information

MDPI

Article

Hybrid System Combination Framework for Uyghur–Chinese Machine Translation

Yajuan Wang [1,2,3,4], Xiao Li [1,3,*], Yating Yang [1,3,*], Azmat Anwar [1,3] and Rui Dong [1,3]

[1] The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; wangyajuan@ms.xjb.ac.cn (Y.W.); azmat@ms.xjb.ac.cn (A.A.); dongrui@ms.xjb.ac.cn (R.D.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China
[4] Department of Information Security Engineering, Xinjiang Police College, Urumqi 830011, China
* Correspondence: xiaoli@ms.xjb.ac.cn (X.L.); yangyt@ms.xjb.ac.cn (Y.Y.); Tel.: +1-502-604-3990 (X.L.)

Abstract: Both the statistical machine translation (SMT) model and neural machine translation (NMT) model are the representative models in Uyghur–Chinese machine translation tasks with their own merits. Thus, it will be a promising direction to combine the advantages of them to further improve the translation performance. In this paper, we present a hybrid framework of developing a system combination for a Uyghur–Chinese machine translation task that works in three layers to achieve better translation results. In the first layer, we construct various machine translation systems including SMT and NMT. In the second layer, the outputs of multiple systems are combined to leverage the advantage of SMT and NMT models by using a multi-source-based system combination approach and the voting-based system combination approaches. Moreover, instead of selecting an individual system's combined outputs as the final results, we transmit the outputs of the first layer and the second layer into the final layer to make a better prediction. Experiment results on the Uyghur–Chinese translation task show that the proposed framework can significantly outperform the baseline systems in terms of both the accuracy and fluency, which achieves a better performance by 1.75 BLEU points compared with the best individual system and by 0.66 BLEU points compared with the conventional system combination methods, respectively.

Keywords: statistical machine translation (SMT); neural machine translation (NMT); multi-source-based combination; voting-based combination; hybrid system combination

Citation: Wang, Y.; Li, X.; Yang, Y.; Anwar, A.; Dong, R. Hybrid System Combination Framework for Uyghur-Chinese Machine Translation . *Information* **2021**, *12*, 98. https://doi.org/10.3390/info12030098

Academic Editor: Kostas Stefanidis

Received: 7 February 2021
Accepted: 20 February 2021
Published: 25 February 2021

1. Introduction

Machine translation (MT) is an important task for the natural language processing (NLP) field. At present, the most popular Uyghur–Chinese machine translation methods can be divided into two categories: statistical machine translation (SMT) [1,2] and neural machine translation (NMT) [3–5]. There are other reasons for SMT to be used since SMT models require large datasets and have the most problems with rare words. However, in some cases (depending on the domain and corpus size), SMT is still in use, mostly in a hybrid approach [6]. Its workflow must be ordered to execute by multiple separately tuned components, such as word alignment, translation rules extractors, and other feature extractors, which seems to be more complex and will bring error propagation problems in the training pipeline [7]. Whereas distributed representation of arbitrary language can be realized through the end-to-end training of the NMT system, the NMT model can prevent the problems during the SMT training process. NMT can produce more fluent results [8–10] but which are often not adequate, while SMT models generally obtain lower results for the criteria of fluency, especially for low-resource languages with relatively free word order [11]. Generally, MT evaluation metrics favor translations that follow a strict word order when compared to the reference translations, which could be the reason for lower

BLEU scores. Inevitably, NMT has a problem in addressing the translation adequacy especially for the rare and unknown words, even using the subword method [12,13]. Therefore, neither the SMT nor NMT model can achieve the expected results for the Uyghur–Chinese machine translation task.

Table 1 shows a Uyghur–Chinese translation example. Given an input Uyghur sentence (denoted as src), there are three erroneous but complementary hypotheses (denoted as hyp1, hyp2, and hyp3) generated by the three different MT systems. Hyp1 can generate some keywords such as "方便面 (Instant noodles)", "鸡蛋汤 (Egg soup)", and "物资 (Supplies)". Hyp2 can translate the verb word "送来了 (Brought)" correctly. Additionally, hyp3 can capture the relationship between "交警 (Traffic police)" and "司机 (Driver)", but it mistranslates the noun words "方便面 (Instant noodles)" and "鸡蛋汤 (Egg soup)". These hypotheses have different strengths and weaknesses, so we suppose that if we can utilize the advantages of the three hypotheses, we can obtain a better translation result. The limited languages such as the Uyghur are far less resourced in available MT services, or even language technologies in general [14]. The system combination method can combine the various machine translation outputs to achieve a better accuracy and fluency. Thus, the system combination can play an important role in the low-resource language translation scenery.

Table 1. Examples of multiple hypotheses. Tsrc, THyp1, THyp2, THyp3 are English translation of the src, hyp1, hyp2, hyp3, respectively.

src	قاتناش ساقچىلىرى شوپۇرلارغا تەييار چۆپ ، مانتا ، تۇخۇم شورپىسى قاتارلىق يىمەكلەرنى ئەكىلىپ بەردى.
Tsrc	The traffic police brought the driver instant noodles, steamed buns, egg soup and other food.
hyp1	交警 方便面，司机 、 鸡蛋 汤 يىمەكلەرنى 等 物资.
Thyp1	Traffic police instant noodles يىمەكلەرنى and other supplies.
hyp2	交警 方便面，、 鸡蛋 汤 等 يىمەكلەرنى 送来 了.
Thyp2	Traffic police, instant noodles, egg soup and so on يىمەكلەرنى brought.
hyp3	交警 给 司机 送 了 方便货，洗菜 、 鸡蛋 、 鸡蛋 等 食品.
Thyp3	Traffic police to the driver to send convenient goods, wash vegetables, eggs, eggs and other food.
tgt	交警 为 司机 送来 了 方便面 、 包子 、 鸡蛋 汤 等 食物.

System combination is a method for combining the outputs of multiple machine translation engines to use the strengths of each of the individual systems. The neural-based system combination has become the dominant paradigm for system combination [15–17]. In the multi-source-based system combination method, the outputs of the multiple MT systems can be regarded as multiple inputs using different encoders to train the NMT system [16]. The voting mechanism is re-introduced into modern system combination methods to improve the performance of neural system combination [17]. The task of combining the various MT systems is not very easy. System combination using either the multi-source-based system combination or voting-based approach cannot provide useful solutions to deal with all the problems. For example, the multi-source-based system combination model (MUSC) can quantify the word in a hypothesis as its attention weight, without using explicit hypothesis alignment. This model can not only generate new words that are not within any hypothesis but also prove to be effective on multiple machine translation tasks. Due to the attention weight between one hypothesis and the output being calculated independently, connections between hypotheses are not taken into consideration in this approach [17]. On the other hand, although the voting-based system combination model (VOSC) can enable connections between hypotheses by finding the consensus among them, the major problem is that the result of the voting is heavily dependent on the performance of the single systems; that is, the closer the vote, the result is better.

Since each of them have their strengths and shortcomings, we attempt to build a hybrid framework, which can take advantage of different system combination models to make a better prediction. In this work, we first built different translation systems, including the SMT system and the NMT system as the first layers, and then implemented the multi-source-based approach and voting-based approach on the outputs of N single systems, respectively. Finally, the proposed framework is used to combine the translation hypotheses of the N systems and the outputs of the previous two models.

In the rest of this article, Section 2 presents some related works about system combination. Section 3 describes the details of our hybrid system combination framework for Uyghur–Chinese machine translation. The experiment details and results are shown in Section 4. Finally, the conclusions are presented in Section 5.

2. Related Work

Nowadays, the NMT model has become the mainstream method for Uyghur–Chinese MT and has obtained more fluent translation results compared to SMT [14]. The neural machine translation (NMT) models still struggle in the translation task on Uyghur–Chinese with complex morphology and limited resources [18]. To tackle this problem, Zhang et al. [19] proposed a novel memory structure to alleviate the rare word problem caused by the agglutinative nature of Uyghur. Pan et al. [20] proposed a multi-source neural model that employs two separate encoders to encode the Uyghur source word sequence and the Uyghur linguistic feature sequences. Zhang et al. [21] used the integrated strategy and reordering strategy to solve the problem that the single NMT model is easy to fall into a locally optimal solution when fitting and training on a low-resource corpus, such as Uyghur to Chinese. Wang et al. [22] used the extracted semantic similarity as a new feature to generate multiple new systems from a single SMT system and to combine multiple systems. Although previous work has improved the performance of Uyghur–Chinese machine translation to some extent, system combination remains huge potential in Uyghur–Chinese machine translation.

The system combination method is an important research branch for machine translation, which was originally derived in 1994 [23]. The statistical-based system combination methods, including the word level, the phrase level, and the sentence level methods, have become the de facto paradigm in the past two decades [24–28]. Among them, the confusion network-based word-level combination method is more successful and several open-source tools such as Jane, MANY, and MEMT have been developed for scientific research [29–31]. This kind of method realizes the system combination via separate modules, such as choosing a backbone, aligning the word between hypotheses, building a confusion network, and generating the translation results [32–37]. Since the method is completed by several relatively independent steps, the statistical-based system combination model also suffers from the error propagation problem [17], which means the errors in the previous step will be passed to the next step to affect the combination result.

In recent years, researchers introduced neural network models into system combination [15–17]. Inspired by the multi-source neural translation, the outputs of the multiple systems are considered as multiple sources, each of which corresponds to an encoder and executes the end-to-end training process with multiple sequence-to-sequence models [38]. The recurrent neural network (RNN)-based system combination model employs the recurrent neural network to encode the translation hypotheses with multiple context vector representations. The weighted sum of these context vectors is used to calculate the probability of the next target word and the final output is generated one by one from the decoder. In the transformer-based system combination model, the multiple encoders are identical to that of the dominant transformer model, which is modeled using the self-attention network [16]. The advantage of this method is that while preserving the original self-attention network, four new combination strategies are introduced by adding a new sub-layer to make full use of the advantages of multiple MT translation hypotheses.

The voting mechanism is proposed to select the best word through voting from multiple translation hypotheses, and then it combines the words selected by voting to form a new hypothesis [39,40]. Huang et al. (2020) transferred the voting mechanism from the statistical framework to the neural network framework, and the experimental results proved that the voting method based on the neural network is the best system combination method at present [17].

Recently, various studies of hybrid system combination methods are available in the literature [41–43]. Some researchers proposed a three-stack architecture for both utilizing the neural-based system combination model and the statistical-based neural system combination model to improve the translation quality [41]. Other researchers proposed a simple reranking system using a smorgasbord of the informative features [42]. Some novel methods are proposed in the literature [43], such as structuring source-side language sentences into the linguistically motivated fragments and combining them using a character-level neural language model, and combining neural machine translation output by employing the source-side translation attention alignments. The main goal is to assemble a set of methods that would be able to improve the quality of the Uyghur–Chinese machine translation, which has rich morphology and limited corpora resources. These characteristics currently make them rather difficult to translate with the tools that are currently available. Therefore, the two best system combination models (the MUSC model and the VOSC model) have been organized together to build a hybrid system combination framework.

3. Methods

Studies have proved that if the individual systems that participated in the combination are more uncorrelated, the more favorable it is for system combination [44]. NMT and SMT are two heterogeneous models. The hybrid system combination framework can combine the merits of these two kinds of translation models. The proposed framework for system combination works in three layers.

- First Layer: A set of translation hypotheses from N systems that are collected for the combination in the consequent layers.
- Second Layer: The outputs from the first layer are combined with the help of the multi-source-based approach and the voting-based approach in the second layer. Here, we combine the N system's translated outputs through the multi-source-based system combination method, which described in Section 3.1. Moreover, we combine the N system's translated outputs through the voting-based system combination, which is described in Section 3.2.
- Third Layer: A hybrid system combination framework will combine the outputs of the previous two combination models to improve the translation quality. For system combination, the proposed hybrid framework chooses the words among the best sentences, which are generated by the various system combination models.

3.1. Multi-Source-Based System Combination Model

Figure 1 illustrates the framework of the multi-source-based system combination model. Sequentially, the outputs of the various MT systems are given to the multiple transformer decoders to capture the semantics of the different translation hypotheses. Four different combination strategies are used in the transformer decoder to combine the multiple context vector representations from the previous encoders. Finally, the combined outputs are generated word by word using the modified attention weights.

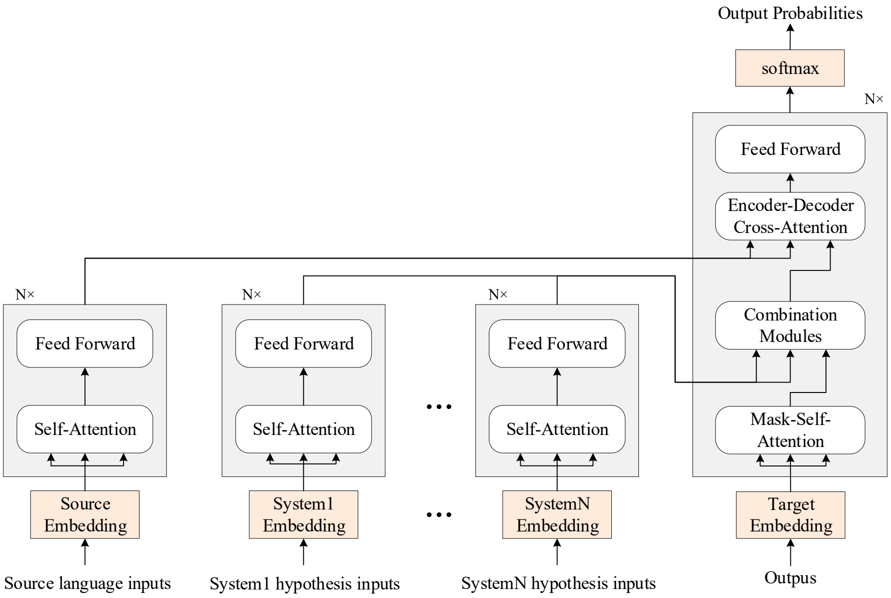


Figure 1. The architecture of the multi-source-based system combination model. The combination module is the core of the model.

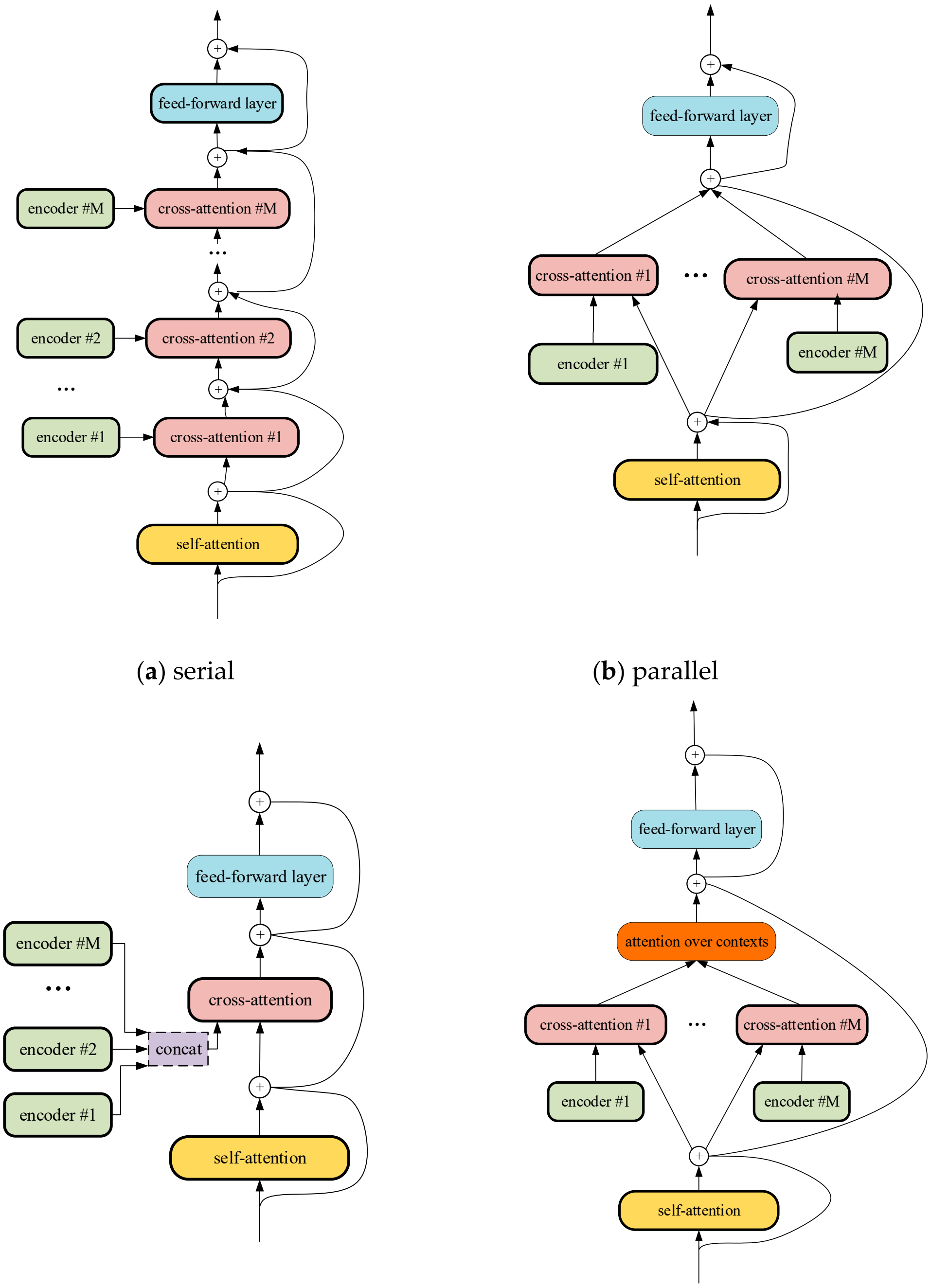
Each encoder in this method has the same structure as the original transformer decoder that has a stack of N identical layers. Each layer has two sub-layers. The first uses a multi-head self-attention mechanism, and the second uses a feed-forward network. The output of each sub-layer is as follows:

$$\tilde{z}_k^l = LayerNorm\left(z_k^{l-1} + MHAtt\left(z_k^{l-1}, z_k^{l-1}, z_k^{l-1}\right)\right) \tag{1}$$

$$z_k^l = LayerNorm\left(\tilde{z}_k^l + FFN(\tilde{z}_k^l)\right) \tag{2}$$

where the l is the layer depth, z_k^l is the hidden states of the l-th layer from the k-th MT system hypotheses. FFN denotes the feed-forward networks, and $MHAtt$ denotes the multi-head attention mechanism.

The decoder in this method is also composed of a stack of N identical layers. However, each layer is added to a new sub-layer which performs the multi-head attention over the output of the encoder stack. Because multiple encoder outputs will be passed to the decoder, how to combine them is a crucial question. Four different input combination strategies for encoder–decoder attention have been proposed to solve this problem [45,46], which are illustrated in Figure 2. The serial strategy, as shown in Figure 2a, calculates the cross-attention one by one for each input encoder. The query set of each cross-attention is the set of the context vectors computed by the preceding sub-layer. All of these sub-layers are interconnected with the residual connections. In the parallel strategy, as shown in Figure 2b, the model attends to each encoder independently and then sums up the context vectors. The flat strategy, as shown in Figure 2c, uses all the states of all input encoders as a single set of keys and values. Thus, the attention models a joint distribution over a flattened set of all encoder states. The encoder–decoder attention in the hierarchical strategy computes the attention independently over each input. The resulting contexts are then treated as states of another input and the attention is computed once again over these states.



(**a**) serial

(**b**) parallel

(**c**) flat

(**d**) hierarchical

Figure 2. Four combination strategies: (**a**) serial, (**b**) parallel, (**c**) flat, and (**d**) hierarchical.

Here, we take a combination module as the second sub-layer of the decoder which can make the decoder receive the context vectors of the multiple encoders. Moreover, the output of the combination module and the context vectors of the source encoder embedded for the source language are transmitted to the encoder–decoder cross-attention sub-layer. These sub-layers work as follows:

$$s_1^l = LayerNorm\big(s^{l-1} + MHAtt(s^{l-1}, s^{l-1}, s^{l-1})\big) \tag{3}$$

$$s_2^l = LayerNorm\left(s_1^l + ComMod(s_1^l, z^N)\right) \tag{4}$$

$$s_3^l = LayerNorm\left(s_2^l + MHAtt(s_2^l, h^N, h^N)\right) \tag{5}$$

$$s^l = LayerNorm\left(s_3^l + FFN(s_3^l)\right) \tag{6}$$

Finally, a linear transformation and a SoftMax activation are used to compute the probability of the next words based on s^N:

$$p(y_j|y < j, x, \theta) = softmax(s^N W) \tag{7}$$

where θ is the model parameters and W is the weight matrix.

3.2. Voting-Based System Combination Model

Figure 3 illustrates the architecture of the multi-source combination model based on the voting mechanism, which uses the source sentence and the outputs of multiple MT systems. Formally, given the source input sentence x, the output sequences $\tilde{y}_{1:N} = \tilde{y}_1 \cdots \tilde{y}_N$ of the N MT system hypotheses for the same source sentence and previously generated target sequence $y = y_1 \cdots y_K$, the system combination model is defined by:

$$P(y|x, \tilde{y}_{1:N}; \theta) = \prod_{k=1}^{K} P(y_k|x, \tilde{y}_{1:N}, y < k; \theta) \tag{8}$$

where y_k is the k-th target word, and θ is a set of model parameters.

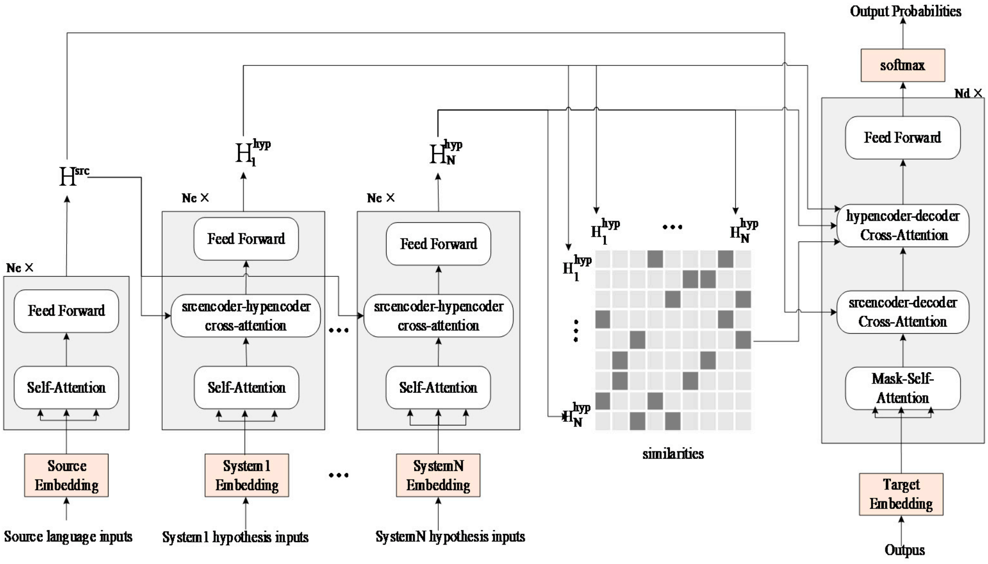


Figure 3. The architecture of the multi-source combination model based on the voting mechanism. The similarities between hypotheses are the core of the model.

In this model, the source encoder is identical to the conventional transformer encoder that consists of two sub-layers:

$$\bar{z}_s^l = LayerNorm\left(z_s^{l-1} + MHAtt(z_s^{l-1}, z_s^{l-1}, z_s^{l-1})\right) \tag{9}$$

$$z_s^l = LayerNorm\left(\bar{z}_s^l + FFN(\bar{z}_s^l)\right) \tag{10}$$

where the l denotes layer depth, z_s^l indicates the hidden state of l-th layer of the source input. FFN and $MHAtt$ are the same as previously mentioned.

The hypothesis encoder is composed of a stack of N identical layers. Each of which has three sub-layers:

$$\bar{\bar{z}}_h^l = LayerNorm\left(z_h^{l-1} + MHAtt\left(z_h^{l-1}, z_h^{l-1}, z_h^{l-1}\right)\right) \tag{11}$$

$$\bar{z}_h^l = LayerNorm\left(\bar{\bar{z}}_h^l + MHAtt\left(\bar{\bar{z}}_h^l, z_s^l, z_s^l\right)\right) \tag{12}$$

$$z_h^l = LayerNorm\left(\bar{z}_h^l + FFN(\bar{z}_h^l)\right) \tag{13}$$

where z_h^l denotes the hidden state of l-th layer of the h-th MT system hypothesis. An additional $src\text{-}hyp$ attention sub-layer is used to capture the relationship between the source and the hypothesis by influencing the energy of words in the hypothesis.

For each layer in the decoder, the lowest sub-layer is the masked multi-head self-attention network:

$$s_1^l = LayerNorm\left(s^{l-1} + MHAtt(s^{l-1}, s^{l-1}, s^{l-1})\right) \tag{14}$$

The second sub-layer is the source encoder–decoder cross attention that bridges the gap between the source and target language by seeking the source language semantic:

$$s_2^l = LayerNorm\left(s_1^l + MHAtt(s_1^l, z_s^l, z_s^l)\right) \tag{15}$$

The third sub-layer is the voting module that considers dependencies between hypotheses by introducing the voting mechanism into the neural network-based system combination model:

$$s_3^l = LayerNorm\left(s_2^l, VotMod(s_2^l, z^N)\right) \tag{16}$$

where $z^N = (z_1^N, z_2^N, \cdots, z_k^N)$. Here we use a new attention weight $\alpha_{n,j}$ to replace the original attention weight. The new attention weight $\alpha_{n,j}$ is the attention weight between the hypotheses and output that is calculated by the voting method. The result of the voting is composed of two parts. The first part is called *influence,* which is used to measure the influence of the voters (the voter can be any of the words in the hypotheses to decide whether the candidate should be included in the output by voting). It can quantify as a real-valued number, which is actually the energy used in calculating attention weight:

$$e_{n,j} = f\left(x, y_{<k}, \tilde{y}_{n,j}, \theta\right) \tag{17}$$

where $f(\cdot)$ is a function that calculates the energy, $\tilde{y}_{n,j}$ is the j-th word of the n-th hypothesis, and $e_{n,j}$ is its corresponding energy that reflects how likely it will be the next word.

The second part, *preference,* is utilized to estimate a voter's preference for a candidate (the candidate usually is the next word to be predicted in the output). It can also be measured as a real-valued number, which is actually the similarity between the voter and the candidate:

$$sim\left(\tilde{y}_{m,i}, \tilde{y}_{n,j}\right) = \frac{exp\left(h_{m,i} h_{n,j}^T\right)}{\sum_{i'=1}^{|\tilde{y}^m|} exp\left(h_{m,i'} h_{n,j}^T\right)} \tag{18}$$

where $\tilde{y}_{m,i}$ is a voter and $h_{m,i}$ is its representation retrieved from H_m^{hyp}. Likewise, $\tilde{y}_{n,j}$ is a candidate and $h_{n,j}$ is its representation. In order to avoid the length bias, the similarities between the voters and a candidate are normalized at the hypothesis level. All the votes can be collected by calculating a weighted sum of the energy of the voters:

$$\tilde{e}_{n,j} = e_{n,j} + \sum_{m=1 \wedge m \neq n}^{N} \sum_{i=1}^{|\tilde{y}_m|} sim\left(\tilde{y}_{m,i}, \tilde{y}_{n,j}\right) \times e_{m,i} \tag{19}$$

As a result, the new attention weight $\alpha_{n,j}$ depends on both the influence and the preference and is computed by:

$$\alpha_{n,j} = \frac{exp\left(\tilde{e}_{n,j}\right)}{\sum_{n'}^{N} \sum_{j'=1}^{|\tilde{y}_{n'}|} exp\left(\tilde{e}_{n',j'}\right)} \tag{20}$$

The last sub-layer of the decoder is the feed-forward network:

$$s^l = LayerNorm\left(s_3^l + FFN(s_3^l)\right) \tag{21}$$

Finally, the same operation with the model described in Section 3.1 is conducted to compute the probability of the next tokens based on s^N:

$$P(y_k|x, \tilde{y}_{1:N}, y < k; \theta) = softmax(s^N W) \tag{22}$$

where θ is the model parameters and W is the weight matrix.

The difference of the voting-based system combination model lies in that the benefits of end-to-end training on the state-of-the-art model are preserved, and the newly added voting modules are able to take the relations between hypotheses into account to find the consensus outputs.

3.3. Hybrid System Combination Framework

Preference of the outputs from the individual system combination models feed as the input into the hybrid system combination framework. The schematic diagram of the proposed framework is shown in Figure 4. We used a replica of the voting-based system combination model (mentioned in Section 3.2) for the hybrid system combination design. In the first layer, we chose N systems as our candidate systems including the SMT systems and the NMT systems. Diversified candidate systems can generate richer translation hypotheses and provide more choices during system combination. The second layer utilizes two popular system combination models, the MUSC and the VOSC. In the MUSC model, by modeling the semantics of source language and translation hypotheses in the multi-source neural network, better combination results can be obtained. While retaining the advantages of the multi-source encoder–decoder model, the relationship between translation hypotheses can be captured when generating the combination results by the VOSC model. Due to the limitation of the methods, neither of the above two models can get the ideal result of the combination, and so we put forward the third layer to take advantage of candidate systems and these two system combination models. The inputs of the third layer are the outputs of the second (i.e., MUSC and VOSC) and the first layer systems (single MT systems). Inputs of MUSC and VOSC are the translated outputs of various MT systems $(T1, T2, \cdots, TN)$. Finally, outputs of the multiple MT systems along with $M1$ and $M2$ (outputs of the second layer) as an individual system's preference are entered into the hybrid system combination framework to produce the better output.

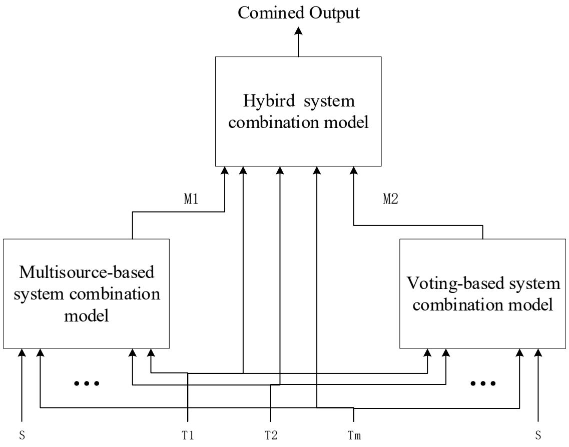

Comined Output
Hybird system combination model
M1
M2
Multisource-based system combination model
Voting-based system combination model
…
…
S
T1
T2
Tm
S

Figure 4. The architecture of the hybrid system combination framework.

4. Experiments

We evaluated the hybrid system combination method on the Uyghur–Chinese machine translation tasks. The evaluation metric was the case-insensitive BLEU [47] and it was measured on the character level. We used the paired bootstrap resampling for statistical significance tests [48].

4.1. Data Preparation

For this experiment, we used two datasets from the CWMT2017 (http://nlp.nju.edu.cn/cwmt2017) (the 13th China Workshop on Machine Translation) and CWMT2019 (http://ccmt2019.jxnu.edu.cn) (the 15th China Conference on Machine Translation) containing about 0.5 M bilingual sentence pairs. Due to the neural-based system combination method based on end-to-end training, the model should be trained on the outputs of multiple translation systems and the gold target translations. In order to keep the consistency in the training and test processes, we also used the training data simulation strategy [12]. We prepared the datasets as follows:

- Merge the training set of the CWMT2017 and CWMT2019, and remove the duplicate sentences;
- Pre-process the raw data, such as Uyghur words tokenization, Chinese words segmentation, and remove the sentences longer than 80 words;
- Divide the pre-processed training data into two parts, then reciprocally train the MT system on one half and translate the source sentences of the other half into target translations. The translated sentences as well as the gold target reference are utilized to train the MUSC and VOSC.
- When training all neural network models including the single systems or system combination models, we used byte pair encoding (BPE) [12] with 30K merges to segment words into subword units both for the Uyghur and Chinese sentences.
  We employed the CWMT2019 development set as the validation data and used the CWMT2019 test set as the test set. Detailed statistics for the dataset are shown in Table 2.

Table 2. Dataset descriptions. Train and Dev represent training and development set, respectively.

System type	Set	#Sentences	#Tokens	
			Uyghur	Chinese
Single MT Systems	Train	239,711	4,497,220	4,174,631
	Dev	1000	25,217	25,262
MUSC	Train	239,763	4,498,108	4,171,378
	Dev	1000	25,217	25,262
VOSC	Train	239,763	4,498,108	4,171,378
	Dev	1000	25,217	25,262
Test Set		1000	23,202	21,559

4.2. Experiment Setup

Previous research had proved that the system combined needs to be almost uncorrelated to be beneficial for system combination [44]. In order to verify whether the conclusion is effective on Uyghur–Chinese machine translation system combination and further verify whether the uncorrelation refers to the model or the performance, we set up three groups of experiments for comparison:

- Firstly, three heterogeneous systems with different performance were selected to observe the effect of system combination;
- Secondly, three heterogeneous systems with similar performance were selected to observe the effect of system combination;
- Finally, three systems with similar performance derived from one model were selected to observe the effect of system combination.

The individual systems used in these three groups of experiments were introduced as follows:

1. PBMT: It is the indispensable phrase-based SMT system still remaining used in Uyghur–Chinese machine translation. We trained the system with a 4-gram language model using modified Kneser–Ney smoothing using SRILM on the target portion of the bilingual training data [49]. The model learns the word alignments with grow-diag-final and heuristics from the parallel training corpus using GIZA++ [50];
2. HPMT: It is a hierarchical phrase-based SMT system, which uses the same default setting and the language model as PBMT. The training and decoding of PBMT and HPMT are both based on the open-sourced Moses toolkit [51]. Due to HPMT using hierarchical phrases which consist of both words and sub-phrases, it has a stronger ability of reordering while preserving the strength of PBMT;
3. RNN-based NMT: It is a recurrent neural network based on encoder–decoder architectures [3]. Here, we used the RNN with an attention mechanism. The weight of the sum is referred to as attention scores and allows the network to focus on different parts of the input sequences as it generates the output sequences;
4. Transformer-based NMT: TNMT has obtained state-of-the-art performance on Uyghur–Chinese machine translation, which predicts target sentences from left-to-right using a self-attention mechanism [5];
5. Transformer-based NMT (1–3): we trained three Transformer-based NMT with three random seeds, which are three inputs for system combination in our experiment.

The BLEU score of these single systems is listed in Table 3.

Table 3. The BLEU score of single systems. DEV and TEST represent the BLEU score on development set and test set, respectively.

#	System Type	DEV	TEST
1	PBMT	27.45	34.78
2	HPMT	27.77	34.89
3	RNMT	28.44	35.24
4	TNMT	39.03	46.71
5	TNMT-1	39.00	46.47
6	TNMT-2	39.27	46.43
7	TNMT-3	39.28	46.61

4.3. Training Details

The Transformer-based baseline was trained on the open-source toolkit fairseq [52]. The MUSC model was implemented on the open-source toolkit neural monkey [53]. The VOSC model was implemented on the open-source toolkit THUMT-TensorFlow [54,55]. We used the same hyperparameter settings of the base Transformer model as [5] for both baselines and the two system combination models. The number of layers was set to 6 for both encoders and decoder. The hidden size was set to 512 and the filter size was set to 2048. The number of individual attention heads was set to 8 for multi-head attention. During training, we used Adam [56] for optimization and the learning rate decay policy described by [5]. Each mini-batch contains 10k tokens for the transformer baseline, 32 sentences for MUSC, and 3 k tokens for VOSC, respectively. In training the Transformer baseline model, we used efficient half floating point computation (FP16) to accelerate the training process. In training the VOSC model, the update cycle was set to 4 to simulate 4GPUs with only 1GPU. The early stopping mechanism was used in training for both single systems and the system combination model, the early-stop learning option was set to 5. In decoding, the beam size was set to 4 for both models. We used the same development set to select the best model.

4.4. Experimental Results

4.4.1. Results of Combination Modules

We first evaluated the four combination strategies including serial, parallel, flat, and hierarchical strategies for the MUSC. Table 4 lists the experimental results of the Uyghur–Chinese development set. Due to the excessively long source-side encoding of join operations in the flat strategy, the results of the flat strategy are significantly lower than those of the other three strategies, and not even better than those of the best single system. In contrast, the hierarchical strategy has shown to be the best strategy, and we will use hierarchical strategy in subsequent experiments.

Table 4. Result of different combination strategies for the multi-source-based system combination model.

#	Strategies	DEV
1	PBMT	27.45
2	HPMT	27.77
3	RNMT	28.44
4	MUSC(Serial)	28.66
5	MUSC(Parallel)	28.51
6	MUSC(Flat)	27.35
7	MUSC(Hierarchical)	28.78

4.4.2. Results on Heterogeneous Systems with Different Performance

Literature has proven that the source language is helpful for the system combination [15], so we will leverage the source language in the subsequent experiment. In this experiment, the single systems involved in the combination are three heterogeneous systems with different performance that are PBMT, HPMT, and TNMT. We named the multi-source-based combination model and voting-based combination model trained on them as MUSC-1 and VOSC-1, respectively. The experimental results of Uyghur–Chinese translation are shown in Table 5. Compared with PBMT and HPMT, TNMT achieves the best translation quality and significantly outperforms HPMT by +11.82 BLEU points. Although both approaches have significant improvements over PBMT (42.37 vs. 34.78, 46.03 vs. 34.78) and HPMT (42.37 vs. 34.89, 46.03 vs. 34.89), neither approach is better than the best single system, TNMT (42.37 vs. 46.71, 46.03 vs. 46.71). Both models have negative growth rates on the test set. One possible reason is that the poor candidate system is counterproductive to system combination. From the experimental result, we can conclude that heterogeneous systems with different system performance are not a wise choice of system combination.

Table 5. Results on heterogeneous systems with different performance for Uyghur–Chinese translation.

System	DEV	TEST	Increase Rate
PBMT	27.45	34.78	X
HPMT	27.77	34.89	X
TNMT	39.03	46.71	X
MUSC-1	39.00	46.47	−9.29%
VOSC-1	39.27	46.43	−1.46%

4.4.3. Results on Heterogeneous Systems with Similar Performance

In this experiment, we took three heterogeneous systems with similar performance—PBMT, HPMT, RNMT—as the input of system combination. We named the multi-source-based combination model and voting-based combination model trained on them as MUSC-2 and VOSC-2, respectively. As listed in Table 6, the BLEU score of MUSC-2 is 1.23

BLEU points higher than the best single model RNMT. VOSC-2 outperforms the best single MT systems by +4.41 points and obtains improvement by +3.18 BLEU points over MUSC-2. Both models are improving positively on the test set. The experimental results show that when the candidate systems are different in structure but similar in performance, both methods are effective in system combination, and the vote-based method is more effective in this scenery.

Table 6. Results on heterogeneous systems with similar performance for Uyghur–Chinese translation.

System	DEV	TEST	Increase Rate
PBMT	27.45	34.78	X
HPMT	27.77	34.89	X
RNMT	28.44	35.24	X
MUSC-2	29.06	36.47	+3.49%
VOSC-2	31.73	39.65	+12.51%

4.4.4. Results on Homogeneous Systems with Similar Performance

In this group of experiments, three transformer candidate systems TNMT-1, TNMT-2, TNMT-3 with different random seeds were regarded as the input of system combination (the random seeds are 1111, 2222, 3333, respectively). We named the multisource model and voting model trained on them as MUSC-3 and VOSC-3, respectively. As shown in Table 7, the best result of VOSC-3 obtains improvement by +0.58 BLEU points over the best single system TNMT-3, while MUSC-3 did not exceed the best single system. The experimental results prove that in system combination, the voting-based method is effective when the candidate systems have similar structure and similar performance.

Table 7. Results on homogeneous systems with similar performance for Uyghur–Chinese translation.

System	DEV	TEST	Increase Rate
TNMT-1	39.00	46.47	X
TNMT-2	39.27	46.43	X
TNMT-3	39.28	46.61	X
MUSC-3	36.36	43.52	−6.63%
VOSC-3	39.63	47.19	+1.24%

It is worth noting that although three stronger candidate systems with similar performance can get the better result, from the proportion of improvement, the VOSC-2 trained on three weaker candidate systems increased more than the VOSC-3 trained on three stronger candidate systems (12.51% vs. 1.24%). To test whether the improvement ratio is an important factor in determining the quality of the voting model, we conducted the following experiment. Table 8 shows the experimental results of using two trained voting models to vote each other, from which we can draw the following conclusions: (1) The system combination for Uyghur–Chinese machine translation prefers candidate systems with different structures but closer performance, because they not only benefit from multiple architectures, but also complement each other in performance. (2) The substantial improvement of BLEU points demonstrates that the greater the improvement ratio, the better the effect of the voting model.

Table 8. Results for different voting models on test set.

Systems	VOSC-2	VOSC-3
PBMT + HPMT + RNMT	39.65	31.66
TNMT-1 + TNMT-2 + TNMT-3	47.29	47.19

4.4.5. Result of Hybrid Framework

Based on previous experiments, we can conclude that due to the limitations of the method, neither MUSC nor VOSC can obtain the expected results. It is necessary to construct a hybrid framework to obtain the advantages of the two methods. Here, we selected the voting method as the third layer of the hybrid framework. The MUSC-2 model, which is the only one with a positive increase rate, and the VOSC-2 model, which has a higher increase rate, were used for the following experiment. In the third layer of the hybrid framework, the outputs of the individual system combination model (the outputs of the second layer) as well as the hypotheses produced by the single systems (the outputs of the first layer) were passed in. From the aspect of accuracy, after hybridization, we observed significant performance improvement over individual MT systems and the individual system combination models in terms of BLEU scores. As listed in Table 9, MUSC, VOSC, and proposed HBSC achieve maximum scores of 44.84, 47.76, 48.46 BLEU points, respectively, for Uyghur–Chinese translation. Thus, the proposed hybrid framework can get the best result of a 48.42 BLEU score and achieved an improvement of +1.75 and +0.66 BLEU points than the best single system (TNMT) and the best individual system combination model (VOSC-2), respectively. We have done significance tests and observed that the results are significant with 95% confidence level (with $\rho = 0.05$ which is < 0.05) for the Uyghur–Chinese translation task. Thus, the proposed method stands to be statistically significant.

Table 9. Translation results (BLEU score) for the hybrid framework.

Systems	MUSC	VOSC	HBSC
PBMT + HPMT + TNMT	42.37	46.03	46.38
PBMT + HPMT + RNMT	36.47	39.65	40.13
TNMT-1 + TNMT-2 + TNMT-3	43.52	47.19	47.25
PBMT + HPMT + RNMT + TNMT	36.79	44.56	44.86
PBMT + HPMT + RNMT + TNMT + TNMT-1	40.14	46.75	47.18
PBMT + HPMT + RNMT + TNMT + TNMT-1 + TNMT-2	44.54	47.33	47.48
PBMT + HPMT + RNMT + TNMT + TNMT-1 + TNMT-2 + TNMT-3	**44.84**	**47.76**	**48.42**

In addition to accuracy, fluency is also an important factor in evaluating the quality of machine translation and we want to know whether the fluency of the proposed model has improved. We evaluated by the automatic evaluation metrics RIBES [56], whose score is a metric based on rank correlation coefficients with word precision. Figure 5 illustrates the experimental results of RIBES scores, which demonstrates that the proposed model outperforms the best result of a single MT system and individual system combination model. The experiment shows that our proposed model can further improve the fluency of machine translation.

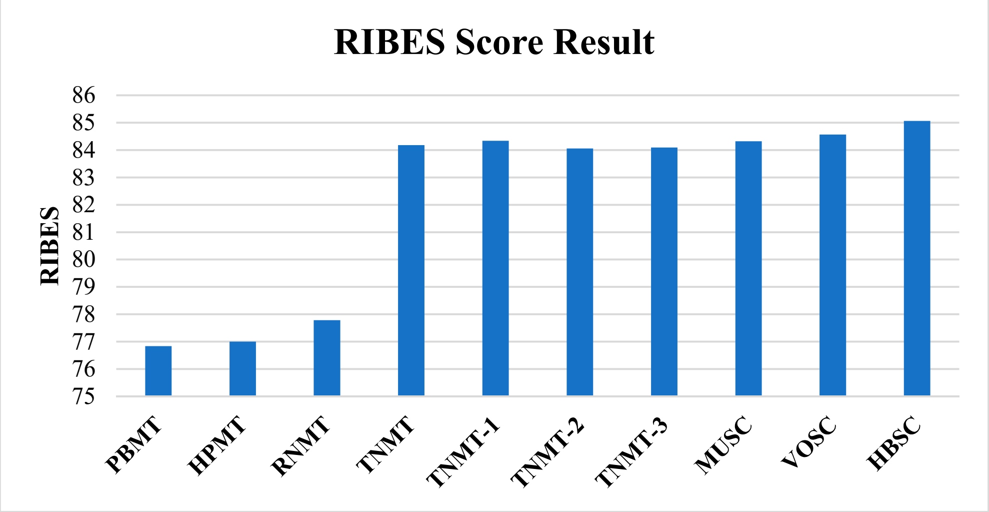


Figure 5. Translation results (RIBES score) for the hybrid framework.

4.5. Case Study

According to the translation results as listed in Table 10, it can be seen that firstly, all the outputs including single systems and the system combination model guarantee the accuracy of the information transmission from the source to the target sentence and the faithfulness of the source sentence. Secondly, the translation hypothesis of the SMT model is slightly less coherent than that of the NMT. Thirdly, the four stronger transformer models either mistranslated the keyword “规模 (Scale)” to “面积 (Area)” or failed to translate the key phrase “زور دەرىجىدە ئاشقان (Far more than)”. In short, no single system can obtain promising translation results. The system combination model based on the multi-source method (MUSC) can obtain the correct keyword “规模 (Scale)” from multiple translation hypotheses but cannot capture the relationship between “政策 (Policy)” and “岗位 (Jobs)”. Meanwhile, not only does the system combination model based on voting mechanism (VOSC) receive the wrong key word “面积 (Area)” due to the essence of this method (the voting is misled by most wrong words “面积”) but also cannot handle the relationship between “政策 (Policy)” and “岗位 (Jobs)”. The output of the proposed hybrid framework can remedy the errors and obtain high-quality translations in this case.

Table 10. Translation examples of single systems and our proposed hybrid framework. Tsrc, TPBMT, THPMT, TRNMT, TTNMT, TTNMT-1, TTNMT-2, TTNMT-3 are English translations of the src, PBMT (phrase-based statistical machine translation (SMT)), HPMT (hierarchical phrase-based SMT), RNMT, TNMT (Transformer-based neural machine translation (NMT)), TNMT-1, TNMT-2, TNMT-3, respectively.

src	**بۇ يىللىق ئىشقا ئورۇنلاشتۇرۇش سىياسىتىدىكى يەنە بىر ئالاھىدىلىك سىياسەت بىلەن ئىش ئورنىنى ماسلاشتۇرۇش بولۇپ ، ئىش ئورنى كۆلىمى ئۆتكەن يىللاردىكىدىن زور دەرىجىدە ئاشقان.**
Ref	今年 就业 政策 的 另 一个 特点 是 政策 与 岗位 配套 出台 ， 就业 岗位 的 规模 数量 大大 超过 了 往年.
Tsrc	Another characteristic of this year's employment policy is that the policy and job matching are issued, and the number of jobs has greatly exceeded that of previous years.
PBMT	今年 的 就业 政策 的 还有 一个 特点 和 就业 岗位 政策 协调 ， 在 岗位 面积 比 往年 大幅 增长.
TPBMT	This year's employment policy still has a characteristic and employment post policy is coordinated, in post area is larger than in previous years.
HPMT	今年 的 就业 政策 还有 一个 特点 和 政策 协调 岗位 ， 是 去年 的 岗位 面积 比 往年 大幅 增长.

THPMT	This year’s employment policy still has a characteristic and policy coordination post, it is last year's post area to increase substantially than in previous years.
RNMT	今年 就业 政策 的 还有 一个 特点 ， 在 岗位 上 协调 协调 ， 岗位 规模 也 大大提高.
TRNMT	The employment policy still has a characteristic this year, coordinate coordinate on post, post scale is also raised greatly.
TNMT	今年 就业 政策 的 另 一个 特点 是 协调 政策 和 岗位 ， 岗位 规模 比 去年 大.
TTNMT	Another feature of this year's employment policy is the coordination of policies and jobs, which are larger than last year’s.
TNMT-1	今年 就业 政策 的 另 一个 特点 是 政策 和 岗位 的 协调 ， 岗位 面积 比 往年 大幅 增长.
TTNMT-1	Another feature of this year's employment policy is the coordination of policies and posts, and the number of posts has increased significantly compared with previous years.
TNMT-2	今年 就业 政策 的 另 一个 特点 就是 政策 和 岗位 协调 ， 岗位 面积 比 往年 大幅 增加.
TTNMT-2	Another feature of this year's employment policy is the coordination between policies and posts, and the number of posts has increased significantly compared with previous years.
TNMT-3	今年 就业 政策 的 另 一个 特点 是 政策 与 岗位 对接 ， 岗位 面积 比 往年 大幅 提升.
TTNMT-3	Another feature of this year's employment policy is the matching of policies with posts. The number of posts has increased significantly compared with previous years.
MUSC	今年 就业 政策 的 另 一个 特点 就是 政策 和 岗位 上 的 协调 ， 岗位 规模 比 往年 大幅 提升.
TMUSC	Another feature of this year's employment policy is the coordination between policies and jobs, with the number of jobs increased significantly compared with previous years.
VOSC	今年 就业 政策 的 另 一个 特点 就是 政策 协调 ， 岗位 面积 比 往年 大幅 增加.
TVOSC	Another feature of this year's employment policy is policy coordination. The number of jobs has increased significantly compared with previous years.
HBSC	今年 就业 政策 的 另 一个 特点 就是 **政策 和 岗位 协调**， 岗位 **规模** 比 往年 **大幅 增加**.
THBSC	Another feature of this year's employment policy is the coordination between policies and posts, and the number of posts has increased significantly compared with previous years.

5. Conclusions

In this work, we proposed a hybrid system combination framework for a Uyghur–Chinese machine translation task. The central idea was to take advantage of various system combination models. Though the proposed model was a little bit more complex than the individual system combination model, the improvement was remarkable. Experiments show that the proposed approaches can obtain significant improvements over the best individual system and the state-of-the-art system combination method. Finally, we can conclude that integration of system combination models can not only address the adequacy of the NMT and the fluency of the SMT, but also can better utilize the advantages of individual system combination models. In the future, we plan to expand our hybrid framework to incorporate the statistics-based system combination method and add a post editing layer on top of the framework to further improve the effect of the combination.

Author Contributions: Y.W., X.L., Y.Y., A.A., and R.D. conceived the approach and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Xinjiang High-level Talent Introduction Project (Xinrenshehan [2017] No. 699); The Western Light of the Chinese Academy of Sciences under Grant No.2017-XBQNXZ-A-005; The National Natural Science Foundation of China under Grant No. U2003303; The Subsidy of the Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant No.2017472

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

References

1. Koehn, P.; Och, F.J.; Marcu, D. Statistical Phrase-Based Translation. In Proceedings of the Conference Combining Human Language Technology Conference Series and the North American Chapter of the Association for Computational Linguistics Conference Series (HLT-NAACL 2003), Edmonton, AB, Canada, 27 May–1 June 2003; pp.48–54.
2. Chiang, D. A hierarchical phrase-based model for statistical machine translation. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics—ACL '05, Ann Arbor, MI, USA, 25–30 June, 2005; pp. 263–270.
3. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceeding of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
4. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1243–1252.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
6. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, 30 July–4 August 2017; pp.28–39.
7. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus Phrase-Based Machine Translation Quality: A Case Study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP2016), Austin, TX, USA, 1–5 November 2016; pp.257–267.
8. Dunder, I.; Seljan, S.; Pavlovski, M. Automatic Machine Translation of Poetry and a Low-Resource Language Pair. In Proceedings of the 43rd International Convention on Information, Communication and Electronic Technology (MI-PRO2020), Opatija, Croatia, 28 September–2 October 2020; pp. 1034–1039.
9. Martindale, M.; Carpuat, M.; Duh, K.; McNamee, P. Identifying fluently inadequate output in neural and statistical machine translation. In Proceedings of the Machine Translation Summit XVII; Volume 1: Research, MTSummit 2019, Dublin, Ireland, 19–23 August 2019; pp. 233–243.
10. Tu, Z.; Liu, Y.; Shang, L.; Liu, X.; Li, H. Neural machine translation with reconstruction. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp.3097–3103.
11. Seljan, S.; Dunder, I. Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain. In Proceedings of the 10th Iberian Conference on Information Systems and Technologies (CISTI), Agueda, Aveiro, Portugal, 17–20 June 2015; pp.128–131.
12. 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, Berlin, Germany, 7–12 August 2016; pp.1715–1725.
13. Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; Li, H. Modeling Coverage for Neural Machine Translation. In Proceedings of the 54th Annual meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 76–85.
14. Abudukelimu, H.; Yang, L.; Maosong, S.U.N. Performance comparison of neural machinetranslation systems in Uyghur-Chinese translation. *J. Tsinghua Univ. Sci. Technol.* **2017**, *57*, 878–883.
15. Zhou, L.; Hu, W.; Zhang, J.; Zong, C.; Barzilay, R.; Kan, M.-Y. Neural System Combination for Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 31 July–4 August 2017; pp. 378–384.
16. Zhou, L.; Zhang, J.; Kang, X.; Zong, C. Deep Neural Network--based Machine Translation System Combination. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2020**, *19*, 1–19, doi:10.1145/3389791.
17. Huang, X.; Zhang, J.; Tan, Z.; Wong, D.F.; Luan, H.; Xu, J.; Sun, M.; Liu, Y. Modeling Voting for System Combination in Machine Translation. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020, Yokohama, Japan, 11–17 July 2020; pp. 3694–3701.
18. Kong, J.; Yang, Y.; Zhou, X.; Wang, L.; Li, X. Research for Uyghur-Chinese Neural Machine Translation. In *Constructive Side-Channel Analysis and Secure Design*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; Volume 10102, pp. 141–152.

19. Zhang, S.; Mahmut, G.; Wang, D.; Hamdulla, A. Memory-augmented Chinese-Uyghur neural machine translation. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 1092–1096.
20. Pan, Y.; Li, X.; Yang, Y.; Dong, R. Multi-Source Neural Model for Machine Translation of Agglutinative Language. *Future Internet* **2020**, *12*, 96, doi:10.3390/fi12060096.
21. Zhang, X.; Li, X.; Yang, Y.; Wang, L.; Dong, R. Analysis of Bi-directional Reranking Model for Uyghur-Chinese Neural Machine Translation. *Beijing Da Xue Xue Bao* **2020**, *56*, 31–38.
22. Wang, Y.; Li, X.; Yang, Y.; Anwar, A.; Dong, R. Research of Uyghur-Chinese Machine Translation System Combination Based on Semantic Information. In Proceedings of the 8th CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC2019), Dunhuang, China, 9–14 October 2019; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 497–507.
23. Frederking, R.; Nirenburg, S. Three heads are better than one. In Proceedings of the fourth conference on Applied natural language processing, Stuttgart, Germany, 13–15 October 1994; pp. 95–100.
24. Bangalore, B.; Bordel, G.; Riccardi, G. Computing consensus translation from multiple machine translation systems. In Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, Italy, 9–13 December 2001; pp. 351–354.
25. Rosti, A.-V.; Matsoukas, S.; Schwartz, R. Improved word-level system combination for machine translation. In Proceedings of Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 24–29 June 2007; pp. 312–319.
26. Ma, W.-Y.; McKeown, K. Phrase-level system combination for machine translation based on target-to-target decoding. In Proceeding of the 10th Biennial Conference of the Association for Machine Translation in the Ameri-cas (AMTA), San Diego, CA, USA, 28 October–1 November 2012.
27. Zhu, J.; Yang, M.; Li, S.; Zhao, T.; Che, W.; Han, Q.; Wang, H.; Jing, W.; Peng, S.; Lin, J.; et al. Sentence-Level Paraphrasing for Machine Translation System Combination. In Proceedings of the Yong Computer Scientists and Educators (ICYCSEE2016), Harbin, China, 20–22 August 2016; pp. 612–620.
28. Ma, W.-Y.; McKeown, K. System Combination for Machine Translation through Paraphrasing. In Proceedings of the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP2015), Lisbon, Portugal, 17–21 September 2015; pp. 1053–1058.
29. Freitag, M.; Huck, M.; Ney, H. Jane: Open Source Machine Translation System Combination. In Proceedings of the Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL2014), Gothenburg, Sweden, 26–30 April 2014; pp. 29–32.
30. Barrault, L. Many: Open Source Machine Translation System Combination. *Prague Bull. Math. Linguist.* **2010**, *93*, 147–155, doi:10.2478/v10108-010-0001-y.
31. Heafield, K.; Lavie, A. Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme. *Prague Bull. Math. Linguist.* **2010**, *93*, 27–36, doi:10.2478/v10108-010-0008-4.
32. Matusov, E.; Ueffing, N.; Ney, H. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006), Trento, Italy, 3–7 April 2006; pp.33–40.
33. 33. Rosti, A.-V.; Ayan, N.F.; Xiang, B.; Matsoukas, S.; Schwartz, R.; Dorr, B. Combining outputs from multiple machine translation systems. In Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, Rochester, New York, NY, USA, 22–27 April 2007; pp. 228–235.
34. Feng, Y.; Liu, Y.; Mi, H.; Liu, Q.; Lu, Y. Lattice-based system combination for statistical machine translation. In Proceedings of the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3—EMNLP ’09, Suntec, Singapore, 6–7 August 2009; pp. 1105–1113.
35. Rosti, A.-V.I.; Zhang, B.; Matsoukas, S.; Schwartz, R. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, OH, USA, 15–20 June 2008; pp. 183–186.
36. Freitag, M.; Peter, J.-T.; Peitz, S.; Feng, M.; Ney, H. Local System Voting Feature for Machine Translation System Combination. In EMNLP2015 Tenth Workshop on Statistical Machine Translation (WMT2015), Lisbon, Portugal, 17–21 September 2015; pp.467–476.
37. Freitag, M.; Ney, H.; Yvon, F. Investigations on Machine Translation System Combination. Ph.D. Thesis, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany, 2017.
38. Zoph, B.; Knight, K. Multi-Source Neural Translation. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT’16), San Diego, CA, USA, 12–17 June 2016; pp. 30–34.
39. He, X.; Yang, M.; Gao, J.; Nguyen, P.; Moore, R. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2008), Honolulu, HI, USA, 25–27 October 2008; pp. 98–107.

40. Banik, D.; Ekbal, A.; Bhattacharyya, P.; Bhattacharyya, S.; Platos, J. Statistical-based system combination approach to gain advantages over different machine translation systems. *Heliyon* **2019**, *5*, e02504, doi:10.1016/j.heliyon. 2019.e02504.
41. Banik, D.; Ekbal, A.; Bhattacharyya, P.; Bhattacharyya, S. Assembling translations from multi-engine machine translation outputs. *Appl. Soft Comput.* **2019**, *78*, 230–239, doi:10.1016/j.asoc.2019.02.031.
42. Marie, B.; Fujita, A. A smorgasbord of features to combine phrase-based and neural machine translation. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA2018), Boston, MA, USA, 17–21 March 2018; pp. 111–124.
43. Rikters, M. Hybrid Machine Translation by Combining Output from Multiple Machine Translation Systems. *Balt. J. Mod. Comput.* **2019**, *7*, 301–341, doi:10.22364/bjmc.2019.7.3.01.
44. Chen, B.; Zhang, M.; Li, H.; Aw, A. A comparative study of hypothesis alignment and its improvement for machine translation system combination. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, 2–7 August 2009; pp. 941–948.
45. Libovický, J.; Helcl, J. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 196–202.
46. Libovický, J.; Helcl, J.; Mareček, D. Input Combination Strategies for Multi-Source Transformer Decoder. In Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 253–260.
47. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002), Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
48. Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004), Barcelona, Spain, 25–26 July 2004; pp. 388–395.
49. Stolcke, A. SRILM-an extensible language modeling toolkit. In Proceedings of the 7th international conference on spoken language processing, Denver, CO, USA, 16–20 September 2002; pp. 901–904.
50. Och F J. Giza++: Training of Statistical translation models. Available online: http://www.statmt.org/moses/giza/GIZA++.html (accessed on 30 January 2001).
51. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, Prague, Czech Republic, 25–27 June 2007; pp. 177–180.
52. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MN, USA, 2–7 June 2019; pp. 48–53.
53. Helcl, J.; Libovický, J. Neural Monkey: An Open-source Tool for Sequence Learning. *Prague Bull. Math. Linguist.* **2017**, *107*, 5–17, doi:10.1515/pralin-2017-0001.
54. Zhang, J.; Ding, Y.; Shen, S.; Cheng, Y.; Sun, M.; Luan, H.; Liu, Y. Thumt: An open source toolkit for neural machine translation. *arXiv* **2017**, arXiv:1706.06415.
55. Tan, Z.; Zhang, J.; Huang, X.; Chen, G.; Wang, S.; Sun, M.; Luan, H.; Liu, Y. THUMT: An Open-Source Toolkit for Neural Machine Translation. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020), Orlando, FL, USA, 8–12 September 2020; pp. 116–122.
56. Isozaki, H.; Hirao, T.; Duh, K.; Sudoh, K.; Tsukada, H. Automatic evaluation of translation quality for distant language pairs. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP2010), Cambridge, MA, USA, 9–11 October 2010; pp. 944–952.