International Journal of Computer Applications Technology and Research
Volume 14–Issue 09, 19 – 22, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1409.1003

Transformer based Tibetan-Chinese Neural Machine Translation

Chao Tang*
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Zehua Lv
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Ximing Yuan
Chengdu University of Information Technology
College of Communication Engineering
ChengDu, China

Abstract: Neural machine translation has demonstrated good performance in many tasks, but its performance in low-resource languages remains unsatisfactory. To address this issue, this paper proposes a Transformer-based Chinese-Tibetan machine translation model. The model introduces a recurrent mechanism and temporal encoding on the basis of Transformer, enhancing the model's generalization ability and computational efficiency. Additionally, Beam Search is employed on the decoding side to optimize the generation process. To address potential homophonic or phonetically similar errors in the output, word-level language model perplexity is used for evaluation. Experiments use BLEU as the evaluation metric, and the results indicate that under the constraint of limited Tibetan-Chinese parallel corpora, the proposed improved Transformer model effectively enhances translation quality, with a 1.63% increase in BLEU scores.

Keywords: Tibetan-Chinese translation; beam search; recurrent mechanism; transformer;

1. INTRODUCTION

Machine translation (MT)[1]refers to the technology of automatically converting one natural language (source language) into another natural language (target language) using the high computing power of machines. Tibetan machine translation has undergone decades of development, evolving from early rule-based and statistical methods[2] to the current stage of neural network-based technologies.

Research on Tibetan machine translation began as early as the beginning of the 21st century. In rule-based methods, Tsai Zangtai[3] combined lexical information with Tibetan grammatical rules to propose a verb-centered binary grammatical analysis technique. The Tibetan machine translation system developed based on this technique has translation functions for dictionaries, official documents, and science and technology, with a system dictionary size of 186,000 entries. After evaluation, the readability of the translated text reached 80%.

Currently, mainstream Tibetan machine translation research focuses on neural network methods. Renqing Dongzhu [4]and others utilized 500,000 Tibetan-Chinese parallel corpora, combined with a LSTM (Long Short-Term Memory Network) model based on bidirectional RNN, to develop a Tibetan-Chinese machine translation system, achieving a BLEU score of 31. Li Yachao[5] and others proposed using transfer learning to address the scarcity of Tibetan-Chinese corpora and validated through comparative experiments with phrase-based statistical machine translation that this method could improve the BLEU score by 3 points.Currently, multiple research teams and institutions, including the Nima Zaxi team, Northeast University's "Xiaoniu Translation Online Open Platform," and Tencent Company, have developed Tibetan-Chinese machine translation systems that all adopt neural network-based methods.

In 2017, Google published the paper "Attention Is All You Need,"[6] formally proposing the Transformer model based entirely on attention mechanisms, achieving a maximum BLEU score of 41.8 in two Latin-based machine translation tasks. In 2019, Sangje Danzhu [7]used the Transformer model to study Tibetan-Chinese machine translation back-translation methods under resource-scarce conditions, achieving a maximum BLEU score of 27.6 using 930,000 Tibetan-Chinese parallel corpora. In comparison, there is still significant room for improvement in Tibetan machine translation performance.

2. RELATED WORK

2.1 Seq2seq structure

The seq2seq framework primarily consists of two components: an encoder and a decoder. In neural machine translation tasks, the encoder and decoder are trained jointly to enhance the model's performance and generalization capabilities. The seq2seq framework specifically includes four components: the word embedding layer, the encoder, the intermediate vector, and the decoder. The word embedding layer typically consists of two components: one for converting the source text into word vectors, and the other for converting the word vectors into the target language text. The encoder modifies the word vectors generated by the word embedding layer to better represent the source language information. The intermediate vector is a context vector containing semantic information, which serves as the input vector for the decoder; the decoder converts the input context vector into target language text. In the seq2seq framework, the encoder and decoder should be as different as possible to increase the model's parameter count, enhance its performance, and improve the quality of the final translation[8].

2.2 Self-attention mechanism

To address the issue of poor performance in handling long texts in seq2seq machine translation models, Bahdanau [9]et al. drew inspiration from the attention mechanism used in image processing in 2014, marking the first application of the attention mechanism in the field of natural language processing. Since the introduction of the attention mechanism, seq2seq models incorporating attention mechanisms have shown improvements across various tasks. In 2017, the Google team proposed the Transformer model, which introduced a fully attention-based architecture to replace

International Journal of Computer Applications Technology and Research
Volume 14–Issue 09, 19 – 22, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1409.1003

LSTM, achieving better performance in translation tasks[10]. This model employs a self-attention mechanism to encode the input sequence. Figure 1 illustrates the self-attention architecture.

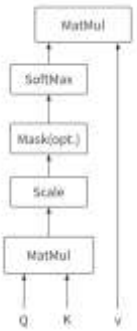

Figure.1 Self-attention Structure

3. MODEL CONSTRUCTION

3.1 Transformer-based machine translation model

This paper combines the recurrent mechanism of RNN (recurrent neural network) with the transformer. The basic structure of the encoder and decoder is consistent with that of the transformer, but the recurrent mechanism of the recurrent neural network is introduced on the basis of the transformer. In the transformer, the input enters the fully connected layer after passing through the multi-head self-attention mechanism, while the recurrent mechanism enters the transition layer, where recurrent calculations continue through a weight-sharing transition function. Figure 2 shows the structure diagram of the recurrent mechanism. Figure 3 shows the model architecture diagram. In Figure 2, h_m^t refers to the positions of various features in a sequence, and the horizontal time primarily represents the computational sequence. For example, in a sequence (a, b, c, d), it first passes through an embedding layer to be represented as (h_a^t , h_b^t , h_c^t , h_d^t), then through an attention layer and transition layer to be represented as (h_a^{t+1} , h_b^{t+1} , h_c^{t+1} , h_d^{t+1}). In an RNN, h_a^t is calculated first, followed by h_a^{t+1} and h_b^t . In contrast, the self-attention mechanism in a transformer can simultaneously compute (h_a^t , h_b^t , h_c^t , h_d^t) and then compute $t+1$. Thus, the output H^t of each self-attention+transition can be represented as:

$$H^t = \text{Layer Norm}\left(A^t + \text{Transition}(A^t)\right) \quad (1)$$

$$A^t = \text{Layer Norm}\left(H^{t-1} + \text{MultiHeadSelf Attention}(H^{t-1} + P^t)\right) \quad (2)$$

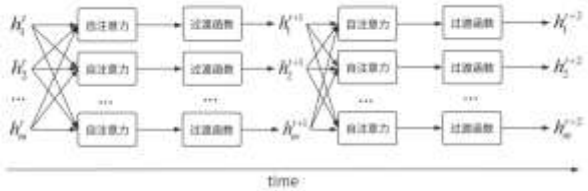

Figure.2 Circular Mechanism Structure Diagram

The Transformer model only considers feature positional information, whereas this paper incorporates a temporal dimension. Each time step performs an embedding of the position coordinates, with the embedding formula defined as:

$$P^t_{\text{pos},2j} = \sin\left(\frac{\text{pos}}{10\,000^{\frac{2j}{d}}}\right) \oplus \sin\left(\frac{t}{10\,000^{\frac{2j}{d}}}\right) \quad (3)$$

$$P^t_{\text{pos},2j+1} = \cos(\text{pos} / 10\,000^{2j/d}) \oplus \cos(t / 10\,000^{2j/d}) \quad (4)$$

3.2 Output result optimization

In machine translation tasks, language models are utilized to evaluate the quality of model-generated sentences. Since machine translation models do not consider all possible outputs but retain only the top k most probable solutions, and since the length of model outputs is predictable, the beam search optimization algorithm is employed to obtain the optimal output result[11]. The computational process for obtaining the optimal output result is as follows:

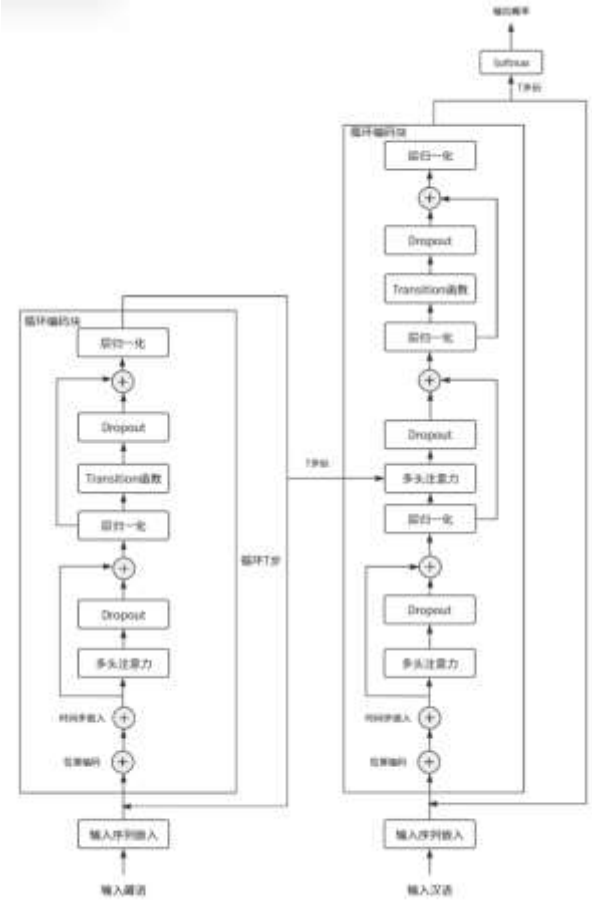

Figure. 3 Architecture Diagram of Tibetan-Chinese Translation Model

$$\text{score}(y_1, y_2, \cdots, y_t) = \sum_{i=1}^{t} \ln P(y_i \mid y_1, y_2, \cdots, y_t, x) \quad (5)$$

3.3 Penalize short sentences and penalize repetition

As shown by the conditional probability formula, beam search tends to select the shortest sentences. However, when multiplying multiple conditional probabilities below 1, issues

International Journal of Computer Applications Technology and Research
Volume 14–Issue 09, 19 – 22, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1409.1003

such as numerical underflow and k parameter adaptation may arise. Therefore, this paper introduces a penalty for short sentences to mitigate this problem[12]. Since attention mechanism coverage can lead to over-translation or incomplete translation, this paper introduces a penalty for repeated tokens to prevent certain tokens from receiving excessive attention[13]. This approach yields a new beam search score.

The specific calculation methods for penalizing short sentences and penalizing repetitions are as follows, where α is used to penalize short sentences and β is used to penalize repetitions:

$$s(Y,X)=\ln(P(Y\mid X))/lp(Y)+cp(X;Y) \quad (6)$$

$$lp(Y)=\frac{(5+|Y|)^{\alpha}}{(5+1)^{\alpha}} \quad (7)$$

$$cp(X;Y)=\beta\times\sum_{i=1}^{|X|}\ln\left(\min\left(\sum_{j=1}^{|Y|}p_{i,j},1.0\right)\right) \quad (8)$$

4. DATA PROCESSING

4.1 Data augmentation

Neural machine translation typically relies on large-scale parallel corpora for training, yet acquiring such data proves challenging for low-resource languages[14]. To expand training datasets, data augmentation techniques are often employed. The machine translation field currently utilizes two primary data augmentation methods: lexical substitution and back-translation. This paper employs back-translation for data augmentation to expand the Tibetan-Chinese parallel corpus. The core idea involves first training a reverse translation model from Chinese to Tibetan to generate pseudo-parallel data, then applying this synthetic data to train the Tibetan-to-Chinese translation model. Figure 4 illustrates the back-translation workflow.

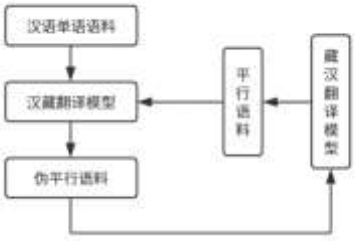

Figure. 4 Back-Translation Flowchart

4.2 BPE encoding

Before training a model, constructing a vocabulary is an extremely important task. Traditional methods typically build the vocabulary using all words that appear in the training corpus or based on individual characters. If the training corpus-based approach is used, it becomes difficult to handle unregistered words. Furthermore, if the training corpus contains a large number of words, the resulting vocabulary becomes enormous, which in turn affects training speed. Conversely, character-based lexicons sacrifice semantic information due to their excessive granularity. The BPE [15]algorithm addresses this by segmenting words into subwords, achieving a middle ground between character and word granularity. This approach reduces lexicon size while preserving as much semantic information as possible.

5. EXPERIMENTAL DESIGN

5.1 Experimental environment and evaluation metrics

The experiments were conducted on a Linux system with an Intel Core i9-12900K CPU and four NVIDIA GeForce RTX 4090 GPUs, running in a GPU environment. To evaluate the Tibetan-Chinese translation model promptly and effectively, the most widely adopted BLEU (Bilingual Evaluation Understudy) metric was selected as the evaluation criterion after comprehensively reviewing existing machine translation performance assessment methods. The specific algorithm is as follows:

$$BP=\begin{cases}1, c>r\\ e^{(1-r/c)}, c,, r\end{cases} \quad (9)$$

$$\text{BLEU}=BP\exp\left(\sum_{n=1}^{N}w_n\ln p_n\right) \quad (10)$$

Here, c denotes the length of the machine translation, r denotes the length of the human translation, and $\exp\left(\sum_{n=1}^{N}w_n\lg p_n\right)$ denotes the log-weighted sum of the accuracies of different n-grams.

5.2 Data augmentation experiment

The experimental training corpus comprised 200,000 parallel data pairs and 200,000 Chinese monolingual sentences. During the back-translation data augmentation phase, the open-source fairseq tool was employed, with the Transformer model serving as the baseline for evaluation. Additionally, results were compared based on whether BPE encoding was utilized. The Transformer model parameters were configured with a maximum sentence length of 50 words, 512-dimensional word embeddings, and a batch size of 32 during training. The network comprised 6 layers, with the multi-head attention mechanism set to 8 heads and dropout at 0.2. The Adam optimizer [16]was employed during training, with an initial learning rate of 2.0 and 10,000 training steps.

The experimental results are shown in Table 1. As indicated in Table 1, both the data augmentation method using back-translation and the results incorporating BPE encoding outperform the baseline model. The back-translation model increased BLUE by 0.88%, while introducing BPE encoding improved it by approximately 1.79%. Concurrently using BPE encoding and data augmentation achieved an enhancement of about 5.11%.

Table 1 BLEU scores for different data processing methods

Processing Method	BLUE(%)
Baseline	36.35
Baseline+ Back-translation	37.23
Baseline+ BPE	38.14
Baseline+ BPE+ Back-translation	41.46

5.3 Improved transformer experiments

The parameters of the improved Transformer model were configured identically to those of the baseline Transformer model. Additionally, RNN, CNN+Attention, and the baseline Transformer model were selected for comparative experiments against the improved Transformer model. The experimental results are shown in Table 2. It is evident that the BLUE score of the Transformer-based model demonstrates a significant improvement compared to both the CNN+Attention and RNN models. Furthermore, the BLUE score of the improved Transformer model proposed in this paper also increased by 1.63%.

Table 2 BLEU Comparison Results Across Different Models

Model	BLUE(%)
RNN	32.06
CNN+Attention	34.45
Transformer	40.13
Improve transformer	41.76

6. CONCLUSION

This paper employs an improved Transformer-based Chinese-to-English translation model. Building upon the Transformer architecture, it incorporates a recurrent mechanism and temporal encoding. This ensures that, except for the initial input which uses raw data, subsequent inputs are derived from the previous time step's output. Additionally, a language model is applied at the output stage to refine the model's results. The approach also introduces back-translation and BPE encoding for data preprocessing. Experimental results demonstrate that incorporating data preprocessing significantly enhances the transformer model's performance. Comparing the standard transformer Tibetan-Chinese translation model with the improved version reveals further substantial gains, with BLUE scores increasing by 1.63% over the baseline transformer model.

REFERENCES

[1] Ren Q ,Li S ,Wei X , et al.Research on Mongolian–Chinese Neural Machine Translation Based on Implicit Linguistic Features and Deliberation Networks[J].

[2] Mohamed S A,Elsayed A A,Hassan Y F,et al.Neural machine translation:past,present,and future[J].Neural Computing and Applications.

[3] Cai Zangtai. Research on Binary Syntactic Analysis Methods in Rule-Based Chinese-Tibetan Machine Translation Systems [C]//National Joint Academic Symposium on Multilingual Knowledge Base Construction.

[4] Renqing Dongzhu, Toudan Tsering, Nyima Tashi. A Review of Chinese-Tibetan Machine Translation Research [J]. Chinese Tibetology.

[5] Li Yachao, Xiong Deyi, Zhang Min, et al. Research on Tibetan-Chinese Neural Machine Translation [J]. Chinese Journal of Information Science.

[6] VASWANI A,SHAZEER N,PARMAR N,etal.Attention is all you need[EB/OL].

[7] Sangye Danzhu. Research on Tibetan-Chinese Machine Translation under Sparse Resource Conditions [D]. Xining: Qinghai Normal University.

[8] Zhang Tianyu. Research on Machine Translation Based on Deep Learning [D]. Beijing: North China Electric Power University.

[9] Ren Huan, Wang Xuguang. Review of Attention Mechanisms [J]. Computer Applications.

[10] Liu H I,Chen W L.Re-transformer:a self-attention based model for machine translation[J].Procedia Computer Science.

[11] Zhang Jianguo. Research and Application of Neural Machine Translation Models Based on Attention Mechanisms [D]. Chengdu: University of Electronic Science and Technology of China.

[12] Zheng Guokai. Research on Machine Learning-Based English Text Summarization and Machine Translation Technologies [D]. Guangzhou: Jinan University.

[13] Li Zhixin, Peng Zhi, Tang Suqin, et al. Text Summarization Integrating Contextual and Key Information [J]. Chinese Journal of Information Science.

[14] You Congcong, Gao Shengxiang, Yu Zhengtao, et al. A Chinese-Vietnamese Neural Machine Translation Method Based on Synonym Data Augmentation [J]. Computer Engineering and Science.

[15] Jia Hao, Wang Xu, Ji Baijun, et al. Non-autoregressive Neural Machine Translation Based on a Masking Mechanism [J]. Journal of Xiamen University (Natural Science Edition).

[16] Aishan Wumaier, Siraj Aihamit Ruzemamit, Xire'ali Hailera, et al. Bidirectional Chinese-English Neural Machine Translation Method with Annotated Syllables [J]. Computer Engineering and Applications.