Low-Resource Neural Machine Translation with Transfer Learning

Tao Feng[1,2]**, Miao Li**[1]**, Lei Chen**[1]
[1]Institute of Intelligent Machines, Chinese Academy of Science, Hefei, China
[2]University of Science and Technology of China, Hefei, China
ft2016@mail.ustc.edu.cn, {mli, chenlei}@iim.ac.cn

Abstract

Neural machine translation has achieved great success under a great deal of bilingual corpora in the past few years. However, it does not work well for low-resource language pairs. In order to solve this problem, we present a transfer learning method which can improve the BLEU scores of the low-resource machine translation. First, we exploit encoder-decoder framework with attention mechanism to train one neural machine translation model with large language pairs, and then employ some parameters of the trained model to initialize another neural machine translation model with less bilingual parallel corpora. Our experiments demonstrate that the proposed method can achieve the excellent performance on low-resource machine translation by weight adjustment and retraining. On the IWSLT2015 Vietnamese-English translation task, our model can improve the translation quality by an average of 1.55 BLEU scores. Besides, we can also get the increase of 0.99 BLEU scores when translating from Mongolian to Chinese. Finally, we analyze the results of experiments and summarize our contribution.

Keywords: Low-resource, Neural machine trannlation, Transfer learning

1. Introduction

Machine translation is one of the most important research field of artificial intelligence and natural language processing, which explores how to use computers to translate automatically between natural languages. In recent years, end-to-end neural machine translation has developed rapidly. The key idea of end-to-end neural machine translation is to achieve automatic translation between natural languages through neural networks. Compared with statistical machine translation, the quality of translation has been significantly improved. In a variety of languages pairs, the performance of neural machine translation has gradually surpassed phrase-based statistical machine translation. (Junczys-Dowmunt et al, 2016) provided comparison of translation quality for phrase-based statistical machine translation and neural machine translation across 30 translation directions with United Nations Parallel Corpus v 1.0. The results showed that neural machine translation surpassed statistical machine translation in 27 languages pairs.

Encoder-decoder (Sutskever et al, 2014) is one of most commonly framework in neural machine translation. The main idea of the framework is to map the input sequence to a fixed-sized vector with encoder, and then map the vector to the target sequence with decoder. Compared to traditional statistical machine translation, encoder-decoder has two major advantages. First, the framework can learn features directly from raw data. The results show, encoder-decoder learns sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice (Sutskever et al, 2014). Second, the framework effectively captures long-range dependencies based on long short-term memory networks (Hochreiter & Schmidhuber, 1997). Therefore, encoder-decoder framework can significantly improves the fluency and readability of the translation. However, encoder-decoder framework needs to map an input sentence of variable length into a fixed-dimensional vector representation, which poses a great challenge for the encoder to deal with long sentences. In order to solve this problem, (Bahdanau el al, 2014) proposed attention mechanism to dynamically computer the context of the source end. Attention mechanism changes the way of information transfer, and significantly improve the performance of neural machine translation. 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 neural machine translation is highly dependent on the size and the quality of parallel corpora. As is known to all, neural machine translation will fail when training data is not big enough (Koehn & Knowles, 2017). In some low-resource translation tasks, the performance of neural machine translation is severely reduced. However, the vast majority of the languages in the world lack large, high-quality parallel corpora (Artetxe et al, 2017). Therefore, research on low-resource neural machine translation is valuable.

In this paper, we propose a simple and effective method to alleviate this problem. First, we train one neural machine translation model with large parallel corpora, and then transfer some parameters of the trained model to initialize another neural machine translation model with less parallel corpora without changing neural network architecture. Whether it is a high-resource language pair or low-resource language pair, we use the encoder-decoder framework with long short-term memory units(LSTM). As illustrated in Figure 1, in the encoder-decoder framework, an encoder at the source compresses the source sentence into a semantic vector, and another decoder at the target side generates a sentence based on this vector. However, a potential issue of this encoder-decoder approach is that neural network needs to be able to compress all the information of a source sentence into a fixed-length vector. This may make it difficult for the neural network to cope with long sentences, especially those that are longer than sentences in the training corpus (Bahdanau et al, 2014). Therefore, we add global attention mechanism (Luong et al, 2015) for each target language. Attention mechanism can better solve the problem of long distance information transmission and significantly improve the performance of neural machine translation.

We follow the transfer learning method proposed by (Zoph et al, 2016). In their work, the high-resource language pair

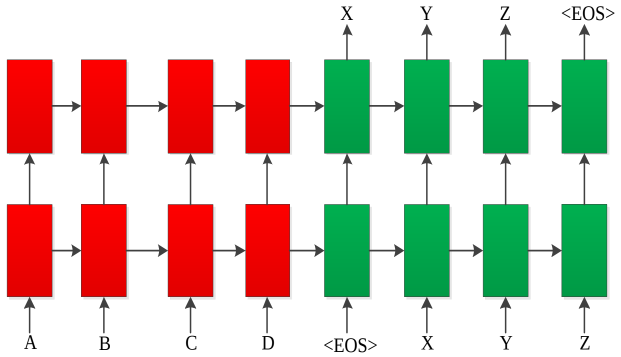

Figure 1: The encoder-decoder framework for neural machine translation. The framework learns to encode a variable-length sequence into a fixed-length vector representation and to decode a given fixed-length vector representation back into a variable-length sequence.

is called the parent model and the low-resource language pair is called child model. The parent model is first trained. Then the parameter values of child model are copied from the parent model and are fine-tuned. Comparison to their work—while our approach is similar in spirit to the model proposed by them, there are several key differences which reflect how we have simplified from the original model .

1. Both in high-resource language pairs and low-resource language pairs, they used uni-directional LSTM at the encoder end. However, we use bi-directional LSTM at the encoder because we would like the annotation of each source word to summarize not only the preceding words, but also the following words. So, our model works better than theirs on long sentence.

2. In their work, the target word embeddings of the child model are copied from the parent model and are frozen during fine-tuning because the target language is same in both parent model and child model. However, in our model, the target word embeddings of child model are initialized randomly and are constantly updated during training. The expression of language is not same in different domains. For example, the expression of sentence in the spoken corpus is more casual, while sentences in news corpus are more formally expressed. However, a word is characterized by the company it keeps (Harris, 1981), so the embedding of same word in different domains is not same. It is not a good choice to remain target word embedding of child model frozen.

2. Related work

Low-resource neural machine translation has attracted a lot of attention in recent years. The performance of neural machine translation is severely reduced when the parallel corpora is not enough. An effective way to alleviate this problem is to extend the scale of parallel corpora. (Sennrich et al, 2016) proposed a method to use the existing machine translation system to translate monolingual data and constructed dummy parallel corpora to relief the issue of lack of corpora. (Currey et al, 2017) utilized neural machine translation system to both translate source language text and copy target-language text, thereby exploiting monolingual corpora in the target language. Besides, for the low frequency words, (Fadaee et al, 2017) proposed a data augmentation method, which is also an effective way to extend the parallel corpora. Zero-resource neural machine translation is another way to deal effectively with insufficient resources, which is usually used in a pivot language. (Johnson et al, 2017) proposed a multilingual neural machine translation method. (Chen et al, 2017) presented a method for zero-resource neural machine translation by assuming that parallel sentences have close probabilities of generating a sentence in a third language called teacher-student framework. (Zheng et al, 2017) showed that maximum expected likelihood estimation can significantly improve the performance of zero-resource neural machine translation.

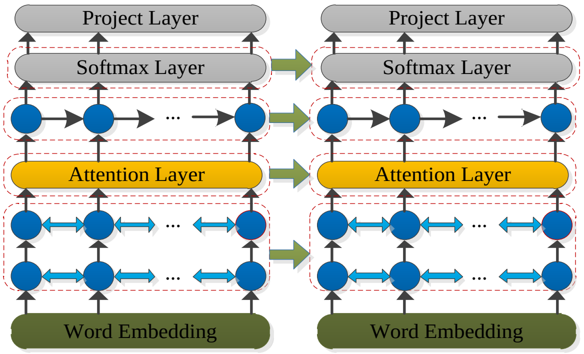

Figure 2: Aritecture of the proposed system. The left one is trained with high-resources, and then we transfer some parameters of trained model to initialize the right one. As is shown above, we transfer the parameters of all layers of the trained model except for the projection layer and word embedding layer.

3. Proposed Method

Figure 2 summarizes this general schema of the proposed system. This section describes the proposed neural machine translation with transfer learning. Section 3.1 first presents the architecture of the basic model, and section 3.2 then describes how to transfer parameters from high-resource language pairs to low-resource language pairs.

3.1 Basic models

Whether it is a high-resource language pair or low-resource language pair, the same neural network architecture is used. As shown in figure 1, the proposed system follows a standard encoder-decoder architecture with an attention mechanism.

In detail, we use a two-layers bi-directional RNN in the encoder, and another two-layers uni-directional RNN in the decoder. All the RNNs use LSTM cells with 1024 units, and the dimensionality of word embedding is set to 1024. As for attention mechanism, we use the global attention method proposed by (Luong et al, 2015). The model are trained using stochastic gradient descent with a minibatch size of 128 and a maximum sentence length of 50. We apply dropout (Gal et al, 2016) in high-resource language pair model with a probability of 0.2 and 0.5 in low-resource language pairs model. For all models, the learning rate decreases as the increase of the number of iterations. We decode using beam search on models with a beam size of 10. We initialize all of the parameters of network with the uniform distribution. We set the maximum value of the gradient to 5 in order to solve gradient explosion.

3.2 Transfer learning model

In short, transfer learning exploits knowledge from a learned task (source task) to improve the performance on a related task (target task), typically reducing the amount of required training data (Pan &Yang, 2010). Generally, the amount of data in the source task is sufficient, and the amount of data in the target area is small. Transfer learning needs to transfer the knowledge learned in the condition of sufficient data to the new environment with small amount of data. Traditional machine learning assumes that training data and testing data share same feature space and the same data distribution. When there is a difference in the data distribution between the training data and testing data, the results of predictive learner can be degraded (Shimodaira, 2000). However, transfer learning relaxes the limitation requirement, and applies the acquired knowledge to different but similar domains, which solves the problem of insufficient training data in the target domain. The transfer learning is usually divided into three types: instance-transfer, feature-representation-transfer, relational-knowledge-transfer.Transfer learning has been applied to many fields of the natural language processing, such as text categorization and machine translation. (Dai et al, 2007) proposed a co-clustering based classification algorithm to classify documents across different domains. (Long et al, 2010) propose Dual Transfer Learning method, which can improve the performance of classification accuracy.

In our paper, we translate Vietnamese into English with the help of French-English. First, we train French-English neural machine translation model, and then Vietnamese-English model is initialized with the parameters of the trained model. We just transfer some parameters to Vietnamese-English model, such as weights and biases of neural network, not all of them.

We follow the transfer learning method proposed by (Zoph et al, 2016). However, we have two improvements over their work. First, our model use bi-directional LSTM at the encoder because we would like each source word to summarize not only the preceding words, but also the following words. Second, we consider that the expression of language is not same in different domains. For example, the expression of sentence in the spoken corpus is more casual, while sentences in news corpus are more formally expressed. Moreover, a word is characterized by the company it keeps, so the embedding of same word in different domains is not same. Therefore, the target word embedding of Vietnamese-English model is initialized randomly instead of being copied from the French-English model. In addition, the projection layer of Vietnamese-English model can not be copied from French to English model, because target vocabulary of these two models is different. In order to verify the effectiveness of the method, we also translate Mongolian into Chinese with the help of English-Chinese.

4. Results and Analysis

4.1 Dataset details

As is shown in Table 1,Vietnamese-English corpora (133K sentence pairs, 2.7 million English words and 3.3 million Vietnamese words) is provided by IWSLT2015 and Mongolian-Chinese (67K sentence pairs, 848K Chinese words and 822K Mongolian words) is provided by CWMT 2009. 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 get English-Chinese corpora (2 million sentence pairs, 22 million English words and 24 million Chinese words) from the WMT2017. We preserve casing for words and replace those whose frequencies are less than 5 by <unk>. As a result, our vocabulary sizes are 17K and 7.7K for English and Vietnamese respectively. And we report BLEU scores on tst2012 and tst2013.For Chinese-Mongolian corpora (67K sentence pairs, 849K Chinese words and 822K Mongolian words), we make the same treatment. Therefore, the vocabulary size of Chinese and Mongolian are 14K and 12K respectively.

Dataset		sentences	words
Fr-En	Frence	2m	52m
	English		50m
Vi-En	Vietnamese	133K	2.7m
	English		3.3m
En-Ch	Chinese	2m	24m
	English		22m
Mn-Ch	Mongolian	67K	822K
	Chinese		894K

Table 1 : Statistics of all datasets

Models	BLEU	
	tst2012	tst2013
Baseline	20.43	23.17
Ours	**21.86**	**24.83**
(Luong & Manning,.2015)	-	23.3

Table 2: The performance of the proposed method on IWSLT2015 Vietnamese to English tst2012 and tst2013 set.

Models	BLEU
Baseline	11.69
Ours	**12.68**

Table 3: The performance of the proposed method on CWMT2009 Mongolian to Chinese test set.

4.2 Results

The results of BLEU scores are presented in Table 2 and Table 3. The architecture of baseline system is similar to the one mentioned in section 3.1. However, in order to prevent overfitting, we use one-layer bi-directional LSTM in the encoder, with 512 cells at each layer and 512 dimensional word embeddings.

As it can be seen, the proposed transfer learning method obtains very competitive results considering that it was trained on nothing but low-resource corpora. Our model can reach 21.86 and 24.83 BLEU scores in Vietnamese-English tst2012 and tst2013 set respectively, we can also achieve 12.86 BLEU scores in Chinese-Mongolian test set, which is much stronger than the baseline system, with improvements of at least 6.9% in all cases, and up to 8.5% in some (e.g. from 11.69 to 12.68 BLEU scores in Mongolian to Chinese). As you can see from the results, the proposed method obtains substantial improvement over baseline system, indicating that transfer learning method is significantly effective. Therefore, our method can improve the performance of low-resource machine translations.

5. Conclusion and future works

In this paper, we propose a novel method to train low-resource neural machine translation system. First we utilize encoder-decoder framework with attention to train one neural machine translation with high-resource language pairs, and then transfer some parameters of the trained model to initialize another neural machine translation model with less bilingual parallel corpora.

The experiments show the effectiveness of our proposal, obtaining significant improvements in the BLEU scores over baseline system. Our model can improve the translation quality on the IWSLT2015 Vietnamese-English translation task. We can achieve 21.86 and 24.83 BLEU scores on Vietnamese-English tst2012 and tst2013 set respectively. Besides, the method in this paper is also effective for Mongolian-Chinese translation. And we can improve the performance CWMT2009 Mongolian-Chinese translation task by 0.99 BLEU scores.

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

6. Acknowledgments

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

7. Bibliographical References

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). Is neural machine translation ready for deployment? a case study on 30 translation directions. arXiv preprint arXiv:1610.01108.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *Computer Science.*

Philipp Koehn and Rebecca Knowles. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28–39)

Artetxe, M., Labaka, G., & Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data.In *Meeting of the Association for Computational Linguistics* (pp.451-462).

Sennrich, R., Haddow, B., & Birch, A. (2015). Improving neural machine translation models with monolingual data. *Computer Science*.

Anna Currey, Antonio Barone and Kenneth Heafield. Copied monolingual data improves low-resource neural machine translation. (2017). In *Proceedings of the Conference on Machine Translation*(pp.148-156)

Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of 55th Annual Meetings of Association for Computational Linguistics*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., & Chen, Z., et al. (2016). Google's multilingual neural machine translation system: enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*(pp.339-351)

Chen, Y., Liu, Y., Cheng, Y., Li, V. O. K., Chen, Y., & Liu, Y., et al. (2017). A Teacher-Student Framework for Zero-Resource Neural Machine Translation. In *Meeting of the Association for Computational Linguistics* (pp.1925-1935).

Zheng, H., Cheng, Y., Liu, Y., Zheng, H., Cheng, Y., & Liu, Y., et al. (2017). Maximum Expected Likelihood Estimation for Zero-resource Neural Machine Translation. In *Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp.4251-4257).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the Neural Information Processing Systems.*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735.

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Harris, Z. S. (1981). Distributional Structure. *Papers on Syntax. Springer Netherlands*.

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Computer Science.*

Gal, Y., & Ghahramani, Z. (2015). A theoretically grounded application of dropout in recurrent neural networks. *Statistics*, 285-290.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning & Inference, 90*(2), 227-244.

Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007). Co-clustering based classification for out-of-domain documents. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.210-219). ACM.

Long, M., Wang, J., Ding, G., Cheng, W., Zhang, X., & Wang, W. (2012, April). Dual transfer learning. In *Proceedings of the 2012 SIAM International Conference on Data Mining* (pp. 540-551).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge & Data Engineering*, 22(10), 1345-1359.

Luong, M. T., & Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.