Optimization Algorithms for Tibetan-Chinese Neural Machine Translation

Jing Jiang\(^1\), Yachao Li\(^1\)*

\(^1\)Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

*Corresponding author e-mail: liyc7711@gmail.com

Abstract. Tibetan-Chinese neural machine translation (NMT) is facing serious resource scarcity problem. This paper compares the application effect of multiple neural network optimization algorithms under the condition of resource scarcity, and proposes an optimization method which is suitable for resource scarcity languages. Then, by which it improving the performance of the Tibetan-Chinese NMT. Experimental results show that when choosing a suitable optimization algorithm, the Tibetan-Chinese NMT can still exceed the traditional statistical machine translation (SMT) and achieve better translation performance.

1. Introduction

In recent years, Neural Machine Translation (NMT) has achieved great success, which surpasses the traditional statistical machine translation in many language pairs\(^{[1-5]}\) and it has become the mainstream machine translation method at present. Neural machine translation directly uses neural networks to complete the translation process from source language to target language, which greatly reduces the time-consuming and laborious operation steps in traditional machine translation such as word alignment, feature selection and phrase extraction\(^{[6]}\). These remarkable advantages have attracted wide attention from academia and industry.

The training neural machine translation model relies on large-scale corpus resources. Under the condition of scarce resources, the effect of neural machine translation is still lower than that of statistical machine translation\(^{[7]}\). Scholars have proposed the Back-Translation method to construct more pseudo-parallel corpus In view of the neural machine translation under the condition of scarcity of resources\(^{[8]}\); Integrating additional language models to use monolingual corpus\(^{[9]}\); and using transfer learning to transfer the model of resource-rich language pairs to resource-scarce language pairs\(^{[10]}\). The basic idea of the above research is to integrate more external resources so that the neural machine translation model can acquire sufficient translation knowledge and improve the translation effect. These methods have achieved good results in practical application, but the disadvantage is that the application effect is limited by the quality of external corpus resources.

Tibetan-Chinese parallel corpus resources are relatively scarce. The research of Tibetan-Chinese machine translation mainly focuses on statistical machine translation and related basic research\(^{[13-17]}\). The overall research foundation is relatively weak. Neural network-based Tibetan-Chinese machine translation has received initial attention. Li Yachao et al.\(^{[18]}\) studied attention-based Tibetan-Chinese neural machine translation, and adopted transfer learning to alleviate the shortage of Tibetan-Chinese parallel corpus. Tibetan-Chinese neural machine translation outperforms statistical machine translation.
in the case of fewer parallel corpus pairs by using transfer learning. This study proves that Tibetan-Chinese neural machine translation can also achieve better results under the condition of scarce resources.

Different from the existing work, this paper compares several optimization methods of neural machine translation models, and proposes a translation model optimization method which suitable for resource scarcity to improve the effectiveness of Tibetan-Chinese neural machine translation. This method does not use additional resources and has certain generality. The following parts of this paper are arranged as follows: the second part briefly introduces the neural machine translation and several commonly used optimization methods of parameters of neural network models; the third part carries out experiments and analysis; the fourth part is the summary of the full text and the next work arrangement.

2. Attention-based Neural Machine Translation

Neural machine translation originates from sequence-to-sequence learning, this paper takes the attention-based neural machine translation model as an example to illustrate\(^\text{[19]}\). Attention-based neural machine translation model can be roughly divided into three parts: encoder, decoder and attention mechanism. Encoders encode source language sentences into vector sequences as source language representations; decoders acquire source language context information through attention mechanism and generate target language word sequences in turn; attention mechanism connects encoders and decoders to make the whole model interrelated.

Tibetan-Chinese neural machine translation adopts a general framework of neural machine translation. In the following, the input Tibetan word sequence is represented by \(x\), the hidden state of encoder is represented by \(h\), and the output Chinese word sequence is represented by \(y\). As follows:

\[2.1 \text{ Encoder}\]

The encoder reads the source language word sequence and encodes as the source language representation vector. Specifically, the encoder reads the input \(x = (x_1, x_2, \ldots, x_I)\), encoding it as a hidden state \(h = (h_1, h_2, \ldots, h_I)\), which is usually implemented by Recurrent Neural Network (RNN). The updating method of hidden state are as follows:

\[h_t = f(x_t, h_{t-1})\]  \hspace{1cm} (1)

\(f\) is a non-linear function. Bi-directional cyclic neural network is used in practical use. For simplicity, this paper only uses forward cyclic neural network to illustrate.

\[2.2 \text{ Decoder}\]

Given the source language representation \(c\) and the precursor output sequence \(\{y_1, \ldots, y_{t-1}\}\), the decoder generates the target word \(y_t\) in turn, as follows:

\[p(y) = \prod_{i=1}^{T} p(y_i \vert \{y_1, \ldots, y_{t-1}\}, c)\]  \hspace{1cm} (2)

\(y = (y_1, y_2, \ldots, y_T)\), the decoder also uses a cyclic neural network, as follows:

\[p(y \vert \{y_1, \ldots, y_{t-1}, x\}) = g(y_{t-1}, s_t, c_t)\]  \hspace{1cm} (3)

\(g\) is a non-linear function to calculate the probability of \(y_t\), and \(s_t\) is the hidden state of the decoder. \(s_t\) is the hidden state of \(t\)-Time of cyclic neural network, which is calculated by the following formula:

\[s_t = f(s_{t-1}, y_{t-1}, c_t)\]  \hspace{1cm} (4)

\(f\) is a non-linear function. Context Vector \(c_t\) is the weighting sum of source language coding sequence \((h_1, h_2, \ldots, h_I)\), which is calculated by the following formula:
\[ c_t = \sum_{j=1}^{l} \alpha_{tj} h_j \]  \hspace{1cm} (5)

\[ \alpha_{tj} \] is the weight of \( h_j \). The calculation method is as follows:

\[ \alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{l} \exp(e_{tk})} \]  \hspace{1cm} (6)

Attention mechanism links the encoder and decoder, which can make the decoder fuse more source language information and improve the translation effect significantly. It is the mainstream neural machine translation method.

2.3 Model Optimization Method

Each part of the neural machine translation model can be trained jointly as follows:

\[ L(\theta) = \max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_{\theta}(y_n|x_n) \]  \hspace{1cm} (7)

\( \theta \) is a model parameter, \((x_n, y_n)\) denotes bilingual training corpus. The parameters of neural machine translation models are mostly trained by Stochastic Gradient Descent (SGD) and related improved algorithms. The performance of neural machine translation is directly determined by the method of model optimization. Following is a brief description of three typical model optimization methods.

**Stochastic Gradient Descent (SGD):** In this paper, the SGD method refers to the Mini-batch Gradient Descent. This method calculates the gradient in each iteration of training corpus, and then updates the model parameters which is the most basic optimization method of neural network model. The disadvantage of SGD optimization method is that it is difficult to select the appropriate learning rate. The advantage of SGD optimization method is that it is simple to implement and the experimental results are more stable and reliable under the appropriate learning rate scheduling scheme.

**Adadelta:** Adadelta is an optimization method of adaptive of learning rate \(^{[20]}\), which is a commonly used method of neural network optimization. The advantage is that the learning rate is adaptive, and the experimental results are stable. The disadvantage is that the convergence speed is slower.

**Adam:** Adam is an optimization method of improved learning rate adaptive \(^{[21]}\). The main advantage is that after offset correction, the learning rate of each iteration has a certain range, which makes the parameters more stable. And different parameters have different adaptive learning rates, which are suitable for large-scale data sets and high-dimensional parameter space.

These are three commonly used optimization methods of neural network model. In the following experiments, this paper compares different model optimization methods and makes a comparative analysis of their application effects in Tibetan-Chinese neural machine translation.

3. Experiments and analysis

3.1 Experimental settings

The Tibetan-Chinese machine translation evaluation corpus of CWMT 2011 was used to the experiment, which have training corpus 100,000 sentences and testing corpus 650 sentences. The length of training corpus sentences is limited to less than 50 words, the vector dimension of bilingual words is 620, the size of hidden layer is 1000, when decoder the size of Beamsize is 10, the size of Mini-batch is 80 sentences, and the Dropout \(^{[22]}\) of output layer is set to 0.5. In order to reduce the problem of unlisted words, the size of Tibetan and Chinese dictionaries is set to 30,000, covering about 99% words. In order to reduce fitting, the maximum number of training rounds is 60 (Epoch). Due to the limitation of test corpus, all experiments are conducted on the same test set, that is, only the experimental results on the development set are reported. Word-based BLEU-4 is adopted as the evaluation criterion for translation. On the whole, it is consistent with the experimental conditions in literature 18.
3.2 Experimental systems

SMT: The baseline system adopted in this paper is a Niutrans phrase statistical machine translation system developed for Northeast University[23], which is represented by SMT.

RNNSearch+Transfer: The Tibetan-Chinese neural machine translation method proposed in literature 18 is compared with the method using by the paper under the same experimental conditions.

DL4MT: In this paper, the open source system dl4mt1 is used as the attention-based neural machine translation system, and the cycle unit is Gated Recurrent Units (GRU) [3-4]. In addition to special instructions, the default parameter configuration is used.

DL4MT+ Adadelta/SGD/Adam: Adadelta/SGD/Adam are used as the optimization method of model parameters for Tibetan-Chinese neural machine translation system respectively. DL4MT+SGD+DC represents learning rate Decay when the iteration exceeds 35 rounds, and the decay rate is 0.5.

3.3 Main experimental results

Table 1. Experimental result.

| SYS | translation system          | Learning Rate | BLEU  |
|-----|----------------------------|---------------|-------|
| 1   | SMT                        | ---           | 33.40 |
| 2   | RNNSearch+Transfer          | ---           | 36.80 |
| 3   | DL4MT+Adadelta             | ---           | 30.50 |
| 4   | DL4MT+SGD                  | 1             | 16.75 |
| 5   | DL4MT+SGD                  | 2             | 36.71 |
| 6   | DL4MT+SGD                  | 3             | 34.21 |
| 7   | DL4MT+SGD+DC               | 2             | 38.23 |
| 8   | DL4MT +Adam                | 0.0001        | 32.95 |
| 9   | DL4MT +Adam                | 0.0002        | 37.45 |
| 10  | DL4MT +Adam                | 0.0003        | 35.73 |

As Table 1 showing, the translation effects of SYS3, SYS4 and SYS8 are lower than those of statistical machine translation. It shows that Tibetan-Chinese neural machine translation is ineffective under the condition of scarce resources. The experimental results conform to the characteristics of the general resource-scarce language neural machine translation system [6].

From the experiments of SYS4-SYS6 and SYS8-SYS10, it can be seen that improving learning rate the effect of Tibetan-Chinese neural machine translation can be greatly improved when the optimization method of SGD and Adam are used. However, when the learning rate is too high, the performance of translation system will be reduced. Therefore, it can be seen that under the condition of scarce resources, the neural machine translation system is sensitive in different model optimization methods and corresponding learning rates. Choosing the appropriate model optimization method and learning rate has a great influence on the final translation results.

To sum up, SYS7 has achieved the best translation effect, which surpasses the statistical machine translation system and the neural machine translation system using transfer learning [18]. It can be seen that Tibetan-Chinese neural machine translation still achieves better translation results by adopting larger learning rate and learning rate scheduling strategies with fewer corpus when choosing appropriate model optimization methods.

3.4 Contrast of Convergence Rate of Model

Figure 1 shows the convergence curves of different system models. Adam and SDG represent respectively the best experimental systems using this optimization method. It can be seen that the system using Adadelta optimization method converges slowly, and the corresponding translation effect is the worst. Adam optimization method converges quickly. The SGD optimization method uses a large learning rate and achieves the effect of Adam optimization method in about 26 rounds. When the
execution learning rate decreases, the translation performance can be further improved, and ultimately the best translation effect can be achieved.

Figure 1. Contrast of Convergence Rates of Different Systems

4. Conclusion and Further Work

In this paper, Tibetan-Chinese neural machine translation method is studied, and the optimization method of neural machine translation model suitable for resource scarcity is adopted to improve the translation effect. The proposed translation model optimization method significantly improves the effectiveness of the Tibetan-Chinese neural machine translation system, and surpasses the phrase statistical machine translation system and the previous similar work, which achieves the best translation results. The method proposed in this paper does not depend on external resources and has language independence. It can be applied to the neural machine translation of other scarce languages in theory.

In the future, this paper will study other methods of neural machine translation to improve effects in resource scarcity, and the application of this method to other languages.

5. Acknowledgments

This research is supported by the Fundamental Research Funds for the Central Universities of Northwest Minzu University (31920190112, 31920190114).

References
[1] Jiajun Zhang, Chengqing Zong. Deep neural networks in machine translation: an overview [J]. Intelligent Systems IEEE, 2015, 30(5):16–25.
[2] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to sequence learning with neural networks[C]// Proceedings of the Neural Information Processing Systems (NIPS 2014), Montreal, Canada, 2014:3104–3112.
[3] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. arXiv preprint/1406.1078v2, 2014.
[4] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, et al. On the properties of neural machine translation: encoder-decoder approaches[C]// Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar, 2014: 103–111.
[5] Sebastien Jean, Kyunghyun Cho, Yoshua Bengio. On using very large target vocabulary for neural machine translation[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL 2015), Beijing, China, 2015:1–10.
[6] Li YaChao, Xiong DeYi and Zhang Min. Summary of Neural mechanism Translation[J]. Journal of Computer Science, published online in 2017.
[7] Rico Sennrich, Barry Haddow, Alexandra Birch. Improving neural machine translation models with monolingual data//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016). Berlin, Germany, 2016:86–96
[8] Tobias Domhan and Felix Hieber. Using target-side monolingual data for neural machine
translation through multi-task learning//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhagen, Denmark, 2017:1501–1506

[9] Barret Zoph, Deniz Yuret, Jonathan May, et al. Transfer learning for low-resource neural machine translation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA, 2016:1568–1575.

[10] Xiaodong Yan, Xiaobing Zhao. Research on Sino-Tibetan Machine Translation Based on the Reusing of Domain Ontology [J]. International Journal of Online Engineering, 2013, 9(4):105.

[11] Liu Huidan, Zhao Weina, Yu Xin et al. A Chinese to Tibetan Machine Translation System with Multiple Translating Strategies [J]. Himalayan Linguistics, 2016, 15(1):148-166.

[12] Dong Xiaofang. Research on Tibetan-Chinese Statistical Machine Translation Phrase Extraction Technology [D]. Northwest Minzu University, 2013.

[13] Huague Cairang. Research on Some Key Technologies of Tree-to-String Tibetan Machine Translation [D]. Shanxi Normal University, 2014.

[14] Huaguo Cairang. Tibetan verbs in Chinese-Tibetan machine translation [D]. Qinghai Normal University, 2014.

[15] Wang Tianhang. Research on Tibetan Functional Block Recognition for Machine Translation [D]. Beijing University of Technology, 2016.

[16] Sun Meng, Huaque Cairang, Liu Kai, et al. Tibetan numeral recognition and translation [J]. Journal of Peking University (Natural Science Edition), 2013, 49 (1): 75-80.

[17] Li Yachao, Xiong Deyi, Zhang Min, et al. Machine Translation Research of Tibetan-Chinese Neural Networks [J]. Journal of Chinese Information Science, 31 (6), 2017:103-109.

[18] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. Arxiv: 1409.0473v6, 2014.

[19] Matthew D. Zeiler. ADADelta: An Adaptive Learning Rate Method [J]. Computer Science, 2012.

[20] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In arXiv:1412.6980.

[21] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research [J], 2014, 15:1929–1958.

[22] Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. NiuTrans: an Open Source Toolkit for Phrase-based and Syntax-based Machine Translation[C]// Proceedings of the ACL 2012 System Demonstrations. 2012:19-24.

[23] Zhaopeng Tu, Zhengdong Lu, Liu Yang, et al. Modeling coverage for neural machine translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), Berlin, Germany, 2016:76-85.