Improving Low-Resource Neural Machine Translation with Weight Sharing

Tao Feng[1,2], Miao Li[1], Xiaojun Liu[3], and Yichao Cao[1,2]

[1] Institute of Intelligent Machines, Chinese Academy of Science, Hefei, China
[2] University of Science and Technology of China, Hefei, China
[3] School of Information and Computer, Anhui Agricultural University, Hefei, China
ft2016@mail.ustc.edu.cn, mli@iim.ac.cn, Lxj2442@ahau.edu.cn,
cycao@mail.ustc.edu.cn

Abstract. Neural machine translation (NMT) has achieved great success under a great deal of bilingual corpora in the past few years. However, it is much less effective for low-resource language. In order to alleviate the problem, we present two approaches which can improve the performance of low-resource NMT system. The first approach employs the weight sharing of decoder to enhance the target language model of low-resource NMT system. The second approach applies cross-lingual embedding and source sentence representation space sharing to strengthen the encoder of low-resource NMT. Our experiments demonstrate that the proposed method can obtain significant improvements on low-resource neural machine translation than baseline system. On the IWSLT2015 Vietnamese-English translation task, our model can improve the translation quality by an average of 1.43 BLEU scores. Besides, we can also get the increase of 0.96 BLEU scores when translating from Mongolian to Chinese.

Keywords: Low-resource, Neural machine translation, Weight sharing

1 Introduction

Machine translation is an important part of artificial intelligence, which explores how to use computers to translate one language into the other. In recent years, neural machine translation (NMT) has achieved great success because of the development of deep learning and the availability of large-scale parallel corpus[1, 2]. In a variety of language pairs, the performance of neural machine translation has gradually surpassed phrase-based statistical machine translation (SMT)[3, 4]. NMT is an end-to-end translation method, which typically consists of an encoder and a decoder[5, 6]. More concretely, the encoder network maps the input sequence to a fixed-length vector and the decoder network gets translation from the vector. However, the defect of encoder-decoder framework is that the encoder only obtains all the information of source sentences through a fixed-length vector. This leads to the poor performance of NMT in long sentences. In order to solve this problem, the attention mechanism was proposed by [7, 8].

The attention mechanism can utilize relevant source side information to help predict the current target word, and significantly improve the performance of NMT. Therefore, the encoder-decoder framework with attention has become the mainstream method of the neural machine translation.

However, as a data-driven approach, the performance of NMT is severely affected by the size and the quality of parallel corpus. As the decrease of parallel corpus, the quality of NMT is greatly reduced[9, 10]. In this case, the NMT lags behind statistical machine translation on low-resource language pairs. However, the vast majority of language pairs lack a large amount of parallel corpus[11]. Therefore, research on low-resource is valuable.

In this paper, we investigate the usage of the similarity and complementarity between different languages to obtain high-quality context vector and strengthen the decoder for low-resource neural machine translation. Intuitively, we employ multi-task learning framework to build two NMT models, one is a low-resource model (e.g., Vietnamese-English) and the other is a high-resource model (e.g., French-English). These two models share the weights of certain layers. To achieve this goal, we propose two approaches. Inspired by [12], the first approach exploits weight sharing of decoder side between low-resource model and high-resource model to improve the performance of target language model of the low-resource NMT system.

The proposed second approach applies multi-lingual translation system to share word embedding space and sentence representation space between different languages in the source side. The motivation behind this is that we can obtain better context vector for low-resource NMT model with the assistance of high-resource parallel corpus. More concretely, the approach builds upon the recent work on cross-lingual embedding[13]. First, we train the embedding for different source languages on monolingual corpora, and then learn a liner transformation to map the embedding from one space to the other. Therefore, we can align the word embedding space in this way. For sentence representation, not only in order to maintain the independence of each language, but also to map sentence representation into same space, we share the weights of last few layers of the encoder, not all of its layers. In this work, we make following contributions:

- To fully investigate weight sharing in low-resource NMT model, we propose and compare two methods. One attempts to reinforce the decoder side for low-resource model by sharing the weights of decoder layers with the high-resource model, and the other tries to share word embedding space and sentence representation space between source languages, so that we can get high-quality context vector for low-resource model.

- The experiments on Vietnamese-to-English and Mongolian-to-Chinese translations show that our proposed methods significantly outperform the NMT baseline model with attention mechanism.

2 Neural Machine Translation Background

The encoder-decoder NMT model has been proposed in recent years[7, 8] and consists of three parts: encoder, attention and decoder. The model takes a source sequence $X = (x_1, x_2, ..., x_{T_x})$ as input and generates corresponding translation $Y = (y_1, y_2, ..., y_{T_y})$, where x_t and y_t are the symbols of source language and the target language respectively.

Encoder: Given a source sentence X, the encoder builds a continuous representation with recurrent neural networks (RNNs). In NMT model, bi-directional neural networks including a forward RNN and a backward RNN are often implemented. The forward RNN reads the input sentence from left to right: $\overrightarrow{h}_t = \overrightarrow{f}_{enc}(E_x(x_t), \overrightarrow{h}_{t-1})$. Similarly, the backward RNN reads the input sentence from right to left: $\overleftarrow{h}_t = \overleftarrow{f}_{enc}(E_x(x_t), \overleftarrow{h}_{t-1})$, where the E_x is the word embedding matrix and the h_t is a hidden state of RNN at time t. $\overrightarrow{f}_{enc}$ and $\overleftarrow{f}_{enc}$ are some nonlinear functions. In encoder side, the RNN can be a Long Short Term Memory Unit(LSTM) or a Gate Recurrent Unit(GRU).

Attention: The attention mechanism[7, 8] was proposed to dynamically compute the context vector of the source end. In general, the current target hidden state is compared with all source states to derive attention weight α_{ts}. Calculate a context vector $c_t = \sum_{s=1}^{T_s} \alpha_{ts} h_s$ as the weighted average of the source states based on the attention weights. Then combine the context vector with the current target hidden state to generate final attention vector a_t. In attention mechanism, the attention weight α_{ts} is used to measure the correlation between the t-th target token and s-th source token, and is calculated as follows:

$$\alpha_{ts} = \frac{exp(score(h_t, h_s))}{\sum_{s'=1}^{T_s} exp(score(h_t, h_{s'}))} \tag{1}$$

where h_t is the hidden state of target at time t and h_s is the hidden state of source at time s. $score()$ is a nonlinear function, usually a feed-forward neural network with a hidden layer.

Decoder: The decoder utilizes recurrent neural networks to predict the target sequence $y = (y_1, y_2, ..., y_{T_y})$. Each word y_i is predicted based on recurrent neural network hidden state h_i, the previously predicted word y_{i-1}, and a context vector c_i. Therefore, each conditional probability is calculated as follows:

$$p(y_t|\{y_1, y_2, ..., y_{t-1}\}, c) = f(y_{t-1}, h_t, c) \tag{2}$$

where the f is a nonlinear function, usually a multi-layered neural network. However, other architectures such as convolutional neural network or hybrid neural network can be used[5]. In summary, the decoder defines the joint probability for translation y:

$$p(y) = \prod_{t=1}^{T} p(y_t|y_1, ..., y_{t-1}, c) \tag{3}$$

where the $y = (y_1, y_2, ..., y_{T_y})$.

3 Architecture

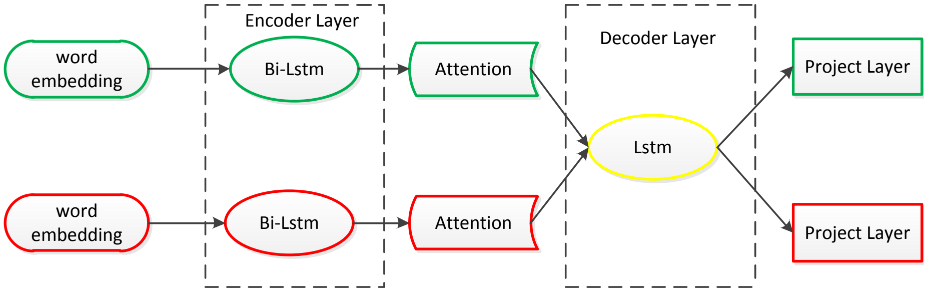

Fig. 1. The framework of SD model. The model consists of three parts, the green curve is a high-resource NMT model, the red curve is a low-resource NMT model, and the yellow curve represents a shared part of the two models. In the training process, the two models independently train the encoder, but share the weights of the decoder. The word embeddings are initialized randomly and updated continuously in iteration. Besides, we investigate the influence of the shared attention mechanism on low-resource NMT model.

The low-resource neural machine translation has attracted lots of attention in recent times. Many authors have conducted in-depth research on this issue, especially how to make use of high-resource parallel corpora to assist low-resource NMT[10, 14]. It is well known that more high-quality related data can lead to better and more robust network models. For example, the amount of Vietnamese-English corpus is not big enough, but the French-English parallel corpus is large, so we can utilize French-English parallel corpus to improve the performance of Vietnamese-English NMT model. In this paper, we exploit high-resource parallel corpus to enhance the encoder and decoder of low-resource NMT model respectively.

3.1 Strengthen Decoder

In neural machine translation model, the decoder plays an important role in improving fluency for translation system. In essence, the decoder is a recurrent neural network language model that is conditioned on source context in encoder-decoder architecture for NMT. In this section, we aim to exploit the signals from high-resource target side corpora to enhance the decoder of low-resource neural machine translation model, which we refer to as SD model. For detail, we share the weights of decoder between high-resource NMT model and low-resource NMT model to achieve the goal. Given the low-resource parallel corpora $D_L = \{(X^{(n,1)}, Y^{(n,1)})\}_{n=1}^{N_1}$, where the N_1 is not big enough. And we also have large-scale high-resource language pairs $D_H = \{(X^{(n,2)}, Y^{(n,2)})\}_{n=1}^{N_2}$ in which the $N_2 >> N_1$. In the SD model, the neural machine translation is trained with maximum likelihood on the mixed language pairs $\{X^{(n,k)}, Y^{(n,k)}\}_{k=\{1,2\}}^{n=1,...,N_k}$:

$$L(\theta) = \frac{1}{2} \sum_{k=1}^{2} \sum_{n=1}^{N_k} logp(Y^{(n,k)}|X^{(n,k)};\theta) \tag{4}$$

where θ is parameter of the neural network.

Since the languages utilized by the two models on the decoder are same, we can make full use of the high-resource target language information to improve the performance of the decoder of the low-resource neural machine translation. In addition, we also investigate the influence of the shared attention layer on low-resource neural machine translation model. The Fig.1 summarizes the general schema of the SD model.

3.2 Strengthen Encoder

Generally speaking, both the source and target word embedding are randomly initialized and then updated as the number of iterators increase in NMT model. This method is feasible for large-scale parallel corpus, but it performs poorly on low-resource corpora. Due to the small size of parallel corpora in low-resource NMT, the translation system can not fully learn the internal structure of sentences and lexical information, which results in the model being unable to generate valid word embedding. This is a major limitation on low-resource neural machine translation system. On the other hand, in many-to-one translation system, it is very significant for low-resource corpora to share sentence representation space with similar languages. For example, it is feasible for low-resource NMT model to share source sentence representation space with high-resource language if the two languages have similar word order. In this section, we use cross-lingual embedding and weight sharing of encoder to alleviate these two issues, which we refer to as SE model.

Share word embedding space For the first problem, we utilize the cross-lingual embedding to align word embedding space. We train the embedding for each language using monolingual corpora independently, and then learn a liner transformation to map the word embedding from one space to the other. For detail, let X and Y represent the word embedding matrix so that X_i is the word embedding of the i-th entry in the source vocabulary table and Y_j corresponds to the j-th target language embedding. The goal is to find a mapping matrix M that satisfies the following condition:

$$M_* = argmin \sum_i \sum_j D_{ij} ||X_{i*}M - Y_{j*}||^2 \tag{5}$$

where $D_{ij} = 1$ if i-th source language word is aligned with the j-th target language word else 0. In this paper, we follow the cross-lingual embedding method proposed by[13].

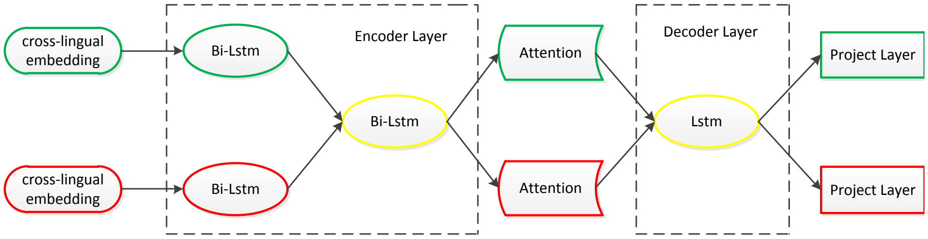

Fig. 2. The framework of SE model. The model can be divided into three parts, the green curve is a high-resource NMT model, the red curve is a low-resource NMT model, and the yellow curve represents a shared part of the two models. In the training process, we utilize cross-lingual embedding to share word embedding space and map the representation of sentences from different languages to same space by sharing weights of the last few layers of the encoder.

Share sentence representation space For sentence representation space sharing, we propose a simple and effective way to implement it. As shown in Fig.2, the proposed method utilizes weight sharing to map representation of sentences from different languages to same space. More concretely, we exploit two independent encoders but sharing the weights of last few layers to extract the high-level representation of the input sentences. Given the input sequence $X = (x_1, x_2, ..., x_n)$ and the initial output sequence of the encoder stack $H = (h_1, h_2, ..., h_n)$, we calculate H_e as follows:

$$H_e = f \odot H + (1 - f) \odot E(X) \tag{6}$$

where H_e is the final output sequence of the encoder and which will be utilized to calculate the context vector by the decoder and the E is cross-lingual word embedding matrix. f is a gate unit and calculated as follows:

$$f = g(WE + UH + b) \tag{7}$$

where the W and U are the weights of neural network, and the b is bias and they are shared by the two encoders. The following reasons support this approach:

- Through the weight sharing of the last few layers, the two encoders can generate approximate output when the sentences with similar semantics from different languages are used as input. Therefore, the proposed method can get high-quality context vector for low-resource NMT model with the assist of high-resource NMT system.
- There are differences between languages, such as syntax and lexical. In our model, the weights of first few layers of the encoder are not shared, which is to obtain the characteristics of each language for the translation system. Accordingly, we share the weights of last few layers rather than the entire encoder.

4 Experiments

In this section, we describe the data set used in our experiments, data processing, the training details and all the translation results we obtain in experiments.

4.1 Dataset

We evaluate our models on four language pairs: French-English, Vietnamese-English, English-Chinese, Mongolian-Chinese. In our experiments, we translate Vietnamese into English with the help of French-English. Similarly, we translate Mongolian into Chinese with the assistance of English-Chinese.

The Vietnamese-English (133K sentence pairs, 2.7million English words and 3.3 million Vietnamese words) is provided by IWSLT2015 and Mongolian-Chinese (67K sentences pairs, 848K Chinese words and 822K Mongolian words) is provided by CWMT2009. We evaluate our approach on the French-English (2 million sentence pairs, 50 million English words and 52 million French words) translation task of the WMT14 workshop. And we obtain English-Chinese parallel corpus(2 million sentence pairs, 22million English words and 24 million Chinese words) from the WMT17. The Chinese sentences are word segmented using Stanford Word Segmenter. We preserve casing for words and replace those whose frequencies are less than 5 by <unk>. As a result, our vocabulary table size is 17K and 7.7K for English and Vietnamese respectively. And we report BLEU scores on tst2012 and tst2013 for Vietnamese-English translation system. And we make the same treatment for Mongolian-Chinese parallel corpora. Therefore, the size of Chinese and Mongolian vocabulary table is 14K and 12K respectively.

4.2 Training setup

In our experiment, we exploit encoder-decoder framework with attention mechanism to train NMT model. More concretely, we employ two-layer bi-directional RNN in the encoder, and another two-layer uni-directional RNN in the decoder. All the RNNs use LSTM[15] cells with 600 units, and the dimensionality of word embedding is set to 512. As for attention mechanism, we use the global attention method proposed by [8]. The models are trained using stochastic gradient descent and the maximum length of the sentence is 50. We apply dropout[16] with a probability of 0.25 during training. For all models, the initial learning rate is 0.2, and then it decreases as the number of iterations increases. We initialize all of the parameters of network with the uniform distribution. The maximum value of the gradient is set to 5 in order to solve gradient explosion.

4.3 Results and analysis

The results of BLEU scores are presented in Table 1. The architecture of baseline system is similar to the one mentioned in section 4.2. However, in order to prevent overfitting, we exploit one-layer bi-directional LSTM in the encoder, with 512 units in each cell.

Table 1. The performance of proposed method on IWSLT2015 Vietnamese to English tst2012 and tst2013 set and CWMT2009 Mongolian to Chinese test set.

Models	BLEU		
	Vi-En(tst2012)	Vi-En(tst2013)	Mn-Ch
Baseline	20.15	23.07	11.69
SD	20.62	23.59	12.07
SD + share attention	20.48	23.34	11.83
SE	**21.43**	**24.65**	**12.65**

As it can be seen from Table 1, the proposed method obtains very competitive results compared to the baseline system. Our model can reach 21.43 and 24.65 BLEU scores in Vietnamese-English tst2012 and tst2013 set respectively, and we can also achieve 12.65 BLEU scores in Mongolian-Chinese test set, which is much stronger than the baseline system, with improvements of at least 6.3% in all cases, and up to 8.2% in some (e.g. from 11.69 to 12.65 BLEU scores in Chinese to Mongolian). The experiment results show that we can improve the performance of low-resource neural machine translation with the help of high-resource language pairs.

In addition, from table 1, we can see that the model can increase the 0.47 and 0.52 BLEU scores in Vietnamese-English tst2012 and tst2013 set respectively by only strengthening the decoder of the low-resource NMT model, and the BLEU scores of Mongolian-Chinese is also improved. It reveals that the low-resource NMT system generates a better target side language model than the baseline system by sharing the weights of decoder with high-resource language pairs. However, the performance of low-resource neural machine translation is reduced when we share the weights of attention layer with the high-resource neural machine translation model. Compared with the SD model, the BLEU scores of Vietnamese-English decrease by 0.14 and 0.25 respectively, and the BLEU scores of Mongolian-Chinese also decline. The attention mechanism is used to capture the source side information dynamically, which allows model learning to align between the target language and source language. For different source languages, the alignment matrix obtained by the model is not same. Therefore, sharing attention layer can lead to a decrease in the performance of low-resource neural machine translation system.

5 Related works

Low-resource neural machine translation has attracted a lot of attention in recent years. [10] presented a transfer learning method to improve the performance of low-resource neural machine translation. Their main idea was to first train a high-resource language pair model, then transfer some of the learned parameters to the low-resource pair to initialize and constrain training. Besides, semi-supervised approach is another way to deal effectively with insufficient resources.

[17] explored strategies to train with monolingual data without changing the neural network architecture. They utilized dummy source sentences and synthetic source sentences to construct pseudo-parallel corpora, which brings substantial improvements to neural machine translation. [18] converted a monolingual corpus in the target language into a parallel corpus by copying it, so that each source sentence is identical to its corresponding target sentence. [19] investigated how to utilize the source-side monolingual data in NMT to enhance encoder network. They applied the multi-task learning framework using two NMTs to predict the translation and the reordered source-side monolingual sentences simultaneously. For zero-resource neural machine translation, [20] made attempt to train a source-to-target NMT model without parallel corpora available, guided by an existing pivot-to-target NMT model on a source-pivot parallel corpus. [21] proposed an approach to zero-resource NMT via maximum expected likelihood estimation. Their results revealed that maximum expected likelihood estimation can greatly improve the performance of NMT. [22] introduced automatic encoder to neural machine translation, and proposed a semi-supervised learning method based on bilingual corpus and monolingual corpus. [23] proposed two methods, which are referred to as shallow fusion and deep fusion, to integrate a language model into NMT. The basic idea is to use the language model to score the candidate words proposed by the translation model at each time step or concatenating the hidden states of the language model and the decoder. [24] proposed a finetuning algorithm for multiway, multilingual neural machine translation that enables zero-resource machine translation. In unsupervised Machine Translation, [25] proposed a method to build unsupervised NMT model. They combined unsupervised cross-lingual embedding, denoising auto-encoder and dual learning together to train NMT system in a unsupervised manner. The method proposed in [26] was similar to [25], but the work in [26] was more complete. Although unsupervised machine translation methods are promising, their performance is far lower than supervised machine translation. In multi-task neural machine translation, [27] proposed a method that can simultaneously translate sentences from one source language to multiple target languages. In detail, their models shared source languages representation and separated the modeling of different target language translation.

6 Conclusion

In this paper, we aim to utilize high-resource languages to improve the performance of low-resource neural machine translation. We propose two methods to achieve this goal. One is to exploit the weight sharing of decoder to enhance the target side language model of low-resource NMT system. The other is to enhance the encoder by using cross-lingual embedding and shared sentence representation space.

The experiments show the effectiveness of our proposal, which has significant improvements in the BLEU scores over baseline system. Our model can improve the translation quality on the IWSLT2015 Vietnamese-English translation task.

In addition, the proposed approaches in this paper is also effective for Mongolian-Chinese translation.

In the future, we plan to combine unsupervised or semi-supervised methods with our model. Besides, we will verify the approach with more datasets from different domains.

7 Acknowledgements

The work is supported by the Nation Natural Science Foundation of China under No.61572462,61502445.

References

1. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv:1609.08144, 2016.
2. Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[J]. arXiv preprint arXiv:1705.03122, 2017.
3. Junczys-Dowmunt M, Dwojak T, Hoang H. Is neural machine translation ready for deployment? a case study on 30 translation directions[J]. arXiv preprint arXiv:1610.01108, 2016.
4. Bojar O, Chatterjee R, Federmann C, et al. Findings of the 2016 Conference on Machine Translation[C]. In: ACL 2016 FIRST CONFERENCE ON MACHINE TRANSLATION (WMT16). The Association for Computational Linguistics, 2016: 131-198.
5. Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1700-1709.
6. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C].In: Advances in neural information processing systems. 2014: 3104-3112.
7. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.
8. Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[J]. arXiv preprint arXiv:1508.04025, 2015.
9. Koehn P, Knowles R. Six challenges for neural machine translation[J]. arXiv preprint arXiv:1706.03872, 2017.
10. Zoph B, Yuret D, May J, et al. Transfer learning for low-resource neural machine translation[J]. arXiv preprint arXiv:1604.02201, 2016.
11. Artetxe M, Labaka G, Agirre E, et al. Unsupervised neural machine translation[J]. arXiv preprint arXiv:1710.11041, 2017.
12. Johnson M, Schuster M, Le Q V, et al. Google's multilingual neural machine translation system: enabling zero-shot translation[J]. arXiv preprint arXiv:1611.04558, 2016.
13. Artetxe M, Labaka G, Agirre E. Learning bilingual word embeddings with (almost) no bilingual data[C]. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017, 1: 451-462.
14. Nguyen T Q, Chiang D. Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation[J]. arXiv preprint arXiv:1708.09803, 2017.

15. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
16. Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks[C]. In: Advances in neural information processing systems. 2016: 1019-1027.
17. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data[J]. arXiv preprint arXiv:1511.06709, 2015.
18. Currey A, Barone A V M, Heafield K. Copied monolingual data improves low-resource neural machine translation. In: Proceedings of the Second Conference on Machine Translation. 2017: 148-156.
19. Zhang J, Zong C. Exploiting source-side monolingual data in neural machine translation[C]. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 1535-1545.
20. Chen Y, Liu Y, Cheng Y, et al. A Teacher-Student Framework for Zero-Resource Neural Machine Translation[J]. arXiv preprint arXiv:1705.00753, 2017.
21. Zheng H, Cheng Y, Liu Y. Maximum expected likelihood estimation for zero-resource neural machine translation[C]. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017). Melbourne, Australia. 2017: 4251-4257.
22. Cheng Y, Xu W, He Z, et al. Semi-supervised learning for neural machine translation[J]. arXiv preprint arXiv:1606.04596, 2016.
23. Gulcehre C, Firat O, Xu K, et al. On using monolingual corpora in neural machine translation[J]. arXiv preprint arXiv:1503.03535, 2015.
24. Firat O, Sankaran B, Al-Onaizan Y, et al. Zero-resource translation with multi-lingual neural machine translation[J]. arXiv preprint arXiv:1606.04164, 2016.
25. Artetxe M, Labaka G, Agirre E, et al. Unsupervised neural machine translation[J]. arXiv preprint arXiv:1710.11041, 2017.
26. Lample G, Denoyer L, Ranzato M A. Unsupervised Machine Translation Using Monolingual Corpora Only[J]. arXiv preprint arXiv:1711.00043, 2017.
27. Dong D, Wu H, He W, et al. Multi-task learning for multiple language translation[C] Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 1723-1732.