Research on Tibetan-Chinese neural network machine translation with few samples

Yidong Sun¹², Cuo Yong¹²*

¹ School of Information Science and Technology, Tibet University, Lhasa, Tibet, 850000, China
² Key Laboratory of Artificial Intelligence for Tibetan Information Technology in Tibet Autonomous Region, Tibet University, Lhasa, Tibet, 850000, China
*Corresponding author’s e-mail: yc@utibet.edu.cn

Abstract. Machine translation is an important task in natural language processing, and the study of Tibetan-Chinese neural machine translation is of profound significance in promoting Tibetan-Chinese scientific and cultural exchanges and the development of education and culture. In this paper, we investigate the performance of these techniques and methods on Tibetan-Chinese NMT with few samples by using deactivated word lists, data augmentation (back translation), pre-training models (ELMO), and attention mechanisms for the techniques and methods widely used in NMT, using seq2seq and Transformer models as the baseline, and finally, the BLEU value of Tibetan-Chinese NMT is increased from the initial 5.53 to 19.03.

1. Introduction
Tibetan-Chinese machine translation refers to the technology of translating between two languages, Tibetan and Chinese, by computer. Traditional machine translation technology mainly refers to rule-based and statistical machine translation, but with the continuous opening of information such as the Internet and the development of computer hardware, low-cost data acquisition and the continuous development of computer hardware make neural machine translation system (NMT) gradually become the mainstream translation technology nowadays by virtue of its high-quality translation. First, NMT systems need to have high-quality large-scale parallel corpus to train efficient translation models, and it is difficult to obtain high-quality large-scale parallel corpus, especially for low-resource languages such as Tibetan. Second, compared with other languages, Tibetan has its own linguistic characteristics: the omission of separators between Tibetan syllables, which leads to tight and adherent words, the difficulty in distinguishing the boundary between imaginary and real words in Tibetan, and the lack of obvious spacing markers between words [1]. These features pose challenges to the improvement of Tibetan-Chinese machine translation quality.

In this paper, we will take the current mainstream Transformer and the classical seq2seq model as the baseline, experiment with technical approaches that have shown excellent results on translation in other languages, and explore the effectiveness of these widely used technical approaches on Tibetan-Chinese neural machine translation with few samples.

2. Related Work
Neural machine translation dates back to 1997, when Spanish scholars Forcada and Ñeco proposed the idea of using an "encoder-decoder" framework for translation [2]. In 2013, Nal Kalchbrenner and Phil
Blunsom proposed a novel end-to-end encoder-decoder architecture for machine translation [3]. This model uses a convolutional neural network (CNN) to encode a given source text as a continuous vector and then uses a recurrent neural network (RNN) as a decoder to convert the state vector into the target language. Theoretically, the RNN is able to capture the information behind long sentences, thus solving the long distance reordering (LDR) problem [4]. However, the gradient disappearance problem [5] makes it practically difficult for RNNs to deal with long distance dependencies.

A year later, Sutskever et al. and Cho et al. developed the sequence-to-sequence (seq2seq) method [6] [7] and introduced the long and short term memory network (LSTM) proposed by Hochreiter and Schmidhuber in NMT [8]. Since the gate mechanism in the LSTM allows episodic memory deletion and updating, the problem of gradient disappearance is controlled, so the model can better capture long-range dependencies in sentences. However, it introduces the fixed-length vector problem - no matter how long the source sentence is, the neural network needs to compress the source sentence into a fixed-length vector, which greatly increases the complexity and uncertainty of the decoding stage [7]. It was only after Yoshua Bengio's team introduced the attention mechanism to NMT in 2014 that the problem of fixed-length vectors started to be solved. The attention mechanism was originally proposed by DeepMind in solving the image classification problem [9], which enables the neural network to pay more attention to the parts that are relevant to the input and less to the irrelevant parts when performing the prediction task. After that, the performance of NMT was significantly improved, and with the Transformer model [10], which was later proposed by Google team to solve the Seq2Seq problem, the LSTM was replaced with a full-attention structure, which made the development of NMT to another level.

The study of Tibetan-Chinese machine translation is of profound significance for promoting Tibetan-Chinese scientific and cultural exchanges and the development of education and culture. In this paper, we investigate the effectiveness of these technical methods, such as data pre-processing (removing discontinued words), data augmentation, pre-trained models, Tibetan syllable-based word separation and subword segmentation, on Tibetan-Chinese NMT translation with few samples, and finally, the best-performing model is analyzed for translation.

3. Construction of the corpus

3.1. Source of the corpus

The experimental data used in this paper are partly from the news and politics domain of "China Tibet News Network" and partly from the Tibetan-Chinese parallel corpus provided by the laboratory, and they are all manually proofread to obtain 12,000 Tibetan-Chinese parallel sentence pairs in the news and politics domain with high quality.

3.2. Data pre-processing

Firstly, data filtering is performed using the rule-based approach, which includes length filtering, length ratio restriction, language identification (langid), and de-duplication. Next, symbol standardization is performed, which refers to the unification of character representation or case in the data, specifically including full angle to half angle, case conversion and conversion of Chinese into simplified and traditional Chinese, etc. Finally, about 10,000 Tibetan-Chinese sentence pairs were obtained after cleaning. Finally, word separation and deactivation are performed. In this paper, the Chinese language is unified using jieba\(^1\) for word separation or processed into subwords by Byte Pair Encoding (BPE\(^2\)) \(^{11}\), and the Tibetan language is used according to the experimental needs, including syllable separation, Xiamen University's open source segtag separation\(^3\), and Byte Pair Encoding (BPE) on the basis of segtag, with words separated by spaces. The Chinese deactivation word list was used from Baidu

---

1 Sun J 2012. Jieba Chinese word segmentation tool. Accessed: Jun, 2012, 25: 2018.
2 https://github.com/rsennrich/subword-nmt/tree/master/subword_nmt
3 http://121.192.180.171:8080/segtag.html
deactivation word list, and the Tibetan deactivation word list was obtained from the internal laboratory, and experiments were conducted on the usefulness of removing deactivated words.

4. Experiment design

4.1. Model architecture
In this paper, the experimental framework uses the PaddleNLP framework based on Baidu Flying Pulp (paddle paddle), trained using the Transformer-base and seq2seq models, where the network parameters of Transformer-base are kept consistent with the parameter settings in Vaswani's paper. Sequence to Sequence (Seq2Seq) uses a multi-layer LSTM-based RNN encoder for encoder, an RNN decoder with Attention mechanism for decoder, and a decoder implementation without Attention mechanism for comparison, and a column search algorithm for prediction. Seq2seq has a hidden layer dimension size of 512 and the same optimization strategy as Transformer-base, both using the Adam optimizer [12]. The evaluation indicator used is the BLEU4 value. All experiments are done under RTX 2060 GPU and Ubuntu system.

4.2. Sequence to Sequence
Using the seq2seq model as the baseline, a joint experiment on the effect of attention mechanism and BPE was conducted, and the experimental data were randomly sliced to obtain 8000 training set utterances, 900 development set utterances, and 900 test sets with epoch=50, where all experimental data were preprocessed (length filtering, length ratio restriction, language identification, de-duplication) and all deactivated words were removed. The size of the Tibetan-Chinese word list after using BPE is 8791 and 10465 respectively, and the size of the word list without BPE is 18812 and 14681. The experimental results are shown in Table 1.

| Ti-Ch | No attention | No attention | With attention | With attention |
|-------|--------------|--------------|----------------|----------------|
|       | No BPE       | With BPE     | No BPE         | With BPE       |
| Val   | 5.53         | 2.04         | **12.10**      | 8.69           |
| Test  | 6.05         | 2.63         | **12.09**      | 8.93           |

From Table 1, we can see that the attention mechanism can improve the translation effect substantially, and it is also the biggest improvement among all experiments. In contrast, using BPE decreases the BLEU value by 3-4 points. To validate the experiments, experiments are conducted on the Transformer upper model for Tibetan-Chinese bidirectional translation using BPE.

4.3. Transformer

4.3.1. Deactivation word list and BPE. In response to the poor results of using BPE on the seq2seq model, we conducted experiments on the Transformer model combined with the deactivation table for Tibetan-Chinese bidirectional translation respectively to verify the conjecture. The experimental data were randomly sliced to obtain 9000 training set utterances and 800 test set utterances, completed with batchsize=2048 and epoch=50, in which all experimental data were preprocessed with data. The experimental results are shown in Table 2 and Table 3.

| Ti-Ch | BPE | NO BPE |
|-------|-----|--------|
| Normal| 9.43| 14.37  |
| Add deactivation list | 11.59 | **16.72** |

4 [https://github.com/moses-smt/mosesdecoder]
Table 3. Experimental results table of deactivating word lists and BPE on Transformer model under Chinese-Tibetan translation.

|                | Ch-Ti | BPE     | NO BPE |
|----------------|-------|---------|--------|
| Normal         | 8.39  | 10.37   |        |
| Add deactivation list | 10.03 | 12.06   |        |

According to Tables 1, 2 and 3, it can be seen that the translation effect of subword syncopation in Tibetan is not good, and the reason for this is that there are many words and even syntactic structures in Tibetan that are not shared between highly related grammars, and secondly, Tibetan contains a large number of adhesions and contractions, such as "ཐ་དེ་རིང་ɍོས་ȭ་ཁོས་གཤེགས་དེ་རིང་ɥན་ཁང་ȭ་ཁོས་གཤེགས་" and "ཁོ་ཁོ་" in the first sentence, "ཐ" means "doctor", and in the second sentence, "ཐ" means "doctor". "ཐ" in the first sentence refers to "doctor" and "ཐ" in the second sentence refers to "patient" [1]. Therefore, simply adopting BPE for Tibetan and Chinese equally will not work, and will only lead to a decrease in the translation effect.

According to Table 1 and Table 2, we can get that relative to the seq2seq model which partially uses the attention mechanism, the Transformer model improves the BLEU value by about 4.6 points again, which is an important reason for Transformer to become the mainstream translation model nowadays.

According to Tables 2 and 3, it can be seen that the translation effect is effectively improved as expected after adding the discontinued word list, and the average increase of BLEU value is nearly 2 points, which is due to the fact that the Tibetan discontinued word list contains a large number of Tibetan dummy words and adherent words, which is equivalent to translation from sentence level down to word level, although it can improve the BLEU value, it will lead to problems such as poor sentence flow and incoherence between words.

4.3.2. Data Augmentation. Data Augmentation is a valuable technique to alleviate the problem of insufficient data volume by generating high-quality pseudo-data, and to improve the robustness of models by preventing overfitting. One common approach is Back Translation, in which a system is trained to translate from the target language to the source language, i.e., a back translation system, which is then used to translate monolingual data in the target language, and finally the monolingual data (target language) and the result of the translation (source language) are used as training data and fed into the source-to-target language translation system. This approach is widely adopted because it can make good use of monolingual data without changing any model structure.

The best effective Chinese-Tibetan NMT (BLEU=12.06) was obtained by previous training, and 8,000 and 18,000 Chinese items in the news domain were used for translation respectively, while the previous data division was disrupted and 9,000 training data and 800 test data were resampled, and finally the obtained Tibetan pseudo-data and Chinese data were combined with the 9,000 training data obtained by resampling The Tibetan-Chinese NMT (BLEU=16.72) with the best result before is compared.

Table 4. Use the back translation experiment results table on Transformer.

|                | 8000 pseudo data | 18000 pseudo data | 18000 pseudo data | 18000 pseudo data |
|----------------|------------------|-------------------|-------------------|-------------------|
| No back translation epoch=50 | 16.72            | 17.07             | 16.58             | 17.68             |
| 8000 pseudo data epoch=50    |                  |                   |                   |                   |
| 18000 pseudo data epoch=50   |                  |                   |                   |                   |
| 18000 pseudo data epoch=70   |                  |                   |                   |                   |
| 18000 pseudo data epoch=100  |                  |                   |                   |                   |

Table 4 shows that the data augmentation by back-translation does result in a better translation of the model to some extent, improving the BLEU value by 1.19. However, to increase the epoch, the value of the loss function at epoch 100 for the back-translation experiment with 18,000 pseudo-data is required to reach a similar value to the loss function for the experiment before back-translation (BLEU=16.72), i.e. The number of iterations must be increased to achieve the desired convergence.
4.3.3. Words by Tibetan syllable. The smallest linguistic unit in Tibetan is the letter, followed by the syllable, which is the most basic formal word-forming unit and linguistic unit of Tibetan. When splitting words by Tibetan syllables, the Tibetan syllable symbols were used as segmentation points, and then the syllable symbols were removed and spaces were used to fill them. Two sets of comparisons were conducted, without back-translation and with 18,000 pseudo-data back-translations. The experimental data size is still done using 9000 training set statements and 800 test set statements with batchsize=2048. The experimental results are shown in Table 5.

Table 5. Table of experimental results on Transformer model by Tibetan syllable splitting with segtag splitting without back translation and with 18000 pseudo data back translation.

|                         | No back translation | No back translation | Back translation | Back translation |
|-------------------------|---------------------|---------------------|------------------|------------------|
|                         | segtag splitting    | Tibetan syllable    | segtag splitting | Tibetan syllable |
|                         | epoch=50            | epoch=50            | epoch=100        | epoch=100        |
|                         | 16.72               | 14.43               | 17.91            | 16.55            |

The experiments keep the same parameters except for the different word splitting strategies (jieba splitting is used uniformly in Chinese). According to Table 5, the effect of Tibetan word division by syllable is not as good as that of segtag word division in Xiamen University, and the BLEU value decreases by 1.36-2.29 points, mainly because when cutting according to the syllable symbol "·" in Tibetan without any other rules or restrictions, it loses its semantic information and cannot retain the original. The reason is that when the Tibetan language is cut according to the syllable symbol "·" without any other rules or restrictions, the semantic information is lost and the information expressed in the original language is not retained. This makes it impossible to correspond to the corresponding translation in Chinese, resulting in invalid bilingual parallel corpus, so that NMT cannot learn the semantic features from it, which eventually leads to poor translation results. The effect of training at word level is significantly better than that of training at syllable level. Because the semantic information of the original text can be retained relatively stably at the word level, and the corresponding parallel words can be found better in the bilingual pairs, which helps NMT to obtain more information and retain more semantic features.

4.3.4. Pre-training model ELMO. ELMO [13] is a bidirectional LSTM language model proposed by Peters et al. in 2018 for solving polysemous word problems. It consists of a forward and a backward language model, and the objective function is to take the maximum likelihood of the language model in both directions. In this paper, we use about 2G size of Tibetan monolingual data in the news domain and 2G size of Chinese monolingual data in the news domain to obtain the Tibetan and Chinese pre-trained language models of ELMO, and generate 512-dimensional Tibetan word vector word lists and Chinese word vector word lists respectively. The Tibetan monolingual data were obtained from the crawler of "China Tibet News", and the Chinese monolingual data were obtained from the cut of the parallel Chinese and English corpus in the field of UN news. Finally, the word vectors without contextual information are replaced with the ELMO-generated word vectors with contextual information and used as new word vectors. The experimental results are shown in Table 6 using the data volume of back translation and training ELMO as control parameters.

Table 6. Table of experimental results of ELMO models trained with different amounts of data without back translation and with 18000 pseudo data back translation.

| Ti-Ch | No ELMO | ELMO (50M) | ELMO (1G) | ELMO (2G) |
|-------|---------|------------|-----------|-----------|

5 The United Nations Parallel Corpus, Language Resources and Evaluation (LREC’16), Portorož, Slovenia, May 2016.
From Table 6, it can be clearly seen that the ELMO trained with 2G monolingual data volume improves the BLEU value by 1.15 on average compared with the previous one without ELMO, but the ELMO trained with 50M data instead makes the translation effect decrease by 1.02 BLEU value on average. It can be seen that the pre-trained model can indeed improve the low-resource translation task by using a large amount of monolingual data, but the amount of trained monolingual data must be large enough, otherwise it will lead to worse translation results instead. Secondly, the effect of using ELMO does not reach the expected conjecture, probably because ELMO only makes a splice of the language model in two directions, which makes the word vector cannot accurately contain the contextual information.

5. Translation analysis

We used a random sampling method for the best-performing model (ELMO, back translation, Transformer, deactivation removal, BLEU=19.03) and randomly sampled four sentences from 800 test utterances for translation analysis and compared them with Sunshine Tibetan-Chinese Translation 6(NMT), as shown in Table 7.

Table 7. Comparison table between the Sunshine Tibetan-Chinese translation model and the model with a BLEU value of 19.03 for Tibetan language prediction.

| Tibetan source language sentences | Model prediction results | Sunshine Tibetan prediction results | Reference text |
|----------------------------------|-------------------------|------------------------------------|----------------|
| སི་རི་ཡིད་གཉིས་ལས་འཛིན་ཀུན་ཚོགས་ཀྱིས་སི་འོད་ཞིབ་པ། | The Syrian Administrative Council Innocent Victims Reference. | Syrian Government Council Provides Information on Innocent Civilian Victims |
| དུས་ཀྱི་ཚུལ་སྒྲིགས་ཀྱིས་ལྷེ་གཞི་བ། | International Chamber of Commerce Management Association. | International Seabed Authority Council |
| མིའི་རིགས་དབང་འབུམ་མིང་བ། | Human Wealth | There is no limit to the number of people, it is endless. | Humanity Wealth Infinite |
| ཆེ་ལེས་ངོ་བོ་གཞན་དག་ལས་འཛིན་ཀུན་ཚོགས་ཀྱིས་སི་འོད་ཞིབ་པ། | The global climate change of climate change - East Turkel-Latuwa - low desert and other developing countries that have gained the security of great damage. | Globalization Impacts Climate Change Sea Level Threats Severe Damage Tuvalu Lowlands Small Islands Developing Countries Children Future Well-being |
From Table 7, we can clearly see that although removing deactivated words can increase the BLEU value, it will lead to incoherent translation and degrade from sentence-level to word-level translation. Therefore, how to maintain the sentence coherence even after removing deactivated words is a direction worth studying and is our next step. Secondly, there are obvious problems of missing translations and word order reversal in the prediction results of the model, which is also a problem faced by the whole machine translation direction at present. In comparison, the poor results of Sunshine Tibetan-Chinese Translation (NMT) model may be caused by the domain adaptation problem, while for most NMT models, new "knowledge" can be learned from the translation, which is an advantage that SMT does not have.

6. Summary and future work
In this paper, the classic seq2seq and Transformer models in machine translation are selected, and most of the current techniques and methods that can improve translation quality are experimentally analyzed on Tibetan-Chinese translation with few samples, among which, except for BPE which cannot be directly applied to Tibetan and Tibetan syllable-based participle which is not as effective as Tibetan lexical participle-based translation, all the other techniques and methods can contribute to Tibetan-Chinese machine translation. In the future, we plan to investigate whether improving the quality of translating Tibetan-Chinese translation while maintaining sentence coherence and applying the pre-training model BERT [14] to Tibetan-Chinese translation is better than ELMO as expected guess.

Acknowledgments
This work is supported by the National Key Research and Development Program of China (No. 2017YFB1402200) and the National Natural Science Foundation of China (No.61262086) as well as the Construction of national team and Key Laboratory of computer and Tibetan information technology ([2018] 81) issued by Department of education of Tibet Autonomous Region.

References
[1] Sha, J., Feng, C., Zhang, T. (2020) A study on multi-strategy slicing granularity for Tibetan-Chinese bidirectional neural machine translation. Journal of Xiamen University (Natural Science Edition). 59(2):213-219.
[2] Forcada, M.L., Ńeco, R.P. (1997) Recursive hetero-associative memories for translation. Springer Berlin Heidelberg, Heidelberg.
[3] Kalchbrenner, N., Blunsom, P. (2013) Recurrent Continuous Translation Models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Washington. pp. 1700-1709.
[4] Sudoh, K., Duh, K., Tsukada, H., Hirao, T., Nagata, M. (2010) Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR. Uppsala. pp. 418-427.
[5] Pascanu, R., Mikolov, T., Bengio, Y. (2013) On the difficulty of training recurrent neural networks. In: International Conference on Machine Learning. Atlanta. pp. 1310-1318.
[6] Sutskever, I., Vinyals, O., Le, Q. V. (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. Montreal. pp. 3104-3112.
[7] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. https://arxiv.org/abs/1406.1078
[8] Sepp, H., Jürgen, S. (1997) Long Short-Term Memory. Neural Comput, vol.9, no:8:1735-1780.
[9] Mnih, V., Heess, N., Graves, A. (2014) Recurrent models of visual attention. In: Advances in neural information processing systems. Montreal. pp. 2204-2212.
[10] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention is all you need. In: 31st Annual Conference on Neural Information Processing Systems. Long Beach. pp. 5998-6008.
[11] Sennrich, R., Haddow, B., Birch, A. (2015) Neural machine translation of rare words with
subword units. In: 54th Annual Meeting of the Association-for-Computational-Linguistics.
Berlin. pp.1715-1725.

[12] Guillaume, K., Yoon, K., Yuntian, D, Jean, S., Alexander, R. (2017) OpenNMT:Opensource
toolkit for neural machine translation. In: Proceedings of ACL 2017, System Demonstrations.
Vancouver. pp. 67-72.

[13] Peters, M.E., Neumann, M., Iyyer, M. et al. (2018) Deep contextualized word representations.
https://www.researchgate.net/publication/323217640_Deep_contextualized_word_representa-
tions

[14] Jacob, D., Chang, M.W., Kenton, L., Kristina, T. (2019) BERT: Pretraining of Deep Bidirectional
Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the
North American Chapter of the Association for Computational Linguistics: Human Language
Technologies, NAACLHLT. Minneapolis. pp. 4171-4186.