Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/53/20241428

Comparison and analysis of two methods for improving the accuracy of OpenNMT in literary and vernacular Chinese translation

Yiwen Li

College of Business Administration, Northeast University, Shenyang, China

20211060@stu.neu.edu.cn

Abstract. As machine translation advances, challenges persist in achieving high accuracy when translating between Literary and Vernacular Chinese. Literary Chinese is known for its concise writing style, often characterized by monosyllabic words. In contrast, Vernacular Chinese incorporates more polysyllabic words. This study utilizes the OpenNMT system to address these challenges and employs different Classical Chinese word segmentation tools to train 2-layer Long Short-Term Memory and Transformer models. These models are then compared and analyzed to measure the improvement in precision. The research findings reveal that adopting a character-level-based word segmentation method for Classical Chinese, coupled with training the Transformer model using OpenNMT, significantly enhances precision. This outcome validates the current observation that existing Classical Chinese word segmentation methods lack sufficient accuracy, consequently impacting the quality of translations between Literary and Vernacular Chinese. By exploring and investigating these approaches, this study contributes to advancing machine translation techniques for improving accuracy in rendering Literary and Vernacular Chinese translations.

Keywords: OpenNMT, Transformer, 2-layer LSTM, word segmentation.

1. Introduction

Machine translation plays a crucial role in the field of artificial intelligence. It can assist humans in cross-lingual communication, reduce communication costs, and find extensive applications in business, politics, and culture [1]. The mutual translation between Literary Chinese and Vernacular Chinese is a specialized field in machine translation. This field greatly assists in exploring the history and culture of China and even the world, carrying significant historical significance. In the current market, the field of machine translation between Literary Chinese and Vernacular Chinese has assisted numerous Chinese scholars in reading and researching documents with greater efficiency and convenience.

Wei developed a Classical Chinese to Vernacular Chinese translation model based on external knowledge collaboration, which effectively improves translation performance [2]. Chang designed a multi-label prediction task, utilizing temporal information from the ancient text as translation assistance and enhancing translation quality based on contextual time order [3]. These approaches have significantly contributed to machine translation between Literary and Vernacular Chinese. They have demonstrated the potential to enhance translation accuracy and efficiency, opening up new possibilities for researchers and scholars in their reading and analysis of historical texts. An increasing number of researchers have been devoting themselves to the field of machine translation between Literary Chinese and Vernacular Chinese.

While machine translation has matured, various challenges remain in Literary Chinese to Vernacular Chinese translation. For example, significant language differences exist between Classical Chinese and Vernacular Chinese. The syntax, semantics, and expression methods of the two forms of Chinese differ considerably, posing unique challenges for machine translation systems. The lack of training data and resources for Literary Chinese to Vernacular Chinese translation further exacerbates this challenge. These issues make it difficult to achieve accurate and natural translations. Therefore, researchers continue to explore innovative approaches and techniques to tackle these challenges and improve the performance of machine translation systems in this domain. Currently, in the translation between Literary Chinese and Vernacular Chinese, there is no standardization for Literary Chinese, which can affect the accuracy of model training. Literary Chinese is known for its linguistic characteristics of conciseness, with words typically being monosyllabic.

On the other hand, vernacular Chinese often consists of polysyllabic words. This difference poses particular difficulties in mutual translation [4]. In the field of Classical Chinese to Vernacular Chinese translation, previous research has predominantly focused on translating from Classical Chinese to Vernacular Chinese to facilitate understanding of its meaning. However, there has been relatively less research on translating from Vernacular Chinese to Classical Chinese, and the accuracy of such translations could be higher. Indeed, word segmentation in Classical Chinese differs significantly from modern Chinese, and accurate word segmentation in Classical Chinese remains a challenge that requires attention and improvement [5].

Therefore, this paper compares and analyzes two methods for improving the accuracy of translating from vernacular Chinese to Classical Chinese. This study utilizes OpenNMT, an open-source neural machine translation system, to improve word segmentation results in Classical Chinese and compare the translation accuracy of two models: a 2-layer LSTM (Two-layer et al.) and the Transformer model. The research aims to achieve better translation accuracy. This study employed the parallel corpus dataset of Literary Chinese and vernacular Chinese from NiuTrans Open Source. The evaluation of the models was based on training accuracy. The research analyzed and compared the effectiveness of different methods in improving translation accuracy. The study achieved excellent machine translation results.

2. Methods

2.1. Data Pre-processing

This study selected the Classical-modern dataset, which is a parallel corpus of literary Chinese and vernacular Chinese maintained by NOS. The dataset consists of 327 books in total. The bilingual data includes 97 books, with a total of 972,467 sentence pairs that are aligned at the sentence level. Table 1 shows the division of the dataset into training, validation, and testing sets. This partitioning is necessary because raw data cannot be directly fed into the model for training.

Table 1. Data set partitioning method.

Settings	Number of sentence pairs
Training Set	962,467
Validation Set	10,000
Test Set	10,000

After selecting the dataset, batch preprocessing was performed on the data. First, data cleaning was conducted to standardize the input strings. Redundant characters and spaces were removed, and only valid sentence pairs were retained. Then, the target language and source language were matched sentence by sentence.

Subsequently, tokenization was performed on the target and source languages. For Chinese text in modern vernacular style, jieba (a Python-based Chinese segmentation tool) was used for tokenization. The jieba library adopts a "prefix dictionary-based" tokenization algorithm, which is efficient, simple, and easy to use. For literary Chinese, tokenization is performed at the character level, separating each character by a space. In this study, the jieba library is used for tokenization of vernacular Chinese. It effectively segments continuous Chinese text into meaningful words and phrases. For literary Chinese, tokenization is performed at the character level, separating each character by a space. The tokenization results are then aligned with the tokenized modern Chinese text on a one-to-one basis.

2.2. Model

Translating from literary Chinese to vernacular Chinese requires the system to accurately understand the correspondence and specific meanings between literary Chinese and vernacular Chinese. OpenNMT is an open-source neural machine translation system that can perform machine translation and various related functions. Its principle architecture is shown in Figure 1.

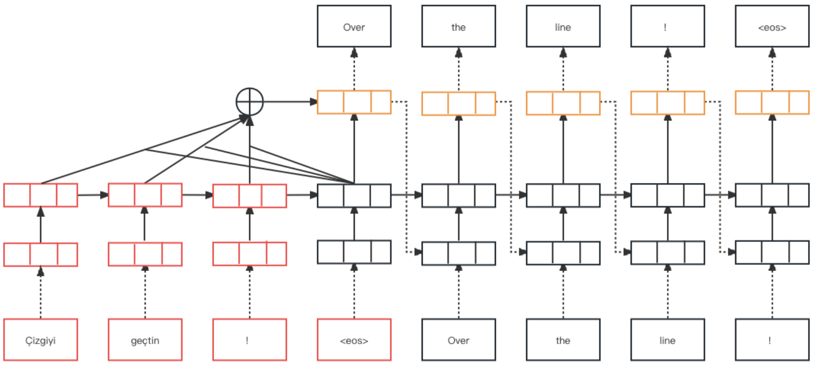

Figure 1. The model structure of OpenNMT [6].

The models commonly used in neural machine translation are LSTM, Seq2Seq (Sequence to Sequence), and the Transformer model [7]. When using OpenNMT for translation between literary Chinese and vernacular Chinese, the system defaults to a 2-layer LSTM model with 500 hidden units in both the encoder and decoder. The input sequence is processed by the first layer LSTM to obtain a set of feature representations, which are then fed into the second layer LSTM. The final output is generated from this process. The architecture of this model is illustrated in Figure 2.

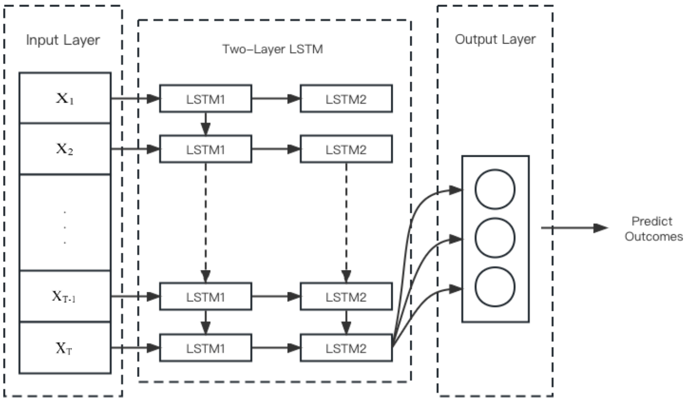

Figure 2. The model structure of 2-layer LSTM [8].

The 2-layer LSTM model enhances the representation and feature extraction capabilities of the system by increasing the network depth [9]. Each layer of LSTM can capture abstract information at different levels. The second layer LSTM can build on the feature representations learned by the first layer, and learn more complex temporal patterns, thus better capturing long-term dependencies.

The Transformer model has shown better performance in machine translation compared to other models, and its overall structure is illustrated in Figure 3. Due to its use of self-attention mechanisms, it can make use of multiple layers of self-attention and feed-forward neural network layers, enabling it to better handle long-range dependencies. Therefore, it has a significant advantage over other models. The internal structure of the Transformer model is shown in Figure 4.

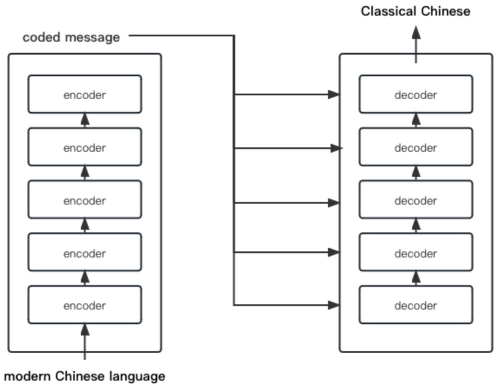

Figure 3. The structure of the Transformer model is used in the literary Chinese to vernacular Chinese translation model [10].

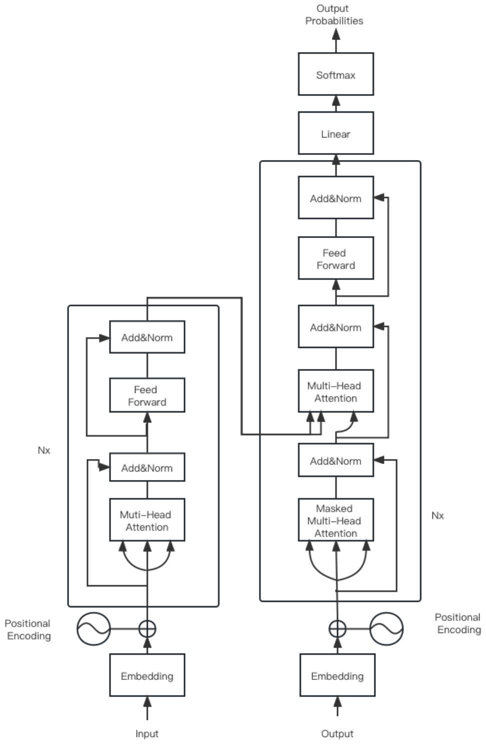

Figure 4. The internal structure of the Transformer model [10].

2.3. Improved method

2.3.1. Improve the literary Chinese word segmentation method. During the data preprocessing stage, both literary Chinese and vernacular Chinese are treated as continuous long texts. For vernacular Chinese, the jieba segmentation library utilizes a prefix-based word dictionary algorithm, which efficiently and accurately splits the vernacular Chinese into words and phrases. For literary Chinese, a character-level segmentation approach is used. This means that proper nouns such as personal names, place names, and years are also segmented into individual characters. As a result, there may be inconsistencies between the segmentation of Classical Chinese and Vernacular Chinese. Therefore, this study attempts to use a word-level segmentation approach for literary Chinese. In this paper, two different literary Chinese segmentation methods are used.

1) The first approach involves using the "Jiayan" word segmentation tool, which is an NLP toolkit specifically designed for processing literary Chinese. It utilizes a bidirectional long short-term memory network (BiLSTM) model combined with a conditional random field (CRF) for sequence labeling to identify the word boundaries of each character in the given text [11, 12].

2) The second approach involves using web crawling to extract all proper nouns from online Classical Chinese dictionaries and forming a custom dictionary. Then, the jieba segmentation library is used to segment Classical Chinese by utilizing a combination of prefix-based trie tree and post-processing algorithms based on the custom dictionary created earlier.

2.3.2. Improve training model. In this paper, there is also an attempt to replace the 2-layer LSTM training model of the OpenNMT system with the Transformer model to improve translation accuracy. The Transformer model is known for its ability to handle long-range dependencies and has shown better performance in various NLP tasks, including machine translation. By incorporating the Transformer model into the training process, it is expected to enhance the translation precision of the OpenNMT system.

Before training, it is necessary to specify the use of the Transformer model in the configuration file. Based on the model parameters specified in the configuration file, the "train" tool in OpenNMT will construct a Transformer model. This includes building the Transformer encoder and decoder, defining components such as word embedding layer, multi-head self-attention mechanism, feed-forward neural network, etc., and combining them into a complete Transformer model. These components work together to enable the Transformer model to effectively learn and capture the dependencies in the data during training. The "train" tool uses a cross-entropy loss function as the optimization objective during training. Cross-entropy loss is typically calculated at each time step for the Transformer model and averaged over the entire batch. This loss function is used to measure the difference between the translation output of the model and the target language during training. The goal is to minimize this loss function so that the model can learn to produce more accurate translations. The model uses the Adam optimization algorithm. The "train" tool takes the input source language and target language sequences during the training process and feeds them into the Transformer model. It performs forward propagation and backward propagation through the model. The gradients are then calculated based on the loss function, and the model parameters are updated accordingly using the Adam optimizer. This iterative process of forward-backward propagation and parameter update helps the model learn and improve its translation capabilities over time.

3. Experiment and result

When training the model using the 2-layer LSTM, the training process is performed by calling the "train" tool in OpenNMT and configuring the corresponding parameters with the 2-layer LSTM. The selected parameters are shown in Table 2.

Table 2. Model parameter setting.

Parameter Type	Value	Explanation of Parameter
src_vocab_size	200000	Vocabulary size of the source language
tgt_vocab_size	200000	Vocabulary size of the target language
queue_size	100	The size of the queue used for batch data processing
bucket_size	2048	The size of the bucket used to group samples by length
world_size	1	The number of GPUs when multiple GPUs are used
gpu_ranks	[0]	Specify the ranking of each GPU in a multi-GPU environment
batch_size	32	Batch size during training
valid_batch_size	16	Batch size at validation time

Based on the combination of the 2-layer LSTM model and three different segmentation methods for Classical Chinese at the character and word levels, the experimental results are shown in Table 3. Thus, based on the 2-layer LSTM model, the character-level segmentation method for Classical Chinese achieves a higher accuracy, reaching up to 60%, and shows the best translation performance. This could be because the accuracy of jiayan tool for Classical Chinese segmentation still needs improvement. It could also be attributed to the limited data volume when constructing custom dictionaries and the reduced segmentation accuracy when jieba segmentation library loads and utilizes custom dictionaries for Classical Chinese segmentation with post-processing algorithms. These factors contribute to the overall decrease in training accuracy. Both improved word-level segmentation methods achieve the highest accuracy of around 55%, as shown in Table 3.

Table 3. Experimental results of different word segmentation methods based on 2-layer LSTM model.

index method	based on character	custom dictionary based on jieba	based on jiayan
Accuracy	60.89	55.23	55.42
Perplexity	7.70	14.50	14.47
Cross-Entropy	2.04	2.90	2.70
Learning Rate	0.00391	0.003709	0.00365

In this work, the OpenNMT system was also utilized to train a Transformer model for improving translation accuracy. The parameter configurations of the model are shown in Table 4.

Table 4. Model parameters of Transformer model trained based on OpenNMT system

Parameter Type	Value	Explanation of Parameter
word_vec_size	512	Number of hidden nodes in the embedding layer
layers	6	The number of encoder and decoder layers
transformer_ff	2048	Number of hidden nodes in the feedforward layer
heads	8	The number of heads of multiple attention
accum_count	8	Cumulative update gradient cumulative steps
optim	adam	Type of optimizer
adam_beta1	0.9	beta1 parameter of the Adam optimizer
adam_beta2	0.998	beta2 parameter of the Adam optimizer
decay_method	noam	Learning rate attenuation method
learning_rate	2.0	learning rate
batch_size	4096	batch size
batch_type	tokens	batch type

Table 4. (continued)

dropout	0.1	dropout rate
label_smoothing	0.1	smoothing rate

The training of the Transformer model using the OpenNMT system has yielded excellent results, as shown in Table 5. Due to its inherent model architecture, the Transformer model demonstrates significant improvements in translation accuracy in the domain of modern Chinese to classical Chinese translation, achieving a maximum accuracy of 76.1%. This represents a 16% increase compared to the accuracy achieved by the 2-layer LSTM model.

Table 5. Experimental results of Transformer model training based on OpenNMT.

Accuracy	Perplexity	Cross-Entropy	Learning Rate
76.1	9.6	2.3	0.0037

4. Conclusion

This study found that using the Transformer model based on character-level segmentation for Classical Chinese in the OpenNMT system can achieve higher accuracy. The research also indicates that there are still some special nouns that cannot be recognized and segmented properly by existing segmentation tools for Classical Chinese. Therefore, in the domain of modern Chinese to classical Chinese translation, the accuracy of models trained with word-level segmentation methods is lower than those trained with character-level segmentation methods.

Similarly, when using character-level segmentation methods for training in the OpenNMT system, the Transformer model outperforms the 2-layer LSTM model in terms of translation accuracy because of its self-attention mechanism and handling of long-distance dependencies. This work has conducted a comprehensive comparative analysis of OpenNMT system's accuracy in the domain of modern Chinese to classical Chinese translation by changing segmentation methods and models, and has identified more suitable models and segmentation methods for this domain that improve translation accuracy.

In future research on the translation between Classical Chinese and Modern Chinese, it is important to establish a standardized evaluation criteria and a part-of-speech contrastive table to enhance the accuracy of segmentation in Classical Chinese and achieve improvements in translation accuracy. Furthermore, combining the Transformer model with the self-attention mechanism and addressing long-distance dependencies in the OpenNMT system can be optimized and adjusted to achieve even higher levels of accuracy.

References

[1] Castilho S, Moorkens J, Gaspari F, Calixto I, Tinsley J and Way A 2017 Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108

[2] Jiaze W 2020 Research on Machine Translation from Classical Chinese to Modern Chinese Based on Collaborative External Knowledge *Beijing: Institute of Scientific and Technological Information of China*

[3] Chang E, Shiue Y T, Yeh H S, and Demberg V 2021 Time-Aware Ancient Chinese Text Translation and Inference *arXiv preprint arXiv 2107.03179*

[4] Chengbin Z H and Zhongbao L 2022 A Machine Translation Method for Classical Chinese Based on Semantic Information Sharing Transformer *Journal of Intelligence Engineering* **8** 6 114-127.

[5] Fugui X and Tingshao Z H 2021 Research on Construction of Classical Chinese Dictionary and Word Segmentation Techniques Based on Large-scale Corpora *Journal of Chinese Information Processing* **35** 7 6

[6] Klein G, Kim Y, Deng Y, Senellart J and Rush A M 2017 arXiv preprint arXiv 1701.02810

[7] Sutskever I, Vinyals O, and Le Q V 2014 OpenNMT: Open-Source Toolkit for Neural Machine Translation *Advances in neural information processing systems* 27

[8] Klein G, Hernandez F, Nguyen V, and Senellart J 2020 The OpenNMT neural machine translation toolkit: 2020 edition *October In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas Volume 1: Research Track* 102-109

[9] Salman A G, Heryadi Y, Abdurahman E, and Suparta W 2018 Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting[J]. Procedia Computer Science *Procedia Computer Science* **135** 89-98

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in neural information processing systems* 30

[11] Siami-Namini S, Tavakoli N and Namin A S 2019 The performance of LSTM and BiLSTM in forecasting time series *December In 2019 IEEE International conference on big data (Big Data)* 3285-3292

[12] Huang Z, Xu W, Yu K 2015 Bidirectional LSTM-CRF models for sequence tagging *arXiv preprint arXiv 1508.01991*