Research on Chinese-Tibetan Neural Machine Translation

Wen Lai, Xiaobing Zhao*, Xiaqing Li

National Language Resource Monitoring & Research Center of Minority Languages, Minzu University of China, Beijing, 100081, China
Lavine_Lai@126.com
nmzxb_cn@163.com
xiaqing0614@foxmail.com

Abstract. At present, the research on Tibetan machine translation is mainly focused on Tibetan-Chinese machine translation and the research on Chinese-Tibetan machine translation is almost blank. In this paper, the neural machine translation model is applied to the Chinese-Tibetan machine translation task for the first time, the syntax tree is also introduced into the Chinese-Tibetan neural machine translation model for the first time, and a good translation effect is achieved. Besides, the preprocessing methods we use are syllable segmentation on Tibetan corpus and character segmentation on Chinese Corpus, which has a better performance than the word segmentation on both Chinese and Tibetan corpus. The experimental results show that performance of the neural network translation model based on the completely self-attention mechanism is the best in the Chinese-Tibetan machine translation task and the BLEU score is increased by one percentage point.

Keywords: Neural Machine Translation, Tibetan, Syntactic tree, Attention.

1 Introduction

Machine translation, studies on how to use computers to achieve the automatic translation between natural languages, is one of the important research directions in artificial intelligence and natural language processing (Liu, 2017). Natural language processing is a discipline that crosses computer science and linguistics. Based on characteristics of this discipline, the system of machine translation can be divided into two categories, which are the rule-based methods and the corpus-based methods. Among them, corpus-based methods can be divided into statistics-based methods and example-based methods (Zhao et al., 2000). In recent years, with the development of internet technology and the improvement of computing speed of computer, machine translation has achieved fruitful results both in academia and industry area.

Since the advent of the neural network in the 1940s, it has experienced a period of rise - low tide - rise. Until 2006, Hinton et al. solved the historic problem of neural net-works (Hinton et al., 2006), and the related researches of deep learning and neural network returned to people's attention again. Since then, with the deepening of theo-

retical research and the improvement of computing speed of computers, neural networks have made great breakthroughs in various fields of artificial intelligence, such as computer vision, machine learning, and speech recognition. Researches about natural language processing have also made a rapid progress along with this tide.

In 2012, the Hinton research group participated in the ImageNet image recognition contest and won the championship, which opened the prelude of deep learning in various fields of artificial intelligence. Neural machine translation (NMT) is also a machine translation method that is gradually emerging at this stage. The main processes of neural machine translation are as follows: Firstly, it uses deep neural networks (RNN, CNN, etc.) to encode the source language into word-embedding. Secondly, the word-embedding generates the target language by decoding.

Tibetan is a kind of pinyin character, and its syllables are composed of 34 vowel consonants, and then Tibetan words are composed of syllables (Wei, 2015). A single character in a Tibetan text is a unit, and it is separated by a syllable separator "·" between characters (Cai, 2016). Based on the characteristics of Tibetan language, the statistical machine translation model is mainly used in the research on Tibetan translation model (Dong, et al., 2012; Luo, et al., 2010; Hua, et al., 2014), and the relevant theoretical research has basically stopped at the stage of word processing and other corpus pre-processing (Cai, et al., 2011; Pang, et al., 2015; Xiang, et al., 2011; Hua, et al., 2014; Nuo, et al., 2011). Overall, compared with other rich languages, the research on Tibetan-Chinese or Chinese-Tibetan machine translation are obviously behind. There are few researches focus on neural network model in Tibetan corpus (Li et al., 2017). Besides, Tibetan texts are all word segmentation pre-processed in traditional Tibetan machine translation tasks (Guan, 2015). In this article, the traditional method of Tibetan word segmentation is completely abandoned, and Tibetan texts are directly divided into syllables. It gets a better result than Tibetan word segmentation.

The study of Chinese-Tibetan machine translation has a far-reaching significance in promoting Chinese-Tibetan technological and cultural exchanges and promoting the development of educational and cultural undertakings. At present, most of the research related to Tibetan machine translation is focused on Tibetan-Chinese machine translation. There are few researches on Chinese-Tibetan machine translation, and the existing Chinese-Tibetan machine translation is mainly based on statistical machine translation methods. For the related research on machine translation, there is no neural machine translation models used in Chinese-Tibetan corpus, Therefore, in this paper it is innovative to apply the neural network model to Chinese-Tibetan corpus for the first time.

In this paper, five common neural network machine translation models and the latest syntax-tree-based neural network machine translation model are used in Chinese-Tibetan machine translation tasks, and the final translation results are analyzed in detail. The experimental results show that the application of neural network machine translation model on Chinese-Tibetan machine translation tasks has a good performance, and the method of syllable segmentation on Tibetan corpus has a better translation performance than the method of word segmentation on Tibetan corpus in Chinese-Tibetan machine translation tasks. Meanwhile, it shows that the syntax tree is

introduced in the neural network machine translation model, which can improve the translation performance.

2 Neural Machine Translation Models

2.1 Seq2Seq

The Seq2Seq model was presented in 2014, and two articles published by the Google Brain team (Ilya et al., 2014) and the Yoshua Bengio team (Cho et al., 2014) illustrate the basic idea of the model. The basic idea of solving the problem of the Seq2Seq model is to map an input sequence to an output sequence through one or more deep neural network models, which is commonly known as LSTM --- Long short-term memories network (D'Informa-tique et al., 2001), and this process consists of two parts of encoding input and decoding output.

2.2 RNNSearch

In 2015, RNNSearch machine translation model was proposed by Bahdanau et al. (Bahdanau et al., 2014). Based on the encoder-decoder structure, the attention mechanism is added to this model, it is used in natural language processing tasks for the first time, and translation performance is greatly improved.

2.3 Fairseq

Fairseq machine translation model was presented by the Facebook team in May 2017 (Gehring, 2017). The traditional method of sequence to sequence learning is to map an input sequence to a variable length output sequence through one or more layers of RNN neural network. In the Fairseq model, a structure is introduced based on convolutional neural networks (CNNs) entirely. Compared with the recurrent neural network model, all calculations of the element sequence are completely parallel while Fairseq model is in training, the number of nonlinear sequences is fixed and independent of the length of the input sequence.

The research shows that in the same environment, the training time of Fairseq model is 9 times faster than the translation model based on RNN network, and its accuracy is also higher than that of the model based on RNN network.

2.4 Transformer

Transformer machine translation model was proposed by the Google team in June 2017 (Vaswani et al. 2017). In the traditional neural network machine translation model, the neural network is mostly used as the model basis of Encoder-Decoder. This model is based on the attention mechanism and completely abandons the inherent mode of the neural machine translation model without any neural network (CNN or RNN) structure. Experiment results show that this model can run fast in parallel, which greatly improves the training speed of the model while improving performance of machine translation.

The Transformer model performs well in natural language processing tasks such as syntactic parsing and semantic understanding, which is also a breakthrough in the system of natural language processing for decades.

2.5 Syntax-NMT

The syntax-based machine translation model introduces syntactic information into machine translation model, and then modeling the language structure. The research shows that the syntax model can reordering long-distance sentences effectively (xue, et al., 2008). Statistical machine translation methods based on linguistic syntax trees can be further divided into three types: string to tree model, tree to string model, and tree to tree model (Xiong, et al., 2008).

Recent years, the syntax-based neural machine translation model is gradually favored by scholars (Chen, et al.,2017; Eriguchi, et al., 2017; Li, et al., 2017; Aharoni, et al., 2017). In the NMT case, the syntax information is introduced, it will be easier for the encoder to incorporate long distance dependencies into better representations, which is especially important for the translation of long sentences.

The research shows that the syntax-based neural machine translation model outperform the sequential attentional model as well as a stronger baseline with a bottom-up tree encoder and word coverage in Chinese-English machine translation task.

3 Experimental Setup

3.1 Experimental Corpus

In this paper, we use the Tibetan-Chinese comprehensive evaluation corpus of the 13th National Machine Translation Symposium (CWMT 2017 in china, http://ee.dlut.edu.cn/CWMT2017/index.html). This corpus is processed into Tibetan-Chinese sentence pairs, which contains word segmentation, syllable segmentation and some alignment process. This corpus is shown in following Table 1.

Table 1. Experimental Corpus

Corpus	Department	Corpus-Area	Scale (sentence pairs)
QHNU-CWMT2013	Qinghai Normal University (in China)	Government	33145
QHNU-CWMT2015	Qinghai Normal University (in China)	Government	17194
XBMU-XMU	Artificial intelligence institute of Xiamen University (in China) Institute of language (technology), Northwestern University of Nationalities (in China)	Synthesize	52078
XBMU-XMU-UTibent	Institute of language (technology), Northwestern University of Nationalities (in China) Tibet University	Government Law	24159

	Artificial intelligence institute of Xiamen University (in China)		
ICT-TC-Corpus	Institute of Computing Technology, Chinese Academy of Sciences (in China)	News	30004

3.2 Corpus Preprocessing

In this paper, Chinese-Tibetan bilingual parallel corpus is preprocessed and then divided into a training set, (141601 sentence pairs), a development set (1000 sentence pairs) and a test set (1000 sentence pairs). Preprocessing tasks include: character segmentation and syllable segmentation on Tibetan corpus, word segmentation and character segmentation on Chinese corpus. For efficient training, we also filter out the sentence pairs whose source or target lengths are longer than 50. Details are shown as Table2.

Table 2. Sentences and words in Corpus

Language	Sentence pairs	Words	Characters
Tibetan	139535	16742	15201
Chinese	139535	23384	4932

3.3 Corpus Preprocessing

In our experiment, to reflect the performance of neural machine translation, phrase-based statistical machine translation model Nitutrans (Xiao T et al., 2012) developed by natural language processing laboratory in northeastern university (in china) is used. The syntax tree-based neural machine translation model framework Syntax-NMT developed by Nanjing University's natural language processing laboratory is used. (Chen, et al., 2017). Chinese and Tibetan corpus are processed by BPE (Sennrich, et al. 2015). The five neural machine translation models used in this paper are consistent in the basic parameter settings, the vocabulary of the sub word table is set to 32000, and the number of training iterations is 200000. Because each model has its own structure, it is difficult to achieve consistent in terms of performance of parameters. In addition, with the language characteristics of the Chinese-Tibetan bilingual corpus, hyperparameters are adjusted to achieve maximum of translation performance on each model. Bilingual evaluation understudy (BLEU) is used as evaluation index in this paper (Papineni, 2007).

4 Experimental Results

4.1 Corpus according to Character Segmentation and Word Segmentation

To verify the translation performance of the character segmentation (Tibetan syllable segmentation and Chinese character segmentation) and the word segmentation (Tibet-

an word segmentation and Chinese word segmentation) on Chinese-Tibetan corpus. Among them, Tibetan word segmentation tool TIP-LAS is used in the Tibetan word segmentation (Li et al., 2015). THU-LAC software opened by Tsinghua university is used to conduct Chinese word segmentation (Li et al., 2009). The experimental results of the Chinese-Tibetan translation based on the statistical machine translation model are shown in Table 3. The experimental results of the Chinese-Tibetan translation based on the neural network translation model are shown in Table 4.

Table 3. Corpus according to Character segmentation and Word segmentation (SMT model)

Model	Corpus processing	BLEU
Niutrans	Character	70.65*
Niutrans	Word	69.08

Table 4. Corpus according to Character segmentation and Word segmentation (NMT model)

Model	Corpus processing	BLEU
Transformer	Character	71.69*
Transformer	Word	71.25

The experimental results show that in Chinese-Tibetan machine translation task, whether SMT model or NMT model, the performance of character segmentation on corpus is obviously better than that of word segmentation on corpus. The reasons are as follows: Firstly, reduce the granularity of corpus can improve the performance of machine translation; Secondly, based on the language characteristics of Tibetan language, the granularity after segmentation is larger unit, which decreased the performance of machine translation.

4.2 Syntax Tree-Based Machine Translation Models

To verify the translation performance of the syntax-based machine translation models. Niutrans is used in our syntax-based statistic machine translation and Syntax-awared-NMT is used in our syntax-based neural machine translation. We compared the seq2seq model with the Syntax-awared-NMT model and both model have a same framework of neural network. And Berkeley parser tools are used (Petrov, et al., 2006) to generate the Chinese syntax tree. Experimental results are show in Table 5.

Table 5. Syntax Tree-based models

Model	Corpus processing	BLEU
Niutrans-syntax	Word	64.53*
Seq2Seq	Word	56.12
Syntax-awared-NMT	Word	61.45

The experimental results show that when the syntax tree is introduced in the seq2seq model, the translation performance has improved. In the Chinese-Tibetan machine

translation task, the tree-to-string model based on statistical machine translation is better than the syntax tree-based neural machine translation model, this may be due to the small-scale corpus. The reasons are as follows: Firstly, adding prior knowledge such as linguistics can improve the performance of machine translation; Secondly, syntactic tree information, which can well parse source and target languages and improve machine translation performance.

4.3 Different Neural Networks with the Same Structure

To verify the performance of different neural networks with the same model structure, experiments were conducted in RNNSearch and Fairseq models respectively. Both RNNSearch and Fairseq models are models based on the neural network and attention mechanism. The only difference is that RNNSearch is a model based on cyclic neural networks, whereas Fairseq is a model based on convolutional neural networks. The experimental results are shown in Table 6.

Table 6. Different Neural Networks with the Same Structure

Model	Framework	Corpus processing	BLEU
Fairseq	CNN + Attention	Character	18.48
RNNSearch	RNN + Attention	Character	67.71

The experimental results show that there is obvious difference in performance of the same translation structure with different neural networks. Based on its frame-work, Fairseq translation model has a pool performance in Chinese-Tibetan translation tasks. The reasons are as follows: Convolutional neural networks and recurrent neural networks are different in network structure. Convolutional neural networks exhibit strong performance in image processing and computer vision, but natural language processing is a sequence of texts, which is difficult to achieve better results in convolutional neural networks.

4.4 Different Neural Machine Translation Models in Chinese-Tibetan Task

To verify the performance of different neural machine translation models on Chinese-Tibetan translation, based on the same corpus, three neural machine translation models are used in this experiment. Niutrans is the SMT system we use in this experiment. The experimental results are shown in Table 7, and Sample translations for each model are shown in Table 8.

Table 7. Different Neural Networks with the Same Structure

Model	Framework	Corpus processing	BLEU
Niutrans	Phrased-based	Character	70.65
		Word	69.08
Seq2seq	RNN	Character	63.07

	RNN	Word	56.12
RNNSearch	RNN+Attention	Character	67.71
		Word	67.20
Niutrans-Syntax	SMT+Syntax	Word	64.53
Syntax-NMT	NMT+Syntax	Word	61.45
Fairseq	CNN+Attention	Character	18.48
		Word	18.37
Transformer	Attention	Character	71.69*
		Word	71.25

Table 8. Sample translations for each model

Example 1	
original text	灾害给经济造成重大损失，
Reference translation	གནོད་འཚེ་ཡིས་དཔལ་འབྱོར་ཐད་ལ་ཐོག་གནོན་ཆེན་བཟོས།
Niutrans	གནོད་འཚེས་དཔལ་འབྱོར་ལ་ཐོང་གྲན་ཚབས་ཆེན་བཟོས་པ་དང་།
Niutrans-syntax	ག་གནོད་འཚེའི་དཔལ་འབྱོར་ལ་ཐོང་གྲན་ཚབས་ཆེན་བཟོས་བ།
Seq2Seq	དཔལ་འབྱོར་ལ་ཐོང་གྲན་བཟོས་པར་ཐོང་གནོན་ཆེན་ཏོ་བཟོས་ཡོང་།
RNNSearch	གནོད་འཚེའི་དཔལ་འབྱོར་ལ་ཐོང་གྲན་ཚབས་ཆེན་བཟོས་པ།
Syntax-NMT	གནོད་སྐྱེན་ལ་བརྟེན་ནས་དཔལ་འབྱོར་ལ་ཐོང་གྲན་ཚབས་ཆེན་བཟོས་པ་དང་
Fairseq	བསྐུར་དུ་འཕྲིག་སྐྱེན་བྱ་དགོས།
Transformer	དཔལ་འབྱོར་ལ་གནོད་སྐྱེན་ཚབས་ཆེན་བཟོས་ཡོང་བས་
Example 2	
original text	给群众生活带来很大困难。
Reference translation	མང་ཚོགས་ཀྱི་འཚོ་བར་དཀའ་ངལ་ཤ་ཅང་ཆེན་ཏོ་བཟོས།
Niutrans	མང་ཚོགས་ཀྱི་འཚོ་བར་དཀའ་ངལ་ཤ་ཅང་ཆེན་ཏོ་བྱུང་།
Niutrans-syntax	མང་ཚོགས་ཀྱི་འཚོ་བར་དཀའ་ངལ་ཤ་ཅང་ཆེན་ཏོ་བཟོ་ངེས་ཡིན།
Seq2Seq	མང་ཚོགས་ཀྱི་འཚོ་བར་དཀའ་ངལ་ཤ་ཅང་ཆེན་ཏོ་བྱུང་ཡོང་།
RNNSearch	མང་ཚོགས་ཀྱི་འཚོ་བར་དཀའ་ངལ་ཤ་ཅང་ཆེན་ཏོ་ཡོང་།
Syntax-NMT	མང་ཚོགས་ཀྱི་འཚོ་བ་ལ་དཀའ་ངལ་ཤ་ཅང་ཆེན་ཏོ་ཡོང་།
Fairseq	ཐོན་བརྟ་འཐུས་སྐོ་ཚང་དུ་གཏོང་བ།
Transformer	མང་ཚོགས་ཀྱི་འཚོ་བར་དཀའ་ངལ་ཤ་ཅང་ཆེན་ཏོ་ཡོང་།

5 Conclusion

In this paper, five influential neural machine translation models are used in Chinese-Tibetan translation tasks, which are Seq2Seq, RNNSearch, Fairseq, Syntax-NMT, and Transformer. Through comparison, findings are as following:

1. In Chinese-Tibetan translation tasks, there is no obvious difference in the translation performance between most of the machine translation models based on neural network and the traditional statistical machine translation model, even the translation performance of most of the neural machine translation models is better than that of the traditional statistical machine translation model.

2. In Chinese-Tibetan translation task, the translation performance of character segmentation processing on corpus (Tibetan syllable segmentation, Chinese Character segmentation) is better than that of word segmentation processing on corpus;

3. The translation performance of the Transformer model based on the completely self-attention mechanism is the best in Chinese-Tibetan translation tasks.

4. Introducing the syntax tree based on the neural machine translation model can improve translation performance on Chinese-Tibetan machine translation tasks.

5. There is significant difference in translation performance of different neural machine translation models with the same structure.

Acknowledgement

This work is supported by the National Science Foundation of China (61331013).

References

1. Liu, Yang. "Recent Advances in Neural Machine Translation." Journal of Computer Research and Development 54.6(2017):1144-1149. (in Chinese)
2. Zhao Tiejun. Machine Translation Theory. Harbin Institute of Technology Press, 1900. (in Chinese)
3. Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.
4. Wei, Sudong. " Research on Tibetan - Chinese Online Translation System Based on Phrases." Diss. Northwest University for Nationalities (in china), 2015. (in Chinese)
5. Cai, Zhijie and Cai, Rangzhuoma. " Research on the Distribution of Tibetan Character Forms." Journal of Chinese Information Processing 30.4(2016):98-105. (in Chinese)
6. Dong, Xiaofang, Cao, Hui, and Jiang, tao. " Phrase Based Tibetan - Chinese Statistical Machine Translation System." Technology Wind 17(2012):60-61. (in Chinese)
7. Luo, Xiaocong. " Research on Syntax-based Chinese-Tibetan Statistical Machine Translation System." Diss. Xiamen university, 2010. (in Chinese)
8. Hua, Quecairang. " Research on Some Key Technologies of Machine Translation Based on Tree-to-String in Tibetan Language." Diss. Shanxi Normal University, 2014. (in Chinese)
9. Cai, Rangjia. "Research on Large-scale Sino-Tibetan Bilingual Corpus Construction for Natural Language Processing." Journal of Chinese Information Processing 25.6 (2011): 157-162.
10. Pang, Wei. " Research on the Construction Technology of Tibetan-Chinese Bilingual Corpus of Corpus Based on Web." Diss. Minzu University of China, 2015. (in Chinese)
11. Xiang, Bao, Zhang, Guoxi. " Research on the Translation of Han Names in Chinese-Tibetan Machine Translation." Journal of Qinghai Normal Universituy (Natural Science) 27.4 (2011): 88-90.
12. Hua, Guocairang. " Tibetan Verb Researching in Chinese Tibetan Machine Translation." Diss. Qinghai Normal University, 2014. (in Chinese)
13. Nuo, Minghua, Wu, Jian, Liu Huidan and Ding, Zhiming. " Research on Phrase Translation Extraction for Chinese-Tibetan Machine Translation." Journal of Chinese Information Processing 25.3 (2011): 112-118.

14. Li Yachao, Xiong Deyi, Zhang Min, Jiang Jing, Ma Ning and Yin Jianmin. " Research on Tibetan-Chinese Neural Machine Translation." Journal of Chinese Information Processing 31.6 (2017).
15. Guan Queduojie. " Research on Tibetan Segmentation for Machine Translation." Electronic Test 11x (2015): 46-48.
16. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.
17. Cho, Kyunghyun, Van Merrienboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger and Bengio, Yoshua. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." Computer Science (2014).
18. D'Informatique D E, Ese N, Esent P, et al. Long Short-Term Memory in Recurrent Neural Networks[J]. Epfl, 2001, 9(8):1735 - 1780.
19. Bahdanau, Dzmitry, Kyunghyun. Cho, and Yoshua. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." Computer Science (2014).
20. Gehring J, Auli M, Grangier D, et al. Convolutional Sequence to Sequence Learning[J]. 2017.
21. Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz and Polosukhin, Illia. "Attention is all you need." Advances in Neural Information Pro-cessing Systems. 2017.
22. Xue Yongzeng, Li sheng, Zhao Tiejun, Yang muyun. " Syntax-based reordering model for phrasal statistical machine translation." Journal on Communications Test 29.1 (2008): 7-14.
23. Xiong Deyi, Liu Qun and Lin Shouxun. " A Survey of Syntax-based Statistic Machine Translation." Journal of Chinese Information Processing 22.2 (2008): 28-39.
24. Chen, Huadong, et al. "Improved Neural Machine Translation with a Syntax-Aware Encoder and Decoder." (2017):1936-1945.
25. Eriguchi, Akiko, Y. Tsuruoka, and K. Cho. "Learning to Parse and Translate Improves Neural Machine Translation." (2017).
26. Li, Junhui, et al. "Modeling Source Syntax for Neural Machine Translation." (2017):688-697.
27. Aharoni, Roee, and Y. Goldberg. "Towards String-to-Tree Neural Machine Translation." (2017).
28. Xiao, Tong, et al. "NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation." Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.
29. Sennrich, Rico, Barry Haddow, and Alexandra Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).
30. Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.
31. Li Yachao, et al. " TIP-LAS: An Open Source Toolkit for Tibetan Word Segmentation and POS Tagging." Journal of Chinese Information Processing 29.6 (2015): 203-207.
32. Li, Zhongguo, and Maosong Sun. "Punctuation as implicit annotations for chinese word segmentation." Computational Linguistics 35.4 (2009): 505-512.
33. Petrov, Slav, et al. "Learning accurate, compact, and interpretable tree annotation." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.