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

Combinatorial Press

A Study of Knowledge Transfer in English Translation Teaching Based on Deep Learning

Guanghui He[1,] ✉

[1] *Zhengzhou Academy of Fine Arts, Zhengzhou, Henan, 451450, China*

ABSTRACT

This study firstly introduces the working principle of deep learning-based neural machine translation model (NMT) and its recurrent neural network translation backbone network, which enhances the semantic characterization capability through Glove word embedding layer. A tree-to-sequence based attention mechanism is innovatively introduced at the encoder side, and a tree-based encoder is appended to the traditional sequence encoder to construct syntax-aware context vectors. On the decoder side, the syntactic tree structure information is integrated into the sequence-to-sequence model (seq2seq), and this model is used to explore the knowledge transfer effect of the English translation teaching process. The results show that the accuracy rates of the neural machine English translation models incorporating syntactic information proposed in this paper are all above 90%. The experiment on the effect of English translation teaching shows that the mean values of students' scores on the post-test of long sentence translation and composition translation in the reading section of the experimental class increased by 11.022 and 12.5388 points respectively compared with those of the control class, with significant differences between the scores of the two groups of students ($p<0.05$), and the same significant differences are presented between the scores on the pre-test and post-test of the students' scores on the long sentence translation and composition translation in the experimental class. It can be seen that the application of the model can effectively promote knowledge transfer and help students better understand and utilize translation skills.

Keywords: Neural machine translation modeling, Glove word embedding method, seq2seq; knowledge transfer, English translation teaching

1. Introduction

English is one of the main courses in high school teaching, which is not only the main way of language knowledge transfer, but also an important way to cultivate students' global perspective and intercultural communication skills [1]. Among them, translation is an important teaching part of

✉ Corresponding author.
E-mail address: m13598830618@163.com (G. He).
Received 25 February 2024; Revised 10 June 2024; Accepted 05 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-181

college English, which involves a lot of knowledge of translation theory and translation skills [2]. In the university English translation classroom, in order to enhance the effect of talent cultivation, the theories and skills often used in English translation should be integrated into the actual teaching activities to ensure that students can fully grasp and apply what they have learned, so as to effectively improve the quality of university English translation teaching [3-5].

Knowledge transfer is the ability of students to effectively apply knowledge, skills and experiences learned in one domain or context to another new domain or context [6]. The core of developing students' knowledge transfer ability lies in understanding and grasping the basic principles and concepts behind the knowledge and being able to apply them flexibly in different environments and situations [7-8]. Knowledge transfer ability involves not only the memorization and reproduction of specific information, but more importantly the in-depth understanding, reorganization and creative application of such information [9-10]. The cultivation of knowledge transfer ability is crucial for students to develop the habit of lifelong learning and adapt to the fast-changing social environment, which not only promotes students' in-depth learning in a single subject area, but also improves their ability to solve real-life problems [11-13]. In the process of knowledge transfer, students will gradually learn how to establish effective connections between old knowledge and new contexts, accurately identify and utilize the similarities and differences between different contexts, in order to better adapt to the ever-changing learning contexts and living environments, and effectively improve their competitiveness and adaptability [14-18]. In conclusion, cultivating students' knowledge transfer ability about English translation teaching is an important and necessary teaching goal in English education, which is of great significance for developing students' English ability and cultivating their habit of lifelong learning [19-20].

This study explores the application of Neural Machine Translation (NMT) model for knowledge transfer in English translation teaching based on deep learning techniques.The NMT model is able to efficiently handle sequence-to-sequence translation tasks through encoder-decoder architecture and attention mechanism. The Glove word embedding technique is introduced in the study to provide richer semantic information for the model by utilizing its global vector property. In addition, this study proposes an NMT model that incorporates syntactic information to further improve the accuracy and naturalness of translation through the introduction of syntactic structures. Finally, this model is applied to English translation teaching in colleges and universities to explore its impact on students' English translation ability enhancement and knowledge transfer.

2. Construction of a Neural Machine English Translation Model Incorporating Syntactic Information

2.1. Deep Learning Based Neural Machine Translation Models

2.1.1. Neural Machine Translation Fundamentals. A language model is the foundation of natural language processing, and its role is to express natural language into a mathematical form that can be processed by a computer. Let V be a word list, and for sentences $Y=\{y_1,y_2,\ldots,y_T\}$, $y_t \in V$, the language model is approximately equivalent to a set of parameters θ such that the following equation is satisfied:

$$P(Y;\theta)=P(y_1,y_2,\ldots,y_T;\theta) \tag{1}$$

The essence of neural machine translation, on the other hand, is autoregressive language modeling, which represents decoding the output words one by one while decoding the generated target language, and the previous decoding results will be used as inputs for decoding the current words. Conditional probability modeling is performed for a possible generated target language sentence $Y=\{y_1,y_2,\ldots,y_T\}$, T representing a target language sentence sequence of length T, given a

source language sentence $X=\{x_1,x_2,\ldots,x_T\}$ condition, T' representing a source language sentence sequence of length T', where $y_{0-t-1}=y_0,y_1,\ldots,y_{t-1}$. ie:

$$P(Y\mid X;\theta)=\prod_{t=1}^{T+1}P\left(y,\mid y_{0-t-1},X;\theta\right) \tag{2}$$

For the autoregressive machine translation model, the maximum likelihood estimation is used for training, and the loss function is the cross-practice loss, where N means that there are N parallel corpus pairs. The training process of neural machine translation i.e. the maximization formula is as follows:

$$L^{ML}(\theta)=\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n+1}\log p\left(y_t^n\mid y_{0\sim t-1}^n,x_{1\sim T}^n;\theta\right) \tag{3}$$

In practice, improved gradient descent algorithms such as Adadelta optimizer or Adam optimizer are used for faster and better training results.

1) Sequence-to-sequence (seq2seq) [21-22] model structure:

(1) Word Embedding Layer. It generally contains two word embedding layers, one acts as the word embedding layer of the source language and is used to convert the words in the text into numeric vectors. The other serves as the word embedding layer for the target language and is used to convert the numeric vectors into text in the target language.

(2) Encoder. The digital vectors obtained from word embedding are dimensionally altered to obtain intermediate vectors that represent the information in the source language, this process can be done by recurrent neural networks.

(3) Intermediate Vector. The intermediate vectors are actually context vectors that contain semantic information and are the initial state of the decoder.

(4) Decoder. The decoder converts the context vectors into text in the target language corresponding to the source language. This process can also consist of a recurrent neural network. The encoder and decoder should be as different as possible so that the number of parameters can be increased to enhance the model performance and improve the quality of translation. The decoder has a softmax layer, which is used to dimensionally change the high-level features and get the probability values for prediction. seq2seq is visualized in Fig. 1:

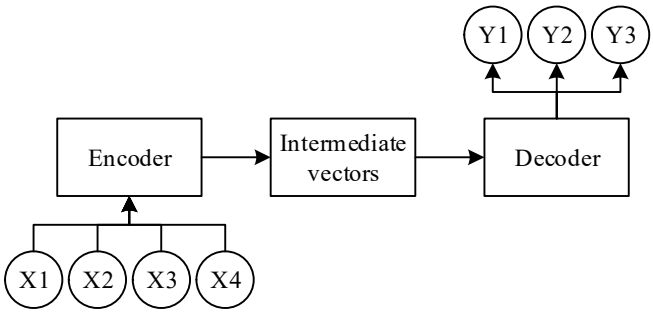

Fig. 1. Seq2seq frame

2) Greedy Search and Cluster Search. Greedy search implies that the model selects the maximum probability for each word in word-by-word decoding, if the output sequence of the decoder is

$\hat{Y}=(\hat{y}_1,\ldots,\hat{y}_T)$, at the moment of t the decoder greedily searches for the output $\hat{y}$, the process is shown in Equation (4), where V denotes the target language word list.

$$\hat{y}_t=\arg\max_{y\in V}\log p\left(y\mid \hat{y}_{0\sim t-1},x_{1\sim T};\theta\right) \tag{4}$$

However, the greedy search method also brings certain problems, the greedy strategy of selecting the maximum probability of generating words at every moment will often fall into the local optimum and fail to reach the global optimum.

The application of cluster search expands the search space by caching each candidate sequence of cluster width and selecting the one with the largest integrated probability as the output, which makes the translation results more diverse and makes the generated translation results close to the global optimum. The process of cluster search decoding is as follows:

$$C_{t-1}=\left\{\tilde{y}^{(1)}_{0\sim t-1},\tilde{y}^{(2)}_{0\sim t-1},\ldots,\tilde{y}^{(K)}_{0\sim t-1}\right\} \tag{5}$$

$$\begin{aligned}C_{t-1}&=\left\{\tilde{y}^{(1)}_{0\sim t-1},\tilde{y}^{(2)}_{0\sim t-1},\ldots,\tilde{y}^{(K)}_{0\sim t-1}\right\}\\&=\underset{y_t\in V,\,y_{0\sim t-1}\in C_{t-1}}{\arg sort}{}^{K}\sum_{t'=0}^{t}\log p\left(y_{t'}\left|y_{0\sim t'-1},x_{1\sim T};\theta\right.\right)\end{aligned} \tag{6}$$

$$\hat{Y}=\arg\max_{\hat{Y}_1\ldots,\hat{Y}_K}\left(\frac{1}{|Y|}\right)^{\alpha}\log p\left(Y\mid X;\theta\right) \tag{7}$$

(1) Let the cluster width be K, then the cluster candidate sequences at the $t-1$ moment are shown in Eq. (5), with a total of K sequences of length $t-1$;

(2) When t moments, K greedy searches are performed for K candidate sequences respectively, as shown in Eq. (6);

(3) Whenever there is a sequence output $\langle eos\rangle$, the decoding of this sequence ends, until all K sequences are decoded;

(4) For the generation of K candidate sequences, such as formula (7), the use of taking the logarithmic regularization method for reordering, where $|Y|$ represents the length of the sequence, when α to take 1, formula (7) calculated that the perplexity of the value of the actual general take α is 0.6, can effectively avoid the generation of too short a sequence.

2.1.2. Neural machine translation backbone network. In general neural networks, each neuron is independently parallelized and cannot process sequential inputs, while recurrent neural networks (RNN) [23] can process sequential data. The schematic of recurrent neural network is shown in Fig. 2, (a) shows the schematic of a simple recurrent neural network, which consists of input, hidden and output layers, respectively; (b) shows its unfolded structure. In the figure, X, S and O correspond to the vector values of the input layer, hidden layer and output layer respectively; U and V correspond to the weight matrices between the input layer and the hidden layer and between the hidden layer and the output layer respectively; and W is the weight matrix of the hidden layer.

The corresponding equation for recurrent neural network is as follows:

$$O_t=g\left(VS_t+b\right) \tag{8}$$

$$S_t=f\left(UX_t+WS_{t-1}+b\right) \tag{9}$$

Equation (8) is the formula to calculate the output layer O, the output layer and the hidden layer are fully connected, i.e., each node is connected correspondingly, g, f denotes the activation function, the commonly used activation functions are Sigmoid and Tanh functions as well as ReLU

function, etc., which is more conducive to model training. Equation (9) is the formula for calculating the state S of the hidden layer, and the input of the current hidden state comes from the input layer X and the hidden state S_{t-1} of the previous moment, and b denotes the offset. Equation (9) can be obtained by repeatedly substituting (8):

$$O_t = g\left(Vf\left(UX_t + Wf\left(UX_{t-1} + Wf\left(UX_{t-2} + \ldots\right)\right)\right)\right) \tag{10}$$

Equation (10) reflects the "memory" function of the recurrent neural network, where the hidden layer S_{t-1} of the previous moment affects the output O_t of the current moment.

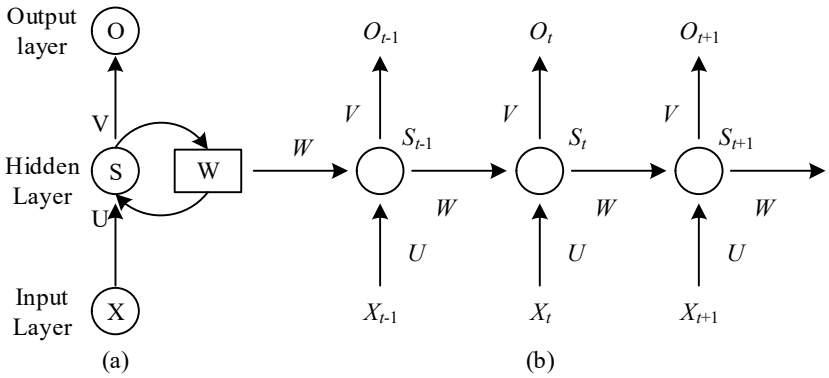

Fig. 2. Cyclic neural network schematic

If a certain eigenvalue λ, its absolute value $|\lambda|>1$, after multiplication will converge to 0, resulting in the gradient disappears; if the absolute value of eigenvalue λ $|\lambda|>1$, after multiplication will surge, resulting in the gradient explosion. That is:

$$W = Q\Lambda Q^T \tag{11}$$

In order to solve the above problem, Long Short-Term Memory (LSTM) network is introduced into recurrent neural network. After weighted by the attention mechanism, the original fixed-length and unchanged semantic vectors will be turned into dynamic semantic vector matrices, which enhances the ability to characterize long sentences.

Figure 3 shows the attention mechanism in the recurrent neural network type. where $X = \{x_1, x_2, \ldots, x_T\}$ represents the source language sentence, s_t represents the hidden state of the decoder at time t, and c_r represents the semantic vector dynamically generated by the model at time t. Dynamic generation comes from the attention mechanism and is calculated as follows:

$$c_t = attention\left(s_{t-1}, h\right) = \sum_{i=1}^{T} a_{t,i} h_i \tag{12}$$

where $a_{t,j}$ denotes the weight coefficient of the encoder output h_t at moment t and $a_{t,i}$ satisfies equation (13):

$$\sum_{j=1}^{T} a_{t,j} = 1 \tag{13}$$

Eqs. (12), (13) reflect that the essence of the attention mechanism is to seek a weighted average, where the dynamic semantic vector c_t is equal to the weighted average of the output vectors of the encoder at moment t. Where the weighting factor $a_{t,i}$ is calculated as:

$$a_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{j=1}^{T} \exp\left(e_{i,j}\right)} \tag{14}$$

where $e_{i,j}$ is computed by a feed-forward neural network, meaning that a softmax layer is added to the tail of this feed-forward neural network to compute the attention weight coefficients, which can also be interpreted as an "alignment" network between the source and target language sequences, and the overall network can be synchronized during the training of the translation model. Training.

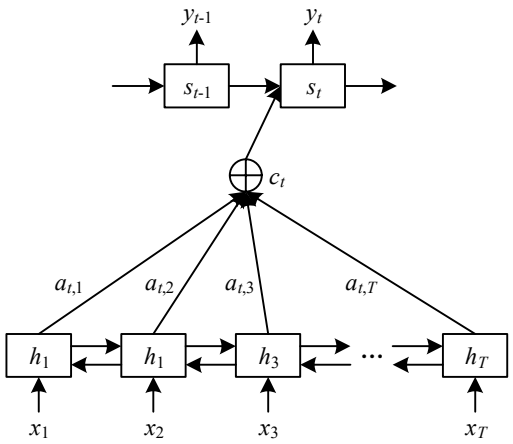

Fig. 3. The focus mechanism of the cyclic neural network

2.2. Introduction of GloVe word vectors

In this paper, we use GloVe word vector representation for word embedding of words.

First, the co-occurrence matrix of a word is memorized as X, where each element X_{ij} represents the sum of the number of times the word j appears in the context of the word i. Let:

$$X_i = \sum_k X_{ik} \tag{15}$$

denotes the sum of the number of all words that appear in the context of word i. Order:

$$P_{ij} = P(j \mid i) = \frac{X_{ij}}{X_i} \tag{16}$$

denotes the probability that word j occurs in the context of word i.

The word vectors are trained so that the probability ratio reflecting the relevance of two words can be obtained after some operation on the word vectors of the two words:

$$F\left(w_i, w_j, \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}} \tag{17}$$

where w_i, w_j, $\tilde{w}_k$ - word vectors corresponding to vocabulary i, j, k, all with dimension d. $\frac{P_{ik}}{P_{jk}}$ in the above equation can be directly calculated using the corpus, and F is an unknown function. Since the word vectors are all located in the same vector space, w_i, w_j can be differenced, i.e.:

$$F\left(w_i - w_j, \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}} \tag{18}$$

$$\Rightarrow F\left[\left(w_i - w_j^T\right)\tilde{w}_k\right] = F\left(w_i^T w_k - w_j^T w_k\right) = \frac{P_{ik}}{P_{jk}} \tag{19}$$

Let F be an exponential function with:

$$\exp\left(w_i^T w_k - w_j^T w_k\right) = \frac{\exp\left(w_i^T w_k\right)}{\exp\left(w_j^T w_k\right)} = \frac{P_{ik}}{P_{jk}} \tag{20}$$

Just have it at this point:

$$\begin{cases} \exp\left(w_i^T w_k\right) = P_{ik} \\ \exp\left(w_j^T w_k\right) = P_{jk} \end{cases} \tag{21}$$

Next for all words in the corpus, by $\exp\left(w_i^T w_k\right) = P_{ik} = \frac{x_{ik}}{X_i}$, there:

$$w_i^T w_k = \log\left(\frac{X_{ik}}{X_i}\right) = \log X_{ik} - \log X_i \tag{22}$$

Swapping the values of i and k does not change the result for $w_i^T w_k$, so introducing two bias terms on the right-hand side of the equation gives it the same symmetry:

$$w_i^T w_k = \log X_{ik} - b_i - b_k \tag{23}$$

At this point 1 is already contained in b_i. The goal of the model translates into making the ends of Eq. (23) as close as possible by learning the representation of the word vectors, so the squared difference between the two can be used as the objective function:

$$J = \sum_{i,k=1}^{V}\left(w_i^T \tilde{w}_k + b_i + b_k - \log X_{ik}\right)^2 \tag{24}$$

where V - number of words.

In order to avoid using the same weights for all co-occurring words, the weights of co-occurring words are adjusted using the statistical information of word co-occurrence in the corpus to make further corrections to the objective function:

$$J = \sum_{i,k=1}^{V} f\left(X_{ik}\right)\left(w_i^T \tilde{w}_k + b_i + b_k - \log X_{ik}\right)^2 \tag{25}$$

Where f - weighting function.

The weight function f needs to have the following properties:

1) $f(0)=0$, when the number of lexical co-occurrences is 0, the corresponding weight should also be 0;

2) $f(x)$ must be a non-decreasing function to ensure that when the number of vocabulary co-occurrence increases its weight will not decrease;

3) For words with a very high frequency of occurrence, $f(x)$ should be able to give them a value that is not too large to avoid overweighting.

Combining these three points, the weighting function can be used:

$$f(x)=\begin{cases}\left(x/x_{\max}\right)^{\alpha} & x<x_{\max}\\ 1 & x\geq x_{\max}\end{cases} \tag{26}$$

2.3. Neural machine English translation model incorporating syntactic information

2.3.1. Construction of the encoder. Start building the encoder for the attention translation model. The encoder uses a bi-directional GRU network, where the forward GRU reads the source language sentence after word embedding from left to right, and the reverse GRU reads the same sentence from right to left, then there:

$$\vec{h}_t = EncoderGRU^{\rightarrow}\left(e\left(\vec{x}_t\right),\vec{h}_{t-1}\right) \tag{27}$$

$$\overleftarrow{h}_t = EncoderGRU^{\leftarrow}\left(e\left(\overleftarrow{x}_t\right),\overleftarrow{h}_{t-1}\right) \tag{28}$$

where $\vec{x}_0=\langle sos\rangle$, $\overleftarrow{x}_0=\langle eos\rangle$.

The input of the bi-directional GRU network [24] is the sentence after word embedding is performed, and PyTorch automatically initializes the hidden states $\vec{h}_0$ and $\overleftarrow{h}_0$ at the initial moment to the all-0 tensor. Since the decoder of the attention translation model does not need to use the bidirectional network, it only needs the context vector c from the encoder as its hidden state s_0 at the initial moment, but the encoder outputs two hidden states, $\vec{c}=\vec{h}_t$ and $\overleftarrow{c}=\overleftarrow{h}_t$, so it uses a linear layer g to stitch these two hidden states together, and then uses $\tanh$ as the activation function of the hidden state:

$$c=\tanh\left(g\left(\vec{h}_t,\overleftarrow{h}_t\right)\right)=\tanh(g(\vec{c},\overleftarrow{c}))=s_0 \tag{29}$$

2.3.2. Attention mechanisms. The attention layer takes as input the hidden state s_{t-1} of the decoder at the previous moment and the hidden state H of the encoder in its entirety, and outputs an attention vector α_t of length the length of the source sentence, which represents how much attention needs to be allocated to each word in the source sentence separately in order to correctly predict the $t+1$th word $\hat{y}_{t+1}$ of the target sentence at the time step t of decoding. Each element of the attention vector has a value between 0 and 1, and all elements sum to 1.

First the energy function between the hidden state s_{t-1} of the decoder at the previous moment and the hidden state H of the encoder is calculated. The hidden state of the encoder is a sequence of T tensors while s_{t-1} is a single tensor, so s_{t-1} needs to be repeated T times before computation. The repeated s_{t-1} and H are spliced and passed into a linear layer $\mathrm{attn}()$, after which $\tanh$ is used as the activation function:

$$E_t = \tanh\left(attn\left(s_{t-1}, H\right)\right) \tag{30}$$

The value of this energy function can be seen as a characterization of how well the hidden state of each encoder matches the hidden state of the decoder at the previous moment. The energy function has a dimension of $\left[dec\_hid\_\dim, src\_len\right]$ for the individual sentences in each batch, i.e., the hidden state dimension multiplied by the length of the source sentence, however, the dimension of the attention vector is $\left[src\_len\right]$, so the energy function is given a left-multiplication by a tensor of dimension $\left[1, dec\_hid\_\dim\right]$, v:

$$\hat{\alpha}_t = vE_t \tag{31}$$

v can be thought of as the weight of the energy-weighted sum of all encoder hidden states. This weight tells the decoder's hidden states how much attention to allocate to each word in the source sentence.

The final normalization of $\hat{\alpha}_t$ yields the attention vector α_t:

$$\alpha_t = soft\max\left(\hat{\alpha}_t\right) \tag{32}$$

2.3.3. Decoder Construction. The decoder contains the attention layer, which computes the weighted sum w_t of H using the encoder hidden state H and the attention vector α_t returned by the attention layer:

$$w_t = \alpha_t H \tag{33}$$

After that the input word $d\left(y_t\right)$ after word embedding, the weighted sum of the encoder hidden states w_t is passed into the decoder along with the hidden states s_{t-1} of the decoder from the previous moment, where $d\left(y_t\right)$ and w_t are made as vector splices, and the decoder is also obtained by using GRUs as activation units:

$$s_t = Decoder\ GRU\left(d\left(y_t\right), w_t, s_{t-1}\right) \tag{34}$$

Afterwards s_t, $d\left(y_t\right)$ and w_t are passed into a linear layer f to predict the next word $\hat{y}_{t+1}$ in the target sentence:

$$\hat{y}_{t+1} = f\left(d\left(y_t\right), w_t, s_t\right) \tag{35}$$

2.3.4. Sequence-to-sequence modeling. After completing the construction of the encoder, decoder, and attention layers, they are integrated into the sequence-to-sequence model by the following steps:

1) Create a $outputs$ tensor used to store all word predictions $\hat{Y}$;

2) Pass the source sentence X into the encoder to obtain outputs C and H;

3) Assign the hidden state of the initial moment of the decoder to a context vector with $s_0 = c = h_T$;

4) Use the $\langle sos \rangle$ identifier as the first input word to the decoder y_1;

5) Perform the decoding in a loop:

(1) Input the tst word y_t of the target sentence, the previous hidden states s_{t-1} and H into the decoder;

(2) Get a prediction word $\hat{y}_{t+1}$ and a new hidden state s_t;

(3) Decide whether to use TEACHER FORCE and rationalize the next input until the output predicted word is $\langle eos \rangle$ or the target word at the current moment is $\langle eos \rangle$.

There exist two training modes for RNN, Free-Running Mode and Teacher-Forcing Mode.Free-Running Mode is the common mode of using the output of the previous time step as the input of the next time step, while Teacher Forcing is a fast and efficient way to train an RNN, the model randomly uses a portion of the true values from the training set as input for the next time step.

2.4. English-Chinese Translation Effects of Different Models

Experimental comparison models:

1) PbSMT: Phrase-based statistical machine translation system.

2) RNNSearch: neural network machine translation model based on attention mechanism.

3) Coverage: a neural network machine translation model based on the RNNSearch model architecture and adding a coverage mechanism to it.

4) ConvS2S: a sequence-to-sequence modeling model based on convolutional neural networks for machine translation tasks.

Experimental dataset:

The corpus used for the experiments is four English monolingual corpora of size 800MB each including textbooks, news, reading and daily language, containing a total of 286,973 English sentences, which have been subworded and used as a word vector training set.

1) English-Chinese translation effect of different models. Table 1 shows the English-Chinese translation effects of different models. Experimental results show that compared with the comparison model, the neural machine English translation model with syntactic information has improved the quality evaluation effect of the test data. Due to the introduction of additional translation knowledge, the Pearson correlation coefficient of the proposed model is increased by 13.5%-41.5% compared with the four models of "PbSMT, RNNSearch, Coverage and ConvS2S" on the four datasets of "Textbooks, News, Read and Daily speaking". MAE and RMSE decreased by 4.2%-26.9%, respectively. Therefore, the translation knowledge can be transferred to the quality assessment task through this model, which can assist the machine translation quality assessment model to achieve better quality evaluation results.

Table 1. The English-Chinese translation of different models

Test index	Date	Pearson's	MAE	RMSE
PbSMT	Textbooks	0.564	0.111	0.116
	News	0.686	0.109	0.116
	Read	0.585	0.109	0.121
	Daily speaking	0.773	0.111	0.12
RNNSearch	Textbooks	0.741	0.218	0.301
	News	0.696	0.108	0.121
	Read	0.659	0.114	0.221
	Daily speaking	0.688	0.104	0.119
Coverage	Textbooks	0.714	0.211	0.121
	News	0.802	0.116	0.213
	Read	0.741	0.109	0.117
	Daily speaking	0.762	0.109	0.312
ConvS2S	Textbooks	0.666	0.115	0.212
	News	0.635	0.112	0.312
	Read	0.785	0.119	0.311

	Daily speaking	0.735	0.113	0.121
Ours	Textbooks	0.968	0.061	0.051
	News	0.956	0.062	0.043
	Read	0.979	0.021	0.062
	Daily speaking	0.937	0.015	0.044

2) Machine Translation Quality Assessment Based on Knowledge Migration. In the stage of English translation teaching based on deep learning, compared with the introduction of exploratory new knowledge translation content, the focus of teaching should be more on cultivating students' ability to construct English translation knowledge system independently. This requires the help of the seq2seq neural machine translation model proposed in this paper, which integrates syntactic information to satisfy the blind spot of knowledge translation produced by students during active thinking, cultivate students' problem-solving consciousness through the help of intelligent machine translation, be good at incorporating new knowledge into the existing cognitive framework of English translation, integrate the fragmented knowledge of English translation into a systematic knowledge, transform the static lexical and syntactic information knowledge into dynamic knowledge, and develop students' ability to independently construct the knowledge system of English translation. The knowledge of static lexical and syntactic information can be transformed into dynamic knowledge, and the knowledge in English books can be internalized into one's own knowledge with the help of the software in this paper. At the same time, it is important to emphasize the use of knowledge transfer to solve translation problems in new English translation situations, and then form higher-order cognition. In addition, based on the principles of self-referencing, other-referencing, and continuity-referencing, teachers need to continuously improve their own knowledge reserves and strengthen their ability to correct students' recognition of the construction of knowledge systems in order to intervene when students' independent learning is at an impasse and to provide them with continuous learning guidance.

This subsection evaluates the quality of machine translation by comparing the models. Table 2 shows the results of the experiments of different machine translation systems on each dataset. The scores of BLEU4, OTEM2 and UTEM4 are presented respectively. From the data in the table, it can be seen that although all the Neural Machine Translation (NMT) models outperform all the Statistical Machine Translation (SMT) models based on the Noisy Channel Model in the evaluation results of the BLEU metrics, almost all the SMT models outperform the results of the NMT models in the OTEM scores. We hypothesize that the reason for this result is that the hard constrained coverage mechanism in SMT models prevents the decoder from translating the same source language phrase multiple times in the process of generating candidate translations. Also based on the advantage of the coverage mechanism in the SMT model, the Coverage model achieves a significant improvement over the RNNSearch model. It is interesting to note that although the ConvS2S model and this paper's model are very similar in terms of BLEU scores, this paper's model outperforms ConvS2S in terms of OTEM. We believe this is because the attentional mechanism used in the model in this paper establishes a strong association between the source language sentence and the target language translation that has been produced in the decoder, whereas the convolutional layer used in the ConvS2S model can only capture localized dependency information.

It can also be observed that although the NMT models all outperform the SMT models in terms of BLEU scores, all the machine translation systems score close to each other in terms of UTEM scores. Through the Coverage mechanism, the SMT model is able to force every word in the source language sentence to be translated, but the advantage of this mechanism is not evident in the Coverage model. To summarize, different machine translation systems have their own ways of dealing with the under-translation and over-translation problems. Relying only on BLEU, which is a metric for translation evaluation based on faithfulness and fluency for overall quality evaluation, cannot lead to the above rich findings. In contrast, the model in this paper, because of the introduction of syntactic and GloVe

word embedding methods, well circumvents the emergence of the above problems, and the metrics of BLEU, OTEM and UTEM performance of the model in this paper are higher than those of the other four comparative models, so it can be seen that it possesses a very high application value.

Table 2. Results of different machine translation systems in each data set

Test index	Model	Textbooks	News	Read	Daily speaking
BLEU4	PbSMT	31.519	35.507	31.051	35.242
	RNNSearch	30.363	32.074	32.204	35.096
	Coverage	35.109	40.722	32.471	38.875
	ConvS2S	36.069	45.046	43.711	41.508
	Ours	39.841	46.965	43.862	49.329
OTEM2	PbSMT	0.762	0.775	0.777	0.784
	RNNSearch	0.742	0.775	0.722	0.851
	Coverage	1.301	1.381	1.387	1.405
	ConvS2S	1.778	1.784	1.769	1.759
	Ours	2.97	2.958	2.859	2.915
UTEM4	PbSMT	55.387	53.345	56.209	54.809
	RNNSearch	61.061	50.949	59.517	54.807
	Coverage	62.286	55.029	56.422	57.588
	ConvS2S	56.147	52.988	55.923	50.598
	Ours	66.491	59.051	56.852	56.358

3. The Effectiveness of Neural Machine Translation Modeling in Teaching English Translation

Knowledge transfer refers to the process of applying knowledge learned in one field to another related field. In English translation teaching, knowledge transfer can help students apply their existing language knowledge and translation skills to new translation tasks, thus improving translation efficiency and quality. While traditional translation teaching often relies on teachers' explanations and students' memorization, knowledge transfer automatically transfers a large amount of translation knowledge and skills to students' translation practice through deep learning models, thus achieving more efficient teaching results.

In this paper, two classes of English majors in a university are selected as the research object, and the teaching process of the experimental group adopts the “Neural Machine Translation Model with Syntactic Information Integration”, while the control group adopts the traditional teaching mode, and the trial period is 8 weeks. There were 50 students in each group, and there was no significant difference in the pre-test scores of long sentence translation, word translation and oral translation between the two classes.

3.1. Questionnaire design

In this paper, we design a questionnaire about the evaluation of the teaching effect of students' seq2seq neural machine translation model based on fused syntactic information on a five-point Likert scale (1=strongly disagree, 5=strongly agree). The questionnaire on the effectiveness of the application of neural machine translation model in English translation teaching is shown in Table 3.

Table 3. The application effect of neural machine translation model in English teaching

Survey target	Question
Investigation of	1. I can make simple communication with my classmates through reading English knowledge.

knowledge migration effect	2. I can understand the author's views and use similar problems in the analysis.
	3. For the same type of English, I know how to read.
	4. I found that in reading comprehension, a new topic might not know what to do.
On translation ability improvement survey	5. After using the seq2seq neural machine translation model, my translation speed has improved.
	6. After using the seq2seq neural machine translation model, my translation accuracy improved.
	7. After using the seq2seq neural machine translation model, my confidence in the translation task increased.

3.2. Post-test results and analysis

The operability and scientificity of the teaching design path are supported by the students' final grades and the analysis of the questionnaire data after the experiment. After the 8-week experiment, this paper conducts a questionnaire survey on the changes in the level of long-sentence translation in the English reading part and the level of English-Chinese translation in the composition part of the experimental class and the control class before and after the experiment. The questionnaires were distributed in 100 copies, and the recovered valid questionnaires were 100, 50 in the control class and 50 in the experimental class. In this paper, the recovered questionnaires are analyzed again by independent samples t-test.

3.2.1. Analysis of final results. The test of students' performance is divided into two parts: long sentence translation in the reading part and English-Chinese translation in the composition, with a total of 50 points for each part.

The test questions of the final exam are consistent with the question types and time schedule of the pre-test. The long sentence translation scores were collected after the test, assuming that "the long sentence translation scores of the two classes after the experiment are normally distributed", and the results of the normality test of the long sentence translation scores in the post-test are shown in Table 4. The statistical results of the post-test long sentence translation achievement groups are shown in Table 5. The p-values of S-W are all greater than 0.05, so the original hypothesis is accepted that the posttest long sentence translation scores of the two classes obey normal distribution, and the t-test analysis can be continued to analyze the data of the two classes. "Assuming that there is no significant difference between the posttest long sentence translation scores of the two classes", the results of the independent samples t-test for posttest long sentence translation scores are shown in Table 6. The mean values of long-sentence translation scores in the reading part of the control class and the experimental class are 26.671 and 37.693, which satisfy variance chi-square. $p=0.0362<0.05$, indicating that there is a difference in the long-sentence translation scores in the reading part of the posttest of the two classes, which suggests that the neural-machine English translation model that incorporates syntactic information is beneficial to the design of English translation instruction in colleges and universities in order to promote the students' long-sentence translation of the English in reading part grammatical mastery and improve their translation level.

Table 4. The results of the positive test of the translation results were measured

Test object	Kormogov smealov (V)[a]			Shapiro wilk (S-W)		
	Statistics	Freedom	Significance	Statistics	Freedom	Significant (P)
Control group	0.098	50	0.1807	0.9658	50	0.1259
Experimental group	0.129	50	0.0514	0.9847	50	0.2276

Table 5. The results of the results were calculated

Test object	Case number	Mean value	Standard deviation	Standard error mean
Control group	50	26.671	5.3137	1.0021

Experimental group	50	37.693	4.7841	0.9826

Table 6. The results of the independent sample T test were measured

	F	Sig.	t	Freedom	Sig (double)	Mean difference	Standard error difference
Assumed equal variance	2.284	0.1535	-2.089	100	0.0362	-3.094	1.461
Unassuming equal variance			-2.086	100	0.0363	-3.094	1.463

The statistical results of the essay translation posttest group are shown in Table 7. The results of the independent samples t-test for the composition translation posttest are shown in Table 8. The mean scores of the control class and the experimental class are 26.7289 and 39.2677 respectively, P=0.0153<0.05, so the variance is not uniform, referring to the Sig. (two-tailed) in "not assuming equal variance", i.e., P=0.000<0.05, which indicates that there is a significant difference between the two classes before and after the questionnaire in terms of the translation of the composition. There is a significant difference between the two classes, and the mean score of the experimental class is significantly higher than that of the control class.

Table 7. The statistical results of the post-translation group

Test object	Case number	Mean value	Standard deviation	Standard error mean
Control group	50	26.7289	0.3547	0.0681
Experimental group	50	39.2677	0.3531	0.0554

Table 8. Test results of the independent sample t test after the translation

	F	Sig.	t	Freedom	Sig (double)	Mean difference	Standard error difference
Assumed equal variance	4.089	0.0153	-4.1089	100	0.000	-0.36094	0.0902
Unassuming equal variance			-4.1086	100	0.000	-0.36094	0.0901

3.2.2. Analysis of the results of the pre- and post-tests on the improvement of translation ability in the experimental classes. In order to test the changes of the experimental class in the reading part of the long sentence translation scores and the composition English-Chinese translation scores, this paper examines whether there is a significant change in the experimental class students' pre-test and post-test reading part of the long sentence translation scores and composition English-Chinese translation scores by the method of paired samples t-test. After the verification of this paper, both groups of scores obey normal distribution.

1) Comparative Analysis of Long Sentence Translation Achievement in Reading Section. Because the students' reading part long sentence translation achievement obeys normal distribution, the paired-sample t-test can be further conducted on the reading part long sentence translation achievement, assuming that "there is no significant difference between the experimental class students' reading part long sentence translation achievement before and after the experiment". The results of the normality test for the long sentence translation achievement of the experimental class in the reading part of the pre and post test are shown in Table 9. As shown in the table, the mean values of the pre-test and post-test long sentence translation scores of the reading part of the experimental class are 25.3671 and 37.693 respectively, t=-4.3027, p=0.000<0.05, which shows that there is a significant difference between the pre and post-test English reading scores of the experimental class.

Table 9. The long sentence translation results of the experimental class are tested

		Mean value	Case number		Standard deviation	Standard error mean
Achievement pairing	Pre-test	25.3671	50		7.6593	1.0928
	Post-test	37.693	50		6.1805	0.8748
	Mean value	Standard deviation	The difference is 95% confidence interval		t	Sig.(double)
			Lower limit	Upper limit		
Pre-measured minus post-test	-12.3259	6.8533	-8.7642	-1.2559	-4.3027	0.000

2) Comparative Analysis of Composition English-Chinese Interpretation Scores. The results of the paired-sample t-test on the pre- and post-tests of the experimental class' performance in composing English-Chinese interpreting are shown in Table 10. The results show that the mean values of the pre- and post-tests of the experimental class in terms of composition English-Chinese translation performance are 26.0142 and 39.2677, respectively, with t=-5.7341, p=0.000 less than 0.05, so the original hypothesis is accepted, and it is considered that there is a significant difference between the experimental class students in terms of their composition English-Chinese translation performance before and after the experiment. It shows that the application of the neural machine English translation model incorporating syntactic information in the English translation teaching classroom in colleges and universities can significantly improve the students' translation level, and promote the students' more solid mastery of the grammatical knowledge and translation skills in translation. It can be seen that the addition of the English machine translation model can better help students apply their existing language knowledge and translation skills to new translation tasks, thus improving translation efficiency and quality.

Table 10. Test results of the match sample of English and English

		Mean value	Case number		Standard deviation	Standard error mean
Achievement pairing	Pre-test	26.0142	50		8.1094	1.0817
	Post-test	39.2677	50		7.6107	0.9143
	Mean value	Standard deviation	The difference is 95% confidence interval		t	Sig.(double)
			Lower limit	Upper limit		
Pre-measured minus post-test	-13.2535	6.7452	-9.57438	-3.22989	-5.7341	0.000

The results of the frequency change survey for the translation ability dimension options are shown in Table 11.

Question 1: My translation speed has increased after using the seq2seq neural machine translation model. The most frequent option in the pre-test data is option 4, “Comparatively agree”, which tests students' views on translation efficiency. The change in the frequency of the option shows that students are conscious of improving the progress of article translation when they read, can think quickly, and think that translation is helpful for their development. Therefore, in the post-test data, the students' options are mostly concentrated in option 4 and option 5.

Question 2: After using the seq2seq neural machine translation model, my translation accuracy has improved.

The frequency of changes in the options in the pre- and post-test of this question shows that there is a significant difference in the responses to this question. The translation level of students in higher

education varies greatly, with fewer students accurately performing reading translations. However, with the assistance of the neural machine English translation model that integrates syntactic information, the translation grammar of obscure and difficult to understand long sentences is clearly presented, and repeated many times allows students to firmly grasp the translation skills and skillfully apply them to writing and other fields, which efficiently completes the transfer of knowledge and improves students' translation accuracy. So this question after the experiment options are equally much focused on option 4 and option 5.

Question 3: My confidence in English translation has increased significantly after using the seq2seq neural machine translation model. The highest frequency in the pre-test data was option 4 “agree more”, but a few students chose options 1-3. After the 8-week trial, the students felt more and more satisfied with the English translation, and their confidence in the translation task increased significantly. Therefore the post-test data showed a higher frequency of option 4 and option 5.

Table 11. Translation ability dimension selection frequency change survey results

Question	1		2		3		4		5	
	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test
Q5	1	0	3	1	19	7	25	14	2	28
Q6	0	0	7	1	15	9	22	17	6	23
Q7	2	0	8	3	17	13	14	20	9	14

3.2.3. Analysis of the results of the pre- and post-tests of knowledge transfer ability of the experimental classes. The paired samples t-test for the pre and post-test knowledge transfer utilization dimensions of the experimental class is shown in Table 12. The mean values of the pre-test and post-test of the experimental class on the dimension of knowledge transfer and utilization are 53.841 and 85.923 respectively, with t=-14.3659, p=0.000<0.05, so it is considered that there is a significant difference between the pre and post-tests of the experimental class on the dimension of knowledge transfer and utilization. It shows that after using the seq2seq neural machine translation model, the English translation teaching model in colleges and universities helps students to use the internalized knowledge.

Table 12. T test of the application of knowledge migration in experimental class

		Mean value	Case number		Standard deviation	Standard error mean
Achievement pairing	Pre-test	53.841	50		9.7401	1.5734
	Post-test	85.923	50		7.935	0.9527
	Mean value	Standard deviation	The difference is 95% confidence interval		t	Sig.(double)
			Lower limit	Upper limit		
Pre-measured minus post-test	-32.082	9.433	-17.3124	-2.5082	-14.3659	0.000

The results of the change in the frequency of knowledge transfer and application options are shown in Table 13.

Question 1 and question 2 are relatively similar, the purpose is to examine the students' ability of transferring and applying after learning, and their high-frequency options in the pre and post-tests change, with pre-test option 3 and option 4 as the higher frequency options, and post-test results with option 4 and option 5 as the higher frequency options. The main reason is that the college English

reading and translation teaching model using the seq2seq neural machine translation model advocates that students pay attention to the core ideas of the content of the discourse, and that students deeply understand the knowledge of the discourse and the essence of the content in the process of reading, so that in the classroom output exercises, students can reasonably utilize the essence of the knowledge. Students can also transfer their knowledge in similar situations.

Question 3: I know how to read the same type of English discourse. The design path of college English reading teaching using the seq2seq neural machine translation model requires students to be able to grasp the structure of the passage and the overall pulse of the discourse. Therefore, when teaching reading, teachers are required to consciously guide students to observe the discourse type and recall the reading methods related to that discourse type, so it is shown in the posttest results that teaching under this design path is favorable to the transfer of students' knowledge related to the discourse type.

Question 4 was a reverse scoring question. In the daily reading lesson type and reading training, it is possible that a question may be asked in a different way, and the students may be hesitant about the answer. After the experimental phase of training, the experimental class students show obvious changes in the post-test data, with option 1 and option 2 appearing most frequently, indicating that the design path of college English reading teaching using the seq2seq neural machine translation model is conducive to training students' ability of synonymous conversion.

Table 13. Knowledge migration USES options frequency variation results

Question	1		2		3		4		5	
	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test
Q1	2	0	2	1	15	5	24	24	7	20
Q2	1	0	6	3	18	8	23	22	2	17
Q3	3	1	4	2	17	7	21	25	5	15
Q4	8	22	15	19	19	7	5	2	3	0

4. Conclusion

This paper proposes a neural machine translation model that incorporates syntactic information and investigates its effect on knowledge transfer in the process of English translation teaching. The results show that:

The post-test scores of the long-sentence translation and essay translation of the English reading part of the students in both the experimental class and the control class show that there is a significant difference between the students' scores and their scores before the experiment after the use of the Neural Machine English Translation Model for Fusing Syntactic Information to Assist Teaching ($p < 0.05$). The analysis results of translation ability improvement and knowledge transfer ability showed that the students in the experimental class showed significant improvement in speed, accuracy and confidence in English translation. After using the neural machine translation model in teaching English translation in colleges and universities, it helps students to use the internalized knowledge. Significant differences were found between the pre-test and post-test scores of the experimental class students in the dimensions of translation proficiency improvement and knowledge transfer and utilization, with a difference in their mean values of 12.3259 and 14.3659 points ($p=0.000 < 0.05$). Obviously, the model can effectively promote knowledge transfer in English translation teaching and significantly improve students' translation ability and understanding of language structure.

References

[1] Zheng, Y. (2021). Strategies to improve the effectiveness of college English translation teaching. Advances in Vocational and Technical Education, 3(2), 86-91.

[2] Huang, S., & Qiu, D. (2022). RESEARCH ON COLLEGE ENGLISH TRANSLATION TEACHING THEORY AND TRANSLATION SKILLS FROM THE PERSPECTIVE OF EDUCATIONAL PSYCHOLOGY. Psychiatria Danubina, 34(suppl 1), 733-735.

[3] Wu, J. (2019). Practice Analysis on the Teaching Theory and Skills of English Translation. Review of Educational Theory, 2(1), 6-10.

[4] Yao, C. (2019). The Practical Strategies of English Translation Education under the Theory of Functional Translation. International Journal of New Developments in Education, 1(2).

[5] Geng, X. (2019). On College English Translation Teaching and Quality Bettering Paths. In Application of Intelligent Systems in Multi-modal Information Analytics (pp. 1424-1429). Springer International Publishing.

[6] Wang, C., Zhang, Y., Moss, J. D., Bonem, E. M., & Levesque-Bristol, C. (2020). Multilevel factors affecting college students' perceived knowledge transferability: From the perspective of self-determination theory. Research in Higher Education, 61, 1002-1026.

[7] James, M. A. (2014). Learning transfer in English-for-academic-purposes contexts: A systematic review of research. Journal of English for Academic Purposes, 14, 1-13.

[8] Jwa, S. (2019). Transfer of knowledge as a mediating tool for learning: Benefits and challenges for ESL writing instruction. Journal of English for Academic Purposes, 39, 109-118.

[9] Granado-Alcón, M. D. C., Gómez-Baya, D., Herrera-Gutiérrez, E., Vélez-Toral, M., Alonso-Martín, P., & Martínez-Frutos, M. T. (2020). Project-based learning and the acquisition of competencies and knowledge transfer in higher education. Sustainability, 12(23), 10062.

[10] Jackson, D., Fleming, J., & Rowe, A. (2019). Enabling the transfer of skills and knowledge across classroom and work contexts. Vocations and learning, 12(3), 459-478.

[11] Wang, C., Yan, W., Guo, F., Li, Y., & Yao, M. (2020). Service-Learning and Chinese College Students' Knowledge Transfer Development. Frontiers in Psychology, 11, 606334.

[12] O'Reilly, N. M., Robbins, P., & Scanlan, J. (2019). Dynamic capabilities and the entrepreneurial university: a perspective on the knowledge transfer capabilities of universities. Journal of Small Business & Entrepreneurship, 31(3), 243-263.

[13] Stasewitsch, E., & Kauffeld, S. (2022). Knowledge Transfer in a Two-Mode Network between Higher Education Teachers and Their Innovative Teaching Projects. Journal of Learning Analytics, 9(1), 93-110.

[14] Larkin, J. H. (2018). What kind of knowledge transfers?. Knowing, learning, and instruction, 283-305.

[15] Guillouet, L., Khandelwal, A. K., Macchiavello, R., Malhotra, M., & Teachout, M. (2024). Language barriers in multinationals and knowledge transfers. Review of Economics and Statistics, 1-56.

[16] Alaoui, A. (2015). Knowledge transfer and the translation of technical texts. International Journal of Humanities and Social Sciences, 9(10), 3380-3386.

[17] Liang, H., & Li, X. (2018). Research on Innovation Method of College English Translation Teaching Under the Concept of Constructivism. Educational Sciences: Theory & Practice, 18(5).

[18] Piperno, R., Bacco, L., Dell'Orletta, F., Merone, M., & Pecchia, L. (2025). Cross-lingual distillation for domain knowledge transfer with sentence transformers. Knowledge-Based Systems, 113079.

[19] Wang, Y. (2022). English Translation Teaching Algorithm in Colleges and Universities Using Data-Driven Deep Learning. Mobile Information Systems, 2022(1), 2712541.

[20] Liu, Y., Li, S., & Cui, D. (2024). Analysis of translation teaching skills in colleges and universities based on deep learning. Computers in Human Behavior, 157, 108212.

[21] Guanghua Zhang,Hua Liu,Junjun Guo & Tianyu Guo. (2025). Distilling BERT knowledge into Seq2Seq with regularized Mixup for low-resource neural machine translation. Expert Systems With Applications125314-125314.

[22] Di Wu,Peng Cheng & Yuying Zheng. (2025). Seq2Seq dynamic planning network for progressive text generation. Computer Speech & Language101687-101687.

[23] Nana Huang,Hongwei Ding,Ruimin Hu,Pengfei Jiao,Zhidong Zhao,Bin Yang & Qi Zheng. (2025). Multi-time-scale with clockwork recurrent neural network modeling for sequential recommendation. The Journal of Supercomputing(2),412-412.

[24] Junzhong Ji,Chuantai Ye & Cuicui Yang. (2024). Convolutional bidirectional GRU for dynamic functional connectivity classification in brain diseases diagnosis. Knowledge-Based Systems111450-.