PROCESSING OF ENGLISH SENTENCES BY MACHINE TRANSLATION BASED ON LANGUAGE FEATURES

Wenhui LU[1], Xiaoxia ZHAI[2]

With the expansion of international communication, higher requirements have been put forward for machine translation. In this paper, a Transformer model was employed to learn the language features of English sentences. A bidirectional long short-term memory (BiLSTM) was added before the encoder to extract bottom-level language features of English sentences. A BiLSTM-Transformer model was established to process English sentences. Experiments were conducted using the collected corpora. It was found that when the two-layer BiLSTM and two-layer Transformer, a batch size of 16, and a learning rate of 0.0001 were used, the BiLSTM-Transformer model achieved the highest bilingual evaluation understudy (BLEU) score (35.01) for the test set. Compared to the recurrent neural network and Transformer models, there was a significant improvement, making English sentences more fluent and coherent. These results demonstrate the reliability of the BiLSTM-Transformer model for English sentence processing and its potential application in practical translation scenarios.

Keywords: machine translation, language feature, English sentence, BiLSTM, translation quality

1. Introduction

Under the influence of multiple factors such as economic development and technological progress, the international economic and cultural exchanges are becoming more and more frequent, and the demand for translation is also expanding. Traditional manual translation has excellent translation quality, but it has high requirements for professionals, high cost, and low efficiency, and it is increasingly unable to meet the growing demand for translation. With the development of computer technology, it has become possible to replace manual translation with machines, and machine translation has developed rapidly [1], becoming an important application of artificial intelligence [2]. Compared with manual translation, machine translation has low cost, faster translation, and simple operation. It plays a huge role in promoting cultural exchanges [3] and promoting cross-border trade [4]. However, there is still a gap between machine translation and manual translation [5]. Therefore, how to further improve the quality of machine translation has become an issue of widespread concern to researchers at

[1] Jinzhong University, Jinzhong, Shanxi, China
[2] Jinzhong University, Jinzhong, Shanxi, China, e-mail: zhaixx_zhai@outlook.com

present. Yirmibesoglu et al. [6] analyzed low-resource Turkish-English translation and studied input segmentation for verbal and non-verbal motivations. They proved the effectiveness of morphology-driven input segmentation for Turkish and the advantages of Transformer architecture in translation. Zhao et al. [7] designed a multimodal neural machine translation method with semantic image regions, integrating visual and textual features. The superiority of the proposed method was verified through experiments on a Multi30k dataset. Uzma et al. [8] proposed a multi-stack recurrent neural network (RNN) model for translation from English to Pakistan sign language and found that using the Bahdanau attention mechanism and GloVe embedding, the multi-stack RNN was able to obtain a bilingual evaluation understudy (BLEU) score of 0.83 and a word error rate of 0.17. Sharma et al. [9] proposed a method to improve translation quality by correctly translating name entities as a pre-processing step. The experiment found that the accuracy rate of this method in the translation of personal name/location name/organization name was 99.86%, 99.63% and 99.05%, respectively, with an overall accuracy of 99.52%. For the translation of English sentences, this paper designed a language feature-based method and combined the Transformer model with a bidirectional long-shot term memory (BiLSTM) to extract the bottom language features. The effectiveness of this method in improving translation quality was verified through experimental analysis, which provides a new and usable method for the actual translation of English sentences. It provides some theoretical support for improving the text processing capability of computers and promoting the progress of machine translation technology.

2. English sentence processing based on language features

2.1 Transformer model

In terms of English sentence processing, the Transformer model is a mainstream method [10], which adopts an encoder-decoder structure, as shown in Fig. 1.

Firstly, a word vector model is used to obtain word embedding in the Transformer model. Currently, the commonly used methods include word-to-vector (Word2vec) [11], Glove [12], etc. In this paper, the bidirectional encoder representation from transformers (BERT) model [13] with a good performance is selected to complete word embedding.

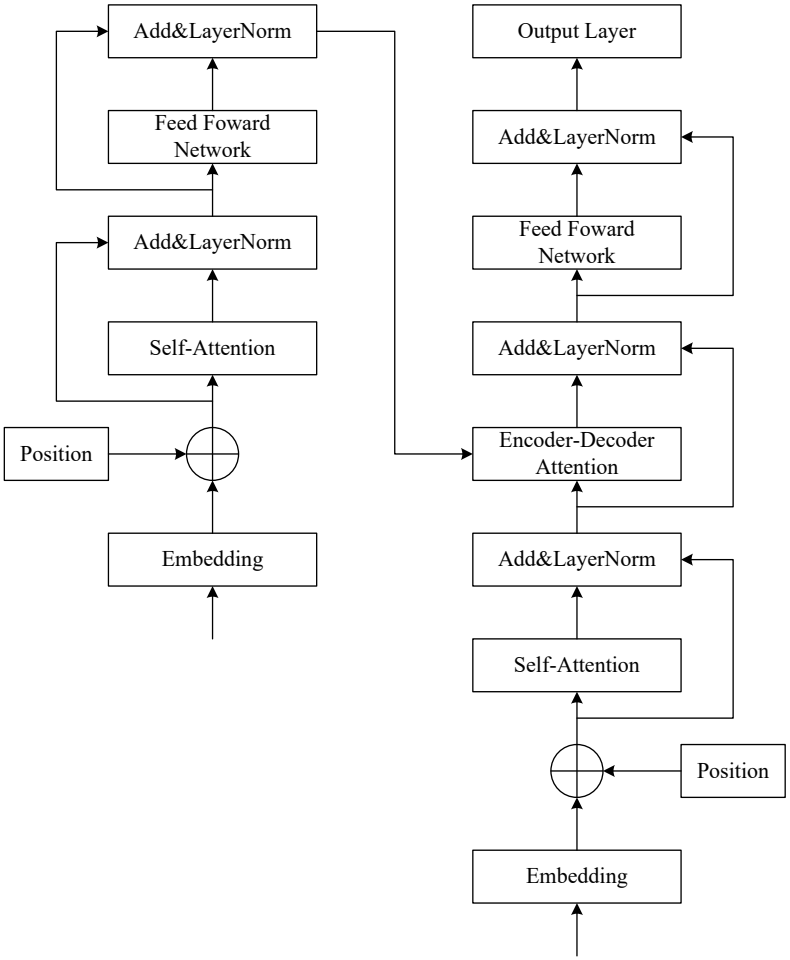

Fig. 1. Transformer model

According to Fig. 1, using the attention mechanism, the Transformer model can model the relationship between bilingual sentence pairs. The operation process is:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q, K, and V are corresponding to query, key and value respectively. The Transformer model uses the multi-head attention mechanism to capture different language features, described as:

$$MultiHaedAttention(Q,K,V) = Concat(head_1, head_2, \cdots, head_h)W^O, \quad (2)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad (3)$$

where h is the number of attention heads.

The feedforward neural network (FFN) in the Transformer model uses two linear fully connected layers and a rectified linear unit (ReLU) function to map the representation after attention calculation into the new space. Then, residual connection (Add) is used to enhance the effectiveness of information transfer. LayerNorm is used to solve the problem of training instability cause by excessively large difference between layers.

2.2 Extraction of bottom language features by BiLSTM

The Transformer model has gained good language feature extraction capability through multi-layer stacking, but the bottom language features are likely to be lost due to the increase in model depth. In order to solve this problem, this paper adds the bottom language feature extraction layer before the encoder in the Transformer model. The obtained bottom language features are transferred to the output of the top encoder. The two vectors are fused through residual connection and output to the decoder for subsequent decoding and translation.

In the selection of the bottom language features, a short term memory network (LSTM) is used [14]. LSTM is a variant of RNN, which has good applications in parameter estimation [15], data prediction [16], etc., and can capture language features in English sentences well. LSTM uses a gating mechanism to determine the forgetting and retention of information, thus alleviating the long-term dependence problem. Its hidden layer includes several memory cells, and the information to be forgotten in the previous layer is determined by the forgetting gate. The information to be reserved is determined by the input gate. Hidden state h_t is generated through the input gate.

The single-layer LSTM can only extract unidirectional language features. In this paper, BiLSTM [17] is used, which can capture language features from the forward and backward directions. Its structure is shown in Fig. 2.

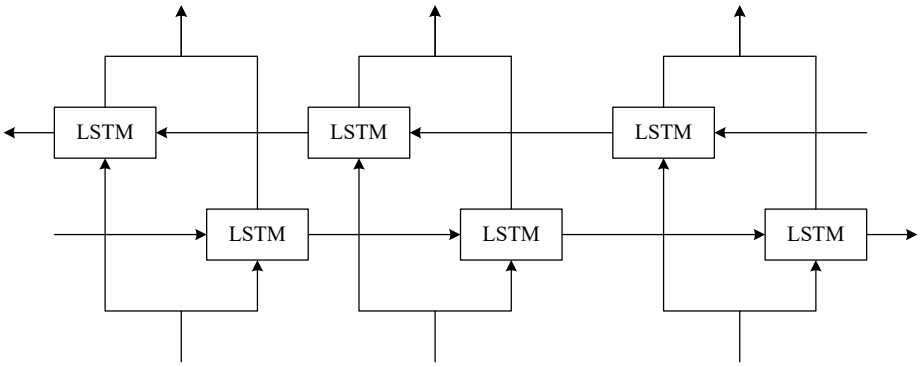

Fig. 2. BiLSTM structure

The output in both directions is concatenated to get:

$$h_t = [\vec{h}_t, \overleftarrow{h}_t], \tag{4}$$

where $\vec{h}_t$ is the forward bottom language features of English sentences at moment t and $\overleftarrow{h}_t$ is the backward bottom language features of English sentences at moment t. The concatenated h_t is input into the Transformer model to process English sentences.

3. Results and analysis

3.1 Experimental setup

The designed BiLSTM-Transformer was built using the PyTorch 1.7 deep learning framework. The programming language was Python 3.7. The specific experimental environment is shown in Table 1.

Table 1

Experimental environment1

Configuration	Parameter
Operating system	Ubuntu 16.04
Central processing unit	Intel(R) Xeon E5-2609 1.70GHz
Graphics processing unit	K40m
Memory	64 G
Hard disk	1T

The experimental data were crawled from Twitter by web crawler, and sentences with length over 100 were filtered. A total of 100,000 English-Chinese corpus information was obtained and divided into a training set and a test set in a ratio of 7:3. The English and Chinese word segmentation was performed using Spaces and Jieba. The word vector dimension was set to 512. An Adam optimizer was used. The number of attention heads was set to 8. The batch size was set to 64. The learning rate was 0.001. The rest were all default parameters.

The BLEU score [18] was used to evaluate the processing effect of English sentences, i.e., the translation quality. The calculation formula is:

$$p_n = \frac{count_{hit}}{count_{output}}, \tag{5}$$

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n \log p_n), \tag{6}$$

$$BP = \begin{cases} 1, c > r \\ exp(1 - r/c), c \le r \end{cases}, \tag{7}$$

where $count_{hit}$ is the number of n-grams from the machine translation in the reference translation, $count_{output}$ is the number of n-grams in the machine translation, BP is the penalty factor, $exp(\sum_{n=1}^{N} w_n \log p_n)$ is the weighted average

of n-grams, c and r are the length of the machine translation and the reference translation.

3.2 Analysis of results

The optimal number of BiLSTM layers and Transformer layers was determined through comparative experiments (Table 2).

Table 2

Effect of the number of layers on the translation quality of the BiLSTM-Transformer model2

BiLSTM	Transformer	BLEU score
1	1	33.36
2	1	32.87
3	1	32.01
1	2	32.87
2	2	34.75
3	2	33.05
1	3	31.45
2	3	30.77
3	3	30.12

If the number of layers of the BiLSTM and Transformer models is too small, they may not be able to fully extract features. On the other hand, if the number of layers is too large, problems such as redundancy and gradient vanishing may occur. Therefore, the optimal number of layers was determined by comparing BLEU scores under different numbers of layers. It can be seen that the number of layers of the BiLSTM and Transformer models had an impact on the processing effect of English sentences. When the number of layers of the Transformer model was 3, the BLEU score was below 32, which indicated that stacking of multiple layers can degrade the model performance. When the number of BiLSTM layers was 2 and the number of Transformer layers was also 2, the resulting BLEU score was the highest, reaching 34.75, which indicated that the BiLSTM-Transformer model was optimal under such conditions. Therefore, this structure was also adopted in the subsequent experiments.

The optimal batch size and learning rate were determined through comparative experiments (Table 3).

Table 3

Effects of batch size and learning rate on the translation quality of the BiLSTM-Transformer model3

Batch size	Learning rate	BLEU score
16	0.001	33.27
32	0.001	33.84
64	0.001	34.75
16	0.0001	35.01

32	0.0001	34.54
64	0.0001	34.17
16	0.00001	32.12
32	0.00001	31.64
64	0.00001	31.55

A smaller batch size is more suitable for low-resource data, but the gradient stability also needs to be taken into account. An excessively high initial learning rate can cause oscillations, while an excessively low one can lead to poor convergence performance. Therefore, comparative experiments were conducted under a batch size of 16 - 64 and a learning rate of 0.001 - 0.00001. It can be seen that the BLEU score also changed with the change of the batch size and learning rate. When the learning rate was 0.00001, the translation quality was worse than that when the learning rate was 0.001 and 0.0001. Specifically, when batch size = 16 and learning rate = 0.0001, the BiLSMT-Transformer model had the best processing effect for English sentences, and the BLEU score reached 35.01. Therefore, this parameter was also adopted in the subsequent experiment.

The BiLSTM-Transformer model was compared with other machine translation methods (Table 4).

Table 4

Comparison with other machine translation methods

	BLEU score	Operation time/s
RNN model	28.79	25,325.12
Transformer model	32.77	22,564.37
Collaborative model [19]	33.05	28,162.85
Lite Transformer model [20]	32.94	25,176.77
The byte-level byte pair encoding model [21]	32.95	29,642.34
BiLSTM-Transformer model	35.01	20,315.62

It can be found that the RNN-based machine translation method performed poorly on English sentence processing, with a BLEU score of only 28.79 and an operation time of 25,325.12 s. The Transformer model had a BLEU score of 32.77, which showed an increase of 3.98 compared with the RNN model, and its operation time was short. The result indicated the advantages of the Transformer model in machine translation. The collaborative model that used the collaborative multi-head attention layer, had an improved BLEU score (33.05), and its operation time was significantly improved. The Lite Transformer model used multiple attentions to calculate the global contextual information and had a compressed size; therefore, it had an improved BLEU score and a slightly extended operation time. The byte-level byte pair encoding model in literature [21] replaced the character representation with byte-level subwords. It obtained a BLEU score of 32.95, but the

operation time had a relatively significant improvement. This paper added a BiLSTM to the Transformer model to extract the bottom language features, making its BLEU score reach 35.01, which increased by 2.24 compared with the Transformer model. The results suggested the reliability of the Transformer improvement in enhancing translation quality. Its operation time was only 20,315.62 s, which suggested that it also ensured computational efficiency while improving the BLEU score. The BiLSTM-Transformer model combines the local sequence modeling ability of BiLSTM and the advantage of Transformer in capturing global dependencies. It makes up for the deficiencies of a single model in language feature extraction. Moreover, the language features extracted by BiLSTM take into account the context information, enabling Transformer to learn semantics more efficiently. As a result, it can achieve faster and better convergence and improve translation performance.

A sentence was extracted from the test set, and the processing effects of several current translation engines were compared with the proposed method. The results are as follows.

English sentences: In the face of heavy traffic during holiday peak periods, hard shoulder running is an important measure to alleviate congestion, as it can function in a short time to improve the traffic situation in bottleneck sections.

Reference translation:
面对节假日的交通高峰，路肩行驶成为一项缓解拥堵的重要措施，因为它可以在短时间内改善瓶颈路段的交通状况。

Engine B:
面对节假日高峰时段的繁忙交通，硬路肩跑是缓解拥堵的重要措施，因为它可以在短时间内改善瓶颈路段的交通状况。

Engine Y:
面对节假日交通高峰，硬肩跑路是缓解拥堵的重要措施，它可以在短时间内起到改善瓶颈路段交通状况的作用。

The BiLSTM-Transformer model: Shoulder running is an important measure to reduce congestion in the face of busy traffic during peak holiday periods, as it can improve the traffic situation in bottleneck sections in a short period of time.

The comparison of various translation results showed that both engine B and engine Y had shortcomings in translating “hard shoulder running”. The phrase was translated rather rigidly as “硬肩跑路”, which did not consider the accuracy of semantics and the specific context, leading to a poor expression. Besides this example sentence, Engines B and Y also incorrectly translated “myocardial infarction” as “心脏攻击” and “kick the bucket” as “提水桶”.

However, the result obtained by the BiLSTM-Transformer model was more similar to the reference translation, highlighting its performance in English sentence processing.

4. Conclusion

On the basis of the Transformer model, this article introduced a BiLSTM to extract the bottom language features, and the BILSTM-Transformer model was designed to process English sentences. Through experiments, it was found that when the number of layers in the BiLSTM-Transformer model was 2, the learning rate was 0.0001, and the batch size was 16, the optimal translation quality could be obtained, with the BLEU score reaching 35.01. Compared with the RNN and Transformer models, the BiLSTM-Transformer model performed better, which verified its reliability in processing English sentences. This model can be further applied in practice.

REFERENCES

[1]. *S. Shi, X. Wu, R. Su and H. Huang*, "Low-resource Neural Machine Translation: Methods and Trends", ACM T. Asian Low-Reso., **vol. 21**, 2022, pp. 1-22.

[2]. *H. Li and H. Chen*, "Human vs. AI: An Assessment of the Translation Quality Between Translators and Machine Translation", Int. J. Transl. Interp. Appl. Linguist., no. 1, 2019, pp. 43-54.

[3]. *L. Shan*, "Research on the External Communication of Chinese Excellent Traditional Culture from the Perspective of Machine Translation", J. Phys.: Conf. Ser., **vol. 1744**, no. 3, 2021, pp. 1-8.

[4]. *H. Wang and H. Wang*, "An Application System for Evaluating and Optimizing the Quality of Neural Machine Translation Corpus", 2022 IEEE International Conference on e-Business Engineering (ICEBE), 2022, pp. 178-183.

[5]. *C. A. Lester, Y. Ding, J. Li, Y. Jiang, B. Rowell and V. G. V. Vydiswaran*, "Comparing Human versus Machine Translation of Electronic Prescription Directions", J. Am. Pharm. Assoc., **vol. 61**, no. 7, 2021, pp. 484-491.

[6]. *Z. Yirmibesoglu and T. Gungor*, "Morphologically Motivated Input Variations and Data Augmentation in Turkish-English Neural Machine Translation", ACM T. Asian Low-reso., **vol. 22**, no. 3, 2023, pp. 1-31.

[7]. *Y. Zhao, M. Komachi, T. Kajiwara and C. Chu*, "Region-attentive multimodal neural machine translation", Neurocomputing, **vol. 476**, 2022, pp. 1-13.

[8]. *F. Uzma, M. R. M. Shafry and A. Adnan*, "A multi-stack RNN-based neural machine translation model for English to Pakistan sign language translation", Neural Comput. Appl., **vol. 35**, 2023, pp. 13225-13238.

[9]. *R. Sharma, P. Katyayan and N. Joshi*, "Improving the Quality of Neural Machine Translation Through Proper Translation of Name Entities", 2023 6th International Conference on Information Systems and Computer Networks (ISCON), 2023, pp. 1-4.

[10]. *A. Slim and A. Melouah*, "Low Resource Arabic Dialects Transformer Neural Machine Translation Improvement through Incremental Transfer of Shared Linguistic Features", Arab. J. Sci. Eng., **vol. 49**, no. 9, 2024, pp. 12393-12409.

[11]. *N. A. Khan, E. H. Zawad and R. M. Rahman*, "Bengali paper classification using ensemble machine learning algorithms", Int. J. Knowl. Eng. Soft Data Paradigms, **vol. 7**, no. 2, 2022, pp. 77-94.
[12]. *R. N. Modi, P. K. Kavya, R. Poddar and S. Natarajan*, "Question Classification Based onCognitive Skills ofBloom's Taxonomy Using TFPOS-IDF andGloVe", Proceedings of Emerging Trends and Technologies on Intelligent Systems, **vol. 1414**, 2023, pp. 25-37.
[13]. *S. Soffer, B. S. Glicksberg, E. Zimlichman and E. Klang*, "BERT for the Processing of Radiological Reports: An Attention-based Natural Language Processing Algorithm", Acad. Radiol., **vol. 29**, no. 4, 2022, pp. 634-635.
[14]. *W. Sun and C. Huang*, "A novel carbon price prediction model combines the secondary decomposition algorithm and the long short-term memory network", Energy, **vol. 207**, 2020, pp. 1-15.
[15]. *Z. Hui, Y. Kong, W. Yao and G. Chen*, "Aircraft parameter estimation using a stacked long short-term memory network and Levenberg-Marquardt method", Chinese J. Aeronaut., **vol. 37**, no. 2, 2024, pp. 123-136.
[16]. *R. Chen, C. Yang, L. Han, W. Wang, Y. Ma and C. Xiang*, "Power reserve predictive control strategy for hybrid electric vehicle using recognition-based long short-term memory network", J. Power Sources, **vol. 520**, no. Feb.1, 2022, pp. 1-13.
[17]. *S. Das, A. Paramane, S. Chatterjee and U. M. Rao*, "Sensing Incipient Faults in Power Transformers Using Bi-Directional Long Short-Term Memory Network", IEEE Sensor. Lett., **vol. 7**, no. 1, 2023, pp. 1-4.
[18]. *H. Cherukuri, A. Ferrari and P. Spoletini*, "Towards Explainable Formal Methods: From LTL to Natural Language with Neural Machine Translation", International Working Conference on Requirements Engineering: Foundation for Software Quality, 2022, pp. 79-86.
[19]. *J. B. Cordonnier, A. Loukas and M. Jaggi*, "Multi-Head Attention: Collaborate Instead of Concatenate", arXiv preprint arXiv: 2006.16362, 2020.
[20]. *Z. Wu, Z. Liu, J. Lin, Y. Lin and S. Han*, "Lite Transformer with Long-Short Range Attention", arXiv, 2020.
[21]. *C. Wang, K. Cho and J. Gu*, "Neural Machine Translation with Byte-Level Subwords", Proc. AAAI Conf. Artif. Intell., **vol. 34**, no. 5, 2020, pp. 9154-9160.