Tibetan-Chinese Neural Machine Translation based on Syllable Segmentation

Wen Lai Lavine_Lai@126.com
National Language Resource Monitoring & Research Center of Minority Languages, Minzu University of China, Beijing, 100081, China
Xiaobing Zhao* nmzxb_cn@163.com
National Language Resource Monitoring & Research Center of Minority Languages, Minzu University of China, Beijing, 100081, China
Wei Bao wei.bao@adaptcentre.ie
National Language Resource Monitoring & Research Center of Minority Languages, Minzu University of China, Beijing, 100081, China

Abstract

Machine translation is one of the important research directions in natural language processing. In recent years, neural machine translation methods have surpassed traditional statistical machine translation methods in translation performance of most of language and have become the mainstream methods of machine translation. In this paper, we proposed syllable segmentation in Tibetan translation tasks for the first time and achieved better results than Tibetan word segmentation. Four kinds of neural machine translation methods, which are influential in recent years, are compared and analyzed in Tibetan-Chinese corpus. Experimental results showed that the translation model based on the complete self-attention mechanism performed best in the translation task of Tibetan-Chinese corpus, and performance of the most of the neural machine translation methods surpassed performance of the traditional statistical machine translation methods.

Key words: Tibetan-Chinese translation; Neural machine translation; Syllabe Segmentation

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 areas of artificial intelligence and natural language processing (Liu, 2017). Natural language processing (including machine translation) 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, machine translation has achieved fruitful results both in academia and industry.

Since the advent of the neural network in the 1940s, it has experienced the different stages of rising, low tide, and rising. Until 2006, Hinton et al. solved the historic problem of neural networks (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 theoretical research and improvement of computing speed of computers, neural networks have been gradually applied to various fields of artificial intelligence and have made major

breakthroughs. Researches about natural language processing have also made a rapid progress along with this tide.

In 2012, With the Hinton research group participated in the ImageNet image recognition contest and won the championship, which opened the prelude of deep learning in the big bang 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 neural networks (RNN, CNN, etc.) to encode the source language into word embedding. Secondly, the word-embedding generates the target language by decoding . Among them, in the neural network training, the problem of long distance dependence can be solved well by the proper joining of long-short term memory (LSTM) networks and attention mechanisms.

Tibetan is a kind of pinyin character, and its syllables are composed of 34 vowel consonants, 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 words (Cai, 2016). Based on the characteristics of Tibetan language, at present, the statistical machine translation model is mainly used in the research on Tibetan translation model, and the relevant theoretical research has basically stopped at the stage of word processing and other corpus preprocessing such as the phrase-based Tibetan-Chinese statistical machine translation system (Dong et al., 2012); besides, related tibetan preprocessing research (Hua, 2014; Wang, 2016; Wan et al., 2015) and so on. On the whole, compared with research on machine translation of other rich languages, the research on Tibetan-Chinese machine translation is obviously behind. There are few researches on using neural network model in Tibetan corpus (Li et al., 2017). Tibetan texts are all word segmentation pre-processed in traditional Tibetan machine translations (Guan, 2015). In this article, the traditional method of Tibetan word segmentation is completely abandoned, and Tibetan texts are directly divided by syllables. It gets a better performance than Tibetan word segmentation.

In this paper, four kinds of influential machine translation models of neural networks are applied to the task of Tibetan-Chinese machine translation, and the final translation results are analyzed in detail. The experimental results show that the application of neural network machine translation model on Tibetan-Chinese machine translation has basically surpassed the performance of the traditional statistical machine translation model. By using the method of syllable segmentation in Tibetan machine translation tasks, it has a better translation performance than the method of word segmentation.

2. Neural Network Machine Translation Models

2.1. Seq2Seq

The Seq2Seq model is a sequence-to-sequence model. In many translation models in the past, a feature matrix was input during training, and each input corresponds to a row in the matrix. Therefore, these rows have the same dimension, which is not in accordance with the task of machine translation. Because, for machine translation tasks, we cannot guarantee that every sentence you input has the same number of words. Conversely, the input of the Seq2Seq model is a sequence, and the output is also a sequence. The length of the input sequence and the output sequence of this model is variable, which is the biggest difference between this model and the previous neural mahine translation model.

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 commonly known as LSTM --- Long short-term memories network (D'In-

formatique et al., 2001), and this process consists of two parts of encoding input and decoding output.

In the encoder section, an input sequence X will be compiled into a vector C via an encoder built with recurrent neural networks (RNNs). The vector C is usually the last hidden node in the RNN, or the weighted sum of multiple hidden nodes.

In the decoder section, vector C will be decoded by the RNN decoder. The decoding process can be simply understood as using a series of algorithms to return the word with the highest probability corresponding to the input vector to get the optimal output sequence.

2.2. RNNSearch

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

The Attention mechanism is added into the translation model, which breaks the limitation that the traditional encoder-decoder structure, such as Seq2Seq model, relies on a fixed length vector in the process of decoding. The attention mechanism is achieved by preserving the intermediate output results of the input sequence encoded by the encoder and then retraining a new model to selectively learn these input sequences and associate the output sequence with the output of the model. In machine translation tasks, the model using the attention mechanism generates a word vector every time, and it finds one of the most relevant word sets in the input sequence, and then the model will predict the next target word generate based on the current context vectors and all previous generated words to achieve the best translation results.

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. The Fairseq model introduces an architecture based entirely on convolutional neural networks (CNNs). Compared with the recurrent neural network model, all calculations of the element sequence of Fairseq model in training are completely parallel, the number of nonlinear sequences is fixed and independent of the length of the input sequence.

An important part of the Fairseq model in model structure is the multihop attention mechanism. The multihop alert mechanism is an enhanced version of attentional mechanics that allows the neural network to scan sentence sequences more often to produce better results and to influence each scan. Another important part of the model is the gating, which controls the flow of information in a neural network. Sentence sequences are transmitted downward through hidden units in a neural network, and the gating mechanism is used to precisely control the sequence information passed to the next unit, and the translation effect is greatly improved.

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). Neural network is mostly used as the model basis of Encoder-Decoder in traditional neural network machine translation model. This model is based on the attention

mechanism and completely abandons the inherent model of the neural machine translation model without any neural network (CNN or RNN) structure. Experiments show that this model can run in parallel and greatly improve the speed of model training while improving machine translation performance.

Transformer model requires only a small number of uninterrupted steps in the training process, at each step, it uses a self-note mechanism that can directly relate to all words in the sentence and the location of each word does not need to be considered, and model efficiency is also greatly enhanced while simplifying the model. Besides the improvement of computational performance and higher semantic understanding, the transformer model also provides a visual display of how words are processed and how the information travels across the network.

Transformer model performs well in natural language processing tasks such as syntax analysis and semantic understanding, which is also a systematic breakthrough for natural language processing communities over decades.

3. Experimental Setup

3.1. Experimental corpus

This paper uses the evaluation corpus of the 13th National machine translation symposium(CWMT 2017 in china, http://ee.dlut.edu.cn/CWMT2017/index.html). These corpora are processed into Tibetan-Chinese sentence pairs, which contains word segmentation, syllable segmentation and some alignment process. These corpora are 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 Artificial intelligence institute of XiaMen University (in China)	Government Law	24159
ICT-TC-Corpus	Institute of Computing Technology, Chinese Academy of Sciences (in China)	News	30004

3.2. Corpus preprocessing

In this paper, Tibetan-Chinese bilingual parallel corpus is pre-processed and then divided into a training set, (141601 sentence pairs), a development set (1000 sentence pairs) and a test set

(1000 sentence pairs). Pre-processing tasks include: word segmentation of the Tibetan corpus, character segmentation, and operation on word segmentation of the Chinese corpus. Details are shown as Table2.

Table 2 Corpus Statistics in Experimental

Language	Sentence pairs	Words	Characters
Tibetan	141601	16547	13701
Chinese	141601	23644	4968

3.3. Experimental setting

In the experiment, in order 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 in the statistical machine translation model. In this paper, four models of neural machine translation are consistent in the basic parameter settings (the vocabulary size of sub-words is set to 32000 and the number of training iterations is 200000). Because each model has its own architecture, it is difficult to achieve consistent in terms of performance of parameters. In addition, with the language characteristics of the Tibetan-Chinese bilingual corpus, in this paper, based on each model, hyperparameters are adjusted to achieve maximum of translation performance. 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

In order to verify the effect of the character segmentation (Tibetan syllables segmentation and Chinese characters segmentation) and the word segmentation (Tibetan word segmentation and Chinese word segmentation) of Tibetan corpus on the translation performance, Syllable segmentation and Tibetan word segmentation of Tibetan-Chinese bilingual parallel corpus was conducted on the basis of the transformer model in the experiment. 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 Tibetan-Chinese machine translation are shown in table 3. The experimental results of Chinese-Tibetan machine are shown in table 4.

Table 3 Corpus according to Character segmentation and word segmentation (Tibetan-Chinese)

Model	Corpus processing	BLEU
Transformer	Character	51.38
Transformer	Word	38.44

Table 4 Corpus according to Character segmentation and word segmentation (Chinese-Tibetan)

Model	Corpus processing	BLEU
Transformer	Character	41.00
Transformer	Word	30.94

The experimental results show that in neural machine translation, whether Tibetan is translated into Chinese or Chinese is translated into Tibetan, the effect of Character segmentation on corpus is obviously higher than that of word segmentation on corpus. This is the big-

gest difference between traditional machine translation corpus processing and machine translation corpus processing in this paper.

4.2. BPE impacting

The problems of OOV (out of vocabulary) in neural machine translation and Rare Words are usually solved by back-off dictionaries. In 2016, Sennrich et al. (Sennrich, 2015) attempted a more simple and effective way (Subword Units) to represent open vocabularies inspired by translation strategies of the same root word, compound word, naming entity, and foreign language. He considered that separating these rare words into a combination of "subword units" effectively alleviate the problem of translating OOV and rare words. The segmentation strategy of subword unit here draws on a data compression algorithm: Byte Pair Encoding (BPE) algorithm (Suarjaya, 2012; Shibata et al., 1999). In order to verify whether the corpus needs to be pre-processed by BPE before Tibetan-Chinese translation, we have a comparation between BPE processing and no BPE processing. The experimental results are shown in Table 5.

Table 5 BPE impacting (Tibetan-Chinese)

Model	BPE	Corpus processing	BLEU
Transformer	Yes	Character	51.38
Transformer	No	Character	48.50

The experimental results show that in the neural machine translation model, the translation effect will be improved when using BPE processing.

4.3. Different Neural Networks with the Same Structure

In order 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 (Tibetan-Chinese)

Model	Network	Corpus processing	BLEU
RNNSearch	RNN	Character	45.63
Fairseq	CNN	Character	46.94

The experimental results show that there are obvious differences in the translation performance for different neural networks models with the same model structure, and because of its characteristics of the model based on CNN, training time greatly reduced and performance exceeds RNN-based neural network model.

4.4. Different Neural Machine Translation Models in Tibetan-Chinese Corpus

In order to verify the performance of different neural machine translation models on Tibetan-Chinese translation，training of Tibetan-Chinese machine translation model was carried out in four different neural network models respectively in this experiment, meanwhile, the same corpus was trained in statistical machine translation model. Machine translation model Niutrans opened by the natural language processing laboratory of Northeastern University(in china) is used in the statistical machine translation model (use Chinese as monolingual data). The experimental results are shown in Table 7.

Table 7 Different Neural Machine Translation Models in Tibetan - Chinese Corpus (Tibetan-Chinese)

Model	Framework	Corpus processing	BLEU
Niutrans	Phrased-based	character	26.98
		word	24.35
Seq2Seq	RNN	character	32.16
		word	22.19
RNNSearch	RNN+Attention	character	32.16
		word	29.21
Fairseq	CNN+Attention	character	46.94
		word	31.66
Transformer	Attention	character	51.38
		word	38.44

The experimental results show that there are obvious differences in the translation performance of different neural machine translation models. Among them, Most neural machine translation models have better translation performance than statistical machine translation models; translation performance of the model Transformer based on complete self-attention mechanism is the best; the same machine translation model, translation performance of character-based processing is better than performance of word segmentation processing; training time of Fairseq model is the fastest.

5. Conclusion

In this paper, the four-influential neural machine translation models: the Seq2Seq model based on the RNN, the RNNSearch model based on RNN+Attention mechanism, the Fairseq model based on CNN + Attention mechanism, and the Transformer model based on self-attention mechanism are compared in Tibetan-Chinese machine translation tasks. Through the comparison, it has the following findings:

1. In Tibetan translation task, most of the translation performance of the machine translation model of neural network is better traditional statistical machine translation model;
2. In the Tibetan translation (Tibetan-Chinese, Chinese-Tibetan) task, the translation performance of character processing on the original corpus (Tibetan syllable segmentation, Chinese word segmentation) is better than that of word segmentation processing on the corpus;
3. In the neural machine translation model, BPE processing on the original corpus can optimize the translation performance;
4. Different neural network with the same structure, the translation performance of CNN-based neural network is better than the translation performance of RNN-based neural network, and the training speed of CNN-based machine translation model of neural network is much faster than that of RNN-based machine translation model of neural network;
5. The translation performance of the Transformer model based on the completely self-attention mechanism is the best in Tibetan translation tasks.

Acknowledgement

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

References

Bahdanau, Dzmitry, Kyunghyun. Cho, and Yoshua. Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." *Computer Science* (2014).

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)

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).

Dong, Xiaofang, Cao, Hui, and Jiang, tao. " Phrase Based Tibetan - Chinese Statistical Machine Translation System." *Technology Wind* 17(2012):60-61. (in Chinese)

Gers, Felix. "Long short-term memory in recurrent neural networks." Unpublished PhD dissertation, *Ecole Polytechnique Fédérale de Lausanne*, Lausanne, Switzerland (2001).

Guan, Queduojie. "Research on Tibetan Segmentation for Machine Translation." *electronic test* 22(2015):46-48. (in Chinese)

Gehring, Jonas, Auli, Michael, Grangier, David, Yarats, Denis, and Dauphin, Yann N. "Convolutional Sequence to Sequence Learning." (2017).

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." *Neural computation* 18.7 (2006): 1527-1554.

Hua, Guocairang. " Tibetan Verb Researching in Chinese Tibetan Machine Translation." Diss. *Qinghai Normal University* (in china), 2014. (in Chinese)

Liu, Yang. "Recent Advances in Neural Machine Translation." *Journal of Computer Research and Development* 54.6(2017):1144-1149. (in Chinese)

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): 103-109. (in Chinese)

Li, Yachao, Jiang, Jing, Jia, Yangji and Yu, Hongzhi. "TIP-LAS: An Open Source Toolkit for Tibetan Word Segmentation and Part of Speech Tagging." *Journal of Chinese Information Processing* 29.6 (2015): 203-207. (in Chinese)

Li, Zhongguo, and M. Sun. Punctuation as implicit annotations for chinese word segmentation. *MIT Press*, 2009.

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu, Weijing. "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.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

Sennrich, Rico, Barry, Haddow, and Alexandra, Birch. "Neural Machine Translation of Rare Words with Subword Units." *Computer Science* (2015).

Suarjaya, I. Made Agus Dwi. "A New Algorithm for Data Compression Optimization." *International Journal of Advanced Computer Science & Applications* 3.8(2012).

Shibata, Yusuke, Kida, Takuya, Fukamachi, Shuichi, Takeda, Masayuki, Shinohara, Ayumi, Shinohara, Takeshi and Arikawa, Setsuo. " Byte Pair encoding: A text compression scheme that accelerates pattern matching." Technical Report DOI-TR-161, *Department of Informatics, Kyushu University*, 1999.

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 Processing Systems*. 2017.

Wei, Sudong. " Research on Tibetan - Chinese Online Translation System Based on Phrases." Diss. *Northwest University for Nationalities (in china)*, 2015. (in Chinese)

Wang, Tianhang. " Research on Tibetan Functional Chunk Identification for Machine Translation." Diss. *Beijing Institute of Technology (in china)*, 2016. (in Chinese)

Wan, Fucheng, Yu, Hongzhi, Wu, Xihong and He, Xiangzhen. " Research of Tibetan syntax for machine translation." *Computer Engineering and Applications* 51.13(2015):211-215. (in Chinese)

Xiao, Tong, Zhu, Jingbo, Zhang, Hao and Li, Qiang. "NiuTrans: an open source toolkit for phrase-based and syntax-based machine translation." *50th Meeting of the Association for Computational Linguistics* 2012:19-24.

Zhao Tiejun. Machine Translation Theory. *Harbin Institute of Technology Press*, 1900. (in Chinese)