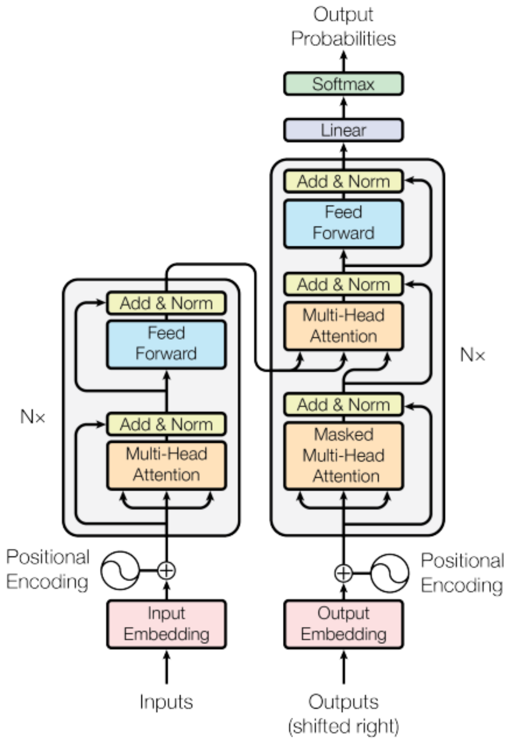

Fig. 1 Transformer model architecture.

As Fig. 1 shows, the Transformer model features a sequence to sequence architecture and consists of an encoder layer and a decoder layer. At the bottom of the encoder and decoder stacks[10], the encoder layer consists of two sublayers: a multi-head self-attention, which allows the model to jointly attend to information from different representation subspaces at different positions, followed by a position-wise fully connected feed-forward layer. The decoder layer consists of three sublayers: a masked multi-head self-attention, encoder-decoder attention, and a position-wise feed-forward layer. Residual connections are used between each sublayer, followed by layer normalization[11], which can help propagate position 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 regarding 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 bottom of the encoder and decoder stacks. The absolute position representations hypothesized that sinusoidal position encodings would help the model to generalize sequence lengths unseen during training.

2-2 Augmented Data

We used the Asian Scientific Paper Excerpt Corpus (ASPEC)[12] as parallel corpora for all language pairs. For the ASPEC dataset, the quality of the first million sentence pairs (1M) is more precise than the other 2 million (2M). Thus, the first 1M sentences pairs are included in our training data. Furthermore, to use more data, we proposed an approach to select pairs from the second 1M data.

We first trained a translation model using the first 1M sentence pairs, and then we generated a predicted sentence based on the trained model for each source sentence in the second 1M. A BLEU score was calculated by comparing the predicted sentence with the reference sentence. If there was no matching n-gram between the predicted sentence and the corresponding reference sentence, the BLEU score was zero, and the source sentence and the corresponding reference sentence were added to the training dataset. Since n is four in the most common scenarios, we selected the 4-gram. Finally, we trained a new model based on the augmented dataset.

The schematic diagram of data augmentation method is shown in Fig. 2. If there was no matching n-gram between the prediction and the reference (indicated by the red line), the input and the corresponding reference formed new parallel sentences (blue line).

It is presumed that predicted sentences with zero BLEU points are not supposed to be fully understood by the trained model. These sentences are added to the training dataset, and the translation model is expected to learn more patterns from the added data.

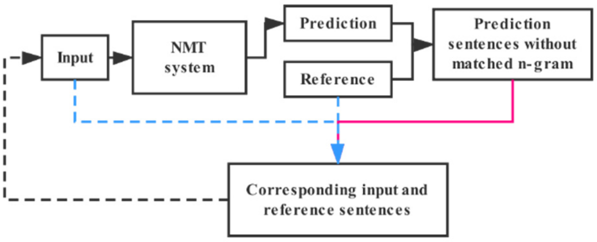
Input
NMT system
Prediction
Reference
Prediction sentences without matched n-gram
Corresponding input and reference sentences

Fig. 2 Data augmentation schematic diagram.

3. Experiments

We experimentally evaluated our NMT system using Japanese-English and English-Japanese scientific paper translation tasks.

3-1 Datasets

For Japanese-English and English-Japanese tasks, we used the first 1M sentence pairs sorted by sentence-alignment similarity as the baseline system, referred to as 1M (system) in Tables 2 and 3. Then we directly added the second 1M sentences, referred to as 2M (system) in Tables 2 and 3, to the first 1M to train the model. Furthermore, for both Japanese-English and English-Japanese tasks, we selected the second 1M data to augment the first 1M training data for a total of nearly 2M, which are referred to as Data augmentation (system) in Tables 2 and 3.

Table 1 shows the number of sentence pairs in each parallel corpus after data augmentation.

Table 1 Number of sentence pairs in parallel corpus.

	Ja-En	En-Ja
Train	1,770,818	1,703,014
Dev	1,790	1,790
Test	1,812	1,812

For all corpora, Japanese sentences were segmented by Juman and English sentences were tokenized by the Moses tokenizer.perl . Sentences with more than 60 words were excluded. We used the subword unit, that is the Joint Byte Pair Encoding (BPE)[13] scheme, to encode vocabulary for both source and target sentences. The Transformer is the base model of OpenNMT which has six encoder layers and six decoder layers. The embedding size is 512 for both the encoder and decoder. The inner dimension for the feed-forward network is 2048. The dropout rate is 0.1. Additional details can be found in the OpenNMT toolkit.

3-2 Results

For both Japanese-English and English-Japanese tasks, we used the same data augmentation method as mentioned above. First, we trained an original model based on the first 1M parallel corpus. Then we predicted the second 1M source sentences to predict target sentences and selected the parallel sentences which don't have matched n-gram adding to the first 1M parallel corpus. Finally, we used the new parallel corpus to train a new model and predict. The two direction tasks both showed improvements.

(1) English-Japanese task

As shown in Table 2, the BLEU score of the first 1M parallel corpus trained model was 40.99. After adding the second 1M data to train the model directly, the BLEU score decreased to 40.17. However, the data augmentation BLEU score based on the second 1M parallel data was 41.85, an improvement of 0.86.

Table 2 Data augmentation results on En-Ja.

System	BLEU	
1M	40.99	
2M	40.17	↓ **0.82**
Data augmentation (~2M)	41.85	↑ **0.86**

(2) Japanese-English task

We trained the original model using the first 1M parallel corpus with a BLEU score of 28.34. The former 2M model's BLEU also decreased to 28.13. The model's BLEU score with the same 2M data using data augmentation improved by 1.34, a significant improvement as shown in Table 3.

Table 3 Data augmentation results on Ja-En.

System	BLEU	
1M	28.86	
2M	28.13	↓ **0.73**
Data augmentation (~2M)	30.20	↑ **1.34**

4. Conclusion

In this paper, we described our NMT system, which is based on the Transformer model. We evaluated our data augmentation method using Japanese-English and English-Japanese scientific paper translation tasks from WAT. The experimental results showed that our method can effectively improve translation quality.

N-gram is a key test standard in an NMT system that also directly affects accuracy and fluency. We plan to explore more methods based on n-gram in future work, such as loss function or beam search. We also plan to explore other methods of back translation as monolingual data is very common in our daily lives. Semi-supervised or unsupervised learning with n-gram is also a future research direction.

References

1) Kalchbrenner N, Blunsom P: Recurrent continuous translation models, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709 (2013).

2) Schuster M, Paliwal K K: Bidirectional recurrent neural networks[J], *IEEE Transactions on Signal Processing*, Vol. 45, Issue 11, pp. 2673–2681 (1997).

3) Hochreiter S, Schmidhuber J: Long short-term memory[J], *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).

4) LeCun Y, Bengio Y: Convolutional networks for images, speech, and time series, *The Handbook of Brain Theory and Neural Networks*, p. 3361(10) (1995).

5) Gehring J et al.: Convolutional sequence to sequence learning, *Proceedings of the 34th International Conference on Machine Learning*, JMLR.org, Vol. 70, pp. 1243–1252 (2017).

6) Vaswani A et al.: Attention is all you need, *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017).

7) Kalchbrenner N, Blunsom P: Recurrent continuous translation models, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1700–1709 (2013).

8) Sennrich R, Haddow B, Birch A: Improving neural machine translation models with monolingual data, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 86–96 (2016).

9) Morishita M, Suzuki J, Nagata M: NTT neural machine translation systems at WAT 2017, *Proceedings of the 4th Workshop on Asian Translation*, pp. 89–94 (2017).

10) Bahdanau D, Cho K, Bengio Y: Neural machine translation by jointly learning to align and translate, ICLR (2015).

11) Ba J L, Kiros J R, Hinton G E: Layer normalization, ICLR (2016).

12) Nakazawa T et al.: ASPEC: Asian Scientific Paper Excerpt Corpus, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pp. 2204–2208 (2016).

13) Sennrich R, Haddow B, Birch A: Neural machine translation of rare words with subword units, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1715–1725 (2016).