





## Diverse English Translation Teaching Strategies from the Perspective of Computer-Aided Technology

Wenjie Li<sup>1</sup> and Huiqin Liu<sup>2\*</sup>

<sup>1</sup>Department of Foreign Languages, Taiyuan University, Taiyuan 030032, China, [13038002918@163.com](mailto:13038002918@163.com)

<sup>2</sup>Department of Foreign Languages, Taiyuan University, Taiyuan 030032, China, [liuhuiqin@tyu.edu.cn](mailto:liuhuiqin@tyu.edu.cn)

Corresponding author: Huiqin Liu, [liuhuiqin@tyu.edu.cn](mailto:liuhuiqin@tyu.edu.cn)

**Abstract.** The era of globalization has accelerated international cultural exchange, and English translation has become an important way of cross-cultural communication, and machine translation has the advantages of low cost, fast speed and high efficiency, and has important application value, especially in the field of English translation teaching, and has great potential for application. However, machine translation technology is still difficult to reach the level of human translation, and it is necessary to apply neural network model to optimize the machine translation process and innovate the English translation teaching mode. In this paper, we firstly identify the problems of five neural network machine translation models, and secondly an application path of migration learning techniques for machine translation tasks with scarce corpus data resources is presented to avoid the overfitting problem and thus improve the capability of end-to-end neural network machine translation models. The research results have important implications for optimizing and expanding the English translation teaching model.

**Keywords:** Computer-aided Technology; English translation teaching; machine translation model; Knowledge distillation

**DOI:** <https://doi.org/10.14733/cadaps.2022.S7.67-78>

### 1 INTRODUCTION

Natural language is the language that people use on a daily basis as agreed upon by human society and is an important tool for human learning and living. The knowledge recorded and transmitted in the form of language and writing accounts for more than 80% of the total knowledge, therefore, the learning and transmission of language is the key to promote the development of human civilization. With the acceleration of world economic integration and the increasingly frequent communication in the international community, the efficiency of English translation teaching is greatly hindered by the traditional manual translation method due to its

high cost. The emergence of natural language processing provides technical support for optimizing and expanding English translation teaching mode. Natural language processing refers to the use of computers as tools to process information in human language, including the manipulation and processing of language. Machine translation is a specific application of natural language processing, in which people use machines to convert a certain language into another language, which is a concrete expression to realize information exchange between human and machine. In recent years, with the development of deep learning technology, neural network machine translation has become mainstream [1]. Among them, the End-to-End neural network translation model has greatly promoted the application and development of deep learning in the field of machine translation because it uses a nonlinear model to solve the problems of linear computation of machine translation under statistical principles and replaces the linear structure with a single complex neural network.

However, it should be pointed out that although machine translation technology is changing rapidly and has the advantages of high-speed processing of linguistic information, improving translation quality and reducing translation cost, it still has not reached the level of fully analyzing and understanding semantic information, and needs continuous improvement of translation model cognition and learning ability. [2]. Importantly, the diffusion of machine translation is greatly limited by the diversity of languages, the slow update of processing techniques for semantic information, and the lack of resources in the corpus. However, cross-language comprehension differences, the upper limit of semantic information processing ability and the lack of corpus resources limit the popularization of machine translation, especially in the teaching of English translation. The optimal logic of teaching English translation requires both teachers and students to teach with minimal cognitive effort, and if the machine translation performance is difficult to reach the naturalness presented by human translation, it will hinder the fluency of teaching and lead to inefficiency. Therefore, this paper identifies the specific problems of the end-to-end neural network translation model and emphasizes that, in the context of scarce corpus resources, the reasonable application of migration learning techniques can effectively deal with the over-fitting problem, and thus improve the performance of neural network machine translation models. The research results are of great significance for improving the efficiency and optimizing the teaching mode of English translation.

## 2 RELATED STUDIES

Based on an extensive literature review, this paper provides a brief review of five classes of mainstream neural network models that are widely used in the field of machine translation: Recurrent Neural Network (RNN), as well as improved Long Short Term Memory Network (LSTM RNN) and Gate Recurrent Network (GRU); Convolutional Neural Network (CNN) model; and Transformer model based on attention mechanism. Recently, many researchers have also used deep learning models such as Reinforcement Learning (RL) ideas and Generative Adversarial Networks (GAN) for machine translation tasks.

### 2.1 Recurrent Neural Network Model

Recurrent Neural Network (RNN) is a neural network with feedback, which has a strong memory and good time series modeling performance. In the NMT translation model based on the encoder-decoder architecture, the encoder-decoder initially uses a recurrent neural network because it is the same as the process originally proposed by Neural Machine Translation (NMT) to imitate human translation, where the RNN first reads the entire sentence to be translated, then understands its meaning in context, and finally produces the translation result.

The training of RNN generally adopts the Back-Propagation Through Time (BPTT) algorithm [3], which back-propagates the error signal in chronological order and recursively calculates the gradient at each moment to update the weight matrix of the network. Due to the time dependence

in the RNN structure, the convergence of the BPTT algorithm is slow during the training process. However, it is difficult to deal with the long-range dependencies during training, and there are problems of "gradient disappearance" and "gradient explosion", so the traditional structure of RNN often falls into local optimal solutions. In order to overcome the problems of traditional RNN models, researchers have proposed many improved forms according to the different connection methods and basic unit types within RNNs, such as LSTM RNN [4], whose implicit layer consists of a set of Memory Blocks, each of which is composed of self-linked memory neurons and three types of memory blocks. The LSTM RNN controls the information transfer through input gates, output gates, and forgetting gates, allowing the gradient to propagate smoothly over a longer period of time.

In the field of machine translation, methods such as RNN and its variants (LSTM RNN and GRU) are not free from temporal limitations and cannot compute in parallel, resulting in slow training speed and speed efficiency problems on large data sets. Moreover, they cannot learn the global information better and solve the long-range dependency problem completely.

## 2.2 Convolutional Neural Network Model

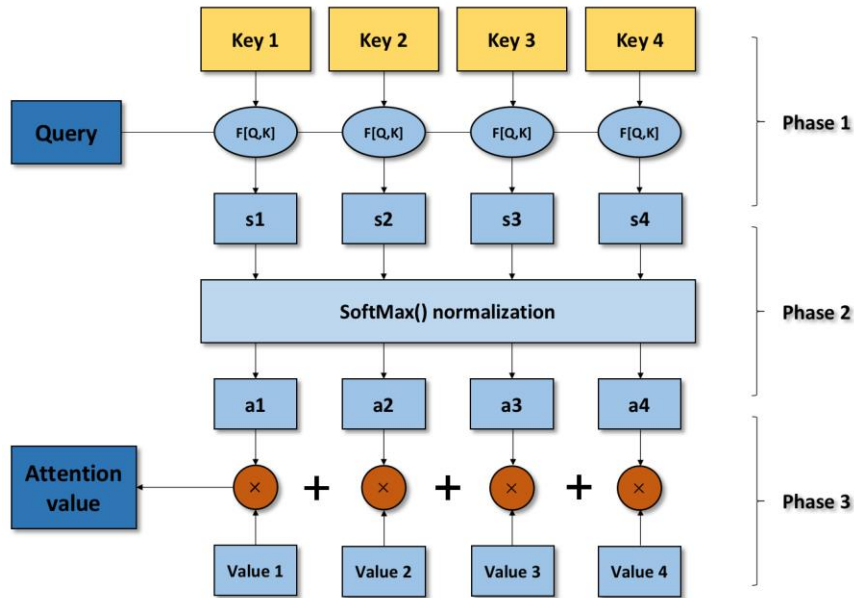
Convolutional neural network (CNN) is a multilayer neural network model that reduces the complexity of the network by using local connectivity and weight sharing. Qi et al. [5] proposed a convolutional coding model applied to the field of neural network translation. The model consists of two convolutional encoders, where the encoder output of one CNN is used to calculate the attention weights and the output of the other CNN is summed up as a conditional input to the decoder, which uses an LSTM structure. The CNN-based neural network machine translation model is more parallelizable than the RNN-based neural network machine translation model, while providing a shorter path to capture long-range dependencies in the source language sentences. Yang et al. [6] proposed a neural network machine translation model with encoder and decoder fully based on CNN, which uses multi-hop attention and gate control unit techniques. Multi-hop attention allows the network to review the target language words multiple times to determine which words are most relevant to the next source language word in the translation, thus improving the translation quality. Gate control units, on the other hand, control the flow of information in the neural network, simplifying gradient propagation.

CNN-based machine translation systems have three advantages: first, text length can be precisely controlled; second, convolution can be processed for multi-task writing and is independent of the previous state, allowing same-time computation for each element in each task; and third, each word input to a complex network undergoes a fixed number of nonlinear computations, while greatly reducing the complexity of the network.

## 2.3 Transformer Model based on Attention Mechanism

Transformer models based on attention mechanism have gradually started to replace RNN and CNN. The attention mechanism (AM) has been widely used in neural network machine translation models, especially in various machine translation tasks such as cross-language translation, and has a wide de in deep learning [7]. The attention mechanism is a resource allocation model introduced by the attention model of human brain. In deep learning, the attention mechanism can quickly select the more important and effective information and then are transformed into knowledge. DeRose et al. [8] pointed out that the AM is an idea that simulates human brain thinking and does not depend on a specific framework itself. The most successful application of the AM is in the field of machine translation, because the Transformer model uses the AM to construct the importance of each word in a sentence, and the word features are updated by weighting their importance and calculating the sum of linear transformations of all words. The attention mechanism is also an important reason for the breakthrough of the Seq2Seq model, and takes the alignment mechanism to a new level. The AM function can be described as a mapping of a query to

a series of key and value pairs. The calculation of the attention value is divided into three main stages, as shown in Figure 1.

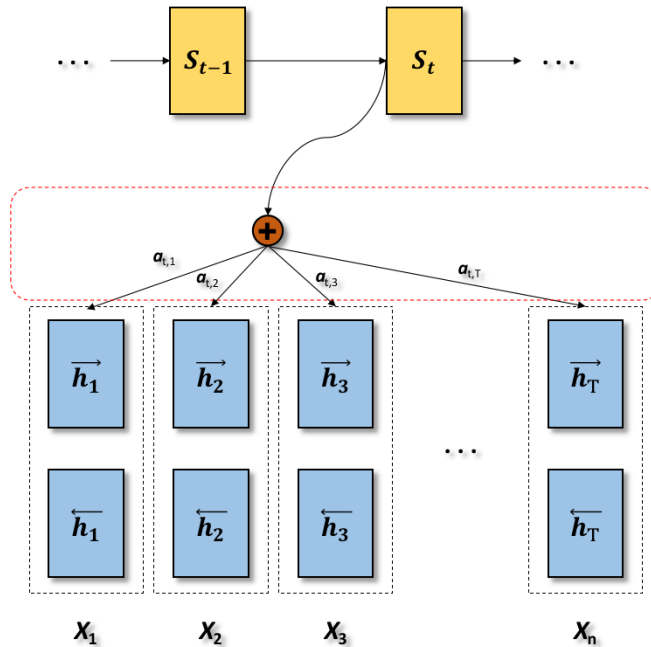


**Figure 1:** Calculation on attention.

Choi et al. [9] applied the attention mechanism to the field of machine translation, proposing joint learning of translation and alignment together in an encoding-decoding framework. In the attention-based machine translation system, firstly, the input module is responsible for reading the information of the source language words and representing them in a distributed manner, with a feature vector associated with each word position; then the system performs retrieval based on the list of feature vectors; finally, the tasks are executed according to the content sequence, focusing on one or several contents at each time depending on the weights. In summary, the matrix provides a clear indication of the probability distribution of the importance corresponding to each English word in the sequence when an English word is translated.

Figure 2 shows that the input is a source sequence  $X$  of length  $n$ , and the output is a target sequence  $y_t$  of length  $m$ . A Bidirectional RNN is used as the encoder model with forward hidden states and backward hidden states. The context information of the words is used in the simple layer state,  $a_{t,i}$  is the alignment between the original sequence and the sequence to be completed, and the alignment scores are obtained using the *softmax* function. To further enhance the prospect of application of attention mechanism in the field of cross-language machine translation, scholars have made a lot of innovations to machine translation models based on the segmentation of deep learning techniques.

Guo et al. [10] proposed the Transformer architecture, where the overall structure of the model is entirely based on the AM, consisting of an encoder and a decoder, as shown in Figure 3. The encoding part of the Transformer is built up of six identical base layers. The base layer contains two sublayer structures, namely the multi-head self-attention layer and the location-sensitive fully connected feed forward layer.



**Figure 2:** Schematic diagram of the joint learning model.

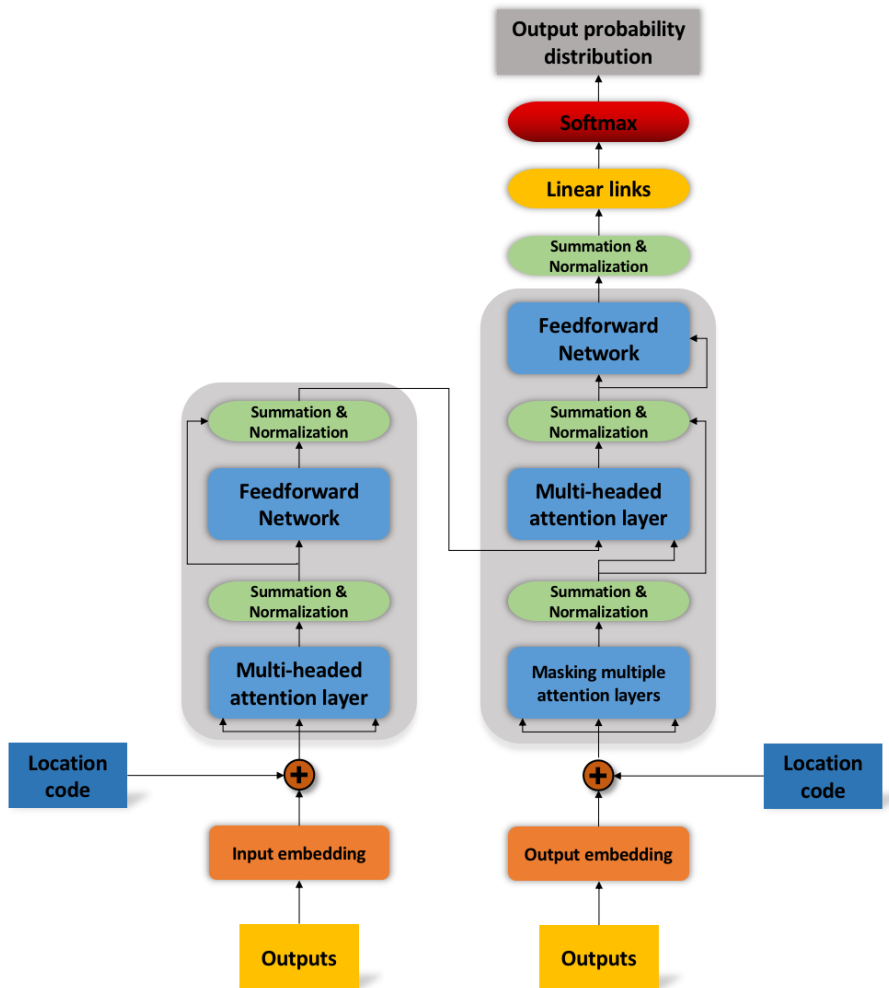
The decoder architecture of the Transformer model is similar to that of the Encoder, which consists of six basic layers stacked on top of each other, with each basic layer adding a new layer of multi-head self-attention layer in addition to the two sub-layers of self-attention layer and fully-connected feed-forward network layer. In addition to the two sub-layers of the multi-headed self-attentive layer and the location-sensitive feedforward network layer, a new multi-headed self-attentive layer is added to the output of the Encoder, also introducing residual connectivity and layer normalization.

However, the structural shortcomings of the Transformer model itself lead to two problems: first, the self-attentive mechanism calculates the attention between all words each time, which is computationally complex as the square of the input length; second, the pre-defined input length of the Transformer model leads to the limitation of capturing long-distance relations, and the segmentation of the input document leads to semantic fragmentation. The second is that the pre-determined input length of the Transformer model leads to limitations in capturing long-distance relations, and the segmentation of the input document leads to semantic fragmentation.

## 2.4 Reinforcement Learning Model

While deep learning and neural networks have led to breakthroughs in machine translation, Reinforcement Learning (RL) has made milestone breakthroughs in areas such as gaming. Reinforcement learning is a different learning paradigm from supervised learning, in which an intelligence interacts with its environment to achieve a learning goal. Its most important concepts include State, Action, and Reward. As shown in Figure 4, an intelligence receives a state from the environment, makes an action in response to that state, and the environment makes a reward in response to that action. The features of reinforcement learning include sequential decision making, trial and error, exploration of low probability events, exploitation of the best current strategy, and the best future payoff. Reinforcement learning is suitable for weakly supervised scenarios without explicit labeling, where probability exploration can be performed through a trial-and-error

adjustment mechanism and reward accumulation can be performed through a reward function. The combination of deep learning and reinforcement learning, applied to natural language understanding, can be used to model sequential problems by characterizing the state, action, and policy functions in reinforcement learning.



**Figure 3:** Transformer model structure schematic.

However, reinforcement learning still has some well-known limitations, such as the sparsity problem of rewards, the design of reward numbers, the high dimensionality of action space, and the instability of targets with large variance in training. Therefore, further research is needed to combine the idea of reinforcement learning and related algorithms with deep learning techniques for effective application in cross-language machine translation.



**Figure 4:** Reinforcement learning schematic.

## 2.5 Generative Adversarial Network Model

Generative Adversarial Network (GAN) is a generative model. The basic idea of GAN is inspired by game theory, which first obtains many training samples from a training library, and then learns these training cases to generate probability distributions [97]. The GAN model consists of a generative model (GM), which is responsible for computing the probability distribution of the samples, and a discriminant model (DM), which is used to estimate the probability that the samples are from different classes of corpus resources. GAN has the function of constantly generating new samples and is widely used in artificial intelligence, and has been widely applied in areas such as visual design and speech processing. Importantly, GAN offers new possibilities for machine translation, especially for processing complex data and computing complex probability distributions.

Both GM and DM are used to process complex, multidimensional and continuous corpus data, but GAN can optimize and adjust the parameter settings of GM to make the generated data more natural, realistic and close to the natural human expression. However, it should be emphasized that the parameters of image processing are traceable and any subtle changes can be captured by pixels, but the data of corpus resources processed by GM are discrete distribution and their changes cannot be changed by parameters, accordingly, the data output by DM is also meaningless information. In short, natural language processing is discrete in terms of probability distribution, and thus requires the mutual collaboration of GM and DM to optimize machine translation.

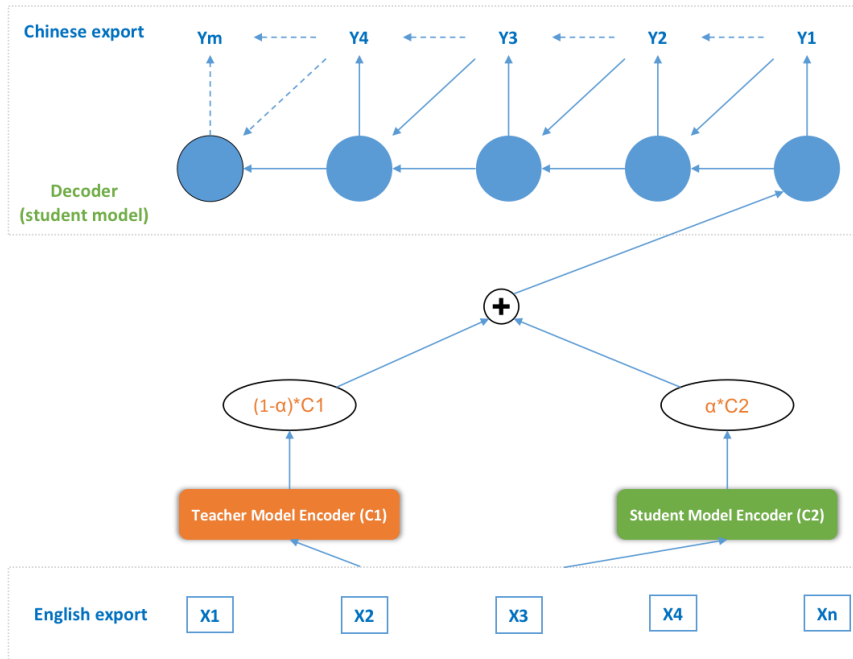
The SeqGAN model provides a solution to the above problem, the output of the discriminative model DM is used as Reward, and then the policy gradient method of reinforcement learning is used to train the generative model GM. Then, the complete sequence is sampled by an LSTM model for the partial generated sequences and then given to the discriminant model DM for scoring, and finally the Reward is averaged. In short, in order to get more stable training, faster convergence and higher quality samples, a lot of research has been done on GAN models to make them more widely used in NLP.

## 3 KNOWLEDGE DISTILLATION-BASED TRANSFER LEARNING MACHINE TRANSLATION TECHNIQUES

The application of migration learning has some promise. For example, when the amount of sample data is small, migration learning can continuously replenish the data and continue to expand the task in the original domain. On the other hand, when the model is more complex, it takes longer time to redo the model computation, and migration learning can accelerate the efficiency of model learning. Importantly, migration learning can port deep learning models from other domains in the context of a lack of corpus resources and fine-tune the models to suit their translation tasks

according to the corresponding task requirements. Thus, the greatest advantage of the transfer learning technique is the ability to continuously update and enrich the database, as well as the continuous optimization of the learning model.

However, there are some limitations of migration learning. For example, there is the problem of overfitting to the data in the process of learning model migration, and it is difficult to converge in computational iterations, especially when the parameters of the model are set too large, which is prone to infinite computation, and migration learning can be most effective only when the function setting of the machine translation task matches the resources of the corpus database. In fact, high-performance machine translation models, although capable of handling high-dimensional data and accomplishing complex tasks, are counterproductive and fail to fit when their performance is higher than the requirements of the translation task.



**Figure 5:** Knowledge distillation machine translation schematic.

Therefore, end-to-end neural network machine translation model design needs to focus on the problem of matching and fitting translation tasks and corpus data, and also on the improvement of translation task processing ability under the condition of corpus data scarcity. To this end, this chapter proposes a transfer learning machine translation technique based on knowledge distillation, in which a translation model (teacher model) is first trained on the resource-rich English parallel corpus data and then used to guide the English-Chinese translation model (student model) for training, taking Chinese (low resource) as an example. The over-fitting problem of neural network machine translation models under low-resource conditions is solved, and the generalization ability of the models is improved.

Knowledge distillation is an algorithm originally proposed in model compression and can be seen as an adaptive method based on regularization. The goal is to train a small and flexible model (the student model) to approximate the output of a large and accurate model (the teacher model). This algorithm can be used not only to reduce the complexity of the model but also to initialize the neural network model. Knowledge Distillation-based Domain Adaptation (KDA) can be implemented by 1) using the configuration and parameters of the teacher model as the initial



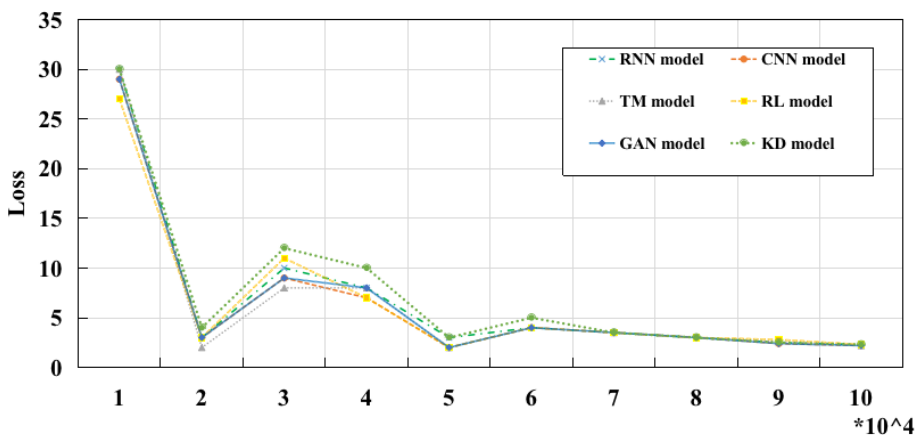
values of the student model, and 2) training the student model by minimizing the loss function on the adaptive data-set.

Applying knowledge distillation techniques to end-to-end neural network translation models has certain data requirements. Specifically, a database with a large number of corpus resources is needed to support the construction of teacher models. Also, a database of sales volume corpus resources is needed to drive deep learning of student models. Knowledge distillation technology, as an adaptive method to optimize and fine-tune the student model, provides technical support for self-feedback and adjustment of the neural network machine translation model. In addition, the optimization of the student model receives a moderation factor that prevents over-fitting problems and allows the knowledge of the teacher model to penetrate more precisely into the student model. Figure 5 illustrates a knowledge distillation-based machine translation model of the Bible network with English-Chinese translation as a case study.

The encoder-decoder machine translation model is first trained on a massively parallel corpus (English-Chinese) as the teacher model, and C1 is the representation of the output of the encoder of the teacher model. The C2 student model (English-Chinese) encoder uses the same network structure and initialization parameters as the teacher model, and is the output representation of the student model encoder. In the training, the same input source language data is substituted into the teacher model and the student model to obtain the encoder outputs C1 and C2, respectively, and  $(1-\alpha)C1 + \alpha C2$  is calculated as the total encoder output and substituted into the decoder of the student model for model training. The model is trained by adding a tuning factor to the decoder output and adjusting different values of  $\alpha$  and T to achieve the optimal model output.

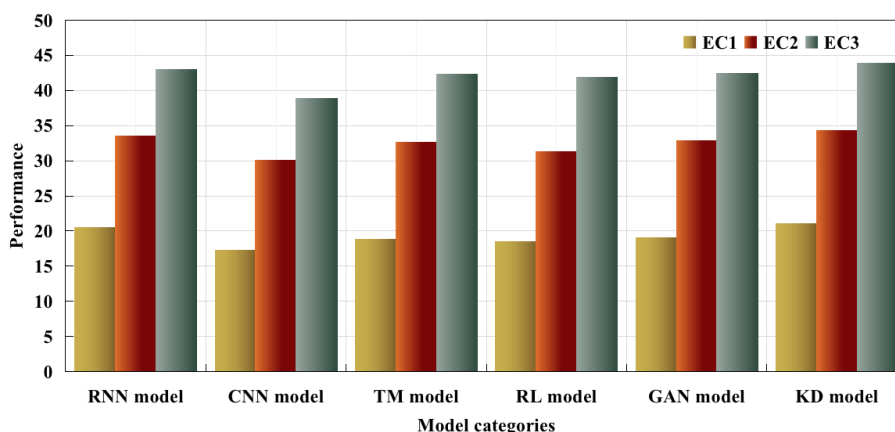
#### 4 ANALYSIS OF RESULTS

In the experiments in this chapter, we apply the knowledge distillation algorithm to an end-to-end neural network machine translation system (Knowledge Distillation Transformer Model Translation (KDTT)) based on the transformer model (Figure 6). The student and teacher models apply the same neutral network structure. The experiment was trained using 3000 tokens as a batch, with up to 1024 words per sentence, discarded in the decoder and encoder hold sections, and positions embedded in the encoder input in order to maintain the temporal information of the sentence. All models were implemented using Pytorch and trained using a single host containing 8 Nvidia P100 GPUs.



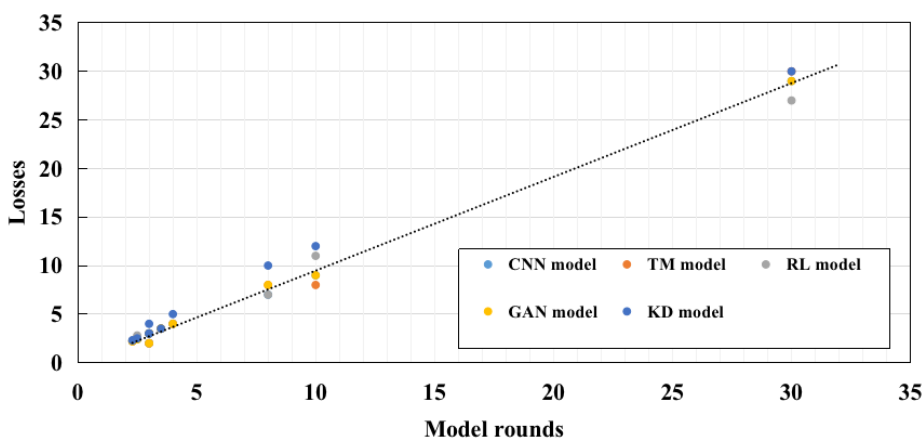
**Figure 6:** The loss curve of knowledge distillation machine translation.

As shown in Figure 7, we tested the English-Chinese translation models in three scenarios. English-Chinese translation model 1 has an adequate English corpus resource base but a scarce Chinese corpus resource base, English-Chinese translation model 2 has a scarce English corpus resource base but a rich Chinese corpus resource base, and English-Chinese translation model 3 has a rich English and Chinese corpus resource base, and it can be found that the analysis results have good cross-model stability.



**Figure 7:** Model performance.

Finally, we tested the fitting ability of different types of models, and it can be found that the loss curve satisfies the linear fit despite the jump from 10 to 30 training units, as shown in Figure 8. This indicates that the knowledge distillation technique does not have an over-fitting problem.



**Figure 8:** Model linear prediction.

## 5 CONCLUSION

The paper first identifies the specific problems of deep learning models such as RNN, LSTM RNN, GRU, CNN, Transformer model, RL and GAN, and proposes a knowledge distillation-based migration learning machine translation model for these problems. The problem of low resource

machine translation system is that it is difficult to train the neural network to get converged models when the parallel corpus is scarce. This paper applies transfer learning techniques to machine translation tasks and construct a teacher model with rich corpus resources and a student model with less corpus resources based on knowledge distillation methods, and propose a logical basis for the teacher model to train the student model. More importantly, our study shows that the application of knowledge distillation technique in the adaptation optimization of neural network machine translation models can effectively prevent the overfitting problem in traditional models and improve the model self-feedback capability. The final experimental analysis results also show that the performance of the knowledge distillation-based machine translation approach is still advantageous under low-resource conditions. It should be emphasized that machine translation technology update is a systematic project, which needs to invest a large amount of corpus resources data support, and the technical improvement of machine translation model in this paper needs more data test, and the current data is not enough to fully illustrate the advantages of the model. Secondly, the technical evaluation of the existing machine translation model in this paper is still not comprehensive enough, and future research can conduct a more extensive literature comb. Finally, this paper lacks discussion on the specific application path of the model in the field of English translation teaching, and future research should continue to discuss in depth the ways of applying different machine translation models in classroom practice.

## 6 ACKNOWLEDGEMENTS

Shanxi Philosophy, Social and Science Planning Office<Research on the Colleges Foreign Language Teachers' Professional Development in Shanxi Province at the Transition Period> (Project Number:2020yj218)

Wenjie Li, <https://orcid.org/0000-0001-7384-2952>  
Huiqin Liu, <https://orcid.org/0000-0003-3471-3618>

## REFERENCES

- [1] Zhang, J.; Zong, C.: Neural Machine Translation: Challenges, Progress and Future, *Science China Technological Sciences*, 63, 2020, 1-23. <https://doi.org/10.1007/s11431-020-1632-x>
- [2] Zamora-Martinez, F.; Castro-Bleda, M.-J.: Efficient Embedded Decoding of Neural Network language models in a machine translation system, *International Journal of Neural Systems*, 28(9), 2018, 1850007. <https://doi.org/10.1142/S0129065718500077>
- [3] Zou, W.; Xia, Y.: Back Propagation Bidirectional Extreme Learning Machine for Traffic Flow Time Series Prediction, *Neural Computing and Applications*, 31(11), 2019, 7401-7414. <https://doi.org/10.1007/s00521-018-3578-y>
- [4] Naseer, A.; Zafar, K.: Meta Features-based Scale Invariant OCR Decision Making Using LSTM-RNN, *Computational and Mathematical Organization Theory*, 25(2), 2019, 165-183. <https://doi.org/10.1007/s10588-018-9265-9>
- [5] Qi, F.; Lin, C.; Shi, G.; & Li, H.: A Convolutional Encoder-Decoder Network with Skip Connections for Saliency Prediction, *IEEE Access*, 7, 2019, 60428-60438. <https://doi.org/10.1109/ACCESS.2019.2915630>
- [6] Yang, D.; Karimi, H.-R.; Sun, K.: Residual Wide-kernel Deep Convolutional Auto-Encoder for Intelligent Rotating Machinery Fault Diagnosis with Limited Samples, *Neural Networks*, 141, 2021, 133-144. <https://doi.org/10.1016/j.neunet.2021.04.003>
- [7] Sun, J.; Han, P.; Cheng, Z.; Wu, E.; Wang, W.: Transformer Based Multi-Grained Attention Network for Aspect-Based Sentiment Analysis, *IEEE Access*, 8, 2020, 211152-211163. <https://doi.org/10.1109/ACCESS.2020.3039470>

- [8] DeRose, J.-F.; Wang, J.; Berger, M.: Attention Flows: Analyzing and Comparing Attention Mechanisms in Language Models, *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 2020, 1160-1170. <https://doi.org/10.1109/TVCG.2020.3028976>
- [9] Choi, H.; Cho, K.; Bengio, Y.: Fine-Grained Attention Mechanism for Neural Machine Translation, *Neurocomputing*, 284, 2018, 171-176. <https://doi.org/10.1016/j.neucom.2018.01.007>
- [10] Guo, Q.; Qiu, X.; Xue, X.; Zhang, Z.: Low-rank and locality constrained self-attention for sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12), 2019, 2213-2222. <https://doi.org/10.1109/TASLP.2019.2944078>