applied sciences

MDPI

Article

A Mongolian-Chinese Neural Machine Translation Model Based on Soft Target Templates and Contextual Knowledge

Qing-Dao-Er-Ji Ren, Ziyu Pang * and Jiajun Lang

School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China; renqingln@imut.edu.cn (Q.-D.-E.-J.R.); langjiajun.he@ccb.com (J.L.)
* Correspondence: 20211800078@imut.edu.cn

Abstract: In recent years, Mongolian-Chinese neural machine translation (MCNMT) technology has made substantial progress. However, the establishment of the Mongolian dataset requires a significant amount of financial and material investment, which has become a major obstacle to the performance of MCNMT. Pre-training and fine-tuning technology have also achieved great success in the field of natural language processing, but how to fully exploit the potential of pre-training language models (PLMs) in MCNMT has become an urgent problem to be solved. Therefore, this paper proposes a novel MCNMT model based on the soft target template and contextual knowledge. Firstly, to learn the grammatical structure of target sentences, a selection-based parsing tree is adopted to generate candidate templates that are used as soft target templates. The template information is merged with the encoder-decoder framework, fully utilizing the templates and source text information to guide the translation process. Secondly, the translation model learns the contextual knowledge of sentences from the BERT pre-training model through the dynamic fusion mechanism and knowledge extraction paradigm, so as to improve the model's utilization rate of language knowledge. Finally, the translation performance of the proposed model is further improved by integrating contextual knowledge and soft target templates by using a scaling factor. The effectiveness of the modified model is verified by a large number of data experiments, and the calculated BLEU (BiLingual Evaluation Understudy) value is increased by 4.032 points compared with the baseline MCNMT model of Transformers.

Keywords: neural machine translation; pre-training; contextual knowledge; soft target template

Citation: Ren, Q.-D.-E.-J.; Pang, Z.; Lang, J. A Mongolian-Chinese Neural Machine Translation Model Based on Soft Target Templates and Contextual Knowledge. *Appl. Sci.* **2023**, *13*, 11845. https://doi.org/10.3390/app132111845

Academic Editor: Andrea Prati

Received: 17 October 2023
Revised: 27 October 2023
Accepted: 28 October 2023
Published: 30 October 2023

1. Introduction

Translation, as a bridge and link between different countries and national cultures, is of great importance. With the economic and social development of the Inner Mongolia Autonomous Region of China and the increasingly close exchanges and cooperation between Mongolia and China, human translation can no longer meet the needs of massively accurate translation. Therefore, machine translation, with its excellent translation quality and speed, has gradually become the mainstream of current information transmission, which has considerable social and practical significance for improving the dissemination and exchange of national culture and breaking down different cultural barriers.

With the advent of the world's first electronic digital computer, the ENIAC (Electronic Numerical Integrator and Computer), Warren first proposed the idea of document translation through computers in 1947 and officially proposed the concept of machine translation in 1949 [1]. In 2003, Université de Montréal's Kandola et al. [2] used neural network models to improve the traditional n-gram model and the statistical language model to develop a new language model that can utilize longer contexts to achieve better results in the text corpus. In 2017, Arbi Haza Nasution et al. [3] proposed a constraint-based bilingual vocabulary induction method in view of the lack of parallel corpora and comparable corpora. This method extends the constraints of recent fulcrum-based induction techniques and further enables multiple Symmetry assumption loops to reach more cognate words in the

transgraph. Compared with previous methods, the constraint-based method proposed in this paper has a statistically significant improvement in accuracy and f-score. The results show that this approach has the potential to complement other bilingual dictionary creation methods. In 2020, the Google team released the Reformer model on ICLR [4], which reduces the spatial complexity by using Locality-Sensitive Hashing (LSH), blocks the feed-forward model of Transformers with a reversible residual network (RevNet), and solves the problem that the memory cost of the residual network is proportional to the number of network units in this model. In 2022, Wang et al. [5] presented a template-based approach to regenerate constraint and non-constrained tokens through templates. The generation and derivation of templates can be learned through a sequence-to-sequence training framework to produce high-quality translations while maintaining decoding speed. In 2022, Li et al. [6] utilized an external parser to parse the source sentence to obtain dependency parsing data, then converted the dependency parsing data into a parent word position vector and a child word weight matrix, and finally integrated the dependency knowledge into the multi-head attention mechanism of the Transformer encoder. Their experimental results indicated that the bidirectional self-attention mechanism can provide richer dependency information to effectively improve the translation performance of the model.

Integrating linguistic information into neural machine translation models and using existing linguistic knowledge to alleviate the inherent difficulties faced by neural machine translation and improve translation quality has become a hot topic in the field of neural machine translation research. In 2022, Guarasci Raffaele et al. [7] studied the ability of the multilingual BERT (mBERT) language model to transfer syntactic knowledge across languages, using structural probes to reconstruct the dependency parse tree of sentences and using context embeddings from the mBERT layer to represent the input sentences. The results of the experimental evaluation show that the grammatical knowledge of the mBERT model can be transferred between languages. Transferring grammatical knowledge not only meets theoretical needs in the case of specific phenomena but also has important practical significance in syntactic tasks (such as dependency parsing). In the same year, Yulia Otmakhova et al. [8] used BERT as a pre-trained model to compare how three different types of languages (English, Korean, and Russian) encode different layers of morphological and grammatical features. The experimental results largely explain that the layers of the model follow the so-called classic NLP pipeline principles, with lower levels specifically processing part-of-speech and other morphological information, middle layers responsible for more complex syntactic relationships, and higher levels dealing with higher-level linguistic phenomena such as anaphora and reference. In 2023, AG Varda et al. [9] studied the inner workings of mBERT and XLM-R to test the performance of single neural units responding to precise grammatical phenomena (i.e., number agreement) in five languages (English, German, French, Hebrew, and Russian). Cross-language consistency. Experimental results show that there is a large overlap in the underlying dimensions of encoding consistency in these languages and that the overlap in XLM-R short-range consistency is larger than that of mBERT and peaks in the middle layers of the network.

The Uighur Mongolian script, which has a history of about 800 years, is the first script mastered by the Mongolian people. However, Mongolian-Chinese machine translation (MCMT) started late, and the parallel corpora are limited. The cutting-edge research on this topic is relatively weak, so there is still great room for improvement.

In response to the scarcity of parallel corpus resources in MCMT, in 2021, Zhang Zhen et al. [10] modeled machine translation by introducing three language pre-training models. In the data pre-processing stage, they introduced two new unsupervised pre-training methods and one supervised pre-training method for cross-border Language modeling is used to learn cross-language representations, and the effects of three language pre-training methods in Mongolian-Chinese translation are studied. Experimental results show that the three cross-language pre-trained models significantly reduce the perplexity of low-resource languages and increase the BLEU value by 20.4 compared with random initialization parameters, improving the quality of Mongolian-Chinese translation. In 2022,

He and Wang [11] took full advantage of agglutination and rich morphological changes in Mongolian, as well as Mongolian grammar rules, to build a neural word segmentation method in which the iLSTM-CNN-CRF neural network was used to filter the Mongolian connecting vowels and unstable "N" as stop words to improve the quality of machine translation, and the BLEU (BiLingual Evaluation Understudy) value in their obtained translations reached 73.30%.

Mongolian is also the same as other languages. Each region has its own dialect characteristics. Approximately 80% of Mongolia is Khalkha Mongolian, so the Mandarin in Outer Mongolia is Khalkha Mongolian. On this basis, many Russian and English words are borrowed. The situation in Inner Mongolia is much more complicated, and the accents between the East and the West are quite different. But generally speaking, their mother tongue is Mongolian (Cyrillic Mongolian: Монгол хэл), which belongs to the Mongolian family of the Altaic language family. It is a pinyin script. Its two writing methods—horizontal writing and vertical writing—do not affect reading. This article is written in Mongolian horizontally.

So far, researchers have made gratifying progress in the study of Mongolian-Chinese machine translation. However, the research conditions are limited, and the construction of the Mongolian-Chinese parallel corpus is not perfect, resulting in a large number of low-frequency words. The traditional Mongolian word formation method is complex, and the direct use of neural networks for translation has poor results. Based on the above problems, this paper presents a Mongolian-Chinese neural machine translation (MCNMT) model based on soft target templates and contextual knowledge and verifies the effectiveness of the proposed algorithm through data experiments, indicating that it can alleviate the problem of less parallel data available in the Mongolian-Chinese neural machine translation tasks.

2. BERT Pre-Training Model

Bidirectional Encoder Representation from Transformers (BERT) [12], as the name suggests, is the Encoder representation of bidirectional Transformers that models polysemy by learning embedding forms of each word through a large number of corpora to learn context-independent semantic vector representations. The architecture of the BERT model is shown in Figure 1, where Trm represents the encoder structure in the Transformer, meaning that the BERT model is mainly composed of multiple encoder parts of Transformers.

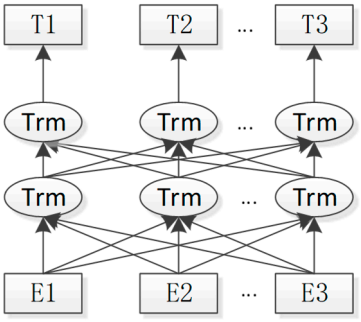


Figure 1. Bert model architecture.

The BERT model differs from the ELMo [13] model and the GPT [14] model in that the ELMo model adopts bidirectional LSTMs, while the BERT model uses bidirectional Transformers, so that the BERT model has superior feature extracting capability; the GPT model's Transformers, which use Decoder models, are unidirectional, but the BERT model's Transformers, which use Encoder models, are bidirectional.

BERT is a context-based pre-training model, and like other pre-training models, it has two steps: pre-training and fine-tuning. In the pre-training stage, the BERT model requires a large amount of text for pre-training. In the fine-tuning stage, it is necessary to modify the model structure according to different downstream tasks and readjust the model parameters through specific task samples. In the pre-training stage, the BERT model utilizes

two pre-training tasks, namely the Masked Language Model (MLM) and Next Sentence Prediction (NSP), to achieve more applications. The MLM model learns word embeddings of samples, and the NSP learns sentence embeddings of samples. In the Transformer model, the word vector and the position vector of the input sentence are added up to obtain a word vector to express the source language information. Inspired by this, the input vector of the BERT model is calculated by adding the three vectors of Token Embedding learned in the MLM task, Segment Embedding learned in the NSP task, and Position Embedding representing sentence position information, as shown in Figure 2:

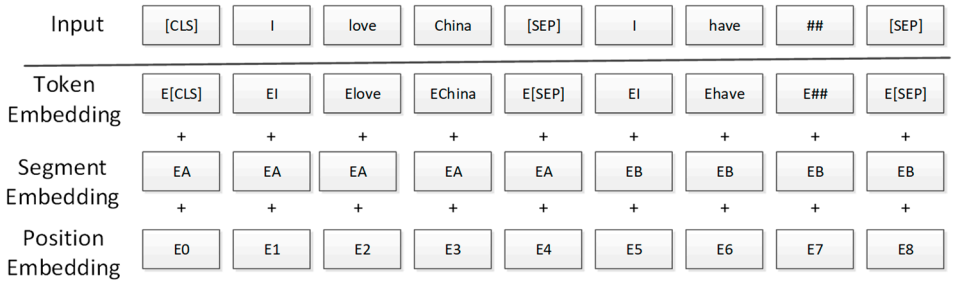


Figure 2. Word vector of bert model.

Where the Position Embedding represents a position vector, and unlike the Transformer model, the position vector in BERT is determined by model learning. The Segment Embedding is employed to distinguish between two sentences in an LSP task in the pre-training stage. The Token Embedding can be used for subsequent classification tasks. The [CLS] token that can be understood as a vector representation of all input features is the first word of the Token Embedding; The [SEP] is an identifier for distinguishing different sentences.

The principle of the MLM tasks in pre-training is to randomly replace some words with [Mask], and then predict them through their contextual information during the training process of the model. For the sentence "He holds a dictionary in his hand" as an example, if the model selects "dictionary" in the sentence, the original sentence will be replaced with "He holds a [MASK]", and the BERT model can predict that the "[MASK]" here is a "dictionary" by training.

The mask rule of LML is that the probability of each word being randomly replaced is 15%, and among the 15% probability, there are three [MASK] cases:

- There is an 80% chance that the selected word will be replaced with [MASK], e.g., "He holds [MASK] in his hand";
- There is a 10% chance that the selected word will be replaced with a random word, e.g., replacing with bench: i.e., "He holds a bench in his hand";
- There is a 10% chance that the selected word will remain unchanged, e.g., "He holds a dictionary in his hand".

The purpose of introducing this rule is to let the model know that the token in the replaced position can be any word and not pay too much attention to the token, so as to promote the model to learn more contextual information. The input of the Next Sentence Prediction task in pre-training is two sentences, and the training purpose is to enable the BERT model to identify whether a sentence is the next sentence of the previous one so that the model can understand the relationship between the two sentences.

3. Soft Object Templates and Contextual Knowledge Based Mongolian-Chinese Neural Machine Translation Model

Most of the neural machine translation (NMT) models rely only on parallel sentence pairs, whereas the performance of such models will drop dramatically in the case of insufficient resources because they cannot mine the language of the corpus. Merging monolingual knowledge such as grammar has proven to be effective for NMT, especially in under-resourced conditions. However, existing methods do not fully exploit the potential

of the NMT architecture for efficiently utilizing external prior knowledge. Therefore, in order to increase the utilization rate of the NMT model on the grammatical knowledge of the pre-training model and improve the translation performance of the NMT model on the Mongolian-Chinese low-resource corpus dataset, this paper constructs soft target templates and contextual knowledge-based MCNMT models, and the process of modeling can be divided into three major steps. Firstly, a baseline Transformer model is used to extract soft target templates; Secondly, the obtained soft target templates are fused by improving the baseline model structure; Finally, the contextual knowledge of the corpus is extracted from the pre-training model by using the dynamic fusion mechanism and knowledge extraction paradigm and integrated into the MCMT model to construct a new MCNMT model.

3.1. Soft Target Template Extraction

The extraction process for soft target templates is divided into two steps: (1) parsing the target language (Chinese) using a syntax parsing tree to obtain the soft template from it; and (2) training the baseline Transformer model using parallel data composed of source language text and the soft template obtained in the previous step to extract the soft target template.

In this paper, the sentence-based selection parsing tree is first extracted by using the natural language processing tool Standford CoreNLP system v4.5.3, and then the nodes beyond the set depth are pruned off, and finally the clipped subtree is restored in the original order to obtain the template data. The process of obtaining a template from the parsing tree is demonstrated in Figure 3. The syntax parsing tree can display the structural and grammatical information of the Chinese text sentence, and the template extracted by the pruning operation consists of terminal nodes and non-terminal nodes. Taking the sentence "There are many people running here" as an example, the tree structure generated by the syntax parsing tree has a terminal node set S = {Here, there are, many, people, running} and a non-terminal node set V = {S, NP, VP, EX, VBP, DT, NNS, VBG}. The final extracted template is that "There are NP and VP here".

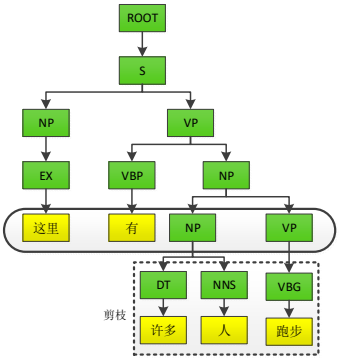


Figure 3. Constituency-Based Parse Trees. The figure shows the tree structure generated by the grammar parsing tree. Given a target sentence and a determined tree depth, we obtain subtrees by pruning nodes. The target sentence is "There are many people running here." The subtree can then be converted from left to right into the soft target template "There is NP VP".

The depth of the parsing tree based on selection areas is determined by the following simple but effective strategy:

$$d = \min(\max(L \times \lambda, \gamma_1), \gamma_2) \quad (1)$$

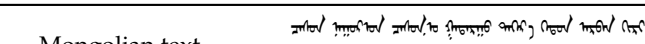
where L is the length of the input sentence, γ_1 is the lower limit depth of the subtree, γ_2 is the upper limit depth of the subtree, and λ is the ratio parameter of the length of the source sentence.

In order to extract the soft target template, this paper uses the Transformer [15] baseline model to train the Mongolian text and extracts the soft target template with soft

temple parallel data obtained by the parsing tree above, with the same parameters as the baseline model. The encoder and decoder read the Mongolian text and soft template data, respectively, and encode them into the vector representation. During the model training process, the soft target template is predicted by using beam search, and the top-k of beam search is set to 1. That is, for each input of a Mongolian sentence, only one corresponding soft target template is extracted and output as the result. The extracted data can be used as training data for the next stage of the experiment.

Some results of the extracted soft target templates through the Transformer model are listed in Table 1:

Table 1. Examples of Soft Target Template.

Mongolian text	 (There are dozens of large and small snow pits on the snow surface of Snow Lake.)
Chinese text	雪湖的雪面上有数十个大大小小的雪坑 (There are dozens of large and small snow pits on the snow surface of Snow Lake.)
Soft target template	雪湖的 LCP 有 CD ADJP (Snow Lake's LCP has CD ADJP.)
Mongolian text	 (The Environmental Protection Agency has recently established limits on some harmful emissions.)
Chinese text	环境保护署最近规定了一些有害的排放物的限制 (The Environmental Protection Agency has recently established limits on some harmful emissions.)
Soft target template	NR 最近 VP 一些VA 排放物的限定 (NR recently VP limits some VA emissions.)

3.2. Dynamic Fusion Mechanism

In the dynamic fusion mechanism proposed in this paper, all layers in the BERT pre-training model are represented by using a multilayer perceptron ($G_l(\cdot)$). Let the source language sequence be $x = (x_1, x_2, \ldots, x_I)$ and the target language sequence be $y = \left(y_1, y_2, \ldots, y_J\right)$, where I and J denote the lengths of x and y respectively.

In the Transformer model, the encoder encodes the source language sequence as R_N^E, which consists of a series of $\left(r_{N,1}^E, r_{N,2}^E, \ldots, r_{N,i}^E, \ldots, r_{N,I}^E\right)$, where N is the depth of the encoder. R_N^E can be calculated by:

$$\begin{aligned} H_N^E &= \mathrm{Att}\left(Q_N^E, K_{N-1}^E, V_{N-1}^E\right) \\ R_N^E &= \mathrm{LN}\left(H_N^E + \mathrm{FNN}\left(R_{N-1}^E\right)\right) \end{aligned} \tag{2}$$

In the BERT pre-training model, the source language sequence is encoded as R^P, with $R^P = \left(R_1^P, \ldots, R_l^P, \ldots, R_L^P\right)$, where L represents the number of layers of the pre-training model. The representation R_l^P of the l-th layer is given by the following multilayer perceptron:

$$R_l^T = G_l\left(R_l^P\right), \tag{3}$$

On the encoder end of the Transformer model, since the attention to information of each layer is different, specific contextual knowledge from all layer representations is obtained through the layer perception attention mechanism, and its calculation formula is as follows:

$$e_l = \mathrm{FFN}\left(\frac{1}{I}\sum_{i=1}^{I} r_{l,i}^T \cdot \frac{1}{I}\sum_{i=1}^{I} r_{n,i}^E\right) \tag{4}$$

$$a_l = \frac{\exp(e_l)}{\sum_{t=1}^{L}\exp(e_t)} \tag{5}$$

$$C_n^T = \sum_{l=1}^{L} a_l R_l^T \tag{6}$$

The layer perception attention mechanism can determine the importance of information in the pre-training model to the current layer, so as to obtain contextual knowledge that is more suitable for the current layer through C_n^T. Similarly, the attention to information of each hidden state in the same layer is different from each other. Therefore, a context gating mechanism is adopted to control the fusion rate of each hidden state, that is, the representation $c_{n,i}^T$ from C_n^T is fused into the corresponding state $r_{n,i}^E$ of R_n^E, and the calculation formulas are as follows:

$$\gamma_{n,i} = \text{sigmoid}\left(\text{FFN}\left(r_{n,i}^E \cdot c_{n,i}^T\right)\right) \tag{7}$$

$$\bar{r}_{n,i}^E = r_{n,i}^E + \gamma_{n,i} * c_{n,i}^T \tag{8}$$

The conceptual diagram of the dynamic fusion mechanism at the encoder end of the transformer model is shown in Figure 4:

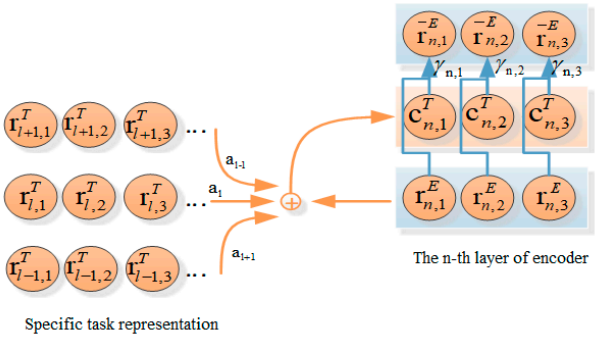


Figure 4. Dynamic Fusion Mechanism at encoder end.

3.3. Knowledge Extraction Paradigm

In our proposed translation model, the contextual knowledge extracted during BERT pre-training is integrated through the dynamic fusion mechanism on the encoder end. In addition, the knowledge extraction paradigm is also added on the decoder end to facilitate the translation model in learning the specific task representation in the pre-training process.

In the decoding stage, the model generates the j-th word by maximizing the conditional probability using the following formulas:

$$\begin{aligned} C_M^D &= \text{Att}\left(Q_M^D, K_N^E, V_N^E\right) \\ R_M^D &= \text{LN}\left(\text{FFN}\left(S_M^D + C_M^D\right)\right) \\ P\left(y_j \middle| y_{<j}, x\right) &= \text{softmax}\left(\text{FFN}\left(r_{M,j}^D\right)\right) \end{aligned} \tag{9}$$

where, D represents the decoder, M is the number of layers of the decoder, Q_M^D is from the output S_M^D of the previous layer of the decoder, S_M^D is calculated by Equation (2), with K, Q, and V being from the R_{M-1}^D of the previous layer of the decoder, and K_N^E and V_N^E are from the output R_N^E of the encoder.

The final optimization function of the translation model is:

$$\mathcal{L}_T = \frac{1}{J}\sum_{i=1}^{J} \log P\left(y_j \middle| y_{<j}, x; \theta_T\right) \tag{10}$$

where θ_T is a parameter of NMT.

In terms of word-level granularity, the knowledge extraction paradigm mainly learns the output distribution of the model from the pre-training model, and the training function can be represented by:

$$\mathcal{L}_W = \frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{V} P\left(y_j = k \middle| y; \theta_P\right) \cdot \log\left(P\left(y_j = k \middle| x, y_{<j}; \theta_T\right)\right) \tag{11}$$

where J is the length of the given target sentence y, and V is the vocabulary. $P\left(y_j = k \middle| x, y_{<j}; \theta_T\right)$ is calculated by Equation (9).

In terms of sentence-level granularity, the contextual knowledge of the sentence is learned by direct fitting, and the training function is written as:

$$\mathcal{L}_S = \frac{1}{J}\left|\left|R_M^D - R_L^P\right|\right|_2^2 = \frac{1}{J}\sum_{j=1}^{J}\left|\left|r_{M,j}^D - r_{L,j}^P\right|\right|_2^2 \tag{12}$$

where M represents the output layer of the decoder, and $r_{M,j}^D$ and $r_{L,j}^P$ are from the decoder and the pre-training model, respectively.

Finally, the translation model is optimized for training by fitting three loss functions through Formula (13):

$$\mathcal{L} = \mathcal{L}_T + \eta\mathcal{L}_S + \beta\mathcal{L}_W \tag{13}$$

where η and β are hyperparameters for balancing the fusion granularity of the word-level and sentence-level knowledge extraction paradigms, and they both are set to 0.5 in this work.

The conceptual diagram of the knowledge extraction paradigm adopted by the Transformer model on the decoder end is presented in Figure 5:

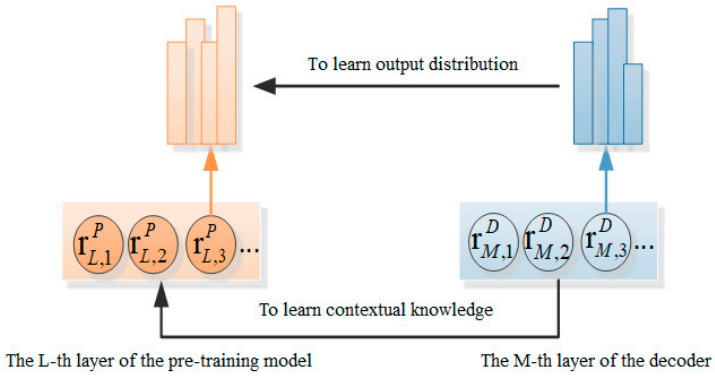


Figure 5. Knowledge Extraction Paradigms at the decoder end.

3.4. MCNMT Model Based on Soft Target Templates and Contextual Knowledge

The concept of the MCNMT model based on soft target templates and contextual knowledge is inspired by the guided translation model, that is, the source language data and the template with the target language grammar information are fused by the decoder, and the template is used to guide the model to improve the utilization rate of grammar knowledge. The contextual knowledge of the target language extracted from the BERT pre-training model is integrated into the model to improve the translation performance.

An example of the template-guided model translation is shown in Figure 6:

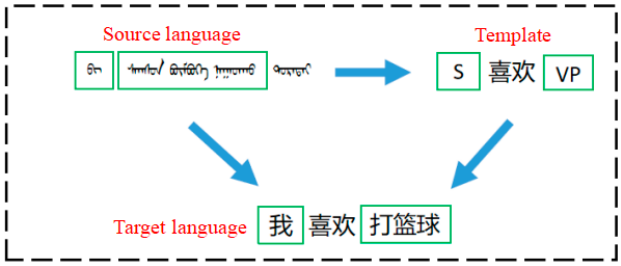


Figure 6. Example of Template-Guided Model Translation. The Mongolian language in the picture means "I like playing basketball". Through the guidance of the template, the target Chinese language is finally obtained. S represents the subject and VP represents the verb phrase.

On the basis of the baseline Transformer model, an additional Transformer encoder structure named soft target template is introduced on the encoder end. It has the same composition structure as the Transformer encoder of the source language and is used to encode soft target template data and convert it into hidden layer vectors. At the same time, the dynamic fusion mechanism and knowledge extraction paradigm are integrated into the model in a fusion manner to improve the utilization of contextual knowledge. Figure 7 clearly illustrates the use of integrating various contextual knowledge from the BERT pre-training model into the NMT model.

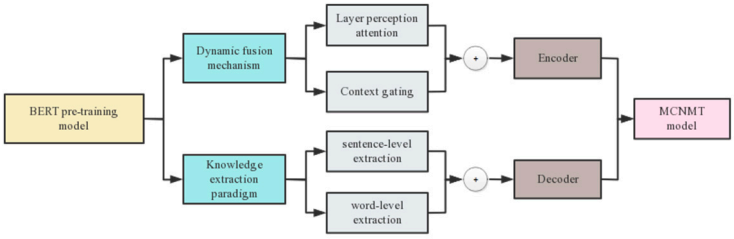


Figure 7. Obtaining Contextual Knowledge Using the BERT Pre-training Model.

From Figure 7, the dynamic fusion mechanism uses the layer perception attention mechanism and context gating mechanism to respectively extract attentions between different layers and the same layer in the pre-training model and applies them to the encoder end; the knowledge extraction paradigm employs word-level extraction and sentence-level extraction to obtain the output distribution and contextual knowledge of the pre-training model, respectively, and applies them to the decoder end.

The overall structure of the model is presented in Figure 8:

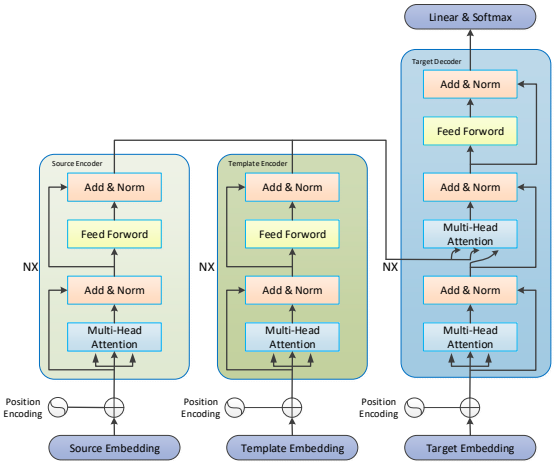


Figure 8. Model Architecture of a Mongolian-Chinese Neural Machine Translation Model Fused with Soft Target Templates.

The formula for generating Chinese text from Mongolian text with soft templates is as follows:

$$P(Y|X) = P_{\theta_{X->T}}(T|X)P_{\theta_{(X,T->Y)}}(Y|X,T) \tag{14}$$

where, $\theta_{X->T}$ is the parameter of the prediction soft target template Transformer model, and $\theta_{(X,T->Y)}$ is the parameter of the fusion soft target template model.

The generation of Chinese translation by the Transformer decoder is based on the hidden state of the source language encoder and the soft target template encoder. Two different sets of attention parameters are used in the decoder's Encoder-Decoder multi-head attention to process two different encoders, respectively, and the hidden layer vectors containing the source language information and these having the soft target template information are fused through the gating unit. The formulas are as follows:

$$\beta = \sigma\left(W_Y Z^{X,Y} + U_T Z^{X,T}\right) \tag{15}$$

$$Z = \beta Z^{X,Y} + (1-\beta) Z^{T,Y} \tag{16}$$

where, W_Y and U_T are the parameter matrices, σ is the Sigmoid activation function, Z is the hidden state of the decoder, and β is the control parameter of the fusion degree of Mongolian text and templates.

In this paper, the model parameters are updated using the maximum likelihood estimation algorithm. When the translation model is trained without a soft target template, the optimization formula for the loss function is:

$$L_{\theta_{X->Y}}(D) = \sum_{X,Y \in D} \log P_{\theta_{X->Y}}(Y|X) \tag{17}$$

When the translation model is trained with a soft target template, the optimization formula for the loss function is:

$$L_{\theta_{(X,T)->Y}}(D) = \sum_{X,Y \in D} \log P_{\theta_{(X,T)->Y}}(Y|X,T) \tag{18}$$

Since low-quality soft templates are inevitable when constructing soft target template data, and their noise will affect model training and reduce the translation quality, a parameter called scale factor is introduced into the training process of the model to optimize these two loss functions at the same time. The soft target template is used in part $(1-\alpha)$ while contextual knowledge is applied in part α, which makes the model stable, and the formula is as follows:

$$L_\theta(D) = \alpha\left(\mathcal{L}_{\theta_{X->Y}}(D) + \eta\mathcal{L}_S(D) + \beta\mathcal{L}_W(D)\right) + (1-\alpha)L_{\theta_{(X,T)->Y}}(D) \tag{19}$$

where, α is the fusion ratio of soft target template and contextual knowledge, taken as 0.5; η and β are the proportion of word-level context and sentence-level context in contextual knowledge, and they both are taken as 0.5; $\mathcal{L}_S(D)$ and $\mathcal{L}_W(D)$ are given by Equations (11) and (12), respectively.

The strategy of the MCNMT model based on soft target templates and contextual knowledge is demonstrated in Figure 9:

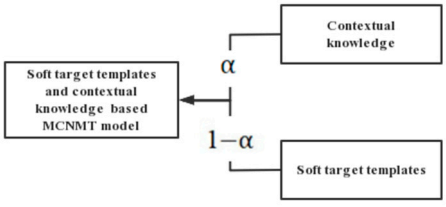


Figure 9. Integration Strategy of the Mongolian-Chinese Neural Machine Translation System.

4. Experiments and Analysis

4.1. Experimental Procedure

The experimental process of this work is shown in Figure 10:

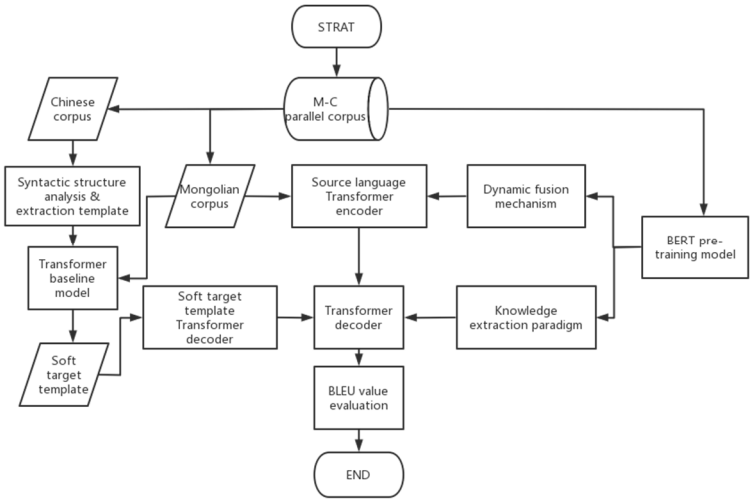


Figure 10. Experiment Flow Chart.

4.2. Evaluation Indexes

The Bilingual Evaluation Understudy (BLEU) method proposed by IBM is used to evaluate machine translation models. BELU is an accuracy calculation model based on N-garm that uses the cumulative BLEU values as the final reference to calculate the weighted average value of BLEU-1, BLEU-2, . . . , BLEU-N. The larger N and the bigger the BLEU value, the higher the accuracy of the translation model. The BLEU value is calculated by:

$$\mathrm{BLEU} = \mathrm{BP} * \exp\left(\sum_{n=1}^{N} w_n \log P_n\right) \tag{20}$$

$$\mathrm{BP} = \begin{cases} 1 & \begin{pmatrix} \text{if } c > r \end{pmatrix} \\ e^{1-\frac{r}{c}} & \begin{pmatrix} \text{if } c < r \end{pmatrix} \end{cases} \tag{21}$$

where w_n is the weight corresponding to different n-grams; P_n is the probability calculated after parsing the sentence using different n-grams; c is the length of the candidate sentence; and r is the length of the reference translation.

To facilitate observation, the BLEU value is generally expanded by 100 times. The BLEU value of 4 g is used as the evaluation index in this paper.

4.3. Data Pre-Processing

4.3.1. Data Cleaning

This work uses the home-made 1.26-million-line Mongolian-Chinese parallel corpus by the Artificial Intelligence and Pattern Recognition Laboratory of Inner Mongolia University of Technology in China, which covers a wide range of fields, including news records, medical reports, proper nouns or phrases, two-part allegorical sayings, daily conversations or online chats, excerpts of literary works, terms and nouns related to the computer field, etc. Many of these bilingual parallel sentence pairs are irregular sentence pairs that need to be modified or deleted. After cleaning and processing, including data denoising, text standardization, and text deduplication, a total of about 20,000 sentence pairs that are irregular and difficult to meet the requirements of the specification are deleted, leaving 1,244,139 lines.

4.3.2. Word Segmentation

The NMT model is limited by objective conditions such as video memory capacity during training, and its model size is usually controlled. Hyperparameters directly related to the model scale include word embedding dimensions, sentence length, corpus dictionary size, the constituent units and number of layers of the encoder and decoder, etc. Take a dictionary as an example. Before construction, a smaller frequency of occurrence of a word is usually set as a limit. Words with a frequency lower than the limit are marked with UNK. The size of the dictionary is generally limited to about 30,000 to 50,000 words. Between them, the translation model cannot recognize the "UNK" information during the training process. If unregistered words—words that occur less frequently—are entered during the testing or validation phase of the model; the model cannot solve this problem. What follows is the difficulty of the NMT model in processing these words, which is reflected in the fact that words marked as UNK cannot participate in model training, which reduces the ability of the NMT model to completely capture the semantic information of the source language and target language, and if certain sentences The first half is the same, but the second half is replaced by UNK, which will also cause ambiguity. The main solutions are: 1. Use an external dictionary or expand the vocabulary. 2. Use fine-grained grammatical units at the character (Character) level and sub-word (Sub-word) level.

The morphological form of Mongolian is similar to that of English. There are roots and affixes, and the same root with different affixes can form words with different meanings, which leads to the rare Mongolian words. Therefore, this article uses the BPE-Dropout algorithm to perform fine-grained segmentation of the Mongolian-Chinese bilingual parallel corpus. In 2016, Sennrich et al. [16] introduced the BPE algorithm into the field of natural language processing to effectively alleviate the problem of unregistered words. The Algorithm 1 [16] process is as follows:

Algorithm 1 BPE word segmentation

Input: corpus data
Output: Multi granularity fusion dictionary V_{BPE}
Translate corpus data into the most basic characters
Count the frequency of occurrence of the characters
For MON (merge operand):
    Merge the character pairs with the highest frequency of occurrence, denoted as $Symbol_{com}$
    Update multi granularity fusion dictionary, $V_{BPE} \cup Symbol_{com}$
End
Return Multi granularity fusion dictionary V_{BPE}

In this algorithm, only the unique hyperparameter, the merge operand, is controlled. The size of the vocabulary list can be controlled by artificially setting the merge operand. In this work, the merge operand is set to 20,000. Words are separated by using @@, the Mongolian text is divided into words or subwords with different granularities; and the Chinese text is segmented into single words.

Examples of Mongolian-Chinese bilingual word segmentation are shown in Figures 11 and 12 respectively:

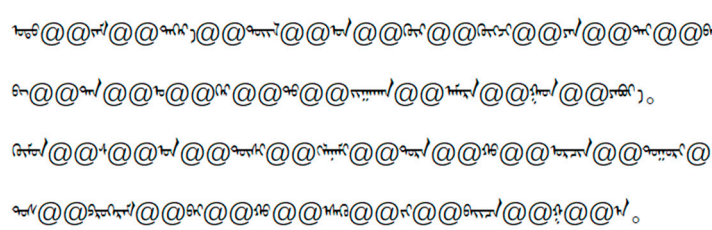


Figure 11. Mongolian word segmentation results through BPE. Mongolian is divided into syllables and phrases. The sentence meanings in the figure are "The streets are very lively now.", "I'll rest at

your house for a while before leaving.", "The environment in which people come to survive.", and "Please check if the program exists." respectively.

如@@今@@街@@面@@上@@非@@常@@热@@闹。
我@@在@@你@@们@@家@@歇@@一@@会@@再@@走。
人@@类@@赖@@以@@生@@存@@的@@环@@境。
请@@检@@查@@该@@程@@序@@是@@否@@存@@在。

Figure 12. Chinese word segmentation results through BPE. Chinese is divided into single characters. The sentence meanings in the figure are "The streets are very lively now.", "I'll rest at your house for a while before leaving.", "The environment in which people come to survive.", and "Please check if the program exists." respectively.

After word segmentation, the Mongolian-Chinese parallel corpus still has a total of 625,697 sentence pairs.

4.4. Experimental Environment and Parameter Configuration

Server system configuration: The server system used is Linux version 3.10.514.el7.x86_64; The CPU processor is an Intel Core i7-6700 CPU@3.40 GHz * 8; The GPU processor is a Nvidia Tesla P100-PCIE.

Transformer model parameter settings: The learning rate of the Adam W optimizer is set to 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and $\lambda = 0.02$; The η_t of learning rate adjustment strategies use a Warm Up strategy of Transformers, with warm_step = 3000; The Word vector dimension Embedding_dim = 256, the number of feedforward neural network layers is 2048, the number of heads for multi head attention is set to 8, and dropout_rate = 0.2; The batch size batch_size = 120; The maximum sentence length max_length = 100; The number of layers for both encoders and decoders is 6.

After a series of pre-processing, the remaining 625,697 sentence pairs in the Mongolian Chinese parallel corpus were randomly shuffled and divided into different datasets, as depicted in Table 2:

Table 2. Dataset division.

Total Sentence	Training Set	Validation Set	Testing Set
1,625,697	605,697	10,000	10,000

4.5. Experimental Results and Analysis

4.5.1. Comparative Tests

In this paper, the symbol of the baseline model was defined as Transformer_Base, and that of the proposed MCNMT model was defined as Transformer_Temp_Kno. To reduce the uncertainty due to a single test, three tests were performed on the conditions of the random seeds of $S_1 = 125$, $S_2 = 1234$, $S_3 = 4096$, as well as the corresponding initialization network parameters. The number of training rounds per session Epoch = 40.

The training times for the Transformer_Base and Transformer_Temp_Kno are listed in Table 3:

Table 3. Model parameters and training time.

Model	Total Parameter Number	S_1	S_2	S_3	Average Time (Min)
Transformer_Base	48,791,375	1536	1531	1518	1528.3
Transformer_Temp_Kno	49,451,601	12,078	2086	2072	2076.8

The changes of BLEU values for each model are shown in Table 4:

Table 4. Comparison of BLEU values of different models.

Model	BLEU Value			Average Time (Min)
	S_1	S_2	S_3	
Transformer_Base	39.735	39.614	39.153	39.501
Transformer_Temp_Kno	43.593	43.144	43.861	43.533

As can be seen from Table 4, compared with the baseline model, our proposed Transformer_Temp_Kno model has a higher BLEU value, with an increase of 4.032. Although the training speed of our model has decreased, its overall translation quality and stability have been improved to a certain extent.

Figures 13 and 14 show the loss function curves and accuracy curves of the Transformer_Temp_Kno model on the training and validation sets, respectively, which were obtained with the aforementioned three seeds.

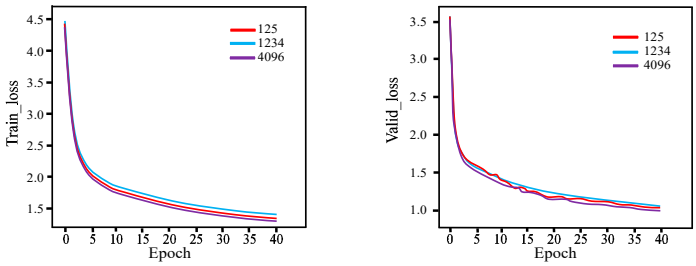


Figure 13. Loss function curves of the Transformer_Temp_Kno model with various seeds on the training set and validation set.

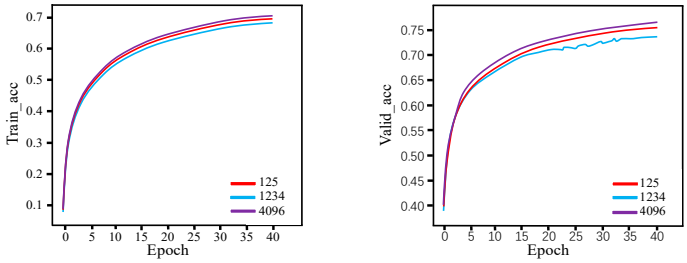


Figure 14. Accuracy curves of the Transformer_Temp_Kno model with various seeds on the training set and validation set.

Table 5 presents the comparison of several translation examples obtained through testing on the test set with the random seed of S_3.

From each translation example in Table 5, due to the limited scale of the Mongolian-Chinese parallel corpus, the baseline model often has the problem of insufficient grammatical information in the translation process, and when translating some long sentences, it can only translate word by word, resulting in the semantics of the translated Chinese script. The integration of soft target templates and contextual knowledge into the model can help the model learn semantic knowledge and guide the translation so that the translated text is richer and more fluent.

Table 5. Translation examples.

Source language sentence	[illegible] (Short-lived passion is worthless, only lasting passion is profitable.)
Reference translation	短暂的激情是不值钱的，只有持久的激情才是赚钱的。 (Short-lived passion is worthless, only lasting passion is profitable.)
Transformer_Base	短暂的激情是没钱的，只有持久的激情是才赚钱。 (There is no money in short-lived passion, only lasting passion can make money.
Transformer_Temp_Kno	短暂的激情是不值钱的，只有持久的激情才能够赚钱。 (Short-lived passion is worthless, only lasting passion can make money.)
Source language sentence	[illegible] (The round moon is like a big mirror hanging in the dark sky, and the stars in the distance are twinkling like beautiful pearls.)
Reference translation	圆圆的月亮就像是一面大镜子挂在黑黑的天空中，远处的星星一闪一闪的就如美丽的珍珠。 (The round moon is like a big mirror hanging in the dark sky, and the stars in the distance are twinkling like beautiful pearls.)
Transformer_Base	圆圆的月亮就像一面镜子黑黑的天空中挂着似的，远处的星星闪闪的就像美丽的珍珠。 (The round moon is like a mirror hanging in the dark sky, and the stars in the distance are shining like beautiful pearls.)
Transformer_Temp_Kno	圆圆的月亮就像一面大镜子挂在黑色的天空中，远处的星星闪闪的就像美丽的珍珠。 (The round moon is like a big mirror hanging in the black sky, and the stars in the distance are shining like beautiful pearls.)
Source language sentence	[illegible] (A person's value should depend on what he has contributed, not what he has achieved.)
Reference translation	一个人的价值，应该看他贡献了什么，而不应当看他取得了什么。 (A person's value should depend on what he has contributed, not what he has achieved.)
Transformer_Base	一个人的价值是他有什么贡献，不是看有什么取得。 (A person's value is based on what he has contributed, not what he has achieved.)
Transformer_Temp_Kno	一个人的价值，不应该只看他取得了什么，还要看他贡献了什么。 (The value of a person should not only depend on what he has achieved, but also what he has contributed.)

4.5.2. Ablation Tests

To investigate the impact of different module modifications on model detection performance and verify the effectiveness of the proposed method, the impact of different fusion methods on model performance was studied by conducting comparative experiments using the random seed of S_3.

The calculated BLEU values are summarized in Tables 3–5:

By comparing the BLEU values of different models in Table 6, it can also be concluded that our MCNMT model can improve the translation quality, and its BLEU value is 2.58 higher than that of the baseline model.

Table 6. Ablation tests.

Model	Fusion Method	BLEU Value
Transformer	None	39.50
Encoder	Dynamic fusion mechanism	40.89
	Knowledge extraction paradigm	40.26
	Dynamic fusion mechanism, Knowledge extraction paradigm	40.53
Decoder	Dynamic fusion mechanism	39.16
	Knowledge extraction paradigm	40.62
	Dynamic fusion mechanism, Knowledge extraction paradigm	39.77
Transformer_Temp_Kno	Encoder: Dynamic fusion mechanism Encoder: Knowledge extraction paradigm	42.08

Different fusion strategies applied on the encoder end are shown in the Transformer_Encoder section in Table 6. Although knowledge extraction can improve the translation quality of the model, it is not as effective as the dynamic fusion mechanism, and the use of both methods together does not improve the performance of the algorithm, indicating that the dynamic fusion method on the encoder end covers the effectiveness of knowledge extraction. In addition, different fusion strategies used on the decoder end are presented in the Transformer_Decoder section in Table 6. Under the same settings, the BLEU value is 0.34 lower than that of the baseline model, indicating that the dynamic fusion mechanism is ineffective at the encoder end.

The comparative experimental results show that it is correct to use different context extraction methods on the encoder and decoder ends separately because the responsibilities of the encoder and decoder are different. Since the encoder obtains contextual information by modeling input sentences, the use of external contextual information can improve its modeling performance. Therefore, even though sentence-level knowledge extraction can be applied on the encoder end, the effect is not as good as that of the dynamic fusion mechanism. The main responsibility of the decoder is to generate the target sentence through the input source representation, which involves the transformation problem of semantic spaces. Therefore, using a knowledge extraction paradigm that can learn the output distribution is more suitable for the decoder, which helps the model generate better sentences.

Figures 13 and 14 show the loss function and accuracy curves of the models with six different fusion methods on the training and validation sets, respectively. (Encoder_Dy represents using the dynamic fusion mechanism on the encoder end; Encoder_Dis represents using the knowledge extraction paradigm on the encoder end; Encoder_Dy_Dis represents using both the dynamic fusion mechanism and the knowledge extraction paradigm on the encoder end; Decoder_Dy represents using the dynamic fusion mechanism on the decoder end; Decoder_Dis represents using the knowledge extraction paradigm on the decoder end; Decoder_Dy_Dis represents using both the dynamic fusion mechanism and the knowledge extraction paradigm on the decoder end. Each model represent an improvement based on the Transformer baseline model).

From Figures 15 and 16, for both the training and validation sets, the loss function and accuracy of Decoder_Dis curves and Encoder_Dy curves are optimal and similar, while Decoder_Dy curves show the worst performance. This also demonstrates that using the dynamic fusion mechanism on the encoder end and the knowledge extraction paradigm on the decoder end is an optimal fusion strategy.

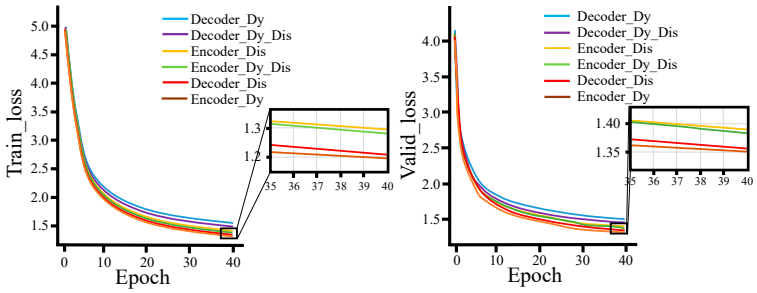


Figure 15. Loss function curves on training set and verification set.

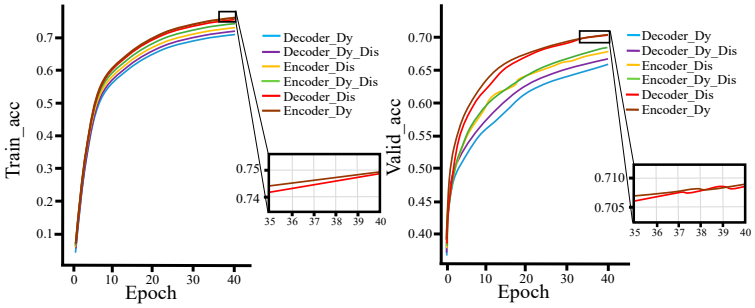


Figure 16. Accuracy curves on training set and verification set.

5. Conclusions

In order to improve the translation quality of Mongolian-Chinese NMT on low-resource corpora, we explored a method to enhance syntactic learning capabilities using soft templates and pre-trained models. By adding an additional soft target encoder on the encoder side to fuse the source language data and the template with the target language grammatical information, the template is used to guide the model. At the same time, in order to improve the utilization of grammatical knowledge, the model will be pre-trained by Bert. The contextual knowledge of the target language is extracted and integrated into the model to further improve the translation effect of the model. Experiments were conducted on the initialization network parameters corresponding to the three random seeds. The BLEU value of Transformer_Temp_Kno was 4.032 points higher than that of Transformer_Base, and the translation quality was significantly improved.

Since each layer of encoder in the Transformer model pays different attention to information, we adopt a dynamic fusion mechanism, which extracts the differences between different layers and the same layer in the pre-training model through the layer-aware attention mechanism and the context gating mechanism. attention. The knowledge extraction paradigm is also added to the decoder side to extract the output distribution of the pre-trained model, allowing it to assist the translation model in learning the specific task representation in the pre-training during the training process. We used ablation experiments to study the impact of different modules on model performance. Finally, by comparing the BLEU values of Encoder_Dy, Encoder_Dis, Encoder_Dy_Dis, Decoder_Dy, Decoder_Dis, and Decoder_Dy_Dis, we found that the BLEU score using the dynamic fusion mechanism on the encoder side is the highest, which is The BLEU score of using the knowledge extraction paradigm on the decoder side is the highest at 40.89, which is 40.62. From the above, it can be seen that the integration strategy of using the dynamic fusion mechanism on the encoder side and the knowledge extraction paradigm on the decoder side is optimal.

Although this study's Mongolian-Chinese neural machine translation model based on soft target templates and contextual knowledge has achieved remarkable achievements in many aspects, there are some obvious limitations that need to be considered in the interpretation of the research results and the planning of future work. First, although the soft target template used in this study improves the accuracy of the translation model, its performance may still be limited in some cases. The effectiveness of soft target templates may vary between language pairs, domains, or contexts, so in some special cases, the translation quality may not be as good as expected. This means that more research is needed to improve model performance in various translation tasks, especially in domain-specific or dialect translation. Moreover, this study is limited to the Mongolian-Chinese translation task. Although this is a challenging translation task, it also limits the application of the model to other language pairs. Future research should explore the performance of this model in different language pairs to understand its generalizability and adaptability.

The pre-training model in this article uses the BERT model, which was chosen for several important reasons. First of all, BERT has achieved significant success in the field of natural language processing and is widely recognized as a high-performance natural

language processing model. Its bidirectional encoding capabilities allow it to better capture contextual information in text, which is particularly important in machine translation tasks. By leveraging pre-training on large-scale text corpora, BERT can learn rich language representations to excel in a variety of natural language processing tasks. Although our research focuses on the Mongolian-to-Chinese translation task, the BERT model has demonstrated its adaptability in multi-language translation and other natural language processing tasks. Second, BERTs open-source nature and widely available pre-trained models make it ideal for research and experimentation. We can easily obtain and utilize BERT models without having to build a completely new neural network model from scratch. This significantly saves research resources and time, allowing research to focus on deeper tasks and questions. We are fully aware that BERT represents only one of many natural language models, which means that we cannot only use BERT as a pre-training model. In future research, we plan to explore the possibility of testing with other NLMs. This includes, but is not limited to, the GPT series, XLNet, T5, and other natural language models. This multi-model testing approach will allow us to more fully understand the performance of different models in the Mongolian to Chinese translation task and will also facilitate cross-model performance comparisons to reveal the differences between various natural language models adaptability.

Author Contributions: Conceptualization, Q.-D.-E.-J.R.; methodology, Q.-D.-E.-J.R. and Z.P.; software, Z.P.; validation, Z.P.; formal analysis, Z.P.; investigation, Z.P.; resources, Q.-D.-E.-J.R.; data curation, Z.P.; writing—original draft preparation, Q.-D.-E.-J.R., Z.P. and J.L.; writing—review and editing, J.L. and Z.P.; visualization, Z.P.; supervision, Q.-D.-E.-J.R. and J.L.; project administration, Q.-D.-E.-J.R. and Z.P.; funding acquisition, Q.-D.-E.-J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62066035, 62206138), the Inner Mongolia Natural Science Foundation (2022MS06013, 2022LHMS06004), the Inner Mongolia Science and Technology Program Project (2021GG0140, 2020GG0104), the Support Program for Young Scientific and Technological Talents in Inner Mongolia Colleges and Universities (NJYT23059), universities directly under the autonomous region Funded by the Fundamental Research Fund Project (JY20220122, JY20220089, RZ2300001739, RZ2300001743, JY20220186), and basic scientific research business expenses of universities directly under the Inner Mongolia Autonomous Region (ZTY2023021, JY20220419).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. *arXiv* **2014**, arXiv:1409.3215.
2. Kandola, E.J.; Hofmann, T.; Poggio, T. A neural probabilistic language model. *Stud. Fuzziness Soft Comput.* **2006**, *194*, 137–186.
3. Nasution, H.A.; Murakami, Y.; Ishida, T. A Generalized Constraint Approach to Bilingual Dictionary Induction for Low-Resource Language Families. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2017**, *17*, 1–29. [CrossRef]
4. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv* **2020**, arXiv:2001.04451.
5. Wang, S.; Li, P.; Tan, Z.; Tu, Z.; Sun, M.; Liu, Y. A template-based method for constrained neural machine translation. *arXiv* **2022**, arXiv:2205.11255.
6. Li, Z.; Lai, H.; Wen, Y.; Gao, S. Neural machine translation integrating bidirectional-dependency self-attention mechanism. *J. Comput. Appl.* **2022**, *42*, 3679–3685.
7. Guarasci, R.; Silvestri, S.; De Pietro, G.; Fujita, H.; Esposito, M. BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Comput. Speech Lang.* **2022**, *71*, 101261. [CrossRef]
8. Otmakhova, J.; Verspoor, K.; Lau, J.H. Cross-linguistic comparison of linguistic feature encoding in BERT models for typologically different languages. In Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, Seattle, WA, USA, 14 July 2022; pp. 27–35.
9. Varda, A.G.; Marelli, M. Data-driven Cross-lingual Syntax: An Agreement Study with Massively Multilingual Models. *Comput. Linguist.* **2023**, *49*, 261–299. [CrossRef]

10. Zhang, Z.; Su, Y.; Ren, Q.; Gao, F.; Wang, Y. Application of cross language multi task learning deep neural network in Mongolian Chinese machine translation. *Comput. Appl. Softw.* **2021**, *38*, 157–160+178.
11. He, W.; Wang, S. Application of neutral word segmentation method in Mongolian-Chinese machine translation. *J. Mizu Univ. China (Nat. Sci. Ed.)* **2022**, *31*, 36–46.
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
13. Sarzynska-Wawer, J.; Wawer, A.; Pawlak, A.; Szymanowska, J.; Stefaniak, I.; Jarkiewicz, M.; Okruszek, L. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* **2021**, *304*, 114135. [PubMed]
14. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Available online: https://blog.openai.com/language-unsupervised (accessed on 22 January 2023).
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762v7.
16. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; pp. 1715–1725.