Research on Transformer-Based Multilingual Machine Translation Methods

Xiaodong Zhao[1], Rouyi Fan[1], Wanyue Liu[2,*]
[1]School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, China
[2]Henan University of Technology, Zhengzhou, China
**Corresponding Author.*

Abstract: Because of the great difference in word order between different languages in machine translation, the translation model has the problem of wrong translation. Translation models with the same target language and different source languages learn different word order information, resulting in different translation quality. Therefore, this paper proposes a multilingual neural machine translation model with multiple languages at the source and one language at the target. Multiple languages with different word orders participate in the model training at the same time, so that the model can learn multiple word order information of sentences with the same meaning. First, 170,000 Russian-Uzbek-Uyghur-English-Chinese parallel corpus is constructed. On this basis, different source languages are added with specified language tags by using the method of adding language tags, and then mixed as a new data set to train a multilingual translation model. In addition, four multilingual neural machine translation models, stacking, parallel, fusion and sublayer fusion, are realized by modifying the Transformer model method. The experimental results show that the method of adding language tags can partially improve the performance of bilingual translation, and the quality of translation can be further improved after the source language is rewritten with Latin letters; The modified four Multilingual models can improve the quality of translation models.

Keywords: Multilingual Neural Machine Translation Model; Multilingual Parallel Corpus; Linguistic Markup; Improved Model

1. Introduction

Li et al. [1] proposed the PreAlign framework for multilingual machine translation, which innovatively focuses on enhancing cross-lingual transfer through pre-training alignment. This framework has demonstrated exceptional performance in handling translation tasks from multiple source languages to a single target language, providing novel concepts and implementation methods for the many-to-one multilingual machine translation paradigm. It allows for the simultaneous use of source statements in various languages to collaboratively contribute to the translation into the target language.

Typically, many-to-one multilingual machine translation models rely on multilingual parallel corpora. These corpora consist of datasets where the same sentences are paired with their translations in various languages. The PreAlign framework, in contrast to traditional models, enhances the pre-training stage through explicit multilingual alignment. It first establishes a semantic alignment space during the initialization phase by incorporating cross-lingual word translation tables (such as those generated by GPT-4). Then, during the training process, it integrates contrastive learning and language modeling for joint optimization while preserving alignment relationships through an improved code-switching strategy.

There are numerous practical applications for many-to-one multilingual translation models. For instance, on multilingual websites, while some content has been translated by human translators into a parallel corpus of multiple languages, further translation into additional languages is often necessary. The PreAlign-based many-to-one multilingual translation model can efficiently manage this task. Another application is in the context of parliamentary proceedings. Since these proceedings are frequently available in multiple languages, the PreAlign-enabled many-to-one model can translate them into various languages, facilitating improved information dissemination and understanding among diverse language-speaking audiences.

Each language possesses distinct characteristics, including variations in word order and grammatical structures, which can lead to numerous ambiguities during the machine translation process. While machine learning techniques can effectively address language ambiguity issues through training on extensive corpora, challenges arise due to the limited availability of large corpora for specific languages. Additionally, the construction of large corpora is often hindered by constraints related to human resources and time. so leveraging multiple low-resource languages can enhance the accuracy of the translation model, particularly in instances where resources are scarce.

The existing methodologies for many-to-one multilingual multi-source machine translation have been achieved through enhancements in translation models, fusion models, and data fusion processes. Knowledge distillation was suggested by Liang et al. [2] as a substitute for the conventional approach of directly concatenating source language data in multilingual machine translation. By transferring information, this method improves the model's performance and solves some of the problems that come with basic data concatenation. In order to improve performance while using fewer resources, knowledge distillation entails teaching a smaller model (the student) to learn from a bigger, more complex model (the teacher). Nevertheless, if the number of languages used for training increases, this approach may result in longer sentences and a larger datasets. As a result, the training data may grow excessively lengthy, thereby impairing the translation model's quality. In order to tackle this challenge, the research advocates for the implementation of linguistic markup that employs specific language labels corresponding to various languages. These labels are integrated into the training datasets, facilitating the development of a novel training set that eliminates the necessity for sentence concatenation, thereby enhancing the processing of multilingual inputs. According to the method they found, Merullo et al. [3] demonstrate that performance on translation tasks can be greatly enhanced by adjusting the internal model representations and model weights. This method greatly enhances its ability to comprehend numerous source languages by displaying attention for individual instances and examining the relationship between attention and syntax across a huge corpus. This methodology not only expands the encoder but also modifies the internal architecture of the decoder; however, the source code for the improved many-to-one translation model remains unpublished. As a result, the present study seeks to replicate the code to implement three many-to-one multilingual translation models, namely, superposition, parallelism, and fusion-based on the specifications outlined in their research. Drawing inspiration from the fusion model, this investigation introduces an innovative many-to-one multilingual translation model referred to as the sub-layer fusion model. In light of the limited availability of multilingual parallel corpora, this research undertakes the manual construction of a datasets consisting of 170,000 parallel sentences in Russian, Uzbek, Uighur, English, and Chinese for the purpose of many-to-one multilingual neural machine translation.

The contribution of this paper is as follows:

(1) The development of 170,000 parallel corpus in Russian-Uzbek-Uyghur-English-Chinese.

(2) A methodology is proposed for the incorporation of language labels that do not necessitate enhancements to the translation model. This approach involves reconstructing a new training datasets by integrating language labels during the preprocessing phase, followed by direct training utilizing a single-source translation model. This method allows for the simultaneous participation of multiple source languages in the training process, thus enhancing the translation quality of the model.

3. Three many-to-one multilingual translation models characterized by stacking, parallelism, and fusion are replicated, and a novel many-to-one multilingual translation model, termed the sublayer fusion model is introduced. Additionally, the source codes for four models are provided.

2. Related Work

Li et al. [1] introduced the PreAlign framework for multilingual translation, which resulted in a significant improvement in translation performance compared to traditional models. This framework not only addressed several limitations of traditional statistical machine translation (SMT) models but also introduced innovative pre-training alignment techniques to enhance cross-lingual transfer capabilities.

However, SMT models generally lack the ability for end-to-end training, which limits their flexibility. To overcome this, Purason and Tättar [4] introduced the concept of shared embeddings in multilingual neural machine translation (NMT) by jointly training language-specific encoder-decoder systems. This approach optimizes the network structure by representing all languages in a common embedding space, which improves both the efficiency of model training and the translation quality. Unlike earlier work relying on Long Short-Term Memory (LSTM) models, shared embedding focus on unifying language representations, which enhances multilingual performance.

A revolutionary approach to utilizing large language models (LLMs) for addressing multilingual tasks is cross-lingual in-context learning (XICL). This method is particularly advantageous when resources are scarce, as it facilitates multilingual translation without the need for extensive parallel corpora. Building on these advancements, Rojas and Carranza [5] introduced an innovative self-supervised technique that leverages the generative capabilities of LLMs to internally select and utilize task-relevant examples. This approach establishes two primary objectives: a semantic coherence loss to ensure cross-lingual consistency and a retrieval-generation alignment loss to enhance the quality of selected examples. The technique developed by Rojas and Carranza [5] enables effective translation outcomes with minimal supervision, paving the way for new research opportunities in multilingual translation.

Wang and Zhang [6] introduce an innovative method based on parameter differentiation, which enables the model to identify parameters that should be tailored to specific languages during the training process. Drawing inspiration from the concept of cellular differentiation, this approach allows each shared parameter to dynamically evolve into more specialized forms. The authors establish the differentiation criterion based on the similarity of gradients across tasks. Consequently, parameters exhibiting conflicting inter-task gradients are more likely to be designated as language-specific, thereby enhancing the model's efficiency and overall translation quality in multilingual contexts.

Subsequently, Cheng et al. [7] introduced a unified framework for multilingual machine translation and cross-lingual language understanding. This framework not only differentiates languages through the addition of language labels but also enables the simultaneous handling of both translation and language understanding tasks. This dual capability broadens the scope of multilingual translation models and significantly enhances their practical applications in diverse linguistic contexts.

Recent developments in multilingual translation have further refined parameter-sharing techniques. Guo et al. [8] introduced an adapter fusion approach to achieve parameter-efficient multilingual machine translation. By integrating two pre-trained BERT models from the source and target language domains into a sequence-to-sequence model through the introduction of lightweight adapter modules, this approach enhances model training efficiency and supports improved scalability in multilingual contexts. Similarly, Escolano [9] proposed a method based on multilingual information fusion, which incorporates multimodal data into the parameter-sharing process. This technique enhances the model's ability to handle complex language scenarios, thereby improving both efficiency and translation accuracy. Additionally, Trankova et al. [10] proposed an enhanced neural machine translation (NMT) framework that integrates cross-sentence context through redesigned positional encoding, hierarchical encoding, and conditional attention mechanisms. This approach strengthens the attention mechanism at various language levels. This targeted focus on specific language structures during translation further enhances the quality of multilingual outputs.

In conclusion, the field of multilingual machine translation has evolved significantly with advancements in parameter sharing techniques, unsupervised methods, and dynamic adaptation to different languages. Each of the proposed methods has contributed uniquely to the improvement of translation efficiency and quality, providing valuable insights for future research and practical implementations of multilingual systems.

3. Construction of Multilingual Parallel Corpus

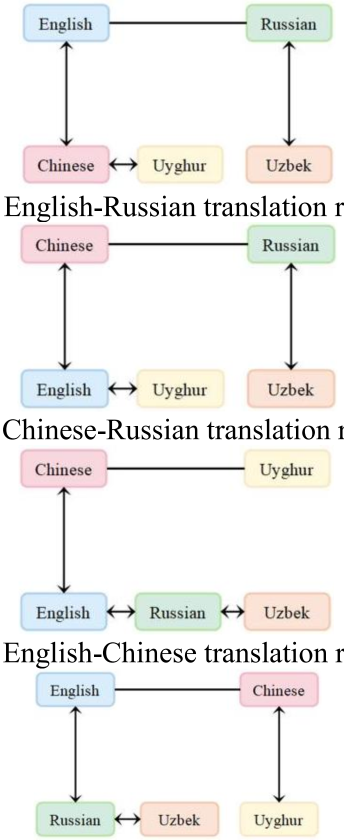
The multilingual parallel corpus encompassing Russian-Uzbek-Uyghur-English-Chinese has been manually constructed, utilizing a bilingual parallel corpus as the foundational data. This corpus was translated into the other three languages using Google Translate and Maverick

Translate, followed by a screening process, resulting in a total of 170,000 multilingual parallel corpora. The construction process is delineated into four distinct phases: the acquisition or creation of the bilingual parallel corpus, translation via translation tools, similarity screening, and manual verification. This section will provide a detailed account of the construction process based on these four steps.

Acquisition or Creation of the Bilingual Parallel Corpus: The extension of bilingual data is employed to facilitate the construction of the multilingual parallel corpus. The bilingual datasets include Russian-English, Chinese-Russian, English-Chinese, and Uyghur-Chinese. The English-Chinese and Chinese-Russian bilingual corpus are generalized datasets sourced from official repositories, while the Uyghur-Chinese parallel corpus has been developed internally in the laboratory and the Russian-English dataset was generated by utilizing the Houyi Collector to extract parallels from a multilingual website.

Translation via Translation Tools: Upon the completion of the bilingual parallel corpus acquisition or creation, the data is translated into the remaining three languages using Google Translate and Maverick Translate. The translation protocols are illustrated in (a)(b)(c)(d) in Figure 1. Each language undergoes bidirectional translation, resulting in a total of eight files, as depicted in Figure 1. The generated files in (a) include English, English-to-Chinese, Chinese-to-English, Chinese-to-Uyghur, Uyghur-to-Chinese, Russian, Russian-to-Uzbek, and Uzbek-to-Russian, which are subsequently utilized for similarity screening.

Similarity Screening: As illustrated in Figure 1(a), a total of eight files were generated utilizing a translation tool, of which five namely English, Chinese-to-English, English-to-Chinese, Uyghur-to-Chinese, Russian, and Uzbek-to-Russian were employed for similarity screening. The presence of substantial error information in sentences produced by machine translation tools often results in inaccuracies, omissions, and other issues, thereby compromising the quality of the multilingual parallel corpus. Consequently, a preliminary screening of sentences is imperative to eliminate those of inferior quality. The metval tool was utilized to compute the similarity of the English-Chinese-to-English, Russian-Uzbek-to-Russian, and English-to-Chinese-Uyghur-to-Chinese translations. In this process, Chinese text was segmented into words, while the other languages were processed using a tokenizer. perl mentioned by Mielke et al. [11]. Following the similarity calculations, parallel sentence pairs with BLEU values exceeding 0.5 were selected from these three groups, and the corresponding parallel sentence pairs for the five languages were subsequently saved.



(a) English-Russian translation rules

(b) Chinese-Russian translation rules

(c) English-Chinese translation rules

(d) Chinese-Uyghur translation rules

Figure 1. Translation Rules of English-Russian Chinese-Russian English-Chinese Chinese-Uyghur Parallel Corpus into Other Three Languages

Manual Filtering: The quality of the multilingual parallel corpus obtained post-machine filtering exhibited a marked improvement compared to the pre-filtering [12] stage, however, certain sentences continue to present issues such as the presence of special symbols, poorly constructed sentences, and incomplete sentences. Therefore, manual filtering was deemed necessary. Sentences characterized by an excessive number of special symbols, poor structure, or a lack of logical coherence were systematically filtered out. The final datasets produced represents the constructed multilingual parallel corpus, with the total volume of multilingual data summarized in Table 1.

Table 1. The Amount and Total Amount of Data Filtered for Each Language Pair

Language	Amount of data	Total amount of data
Russian-English	40 000	17 000
Chinese-Russian	40 000	
English-Chinese	40 000	
Uyghur-Chinese	50 000	

4. Multi-Source Neural Machine Translation Model

This thesis presents the implementation of a many-to-one multilingual neural machine translation model through two primary approaches: data preprocessing and structural modifications to the Transformer model. Data preprocessing involves applying linguistic markup to incorporate language labels for various languages, which are then combined to create the dataset for the single encoder-single decoder translation model, thereby facilitating the training of the many-to-one multilingual translation model. The modification of the Transformer model structure initially replicates the three models of stacking, parallelism, and fusion as proposed by Liang et al. [2], and subsequently introduces a novel sub-layer fusion model to achieve the many-to-one multilingual translation model. This section will first outline the labeling rules associated with the linguistic markup, followed by a detailed discussion of the four translation model structures: stacking, parallelism, fusion, and sub-layer fusion.

4.1 Linguistic Markup

The many-to-one multilingual machine translation model is achieved through a straightforward modification of multilingual data, maintaining the original structure of the translation model. The linguistic markup involves annotating words in multilingual sentences at the source level with the format #lang#, where "lang" represents various languages, such as "ru" for Russian, "uy" for Uyghur, "en" for English, and "uz" for Uzbek. This approach facilitates the direct amalgamation of different languages into a new source datasets, which inherently contains multiple languages, thereby enabling many-to-one Multilingual Neural Machine Translation.

Table 2. Multilingual Training Data with Label

	Sentence	Original sentence	Tagged sentences
src	ru	Ясчастлив.	#ru#Я#ru#счастлив#ru#.
	uz	Men baxtliman.	#uz#Men#uz#baxtliman#uz#.
	uy	mEn bEk huxal.	#uy#mEn#uy#bEk#uy#huxal#uy#.
	en	I am happy.	#en#I#en#am#en#happy#en#.
tgt	zh	Wohenkaixin	Wohenkaixin

In this study, the source languages include ru, en, uy, uz, while the target language is Chinese. Language labels are systematically applied to four semantically equivalent sentences in the source languages, with the results of this labeling process presented in Table 2.

4.2 Modifying the Model Structure

The structure of the Transformer model is illustrated in Figure 2. This research builds upon the Transformer model by increasing the number of multi-head non-self-attention mechanisms in both the encoder and decoder, tailored to the type of source language. This extension replicates the three models of stacking, parallelism, and fusion as proposed by Liang et al. [2]. Drawing inspiration from the network architecture of the fusion model, we decompose the internal components of the encoder and enhance the inner sub-layers based on the source language type, thereby introducing a novel model-sublayer fusion approach.

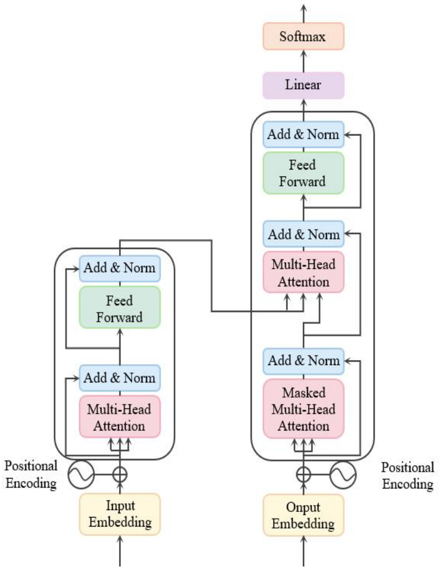


Figure 2. Transformer Model Structure Diagram

3.3.1 Stacking model

The stacking model operates on the principle of increasing the number of encoders and the multi-head non-self-attention mechanisms in the decoder, corresponding to the type of source language. Each source language is associated with a distinct encoder, and each encoder is linked to a specific multi-headed non-self-attention mechanism in the decoder. The output from the preceding multi-headed non-self-attention mechanism acts as the input for the subsequent mechanism. The structural representation of the stacking model is illustrated in Figure 3.

3.3.2 Parallel model

The parallel model shares a fundamental similarity with the stacking model, as both approaches expand the number of encoders and multi-head non-self-attention mechanisms in the encoder based on the source language type. However, the parallel model diverges in its processing within the decoder, where multiple multi-head non-self-attention mechanisms are computed concurrently, and their outputs are subsequently integrated as inputs to the feed-forward network in the decoder. A schematic of the parallel model structure is depicted in Figure 4.

The parallel model fuses multiple multi-head non-self-attention in the decoder according to formula (1), where D_out is the fusion result of multiple multi-head non-self-attention,m is the language type, Mul1 is the result of the first multi-head non-self-attention in the decoder, and Add is the direct addition of the results according to the same dimension.

$$\mathrm{D}_{\_\mathrm{out}} = \mathrm{Add}(\mathrm{Mul}_1, \ldots, \mathrm{Mul}_\mathrm{m}) \qquad (1)$$

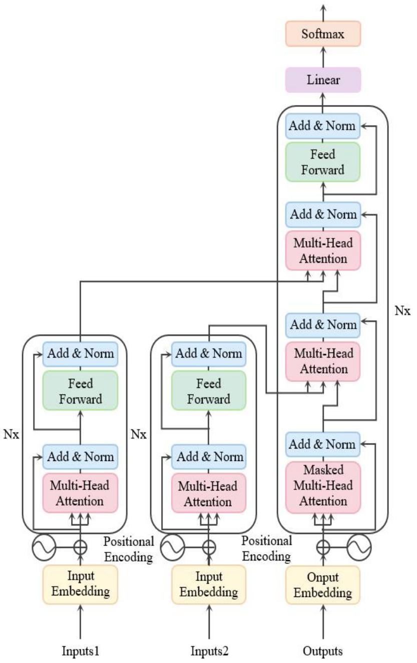


Figure 3. Stacking Model Structure Diagram

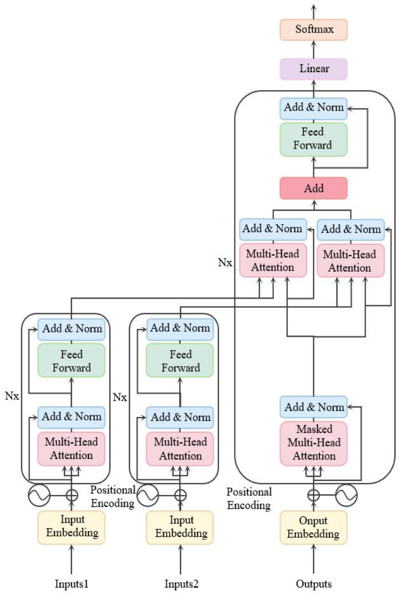


Figure 4. Parallel Model Structure Diagram

3.3.3 Fusion model

The fusion model operates on the principle of augmenting the number of encoders by the specific source language while maintaining a constant decoder structure. Given the variability in languages and sentence lengths, the output dimensions produced by the encoders for different languages exhibit discrepancies, rendering a straightforward summation or averaging of the outputs from multiple encoders inadequate. In this study, we propose a method to integrate the outputs from various encoders based on sentence length, which will subsequently serve as the input for the decoder. The architecture of the fusion model is illustrated in Figure 5.

The fusion model synthesizes the outputs of multiple encoders as delineated in equation (2), where E_con represents the consolidated output from the various encoders, m denotes the language type, E1 signifies the output from the first encoder, and Concat refers to the concatenation process executed by the sentence length.

$$E_{\_con} = Conct(E_1, \ldots, E_m) \qquad (2)$$

3.3.4 Sublayer Fusion Model

The fundamental concept underlying the sublayer fusion model is that the decoder remains constant, while the encoder is bifurcated into upper and lower layers. The number of upper layers is adjusted based on the type of source language, and the outputs from multiple upper layers are concatenated following the sentence lengths. These concatenated outputs serve as inputs for the lower layers. There are six encoder layers in the context of the Transformer model, these layers are categorized into upper and lower segments. The upper layers are augmented in number relative to the variety of source languages, whereas the number of lower layers remains fixed. Specifically, the upper layer can consist of one to five layers, while the corresponding lower layer will consist of five to one layers, respectively. The outcomes of the sublayer fusion model are illustrated in Figure 6.

The fusion of the upper layer of the encoder is articulated through equation (3), where E_mid represents the result of the upper layer fusion, m denotes the language type, o1 signifies the output of the first upper layer, and Concat refers to the direct concatenation based on sentence length.

$$E_{\_mid} = Concat(o_1, \dots, o_m) \quad (3)$$

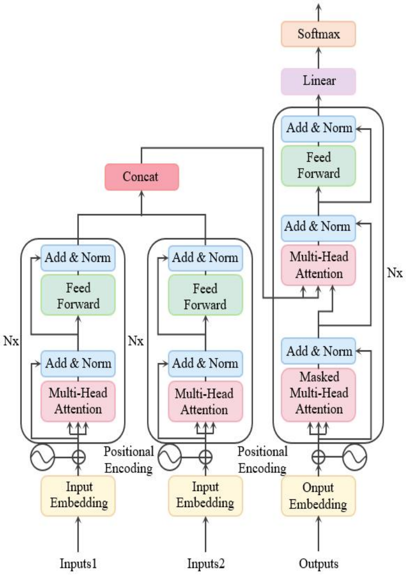


Figure 5. Fusion Model Structure Diagram

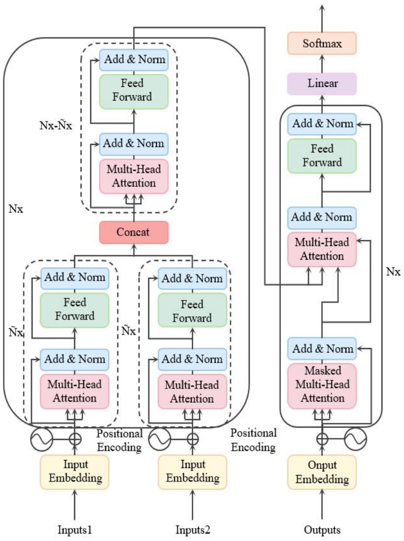


Figure 6. Sublayer Fusion Model Structure Diagram

5. Experimental

In this paper, a self-constructed multilingual parallel corpus is used as the experimental dataset. Firstly, we study the enhancement of low resource translation models by linguistic markup experiments, during which the differences between Latinized and un-Latinized source languages are compared. Secondly, experiments are conducted on the reproduced stacked, parallel, fusion model and the sub-layer fusion model proposed in this paper, comparing the different models and comparing them with the baseline model. The effect of the number of language varieties in the test set on the multi-source translation model is then explored. The extent of model enhancement is then investigated in terms of the number of source languages involved in training. Finally, one's translation model is compared with traditional translation tools.

5.1 Experimental Data

The self-constructed parallel resources of Russian-Uzbek-Uyghur-English-Chinese are used as the experimental data, and 1,000 are randomly selected from all the corpus as the test set and validation set respectively, and the remaining data are used as the training set, as shown in Table 3 for the data volume of the multilingual parallel corpus. The Chinese language in the multilingual corpus is segmented using the THULAC (Tsinghua University Language and Computing) segmentation tool, and the other languages are preprocessed using Moses for data preprocessing. Chinese is used as the target language in the multilingual parallel corpus, and the other four languages are used as the source languages. Chinese, English, Uzbek, Russian, and Uyghur have 16 k, 14 k, 14 k, 12 k, and 14 k BPE fusions times.

Table 3. Multilingual Data Volume

Dataset	Number of sentences
Training set	176 009
Test set	1 000
Validation Set	1 000
All	178 009

5.2 Experimental Setup

The baseline model, along with the language labeling experiments presented in this study, utilizes the PyTorch implementation of FACEBOOK's open-source FairSeq framework. The enhanced multi-source translation model is constructed upon the Transformer architecture inherent to the FairSeq system. All experimental setups adhere to the Transformer Base configuration, maintaining consistent parameters for the hidden layer nodes and word vector dimensions, both set at 512. The feedforward neural network comprises 2,048 intermediate layer nodes, while the architecture includes 6 encoder-decoder pairs and 8 multi-head attention mechanisms. The Drop-out ratio applied within the model is established at 0.3, with an additional Drop-out rate of 0.1 implemented following the activation function in the feedforward network. The activation function employed in this model is the glue. Parameter optimization is conducted using the batch stochastic gradient descent method, with a data batch size configured to 4,096. The learning rate is adjusted through the adaptive Adam algorithm (10-9, 0.9, 0.997), and a warm-up period of 4,000 iterations is incorporated.

In the decoding phase of the model, the Beam-search parameter is established at a value of 4, with a data batch size of 64 without taking the influence of the reference translation's sentence length on the scoring mechanism into account. The length penalty coefficient is assigned a value of 0. Additionally, post-processing of the translation outputs involves the removal of the BPE (Byte Pair Encoding) flag. The evaluation of the model's results is conducted utilizing the multi-bleu.perl.

5.3 Baseline Model

Table 4. Multilingual Data Volume

Language	BLEU
English→Chinese	37.13
Russian→Chinese	36.32
Uzbek→Chinese	36.90
Uyghur→Chinese	44.90

In this paper's many-to-one multilingual neural machine translation experiment, the source languages are English, Uzbek, Russian, and Uyghur, so this experiment adopts the English→Chinese, Russian→Chinese, Uzbek→Chinese, and Uyghur→Chinese translation models as the baseline model, and the experimental results are shown in Table 4.

5.4 Experimental Results

(1) Results of Linguistic Markup in Experiments

The findings from the experiments utilizing the linguistic markup are presented in Table 5. The term "unLatinized" refers to the traditional experimental outcomes derived from the direct mixed training of a multi-source translation model involving English, Russian, Uzbek, and Uyghur, supplemented with linguistic labels at the source level. The results indicate that in comparison to the baseline model, there is no enhancement in translation performance from English→Chinese and Uyghur→Chinese. However, improvements were observed in the Russian→Chinese and Uzbek→Chinese translations, with increases of 1.95 and 0.8 BLEU (Bilingual Evaluation Understudy) scores. To reduce source-side multilingual complexity and enhance lexical sharing, source-side multilingual sentences were Latinized. This approach aimed to increase commonalities and minimize differences among these languages. Latinization involved employing a Latinization tool to convert the four languages at the source level into Latin script, followed by the addition of language labels. Specifically, the uroman tool was utilized to transform all four languages into their Latin-script equivalents, thereby standardizing the diverse writing systems. This approach enabled the fusion of these languages to create a novel source language for training the many-to-one multi-language machine translation model. The results of the Latinization experiment demonstrated further improvements in the Russian→Chinese, Uzbek→Chinese, and Uyghur→Chinese translations compared to their un-Latinized counterparts.

Table 5. Experimental Results of Latinization and Non-Latinization with Labels

Model	English→Chinese	Russian→Chinese	Uzbek→Chinese	Uyghur→Chinese
Non-Latinization	36.74	38.27	37.70	42.58
Latinization	36.24	39.10	38.57	44.37

(2) Experimental Results of Modifying the Model Structure

In the conducted experiments involving four many-to-one multilingual machine translation models namely stacking, parallelism, fusion, and sublayer fusion, the training process incorporates multiple source languages. Each language is associated with a specific encoder or a designated upper layer of the encoder. Consequently, the model's performance is influenced collectively by all the languages involved. The parameter updates are oriented towards optimizing the overall model rather than focusing on achieving optimal performance for any single language or encoder layer. Thus, the model's quality is a function of the combined contributions from the various source languages. During the testing phase, a test set comprising four source languages is utilized simultaneously, resulting in a singular output from the model, which is then compared against baseline models for each language.

As illustrated in Table 6, the experimental outcomes for the three many-to-one multilingual translation models—fusion, stacking, and parallelism indicate that the stacking model outperforms the others, achieving a BLEU score of 47.30. The performance of the fusion and parallelism models is relatively similar. When comparing the stacking model to the translation models for English-Chinese, Russian-Chinese, Uzbek-Chinese, and Uyghur-Chinese, it demonstrates improvements in BLEU scores of 9.63, 10.44, 9.86, and 1.86, respectively.

Table 6. Fusion, Stacking, and Parallel Model Experiments

Model	BLEU
Fusion	46.76
Stacking	47.30
Parallel	46.87

Table 7. Sublayer Fusion Model Experiment

The layer number of the first part of the sublayer	BLEU
1	46.91
2	47.11
3	47.16
4	47.42
5	47.02

Table 7 presents the experimental results for the sublayer fusion model proposed in this study. In this model, the encoder's sublayers are divided into upper and lower layers, with the total number of layers equating to the sum of the upper and lower layers. When the upper layers are set to four, the model achieves its optimal performance with a BLEU score of 47.42, reflecting enhancements of 10.29, 11.1, 10.52, and 2.52 BLEU points compared to the English-Chinese, Russian-Chinese, Uzbek-Chinese, and Uyghur-Chinese translation models. Notably, irrespective of the number of upper layers, the quality of this model significantly surpasses that of the aforementioned translation models. A comparison between the sublayer fusion model and the three reproduced models reveals improvements in BLEU scores of 0.66, 0.12, and 0.55.

5.5 Multi-Source Testing

The four many-to-one multilingual translation model experiments were implemented by modifying the model structure, the model uses test sets of English, Russian, Uzbek, and Uyghur input into the model at the same time to test the translation quality of the model, and does not consider testing the quality of the model with a single test set or any other combinations of test sets, so this subsection uses a different number of test sets of languages involved in the model translation to study the changes in the quality of the model's translations.

Using the fusion model as a case study, the test outcomes for various languages in the test set are presented in Table 8. These outcomes were derived from evaluating a many-to-one multilingual translation model utilizing test sets comprising Uyghur, Uyghur-English, Uyghur-English-Uzbek, as well as Uyghur-English-Uzbek-Russian. Notably, when the test set consisted of Uyghur data, the translation result was recorded at 34.98, which is inferior to the results obtained from the bilingual training model. However, it is observed that the translation performance of the model improves as the number of languages in the test set increases. Specifically, when the model is assessed using all available test set languages, it achieves its highest performance score of 46.76.

Table 8. Test Results of Test Sets with Different Number of Languages

Test set	BLEU
Uyghur	34.98
Uyghur-English	45.20
Uyghur-English-Uzbek	46.68
Uyghur-English-Uzbek-Russian	46.76

5.6 Multi-Source Training

Subsequently, this study performs a multi-source evaluation to investigate the impact of the quantity of source language types utilized during training on the translation model. As illustrated in Table 9, the fusion model demonstrates that incorporating two, three, and four source languages in the training process correlates with an enhancement in the quality of the translation model, indicating that an increase in the number of source languages contributes positively to translation performance.

Table 9. The Degree of Improvement of the Research Model from the Perspective of the Number of Source Languages

Source Languages	BLEU
Uyghur	44.90
Uyghur-English	46.63
Uyghur-English-Uzbek	46.82
Uyghur-English-Uzbek-Russian	47.30

5.7 Comparative Experiments

As illustrated in Table 10, this subsection conducts a comparative analysis of the fusion, stacking, parallel, and sublayer fusion models against the methodologies put forth by Zoph, Garmash, Dabre, et al. The findings indicate that the translation models derived from the sublayer fusion approach exhibit superior quality compared to those developed by prior researchers.

Table 10. Result of Comparative Experimental

Model	BLEU
Fusion	46.76
Stacking	47.30
Parallel	46.87
Sublayer fusion	47.42
Zoph	33.47
Garmash	46.32
Dabre	45.63

5.8 Translation Tools Translation

In this study, we present a baseline model that we have developed, which is subsequently compared with widely used translation tools. The evaluation is conducted using a test set comprising 1,000 self-constructed multilingual sentences. This allows us to analyze the performance of our model about popular translation tools across various languages, identifying those languages in which our model demonstrates superior translation capabilities as well as those in which it performs less effectively. Furthermore, we compare the translation outcomes of the linguistic markup and the stacking, parallelism, fusion, and sub-layer fusion models proposed in this paper against established translation tools, thereby assessing the enhancements made to the multilingual translation model introduced herein.

As presented in Table 11, various translation tools, including Google Translate, Baidu Translate, and Maverick Translate were employed to translate the test set, with BLEU scores computed based on the translation outcomes and subsequently compared to the baseline model. The translation quality of Google Translate surpassed that of the baseline model solely for the Russian → Chinese translation. In contrast, Baidu Translate demonstrated significantly superior performance compared to the baseline model for English→Chinese and Uyghur→Chinese translations, although it performed worse than the baseline for the other two language pairs. Maverick Translate yielded commendable results for both English→Chinese and Russia→Chinese translations. However, its performance in the Uzbek→Chinese translation was marginally inferior to that of the baseline model.

The findings illustrated in Tables 5, 6, 7, and 11 indicate that, within the context of the linguistic markup methodology employed in this study, the results for English→Chinese and Russian→Chinese translations were generally slightly inferior to those produced by the translation tools. Conversely, for the Uzbek→Chinese translation, the results for English→Chinese were markedly superior to those generated by the three translation tools. In the experiments involving stacking, parallelism, fusion, and sub-layer fusion models conducted in this study, the optimal result achieved was 47.42, which outperformed all other results, except for the Russian→Chinese translation result obtained by Maverick Translate.

Table 11. Translation Tools translation

Translation tools	English-Chinese	Russian-Chinese	Uzbek-Chinese	Uyghur-Chinese
Google	36.43	39.98	35.46	31.24
Baidu	43.98	32.59	11.35	46.66
Maverick	36.82	55.28	32.31	-

6. Conclusions

This research focuses on developing a many-to-one multilingual neural machine translation model. It constructs a parallel corpus covering Russian, Uzbek, Uyghur, English, and Chinese. Meanwhile, it proposes linguistic markup and model structure modification methods. Linguistic markup adds language labels to the source-side mixed parallel corpus to create new datasets for single-source, single-target training. The structure modification adjusts the encoder and decoder of the Transformer model to allow multiple source languages to be input simultaneously for translation into the target language.

Effect of Linguistic Markup: Linguistic markup can significantly improve the translation quality of certain languages. After the multilingual Latinization of source data, the translation quality is further enhanced.

Effect of Model Structure Modification: Four modeling strategies, namely stacking, parallelism, fusion, and sub-layer fusion, can greatly improve the translation quality of low-resource models. Among them, the sub-layer fusion model performs best, with a BLEU score of 47.42, showing improvements of 10.29, 11.1, 10.52, and 2.52 compared to the English-Chinese, Russian-Chinese, Uzbek-Chinese, and Uyghur-Chinese translation models respectively.

Impact of Multi-source Training and Testing: Through multi-source training and testing methods, it is found that more involvement of source languages in the model reasoning process and incorporating more source languages during the training phase are positively correlated with the improvement of translation quality.

Model Performance Comparison: A comparative evaluation of the four modeling techniques shows that the sub-layer fusion method outperforms previous approaches in performance. Compared with traditional translation tools, the stacking, parallelism, fusion, and sub-layer fusion models are significantly more effective in translating from Uzbek to Chinese.

References

[1] Li Y, Wang H, Zhang M, et al. PreAlign: Enhancing Cross-Lingual Transfer via Pre-training Alignment for Multilingual Large Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Taipei, Taiwan, 2024. Association for Computational Linguistics.

[2] Liang X, Wu L, Li J, et al. Multi-teacher distillation with single model for neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 992-1002.

[3] Merullo J, Eickhoff C, Pavlick E. Talking heads: Understanding inter-layer communication in transformer language models. Advances in Neural Information Processing Systems, 2024, 37: 61372-61418.

[4] Purason T, Tättar A. Multilingual neural machine translation with the right amount of sharing//Proceedings of the 23rd Annual Conference of the European Association for Machine Translation. 2022: 91-100.

[5] Rojas M A, Carranza R. Align, Generate, Learn: A Novel Closed-Loop Framework for Cross-Lingual In-Context Learning. arXiv preprint arXiv:2412.08955, 2024.

[6] Wang Q, Zhang J. Parameter differentiation based multilingual neural machine translation//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(10): 11440-11448.

[7] Cheng S, Chen W, Tang Y, et al. Unified Training for Cross-Lingual Abstractive Summarization by Aligning Parallel Machine Translation Pairs. Mathematics, 2024, 12(13): 2107.

[8] Guo J, Zhang Z, Xu L, et al. Adaptive adapters: An efficient way to incorporate BERT into neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1740-1751.

[9] Escolano Peinado C. Learning multilingual and multimodal representations with language-specific encoders and decoders for machine translation. 2022.

[10]Trankova N, Rykunov D, Serov I, et al. Hierarchical Encoding and Conditional Attention in Neural Machine Translation. The American Journal of Engineering and Technology, 2024, 6(09): 45-55.

[11]Mielke S J, Alyafeai Z, Salesky E, et al. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. arXiv preprint arXiv:2112.10508, 2021.

[12]Al Jawarneh, IM, Bellavista, P, Murillo, JM. A Pre-Filtering Approach for Incorporating Contextual Information into Deep Learning Based Recommender Systems//Proceedings of IEEE ACCESS. 2020: 40485-40498.