Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Paradigm Construction and Improvement Strategy of English Long Text Translation Based on Self-Attention Mechanism

Xiaoyan Li[1,2], Wei Chen[2,✉], Jingjing Zhang[1]

[1] *School of Education, Hefei University, Hefei, Anhui, 230616, China*
[2] *School of Foreign Languages, Bengbu University, Bengbu, Anhui, 233030, China*

ABSTRACT
The rapid development of natural language processing technology makes machine translation play an increasingly important role in cross-lingual information exchange. In this paper, we propose an English long text translation paradigm based on the self-attention mechanism and introduce various improvement strategies to enhance the model performance. The model's ability to process English long text is improved by introducing multi-head attention and hierarchical self-attention modules. The long text translation paradigm is optimized by using techniques such as residual linkage, layer normalization and dynamic memory network. A series of experiments are conducted to verify the effectiveness of the improved model on the English long text translation task. The English long text translation paradigm constructed in this paper outperforms the Transformer model and other related variants on both CPU and GPU. And Transformer outperforms this paper's model in terms of n-gram accuracy in real translation experiments. The BLEU scores of the improved model on News and other datasets are significantly improved compared with the original baseline model, which verifies the effectiveness of the improvement strategy of this paper and provides a reference for the solution of the problem of English long text translation.

Keywords: self-attention mechanism, multi-attention, residual connection, layer normalization, long text translation

1. Introduction

The term “paradigm” first appeared in the book The Structure of Scientific Revolutions, published in 1964 by the famous American philosopher Thomas Samuel Kuhn. A paradigm is a recognized example of some specific scientific practice, including laws, theories, applications, and instruments in a

✉ Corresponding author.
E-mail address: davidchenlxy7579@163.com (W. Chen).
Received 25 March 2024; Revised 05 June 2024; Accepted 15 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-497
© 2025 The Author(s). Published by Combinatorial Press. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

particular field [1-2]. It can be seen that paradigms construct guiding theories and practical methods in specific scientific fields at specific times. In the field of science, the breakthrough of paradigm leads to the revolution of science, which can make the science get a brand-new face, and the same is true in the field of literature. The study of translation paradigm is of great significance for the development of translation studies [3-4].

In the 1980s, the concept of philosophy of science was used in translation studies, and translation studies began to form its paradigm. Some scholars summarize Western translation studies into classical translation paradigm, modern translation linguistics paradigm and contemporary translation cultural integration paradigm, while translation paradigm has gone through the development process of author-centered paradigm, text-centered paradigm and translator-centered paradigm, in which translation paradigm includes philological paradigm, structuralist paradigm, deconstructionist paradigm and constructivist paradigm. Obviously, the evolution of translation history has contributed to the evolution of translation paradigms [5-7].

Business English is a branch of English for Specialized Purposes, which involves foreign propaganda, social life, production activities, business communication, laws and regulations and many other fields [8-9]. The research on business English translation is in full swing at present. Some scholars focus on the study of translation skills, some scholars focus on the study of vocabulary features and sentence features, and some scholars discuss the problem of translation standards [10-11]. However, it is understandable that many scholars cite language texts from different business activities to prove their point, but it is necessary to clarify a basic concept when discussing business English translation standards, that is, business English itself is a superordinate word (with the help of lexicological concepts), and under the category of "business English", it also includes many "subordinate words" of different styles, such as advertising English, contract English, etc. [12-15]. In this paper, we think that when discussing the translation standard, we can't use the standard of sub-category to replace the superior category, because although these texts can be collectively called business English, and there are many similarities between them, but ignoring the huge objective differences between them to arrive at the translation standard of “business English” will inevitably be biased to cover the whole picture. [16-17]. Therefore, based on the perspective of text type theory, combined with the wide range of business English texts, the formal degree of the genre spanning a large range of characteristics, according to the characteristics of its subordinate text types, specific issues and specific analysis, the implementation of diversified standards [18-19].

Castilho, S et al. articulated the neural network machine translation strategy as a new paradigm in the field of machine translation and carried out a study to compare the translation quality of NMT systems and statistical MT, concluding that NMT systems have pushed machine translation forward [20]. Toral, A et al. examined the effectiveness of neural network machine translation practiced on various different texts, and by comparing it for the MT-dominated PBSMT translation paradigm, it was confirmed that the quality and accuracy of NMT's translations were significantly improved [21]. Dabre, R et al. reviewed the research related to neurotranslation strategies and found that many researchers have facilitated translation quality through multilingual parallel corpora, while various specific methods of MNMT were classified and analyzed to identify the advantages and disadvantages of these techniques [22]. Bowker, L reviewed the related journal papers on AI translation tools assisting English writing and combined with a pilot practice at a university, argued the feasibility and effectiveness of machine translation to promote the learning of English translation knowledge, and concluded that the future direction of the future development of the AI translation machine tools is the translation and writing skills and so on [23]. Rishita, M. V. S et al. introduced machine translation tools and focused on analyzing the problem of building deep neural networks as part of an end-to-end translation pipeline, which includes preprocessing, model building, and running the model on English text, deepening the knowledge of machine translation [24]. In summary, the above is a strategy for English text translation from the perspective of neural network machine translation technology and tools.

Hkiri, E. et al. explored the problems related to named entities in the translation process and concluded that improper translation of NER can lead to a significant decrease in sentence readability, so they tried to utilize machine learning strategies and rule-based techniques in order to deal with the problem of NER in translating English from Arabic, and obtained good results [25]. Polyakova, L. S et al. discussed the structural and semantic characteristics of English terminology of texts in the field of science and technology, as well as methods of translation from Russian, noting that the development of science and technology constantly creates a new vocabulary of scientific and technical terms, and that there are difficulties in the process of translating English terminology [26]. Sandra, R. A based on the critical review methodology examines the need to optimize the translation of English source text into Indonesian language by focusing on the vocabulary, concepts of the source text and considering the cultural context of the translated language in order to improve the readability of the language translation [27]. The above studies have been studied and discussed around the optimization of English translation from the perspective of cultural context readability of the language translated from English, naming of objects, and structural and semantic features of terminology.

Aiming at the problems of loss of contextual information and difficulty in capturing long-term dependencies faced by traditional sequence-to-sequence models in processing long texts, this paper proposes a translation paradigm based on the self-attention mechanism for English long texts to enhance the model's ability to process long texts. The encoder part of the translation paradigm constructed in this paper adopts a multi-layer self-attention mechanism, and each layer includes multiple attention heads to capture different levels of linguistic features. The decoder part combines the self-attention mechanism and the multi-head attention mechanism to ensure that the model makes full use of the source sequence information in the process of generating the target sequence. WMT14 and WMT18 are selected as the datasets, and BLEU is used as the evaluation index to compare and analyze the performance of this paper's model in theoretical and practical translation. To address the problems of the model in the experiments, this paper proposes improvement strategies such as enhancing the self-attention mechanism, introducing residual connectivity and layer normalization, and dynamic memory network, and selects the publicly available English long text dataset for further experiments. Based on the experimental results, the effectiveness of the improvement strategies in this paper and the good performance of the model in English long text translation are illustrated.

2. Relevant theories and technologies

2.1. Overview of machine translation

Machine translation is a technology that uses computers to simulate the human translation process to achieve translation between different languages, i.e., the conversion of one natural language to another natural language [28]. The natural language being converted is called the source utterance and the converted natural language is called the target utterance. Although the process and concepts are simple and easy to understand, the scientific fields involved are cutting-edge and complex, including artificial intelligence, big data processing, and linguistics. Therefore, machine translation is an important branch of deep learning, as well as one of the important research areas of natural language processing.

Nowadays, neural network-based machine translation models make the translation level almost comparable to human translation, in which the most mainstream method is to use neural networks to directly map the source utterance to the target utterance, which is also known as the end-to-end neural network machine translation model, but there are drawbacks to this model, and there is a loss of contextual information when dealing with long texts. The problems of translation selection, introduction of knowledge, interpretability, and part-of-speech translation make the machine translation still not in line with contemporary translation needs, and the machine translation model needs to be further improved.

2.2. *Self-attention mechanisms*

Self-attention mechanism is an approach that mimics the human attention mechanism for processing information in the field of machine learning and artificial intelligence. It is based on the principle of how the human brain works by selectively focusing attention to process specific information, thus improving the efficiency and accuracy of information processing. The principle of self-attention mechanism for text processing in the field of NLP is that the input word embedding vector is solved for its query matrix (Q), key matrix (K), and value matrix (V) respectively. The similarity is obtained by dot product method, as well as processing such as normalizing the exponential function and dividing by the square root of the dimension in order to get different weights and attention distributions, so as to get the input information that has a higher weight of influence on the current output information, so that the model can focus its attention on inputs that have an important influence on the next target output. The development of the self-attention mechanism can be traced back to research in the field of neuroscience, especially to the human visual system. In the visual system, human attention can be focused on the target of interest, ignoring other irrelevant information. This selective attention mechanism allows humans to process complex visual information efficiently.

2.3. *Sequence-to-sequence modeling*

The Seq2Seq model is a new end-to-end mapping approach with the structure shown in Fig. 1, which centers on the encoder-decoder architecture [29]. The encoder is responsible for converting an input sequence of variable length into a context vector of fixed shape and encoding the input sequence information in that context variable, which can be viewed as the semantic vector of this sequence C. The decoder is responsible for generating the specified sequence based on the semantic vector C, a process also known as decoding.

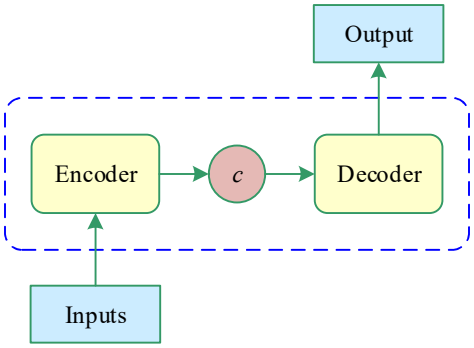


Fig. 1. Seq2Seq model

If the input sequence is $\{X_1, X_2, \cdots, X_m\}$ and X_m is the m rd marker of the input sequence, at time step m, the recurrent neural network converts X_m and the hidden state h_{m-1} of the previous time step into the current hidden state h_m through function f as shown in equation (1). And then the encoder converts the hidden states of all time steps into semantic vector C through function q, as shown in equation (2). If the output sequence of the training set is $\{Y_1, Y_2, \cdots, Y_t\}$, for each time step t, the conditional probability distribution P of the decoder output Y_t depends on the subsequence $\{Y_1, Y_2, \cdots, Y_{t-1}\}$ of the previous output and the semantic vector C, as shown in equation (3):

$$h_m = f\left(X_m, h_{m-1}\right) \tag{1}$$

$$C = q\left(h_1, h_2, \cdots, h_m\right) \tag{2}$$

$$P\left(Y_t \middle| Y_1, Y_2, \cdots, Y_{t-1}, C\right) \tag{3}$$

Seq2Seq also uses a recurrent neural network as a decoder to model the conditional probability distribution of Y_t. At any time step t on the output sequence, the recurrent neural network takes as inputs the outputs Y_{t-1} and semantic vectors C from the previous time step, and then transforms these inputs from the current time step t with the previous hidden state h_{t-1} into a new hidden state h_t through a function g as shown in Eq. (4), with a function g to represent the decoder hidden layer transformation:

$$h_t = g\left(Y_{t-1}, h_{t-1}, C\right) \tag{4}$$

Seq2Seq usually uses GRU or Long Short-Term Memory Network as the encoder and decoder, and the convergence of GRU is significantly better than LSTM under different tasks, so in this paper, GRU is used as the encoder and decoder of Seq2Seq, and its unit structure is shown in Fig. 2.

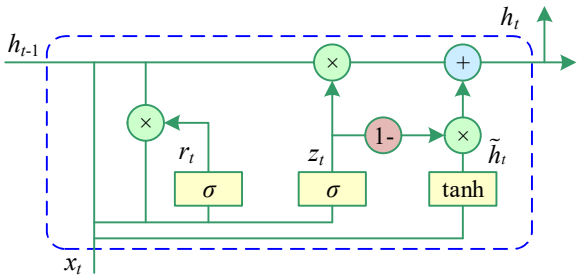


Fig. 2. GRU cell structure

In Fig. 2, z_i is the update gate of GRU, which is used to control how much information the current h_t state needs to retain from the previous moment h_{t-1} state and how much information it can receive from the candidate state $\tilde{h}_t$; the reset gate r is used to control whether the calculation of the candidate state $\tilde{h}_t$ is dependent on the previous moment state h_{t-1}; σ is the activation function; b is the bias value; the corresponding weight matrices of the time-step parameters are W_z, W_t, and W_h; and the corresponding parameters of the hidden state are U_z, U_t, U_h The calculation process is shown in Eq. (5) to Eq. (8):

$$z_t = \sigma\left(W_z X_t + U_z h_{t-1} + b_z\right) \tag{5}$$

$$r_t = \sigma\left(W_t X_t + U_t h_{t-1} + b_t\right) \tag{6}$$

$$\tilde{h}_t = \tanh\left(W_h X_t + U_h\left(r_t \square\ h_{t-1}\right) + b_h\right) \tag{7}$$

$$h_t = z_t \square\ h_{t-1} + \left(1 - z_t\right) \square\ \tilde{h}_t \tag{8}$$

The application of Seq2Seq model in time series forecasting is to take the historical observation sequences as inputs and generate the prediction sequences for a future period as outputs. The model captures the dependencies and patterns between sequences through the encoder and decoder structure, and is able to effectively learn the evolutionary trend of sequences.The Seq2Seq model can adapt to input and output sequences of different lengths, and has the capability of being able to deal with sequences of variable lengths, which makes it suitable for a variety of time series prediction tasks. However, the Seq2Seq model based on traditional recurrent neural networks still retains the shortcomings of recurrent neural networks, and the use of the attention mechanism in the decoding process of the model can effectively increase the ability of the model to extract the dependencies of long sequences.

2.4. Challenges and opportunities in translating long texts

Long text translation faces more challenges compared to short text translation. First, long texts contain more contextual information and complex logical structures, which puts higher requirements on the modeling comprehension ability. Second, the long-term dependencies in long texts are more complex and require stronger modeling capabilities to capture them accurately. In addition, long text translation needs to cope with the consistency problem. That is, maintaining terminology and style consistency throughout the translation process. However, with the increased availability of big data and computational resources, and the continuous advancement of deep learning techniques at the first level, new opportunities are opening up for long text translation. The quality of long text translation can be significantly improved by constructing more effective model structures and adopting advanced training strategies. The long text translation paradigm with opportunity self-attention mechanism proposed in this paper formally arises in this context, aiming to improve the performance of long text translation by improving the self-attention mechanism and optimizing the model architecture.

3. Paradigm construction of translation of long text based on self-attention mechanism

3.1. Characteristics and requirements for translating long texts

Long text translation has unique characteristics and requirements compared to short text translation. Long texts usually involve more contextual information and complex logical structures, which require translation models not only to have strong comprehension ability, but also to effectively capture long-term dependencies. In addition, long text translation needs to maintain the consistency of the overall style and terminology to avoid inconsistencies or sudden changes in style in long content. Therefore, long text translation not only requires the model to have strong language comprehension ability, but also needs to have good memory and consistency maintenance mechanism.

3.2. Application of Self-Attention Mechanism in Long Text Translation

Although the Seq2Seq model is already very much in line with the idea of human translation behavior, i.e., reading a sentence in the original language, understanding its meaning first, and then translating it into a sentence in the target language. But Seq2Seq still has two drawbacks: one is that the encoder has difficulty in compressing a long sentence into a fixed-dimension vector. The second is that the decoder also forgets the information in the output vector encoded by the encoder when decoding, this is because for every word translated in front of it during decoding, the information of the word will be passed into the next time step as an input along with the encoder output vector, at this time, there will be less and less information in the encoder output vector before, so the later the time step, the worse the translation will be. Therefore, when neural machine translation is performed, often the first half of the sentence is translated well, and the second half is translated poorly.

In order to solve the problems of Seq2Seq model, related scholars proposed a self-attention mechanism and used it in Seq2Seq model. The main idea introduced by the self-attention mechanism is to mimic the behavior of human beings in translating long texts, i.e., at each step of decoding, not only to combine the fixed-size vectors encoded by the encoder (read-through the whole text), but also to check back to the original each word (precision localization), both of which work together to determine the output of the current step [30]. The structure of the self-attention mechanism is shown in Figure 3.

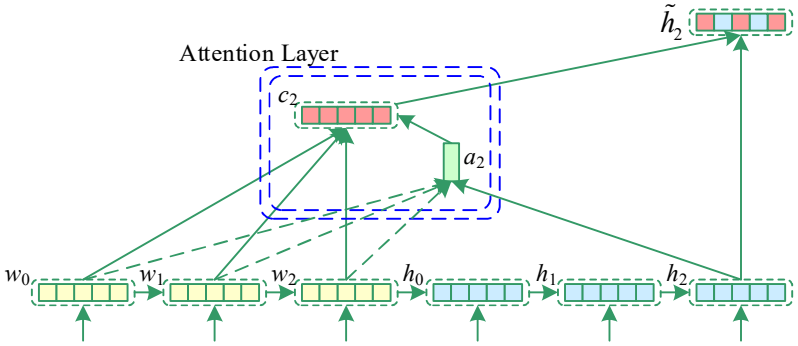


Fig. 3. Self-attention mechanism

The self-attention mechanism used in this paper does not require the use of two-way GRU to perform the calculation of global attention, and this self-attention mechanism provides is the kind of algorithms to calculate the weight of attention, all of them are simpler to calculate than the other self-attention mechanisms, while the results are not much different, the three calculations are shown in Equation (9):

$$score\left(h_i,\bar{h}_s\right)=\begin{cases} h_i\cdot\bar{h}_s & dot \\ h_i\cdot W_a\bar{h}_s & general \\ v_a\cdot\tanh\left(W_a\left[h_i;\bar{h}_s\right]\right) & concat \end{cases} \tag{9}$$

where h_t denotes the current time-step hiding state, $\bar{h}_s$ denotes the source language hiding state, and both W_a and V_a are the parameter matrices to be learned.

3.3. Architectural Design of Long Text Translation Paradigm

In order to adapt to the characteristics and requirements of long text translation, this paper designs a novel long text translation architecture. The architecture introduces a self-attention mechanism and a multi-level attention module on the basis of the traditional Seq2Seq model to enhance the model's ability to process long texts. The specific architecture is shown in Fig. 4. The encoder part employs a multi-layered self-attention mechanism, each layer including multiple attention heads to capture linguistic features at different levels. The decoder part, on the other hand, combines the self-attention mechanism and the multi-head attention mechanism to ensure that the information of the source sequence is fully utilized when generating the target sequence.

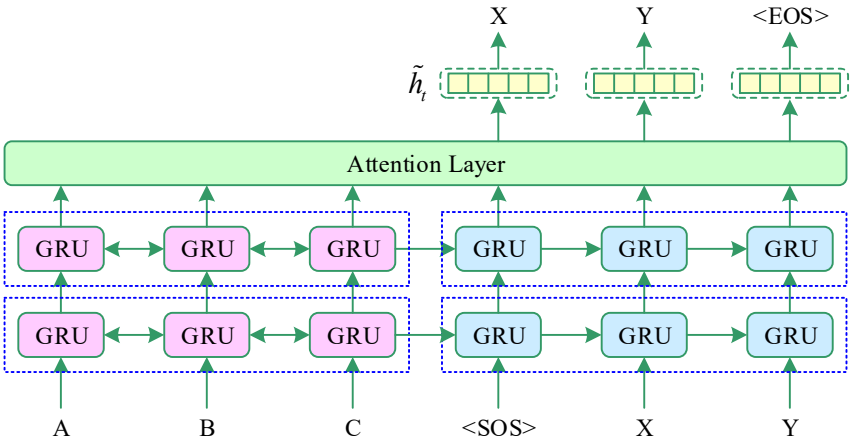


Fig. 4. The paradigm of the translation paradigm

3.3.1. Encoder. The encoder is responsible for converting the input sequence into a fixed length vector representation. Assuming that the input sequence is $X = \{x_1, x_2, \cdots, x_n\}$, the encoder is given by:

$$h_i = SA(h_{i-1}, x_i) \tag{10}$$

where SA denotes the self-attention mechanism function and h_i denotes the i rd hidden state vector.

3.3.2. Multilayer Self-Attention Module. The multi-layer self-attention module consists of multiple self-attention layers, each containing multiple attention heads. The formula for layer 1 is:

$$H_1 = MultiHead(H_{1-1}, Q_1, K_1, V_1) \tag{11}$$

where $MultiHead$ denotes the multi-head attention function and Q_1, K_1, V_1 denotes the query vector, key vector and value vector of layer 1, respectively.

3.3.3. Decoders. The decoder is responsible for generating the target sequence based on the output of the encoder. Each layer of the decoder equally contains self-attention mechanism and multi-attention mechanism. The formula for layer j is:

$$y_j = Decoder(y_{j-1}, H_n, Q_j, K_j, V_j) \tag{12}$$

where $Decoder$ denotes the decoder function and H_n denotes the final output vector of the encoder.

3.3.4. Loss function. In order to optimize the performance of the model, the cross-entropy loss function is used in this paper. The loss function is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log P(y_i \mid x_i) \tag{13}$$

where $P(y_i \mid x_i)$ denotes the probability of target output y_i given input x_i.

Through the above algorithmic process and formula derivation, it can be seen that the proposed long text translation paradigm is able to effectively capture the complex structures and long-term dependencies in long texts, thus improving translation accuracy and fluency.

3.4. Experimental analysis

3.4.1. Introduction to the experimental data set. The WMT14 English-German bilingual dataset focuses on translation between English and German and provides a large parallel corpus about long texts from a variety of sources, including news, websites, and books, etc. The WMT18 dataset is an important dataset used for research on machine translation, which was released by the WMT series of conferences organized by the International Natural Language Processing Community in 2018. This dataset contains a parallel corpus of long texts in multiple language pairs and is widely used to evaluate and compare different machine translation systems.

To facilitate comparison with other models, this paper will use the WMT14 English-German bilingual dataset for BLEU and efficiency comparison. To facilitate checking the translation quality, this paper will use the Chinese-English language in WMT18 to train and test the dataset.

3.4.2. Evaluation indicators. BLEU is a method for measuring the quality of machine translation, which assesses the quality of translation mainly by comparing the agreement between machine translation and human translated text. The underlying assumption of the method is that the better the

performance of a machine translation is if its results more closely resemble those of a professional translation.

BLEU is calculated by comparing the translated text (usually sentences) with a set of high-quality reference translations and assigning scores to each text segment. These scores are then averaged over the entire text set to estimate the overall translation quality. The formula for the calculation is as follows:

$$BLEU = BP \cdot \exp\left(\sum_{i=1}^{N} w_n \log P_n\right) \tag{14}$$

where N is the upper limit of the value of n-grams (segments consisting of n words), which is taken as 4, i.e., only 4-grams are counted at most. w_n is the weight of the n-grams, and usually the same weight is taken for each n-gram. BP is the shortness penalty factor, which prevents the machine-translated sentences from being too short instead of resulting in higher scores:

$$BP = \begin{cases} 1, & if L_{MT} > L_{ref} \\ \exp\left(1 - \dfrac{L_{MT}}{L_{ref}}\right) & if L_{MT} \le L_{ref} \end{cases} \tag{15}$$

P_n for n-gram based accuracy:

$$P_n = \frac{\sum_{n-gram \in MT} Counter\ Clip(n-gram)}{\sum_{n-gram \in MT} Counter(n-gram)} \tag{16}$$

where Counter Clip is to count the number of n-grams appearing in the reference translation and Counter is to count the number of n-grams in the machine translated text.

In this paper, the metric used to evaluate translation quality is SacreBLEU, which is an important tool for machine translation evaluation, aiming to provide a more standardized and reproducible method for calculating BLEU scores.SacreBLEU solves many problems encountered when calculating BLEU scores, such as the lack of space as a word segmentation means in comparing Chinese translations, which makes conventional BLEU difficult to be used for evaluating Chinese translations. For example, when comparing Chinese translations, due to the lack of spaces as a means of word separation, it is difficult to use conventional BLEU to assess the quality of Chinese translations.

3.4.3. Reference models. In this paper, we will use the Anglo-German and Anglo-Chinese models in Opus as the baseline models, both of which use the standard Transformer architecture, and the model parameters are shown in Table 1. Among them, the Anglo-German model achieved a BLEU score of 28.31 on the WMT14 dataset and the Anglo-Chinese model achieved a BLEU score of 26.78 on the WMT18 dataset.

Table 1. Opus model parameters

Parameter name	Parameter value
Encoder layer number	8
Decoder layer number	8
Attention number	10
Characteristic dimension	508
Feedforward dimension	2142

3.4.4. Comparison of results. The experiments in this section are designed to verify whether the model in this paper can improve the translation speed of English long text while ensuring the quality of the translated text. The hardware and software used for the experiment are CPU: AMD Ryzen(TM) 55600, GPU: NVIDIA GeForce RTX 3090Ti, OS: Ubuntu22.04, Python: 3.10.12, PyTorch: 2.0.0, CUDA: 11.8.

1) Experimental data preprocessing. In order to train the models designed in this chapter, the training, validation and test sets of WMT14 and WMT18 were selected. For WMT18, only the comment section of news reports was selected. The following data cleaning measures are applied to the above datasets:

(1) Remove statement pairs containing HTML tags.

(2) Remove statement pairs containing the same target text.

(3) Remove statement pairs with the same language of source and target statements.

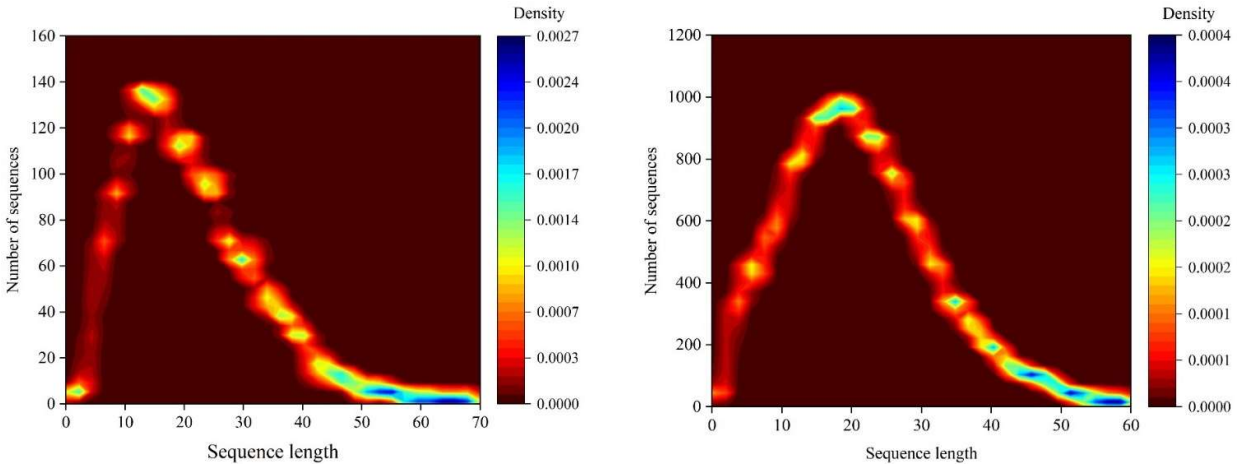
(4) For the utterance pairs after disambiguation, if the sequence length ratio is less than 1:5 or greater than 5:1, the utterance pair is discarded.

After the above processing, the utterance pairs of each collection of the final corpus are shown in Table 2.

Table 2. The logarithmic amount of the statement of the data set after washing

Data set	Training set	Verification set	Test set
WMT14	4326984	3046	3218
WMT18	178632	51378	27694

where the sequence length distribution of the test set for both is shown in Fig. 5, and (a) and (b) denote the WMT14 dataset and WMT18 dataset, respectively.



(a) WMT14 (b) WMT18

Fig. 5. The sequence length distribution of the test set

2) Model Theory Performance Experiments. In order to compare the models' English long text translation ability, this subsection will initialize the standard Transformer and its various variants and the long text translation paradigm proposed in this paper with the parameters in Table 1 and initialize them with the parameter weights of these models, respectively. Afterwards, the same sequences are input to these models, and sequences of equal length to the input sequences are generated in an autoregressive manner using a greedy search strategy, and the time taken to generate the sequences is recorded. For each length of sequence, each model will be generated 7 times and the final result will

be the average of the 7 times. To test the inference performance of the models on CPU and GPU, one experiment will be done on CPU and one on GPU.

The results of the experiments on GPU are shown in Fig. 6, where the horizontal and vertical coordinates are logarithmic axes. When using GPU for inference, it can be seen that the model in this paper is basically comparable to the standard Transformer in terms of new element generation speed when the length of the sequence is short, but it shows in addition to a faster advantage in long text. The translation time of this paper's model is basically linear with the text length. In the text length of about 450, the standard Transformer in the translation time appeared to rise significantly, consuming time from the more obvious quadratic upward trend. When the text length is 500 and above, the standard Transformer has exceeded the upper limit of video memory and cannot complete the test in this experimental environment. Compared to other linear complexity Transformer variants, the model in this paper also shows an advantage, in most cases can achieve better performance.

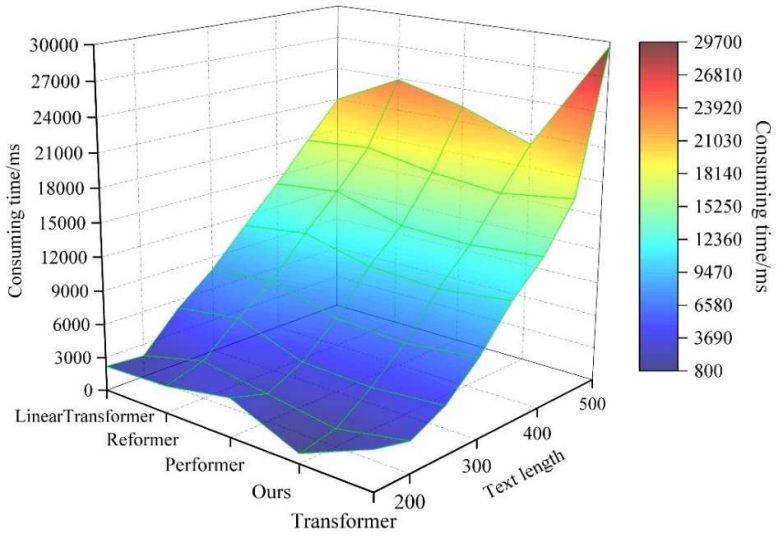


Fig. 6. Model theoretical ability experiment results (GPU)

The results of the experiments on CPU are shown in Fig. 7. When reasoning with CPU, thanks to the introduction of the self-attention mechanism, the model in this paper is especially suitable for CPU reasoning and can achieve better performance under all lengths of text experiments. Especially for long texts, the translation time of this paper's model is proportional to the text length, while the Transformer is proportional to the square of the text length. Compared to other variants, the model in this paper is able to achieve better performance in experiments with all input text lengths.

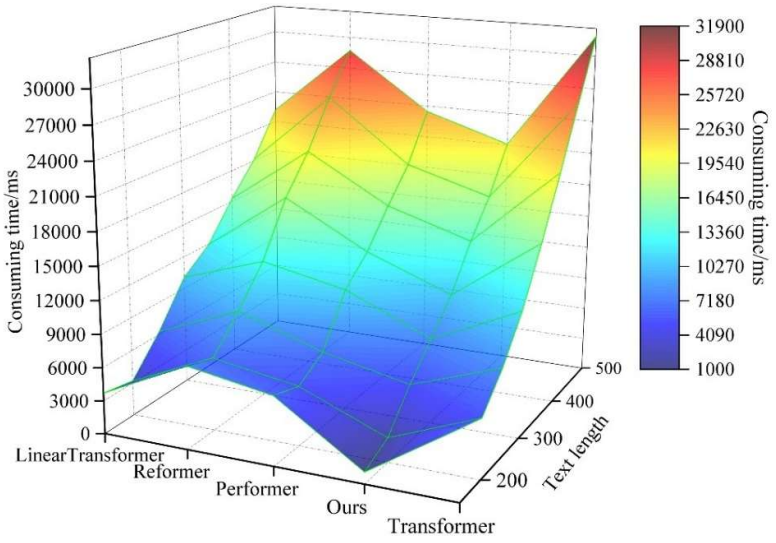


Fig. 7. Model theoretical ability experiment results (CPU)

3) Model actual translation experiment. The trained model of this paper and Opus English-Chinese model are translated on the test set, and the BLEU scores are calculated and the translation time is recorded. The final results are shown in Table 3. The results shown in Table 3 are contrary to the theoretical performance experimental results, based on the components of BLEU as shown in Table 4.Transformer still outperforms this paper's model in n-gram accuracy, and its words are closer to the reference translation compared to this paper's model. However, on this test machine, the length of the text it translates is on the short side, resulting in a lower BP, making the final score lower than that of this paper's model. Similarly, since the translated text translated by this paper's model is longer in length, this makes the final time longer.

Table 3. Translation results on WMT18

	BLUE	Time/ms
Transformer	25.34	30584.74
Ours	29.71	40196.83

Table 4. The various components of BLEU score

	1-gram	2-gram	3-gram	4-gram	BP	Machine translation length	Reference length
Transformer	64.83	42.54	28.49	20.36	0.71	709643	1036748
Ours	52.74	33.87	21.58	15.48	0.98	998746	

4. Improvement strategies and experimental analysis

4.1. Improvement strategies

In order to further enhance the performance of the English long text translation paradigm based on the self-attention mechanism, the following improvement strategies are proposed in this paper:

1) Enhancing the self-attention mechanism. By increasing the depth and width of the attention heads, the model's ability to capture long-term dependencies is improved. Specifically, this paper extends the multi-head attention mechanism into a hierarchical multi-head attention mechanism, which enables the model to focus on different levels of linguistic features at the same time.

2) Introducing residual connectivity and layer normalization. In order to improve the stability and training efficiency of the model, this paper introduces residual connectivity and layer normalization techniques in the encoder and decoder. Residual concatenation allows the signal to run directly through multiple layers, reducing the risk of gradient vanishing. Layer normalization, on the other hand, helps to accelerate the training process and improve the convergence speed of the model.

3) Dynamic Memory Networks. In order to solve the consistency problem in long text translation, this paper introduces a dynamic memory network. The network can store and update key information to ensure consistency and coherence throughout the translation process.

4.2. Experimentation and Analysis

4.2.1. Data sets. In this section of the experiments the commonly used Anglo-German dataset in the domain is used for validation, the processing of the dataset used is kept consistent with the previous work to allow for fair comparisons, and Moses is used for the segmentation process and OpenNMT is used for the BPE segmentation process. The statistics for the corpus used are shown in Table 5. The experiments in this section use the same hyperparameters as the previous work in the model training session for fair comparison.

Table 5. Data set statistics

Data set	Training set size	Validation set size	Test set size
News	10000	1000	1000
TED	50000	5000	5000
Novels	80000	8000	8000

4.2.2. Training parameters. The experiments used hyperparameters consistent with the previous work in the model training session for fair comparison. Where batch size was set to 5000 tokens, the hidden layer unit of the multi-head attention mechanism was set to 1048, the feed-forward neural network layer was set to 1048, the embedding dimension was set to 256, and the number of heads was set to 8. For the TED and Novels datasets due to their small size, the dropout ratio was set to 0.1, and for the News dataset, the dropout was set to 0.05. The top module in the unified encoder was set to 7 layers and the bottom module was designed to 2 layers, for a total of 5 layers in the encoder. In order to balance the training overhead and model performance, the context information used is limited to the previous and next sentences of the current statement to be translated. The model was trained using the Adam optimizer with the parameters set to $\beta_1 = 0.95, \beta_2 = 0.97, \varepsilon = 1 \times 10^{-7}$, and the learning rate was set to increase linearly from 0 to 3 × 10^{-4} for 5000 steps at the beginning of model training, and then decrease to inversely proportional to the square root. Cross moisture with label smoothing was used as the loss function, and the smoothing rate was set to 0.3. The model was trained until 20 consecutive epochs when the loss function no longer decreased.

4.2.3. Experimental results. In this experiment, the experimental results were compared with the initially constructed baseline model of the English long text translation paradigm the experimental results are shown in Table 6. From the table, it can be seen that the enhanced self-attention mechanism, the introduction of residual connectivity and layer normalization, and the dynamic memory network can significantly improve the BLEU scores, verifying the effectiveness of the improvement strategies.

Table 6. BLEU fraction comparison

Model name	News	TED	Novels
Baseline model	28.5	26.3	27.8
Enhanced self-focus model	30.2	28.9	29.4
Add the residual connection and the layer normalization model	31.7	30.1	30.8
Dynamic memory network model	32.4	31.5	31.9

5. Conclusion

This paper introduces the self-attention mechanism on the basis of the traditional Seq2Seq model to construct a baseline English long text translation paradigm. Various improvement strategies are proposed to optimize the model according to the problems of the model in the experiments to improve the model's performance in translating English long texts. The model before optimization outperforms the standard Transformer model and other related variants in the theoretical experiments. However, in the actual translation experiments, the translation time of the model is 40196.83ms, which is 14.29% more than the standard Transformer model, and it is also inferior to the standard Transformer model in terms of n-gram accuracy. The BLEU scores of the optimized model on the News, TED and Novels datasets are 30.2, 31.7, 32.4, 28.9, 30.1, 31.5, and 29.4, 30.8, 31.9, which are significantly improved compared to those of the pre-optimization baseline model, indicating that the optimization strategy used in this paper is able to effectively overcome the deficiencies of the baseline model in the actual translation of English long texts. This indicates that the optimization strategy used in this paper can effectively overcome the shortcomings of the baseline model in the actual translation of English long

text, and provide strong technical support for solving the problem of English long text translation.

Funding

This work was supported by Anhui Philosophy and Social Sciences Planning Project: Comparative Study of the Chinese Translation of the Western Philosophical Classic "The Enneads" (AHSKY2020D83).

References

[1] Esteban, M., Freire, E., Ponce, E., & Torres, F. (2022). On normal forms and return maps for pseudo-focus points. Journal of Mathematical Analysis and Applications, 507(1), 125774.

[2] Knoblauch, H. (2018). Conclusion: The Social Construction of Reality as a paradigm? 1. In Social Constructivism as Paradigm? (pp. 325-338). Routledge.

[3] Kardiansyah, M. Y., & Salam, A. (2021, March). Reassuring feasibility of using Bourdieusian sociocultural paradigm for literary translation study. In Ninth International Conference on Language and Arts (ICLA 2020) (pp. 135-139). Atlantis Press.

[4] Mustafa, A., & Mantiuk, R. K. (2020). Transformation consistency regularization–a semi-supervised paradigm for image-to-image translation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16 (pp. 599-615). Springer International Publishing.

[5] Poibeau, T. (2017). Machine translation. MIT Press.

[6] Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In Extended abstracts of the 2021 CHI conference on human factors in computing systems (pp. 1-7).

[7] Laviosa, S. (2021). Corpus-based translation studies: Theory, findings, applications (Vol. 17). Brill.

[8] Indurthi, S., Han, H., Lakumarapu, N. K., Lee, B., Chung, I., Kim, S., & Kim, C. (2020, May). End-end speech-to-text translation with modality agnostic meta-learning. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7904-7908). IEEE.

[9] Botirova, P., & Sobirova, R. (2019). FEATURES OF THE TRANSLATION OF POETRY INTO ENGLISH. Theoretical & Applied Science, (6), 383-387.

[10] Song, X. (2021). Intelligent English translation system based on evolutionary multi-objective optimization algorithm. Journal of Intelligent & Fuzzy Systems, 40(4), 6327-6337.

[11] Fantham, E. (2019). Seneca's Troades: A literary introduction with text, translation and commentary (Vol. 5385). Princeton University Press.

[12] Taxirovna, A. S. (2023). Lingua-cultural aspects of the translation of English and Uzbek stories. Current Issues of Bio Economics and Digitalization in the Sustainable Development of Regions (Germany), 536-540.

[13] Fatima, N., Durdona, I., & Mashkhura, J. (2019). Theoretical foundations of the translation of scientific literature from Russian into English. ACADEMICIA: An International Multidisciplinary Research Journal, 9(4), 112-116.

[14] Hijjo, N. F., & Kadhim, K. A. (2017). The analysis of grammatical shift in English-Arabic translation of BBC media news text. Language in India, 17(10), 79-104.

[15] Ahmed, M. A., Baharin, H., & Nohuddin, P. E. (2023). k-means variations analysis for translation of English

Tafseer Al-Quran text. Int. J. Electr. Comput. Eng, 13(3), 3255.

[16] Pham, A. T., Nguyen, L. T. D., & Pham, V. T. T. (2022). English language students' perspectives on the difficulties in translation: Implications for language education. Journal of Language and Linguistic Studies, 18(1), 180-189.

[17] Ma, C., Zhang, Y., Tu, M., Han, X., Wu, L., Zhao, Y., & Zhou, Y. (2022, August). Improving end-to-end text image translation from the auxiliary text translation task. In 2022 26th International Conference on Pattern Recognition (ICPR) (pp. 1664-1670). IEEE.

[18] Turg'unova, F. R. (2022). THE TITLE OF THE ARTICLE IN ENGLISH AS THE SUBJECT OF RESEARCH. British Journal of Global Ecology and Sustainable Development, 10, 174-177.

[19] Nasution, D. K., & Syahputra, R. (2020, October). The impact of the translation techniques and ideologies on the quality of the translated text of Mantra Jamuan Laut from Malay language into English. In Proceeding International Conference on Language and Literature (IC2LC) (pp. 165-171).

[20] Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art?. The Prague Bulletin of Mathematical Linguistics, (108).

[21] Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text?. Translation quality assessment: From principles to practice, 263-287.

[22] Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR), 53(5), 1-38.

[23] Bowker, L. (2020). Chinese speakers' use of machine translation as an aid for scholarly writing in English: a review of the literature and a report on a pilot workshop on machine translation literacy. Asia Pacific Translation and Intercultural Studies, 7(3), 288-298.

[24] Rishita, M. V. S., Raju, M. A., & Harris, T. A. (2019). Machine translation using natural language processing. In MATEC Web of Conferences (Vol. 277, p. 02004). EDP Sciences.

[25] Hkiri, E., Mallat, S., & Zrigui, M. (2017, November). Arabic-English text translation leveraging hybrid NER. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (pp. 124-131).

[26] Polyakova, L. S., Yuzakova, Y. V., Suvorova, E. V., & Zharova, K. E. (2019, March). Peculiarities of translation of English technical terms. In IOP Conference Series: Materials Science and Engineering (Vol. 483, No. 1, p. 012085). IOP Publishing.

[27] Sandra, R. A. (2018). From English to Indonesia: Translation problems and strategies of EFL student teachers-A Literature Review. International Journal of Language Teaching and Education, 2(1), 13-18.

[28] Xinran Chen,Sufeng Duan & Gongshen Liu. (2025). Improving semi-autoregressive machine translation with the guidance of syntactic dependency parsing structure. Neurocomputing128828-128828.

[29] Xiaolong Wang & Yang Li. (2024). Prediction of mine water quality by the Seq2Seq model based on attention mechanism. Heliyon(18),e37916-e37916.

[30] Yeongjoon Kim,Sunkyu Kwon,Donggoo Kang,Hyunmin Lee & Joonki Paik. (2025). Enhancing video frame interpolation with region of motion loss and self-attention mechanisms: A dual approach to address large, nonlinear motions. Neurocomputing128728-128728.