doi:10.3772/j.issn.1006-6748.2023.03.010

Research on system combination of machine translation based on Transformer①

LIU Wenbin(刘文斌), HE Yanqing②, LAN Tian, WU Zhenfeng
(Research Center for Information Science Theory and Methodology, Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract

Influenced by its training corpus, the performance of different machine translation systems varies greatly. Aiming at achieving higher quality translations, system combination methods combine the translation results of multiple systems through statistical combination or neural network combination. This paper proposes a new multi-system translation combination method based on the Transformer architecture, which uses a multi-encoder to encode source sentences and the translation results of each system in order to realize encoder combination and decoder combination. The experimental verification on the Chinese-English translation task shows that this method has 1.2 – 2.35 more bilingual evaluation understudy (BLEU) points compared with the best single system results, 0.71 – 3.12 more BLEU points compared with the statistical combination method, and 0.14 – 0.62 more BLEU points compared with the state-of-the-art neural network combination method. The experimental results demonstrate the effectiveness of the proposed system combination method based on Transformer.

Keywords: Transformer, system combination, neural machine translation (NMT), attention mechanism, multi-encoder.

0 Introduction

Machine translation refers to the process of translating one language into another with a computer. Both statistical machine translation (SMT) and neural machine translation (NMT) are particularly dependent on their training corpus. Translation models show varying performance with training data of different sources, and their translation effects are rather spotty. Combining the translation results of multiple systems can not only improve the generalization ability of the translation model, but also expand the translation hypothesis space. Therefore, system combination usually produces a performance that is commensurate with or even better than a single translation model.

At present, combining machine translation systems is mainly achieved in one of two ways: one is based on statistics, and the other is based on neural networks. The statistical method designs multiple features and uses a voting mechanism to combine the translation results from multiple systems at the sentence level, phrase level, or word level. These methods focus on the potential mapping relationship between the translation hypotheses of multiple systems and the search space of a single system[1-5]. However, these methods are not end-to-end modeling methods, and there are some error propagation problems in the combination process. Neural network combination includes two mechanisms: model-level combination and parameter-level combination. The model-level combination method involves the translation results of multiple machine translation systems, adopts a recurrent neural network (RNN) encoding-decoding neural network architecture and end-to-end modeling, and fuses the source context information in the encoder and decoder. The parameter-level combination method combined with the prediction probabilities of multiple decoders can predict the translation at the next moment, instead of absorbing the translation results of multiple machine translation systems; only the decoding strategy of the model average or model ensemble within the system is adopted to fuse the system.

Different from the above methods, this paper draws on the research methods of paragraph translation[6-7], and proposes a multi-system translation com-

① Supported by the National Key Research and Development Program of China (No. 2019YFA0707201) and the Fund of the Institute of Scientific and Technical Information of China (No. ZD2021-17).

② To whom correspondence should be addressed. E-mail: heyq@istic.ac.cn.

Received on June 28, 2022

bination method based on the Transformer architecture, which uses a multi-encoder to encode the source language sentences and the translation results of each system to realize encoder combination and decoder combination. Encoder combination transforms the hidden information of the multi-system translation into a new representation through an attention network, and the hidden layer information of synonymous sentences is fused through a gating mechanism at the encoder side. Decoder combination calculates the attention of the hidden layer information of the multi-system translation and the source language sentence at the decoder side to obtain the fusion vectors, thus obtaining a combination translation with higher quality.

In this work, experimental verification of the proposed method on a Chinese-English translation task shows that compared with the best single system results, this method has 1.2 – 2.35 more bilingual evaluation understudy (BLEU) points, 0.71 – 3.12 more BLEU points compared with the statistical combination method, and 0.14 – 0.62 more BLEU points compared with the neural network combination method. The experimental results show that the combination method of machine translation system based on Transformer can effectively improve the translation quality.

The main contributions of this study can be summarized as follows.

(1) The proposed neural network system combination method based on Transformer introduces the source language and the translation assumptions of multiple systems into the Transformer architecture for combination.

(2) The multi-encoder method is proposed to encode the translation results of multiple systems, and achieve two different combination models of system translation, namely encoder combination and decoder combination.

1 Related work

Research on the combination of machine translation systems started in the 1990s, and the early combination technology was mainly based on statistical methods. According to the different levels of target translation in the combination process, the statistical combination method can be divided into three categories: (1) Sentence-level system combination. Taking sentences as the smallest unit, sentence-level system combination recalculates the translation result score of multiple systems of the same source language sentences by using minimum Bayes risk decoding or a logarithmic – linear model[8]. Theoretically, the method of sentence-level system combination will not produce new translation hypotheses, but only choose the best one among existing translation hypotheses. (2) Phrase-level system combination. Different from sentence-level system combination, the core idea of phrase-level system combination method is to use a phrase table for further decoding[9]. (3) Word-level system combination, which takes words as the minimum unit. First, a confusion network is constructed by using the word alignment method of sentence pairs in the same language[10], the confidence of candidate words in each position of the confusion network is estimated, and then the confusion network is decoded. Compared with sentence-level system combination and phrase-level system combination, word-level system combination is more effective because of its finer combination granularity and more accurate word alignment, thus showing huge performance advantages. However, the statistical combination method is not an end-to-end modeling method, and there are some error propagation problems in the combination process.

Due to the growing popularity of NMT since it was proposed in 2015, the system combination method has also turned into a new neural network pattern that contains two mechanisms of neural network combination at the model-level and parameter-level. Model-level combination incorporates the translation results of multiple machine translation systems, and then builds a neural network architecture to realize end-to-end modeling. Ref. [11] adopted a long short-term memory (LSTM) network to employ multiple encoders on the source end, where each encoder corresponds to source language sentences in different languages, and the target end contains a decoder which uses the sum of the last layer states of multiple encoders on the source end to initialize the state hidden layer vector on the target end. Ref. [12] adopted a multisource translation strategy, inputting source statements in different languages at the input level, and averaging the probability distributions of target words generated in different languages at the output level, to reduce the errors of the model's predicted probabilities. Based on the research of Ref. [12], Ref. [13] achieved combination by dynamically controlling the contribution of different translation models to the target-side probability prediction through a gating mechanism. Ref. [14] proposed a neural system combination framework leveraging multi-source NMT. Ref. [15] used an RNN to classify and expand the system combination method and carried out experimental comparisons five ways, including average combination, weight combination, splicing combination, gate mechanism combination, and attention combination. Ref. [16] proposed a deep-neural-network-based machine

translation system combination. Ref. [17] proposed an approach to model voting for system combination in machine translation.

Parameter-level combination combines the prediction probability of multiple decoders within the same codec framework to predict the translation at the next moment. Without storing translation results of multiple machine translation systems, it only uses the decoding strategy of a model ensemble[18-19] and model average[20] in the system to fuse at the model parameter level. Ref. [21] improved the Transformer model by using a convolutional neural network (CNN) and gating mechanism and guided the optimization of model parameters using confrontation training, then reorganized and merged the output from multiple machine translations into a single improved translation result through multi-model fusion. Ref. [22] presented a hybrid framework for developing a system combination for the Uyghur-Chinese machine translation task that works in three layers to transmit the outputs of the first layer and the second layer into the final layer to make better predictions. Ref. [23] proposed an NMT model that treated the generated machine translation outputs as an approximate contextual environment of the target language, and then re-decoded each token in the machine translation output successively.

In this paper, a model-level neural network combination method based on Transformer is adopted. Different from previous model-level work, this paper takes the multi-encoder method to encode the source language sentences, merges the translation results of various systems, and selects the Transformer neural network architecture instead of the previous RNN and LSTM neural networks. Compared with parameter-level combination, the proposed method dynamically constructs the attention between the multi-system translation and the source language sentence, as well as the attention between the multi-system translation and the target language sentence, providing additional contextual information and finally realizing the combination of multi-system translation.

2 Model

2.1 Problem definition

Given source language sentence x and multiple system translations $Tr = \{Tr_k \mid 1 \leqslant k \leqslant K\}$, the purpose of system combination is to find the target language translation $\hat{y}$ with the highest combination probability. The calculation formula is as follows.

$$\hat{y} = \operatorname{argmax} P(y \mid x, T_r), \tag{1}$$

where $x = \{x_1, \cdots, x_m\}$ represents the sequence of words in the source sentence, $Tr_k = \{Tr_{k1}, \cdots, Tr_{km}\}$ represents the sequence of words in the k^{th} system translation, and $y = \{y_1, \cdots, y_n\}$ represents the sequence of generated target words. The combination system generates the translation word by word from left to right, and as word y_i is translated, the generated results are taken into account. Therefore, the following formula could be derived.

$$P(y \mid x, T_r) = \prod_{j=1}^{n} P(y_j \mid y < j, x, T_r), \tag{2}$$

where $y < j$ indicates the word sequence of the combination translation that has been generated before the j^{th} position in the target language $\{y_1, \cdots y_{j-1}\}$. $P(y_i \mid y < j, x, T_r)$ is the probability of generating the j target word based on the source language sentence x, the multi system translation T_r, and the generated target language combination translation segment $\{y_1, \cdots y_{j-1}\}$.

2.2 System combination based on transformer

In 2017, Ref. [24] proposed a complete attention mechanism encoder-decoder structure, Transformer, to realize machine translation. In the Transformer structure, the encoder consists of six layers of networks with the same structure. Each layer is composed of two parts. The first part is multi-head self-attention, and the second part is a position-wise feed-forward network, which is a fully connected layer. Both parts have a residual connection and layer normalization. The decoder also consists of six layers of the same network stack, where each layer is composed of three parts, namely multi-head self-attention, multi-head encoder – decoder attention, and a position-wise feed-forward network. Like the encoder, each part has a residual connection and layer normalization.

In this paper, research on translation with document-level context is based on the work of Ref. [25]. Transformer is also used in multi-system translation combination, and the overall model architecture is shown in Fig. 1. The model still uses an encoder-decoder structure, and the single encoding layer and decoding layer are the same as that in the original structure. In each sublayer, a residual connection and layer normalization are used. The encoder is a multi-encoder, which accepts the input of source language sentences and multiple system translations at the same time, and encodes the multi-system translations and source sentences as intermediate hidden layer vectors. That is, a source-sentence encoder and a multiple-translation encoder are obtained. The decoder decodes the target language words one by one according to the intermediate vectors to form a combined translation. Translation attention is introduced on the source lan-

guage encoder side and the target language decoder side, aiming at making full use of the encoding information of the multi-system translation in the encoder and decoder for attention combination.

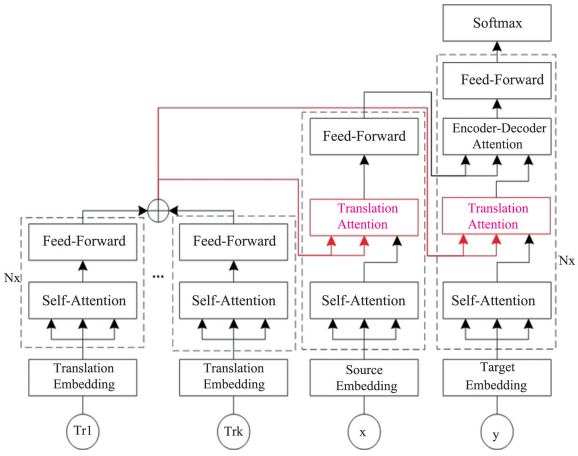

Fig. 1 Model architecture

In Transformer, the attention mechanism can be regarded as a process of calculating a given series of queries Q and a series of key-value pairs K and V, obtaining the weight of V through the calculation of Q and K, and then carrying out the weighted sum of V:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{Q K^{\mathrm{T}}}{\sqrt{dk}}\right)V \tag{3}$$

where, d_K represents the dimension of K, and 64 is used by default. The multi-head attention mechanism used in the Transformer model can simultaneously 'notice' multiple different locations and capture different levels of information. In the encoder's self-attention, Q, K, and V are all from the output of the previous layer of the encoder, and in the decoder's self-attention, Q, K, and V are all from the output of the previous layer of the decoder. In the encoder-decoder attention, Q comes from the output of the previous layer of the decoder, while K and V come from the encoder.

Using Transformer-based system combination, this paper introduces two ways of combining translation information into Transformer.

2.2.1 Encoder combination

Encoder combination model is as shown in Fig. 2, the encoder combination uses multiple system translations, and then converts the system translations into new representations through the attention network, integrating the hidden layer information of homologous language sentences for attention fusion through the gating mechanism in the encoder. In the encoder combination mode and in the self-attentionof the multi-system translation encoder, Q, K, and V are all from the upper layer output of the multi-system translation encoder;

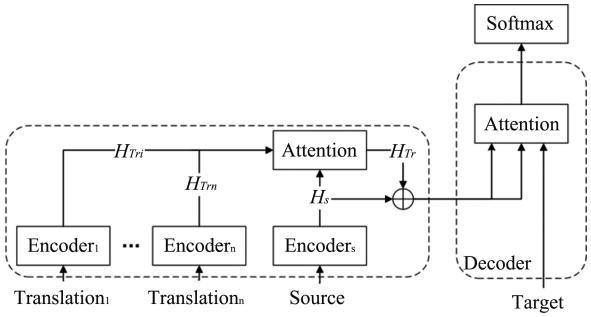

Fig. 2 Encoder combination model

in the self-attentionof the source language encoder, Q, K, and V are all from the upper layer output of the source language encoder. In the translation attention of the source language encoder, both K and V come from the upper hidden layer state H_{T_r} of the multi-system translation encoder, and Q comes from the upper layer hidden state H_s of the source language encoder. The hidden state of the translation attention part of the encoder, H, is given as

$$H_{T_r} = \text{Concat}(H_{T_{r1}}, \cdots, H_{T_{rn}}) \tag{4}$$

$$H = \text{MultiHead}(H_s, H_{T_r}) \tag{5}$$

where, H_s represents the hidden state of the source language sentence, and H_{T_r} represents the hidden state of the multi-system translation.

2.2.2 Decoder combination

As shown in Fig. 3, the decoder combination method combines the hidden layer information of multiple encoders with the attention in the decoder. The decoder can process multiple encoders separately, and then fuse them using the gating mechanism inside the decoder to obtain the combined vector. In the decoder combination mode and in the self-attention of the target language decoder, Q, K, and V are all from the output of the previous layer of the target language decoder; in the translation attention of the target language decoder, Q comes from the output of the upper layer of the target language decoder, K comes from the upper hidden layer state H_s of the source language encoder, and V comes from the

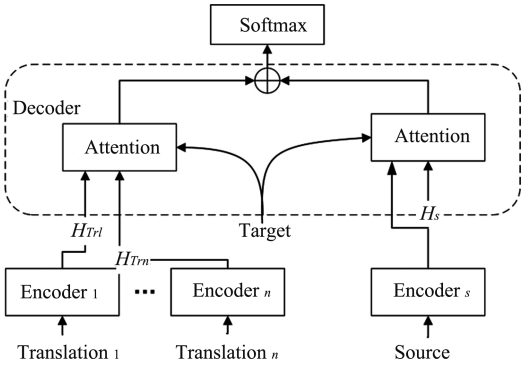

Fig. 3 Decoder combination model

upper hidden layer state H_{T_r} of the multi-system translation encoder. In the encoder-decoder attention of the target language decoder, Q comes from the upper layer output of the target language decoder and K and V come from the previous output of the source language encoder. H, the hidden state of the translation attention part of the decoder, is given as

$$H = \text{MultiHead}(H_s, H_{T_r}, H_{\text{Decoder}}), \quad (6)$$

where H_s represents the hidden layer state of the source language sentence, H_{T_r} represents the hidden layer state of the multi-system translation, and H_{Decoder} represents the hidden layer state of the upper layer output of the decoder.

3 Experiments

The experiments on the Chinese-English translation task are performed, where the implementation was based on Fairseq[26] and the evaluation metric was the case-insensitive BLEU[27].

3.1 Data and settings

This paper evaluated the combination approach using several publicly available data sets. The training data consisted of 20×10^4 pairs of sentences randomly extracted from NEU2017 of the Chinese-English translation tracks in CCMT2020. The NJU-newsdev 2017 data set and NJU-newstest2017 data set are also used as a validation set and test set, respectively. The NIST 2003 – 2006 Chinese-English data set was also used as a test set. All sentences in the data were preprocessed with the Urheen[28] tokenizer, which used byte pair encoding[29] with 32k merged operations to segment words into subword units. The data statistics of each data set can be seen in Table 1. Adam[30] was used for optimization, and the systems are trained by using one to four GPUs. The learning rate strategy was the same as that used in Ref. [24].

Table 1 Details of the datasets

Data set	Source	Scale
Train	NEU2017	20×10^4
Valid	NJU-newsdev2017	2002
	NJU-newstest2017	1000
Test	NIST 2003	919
	NIST 2004	1788
	NIST 2005	1082
	NIST 2006	1664

3.2 Training Details

In order to verify the effectiveness of this method, three groups of experiments were set up for comparison: combination model versus single system model, combination model versus statistical combination model, and combination model versus neural network combination model. For the single system model, the same data set was used to train the two Chinese-English NMT systems, Sys1 and Sys2, based on Fairseq with the same model but different initialization seeds. Some parameter settings of the training model are shown in Table 2. For the statistical combination model, it can be implemented the combination of five different word alignment methods based on the trained single system model[2], namely word alignment based on the word error rate (WER), word alignment based on the translation error rate (TER), word alignment based on word ordering (WRA), word alignment based on an indirect hidden Markov model (IHMM), and word alignment based on an incremental hidden Markov model (INCIHMM). For the neural network combination model, the decoding strategy of model averaging and model integration based on the trained single system model for combination were implemented.

Table 2 Settings of some parameters

Parameter	Value
lr	0.0007
dropout	0.3
max_tokens	3000
max_epoch	30
adam_betas	(0.9, 0.997)
warmup	4000

3.3 Results and discussion

Compare the proposed neural combination system with the best individual engines, the statistical combination system, and the neural network combination system. The BLEU points of the different models for the development data and test data are shown in Tables 3 to 5.

As shown in Table 3, on the one hand, the performance of single system Sys2 was superior to Sys1, with an average of 0.69 BLEU points higher. Compared with Sys1, the encoder combination model had 1.73 – 3.19 more BLEU points, with an average increase of 2.41 BLEU points, while the decoder combination model had 1.09 – 2.18 more BLEU points, with an average increase of 1.61 BLEU points. Compared with Sys2, the encoder combination model had 1.84 – 3.36 more BLEU points, with an average increase of 2.58 BLEU points, while the decoder combination model had 1.2 – 2.35 more BLEU points, with an average increase of 1.78 BLEU points. On the other hand, except for the test set

NIST03, the BLEU scores of the five statistical combination models were all better than those of the two single system models. Among the five statistical combination models, the INCIHMM model performed the best, while the WRA model performed the worst. The INCIHMM model had 0.53 more BLEU points than the Sys2 model, on average. For the test set NIST03, the combination results of the five methods of statistical combination were all lower than the results of the single system before combination.

After analyzing the results, it was found that the distributions of the NIST 03 dataset and the development set were far from each other. Therefore, in the statistical combination model, the parameters of the development set failed to guide NIST 03 to combine well. Compared with the INCIHMM model, the encoder combination model had significantly more BLEU points (0.6 – 1.67), with an average increase of 1.06 BLEU points. The decoder combination model had 0.71 – 1.84 more BLEU points, with an average increase of 1.24 BLEU points. Compared with the WRA model, the encoder combination model had significantly more

Table 3 Translation results (BLEU scores) compared with the single system and the statistical combination method

System	Valid	Test	NIST03	NIST04	NIST05	NIST06	Ave
Sys1	10.35	12.13	19.83	22.02	19.48	18.53	14.62
Sys2	10.99	12.40	20.16	23.39	20.66	19.54	15.31
WRA	11.08	13.00	17.66	22.90	20.01	19.39	14.86
WER	11.38	13.32	18.43	23.70	20.70	19.92	15.35
IHMM	11.42	13.49	17.53	22.93	20.36	20.09	15.12
INCIHMM	11.48	13.46	18.86	23.82	20.82	20.05	15.50
TER	11.32	13.25	17.67	22.40	20.02	19.83	14.93
Encoder	12.08	14.28	21.86	25.04	21.81	21.72	16.68
Decoder	12.19	14.46	21.98	25.17	22.14	21.89	16.83

Table 4 Translation results (BLEU score) compared with the neural combination method

System	Valid	Test	NIST03	NIST04	NIST05	NIST06	Ave
Sys1	10.35	12.13	19.83	22.02	19.48	18.53	14.62
Sys2	10.99	12.40	20.16	23.39	20.66	19.54	15.31
Average	12.03	14.04	21.10	24.49	21.61	21.25	16.36
Ensemble	12.05	14.11	21.47	24.55	21.86	21.32	16.48
Encoder	12.08	14.28	21.86	25.04	21.81	21.72	16.68
+ Average	12.16	14.47	21.92	25.13	22.45	21.96	16.87
+ Ensemble	12.10	14.22	22.09	24.99	22.27	21.77	16.78
Decoder	12.19	14.46	21.98	25.17	22.14	21.89	16.83
+ Average	12.29	14.53	21.95	25.20	22.14	21.94	16.86
+ Ensemble	12.18	14.45	21.92	25.22	22.19	21.74	16.81

Table 5 Translation examples

Source	格雷西亚是一只萌萌的雌性巨嘴鸟,浑身散发着热带雨林的气息。
Pinyin	gé léi xī yà shì yī zhī méng méng de cí xìng jù zuǐ niǎo,hún shēn sàn fā zhe rè dài yǔ lín de qì xī。
Reference	Gracia is a cute female toucan, exuding the breath oftropical rain forest.
Sys1	Grace is an essential female mouth bird, which receives the scent of rainforest.
Sys2	Grace is the starting point of the female giant mouth birds, approaching the scent of rainforest.
INCIHMM	grace is the starting point of the female giant mouth bird, which receives the scent of rainforest.
Ensemble	Graeme was a beginning female big mouth bird, which smelled of tropical rain.
Encoder	Gracia is a beginning female big mouth bird, taking the breath of tropical forest.
Decoder	Gracia is a beginning of the female big mouth bird, taking the breath of tropical forest.

BLEU points (1 – 2.33), with an average increase of 1.71 BLEU points. The decoder combination model had 1.11 – 2.5 more BLEU points, with an average increase of 1.89 BLEU points. Overall, the method proposed in this paper had 0.71 – 3.12 more BLEU points compared with the statistical combination meth-

od, indicating that the encoder combination and decoder combination models were better than the statistical combination models.

As shown in Table 4, in the neural network combination model, the decoding strategies of the model average and the model ensemble were also superior to the single system model for the development set and the test set. The average result of the model was 0.94 – 1.71 BLEU points higher than that of Sys2, with an average increase of 1.23 BLEU points. The result of the model ensemble was 1.06 – 1.78 BLEU points higher than that of Sys2, with an average increase of 1.37 BLEU points. In this paper, the results of the encoder combination model were 0.05 – 0.76 BLEU points higher than the average result of the model, and the average increase was 0.38 BLEU points. The results of the decoder combination model were 0.16 – 0.88 BLEU points higher than the average result of the model, with an average increase of 0.55 BLEU points. The results of the decoder combination model were 0.14 – 0.62 BLEU points higher than the result of the model ensemble, with an average increase of 0.41 BLEU points. These results show that the encoder combination and decoder combination models are better than the decoding strategy of the model average and model ensemble to a certain extent. In addition, for the encoder combination and decoder combination methods, this word adopted model average and ensemble decoding strategies, respectively, and the results were further improved. It can be seen that the encoder combination method and decoder combination method can be combined with traditional integrated learning combination methods to obtain a more robust machine translation model.

Table 5 shows an example of system combination. The Chinese word rèdài is out-of-vocabulary (OOV) for Sys1 and Sys2, so the statistical combination model could not correctly translate this word. Although the ensemble model could translate this word well, it does not conform to the required grammar. By combining the merits of Sys1 and Sys2 based on Transformer, the model gets the correct translation. All in all, compared with the best results of statistical combination and neural network combination, the Transformer-based neural system combination method proposed in this paper can effectively integrate the translation results of multiple systems to obtain higher quality translations.

4 Conclusion

In this paper, a novel neural network framework based on Transformer was proposed for system combination of machine translation. The neural combination method is not only able to adopt multi-references and multi-source sentences in the combination process, but also addresses the attention mechanism of Transformer to get fluent translations. Furthermore, the approach can use average and ensemble decoding to boost the performance compared with traditional system combination methods. Experiments on Chinese-English data sets showed that the approaches obtained significant improvements over the best individual system and the traditional system combination methods. In future work, more translation results will be combined to improve the system combination quality further.

References

[1] LI H Z, ZHAO K, HU R F, et al. A hybrid system for Chinese-English patent machine translation[J]. Technology Intelligence Engineering, 2017, 6(3):105-115.

[2] LI M X, ZONG C Q. A survey of system combination for machine translation [J]. Journal of Chinese Information Processing, 2010, 7(4):74-85..

[3] MATUSOV E, UEFFING N, NEY H. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment[C]//Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento: Association for Computational Linguistics, 2006:33-40.

[4] OCH F J, NEY H. Statistical multi-source translation[C]//Proceedings of Machine Translation Summit VIII. Santiago de Compostela:Spain, 2001:253-258.

[5] SCHWARTZ L. Multi-source translation method[C]//Proceedings of the 8th Conference of the Association for Machine Translation in the Americas:Student Research Workshop. Waikiki:USA, 2008:279-288.

[6] LI B, LIU H, WANG Z Y, et al. Does multi-encoder help? a case study on context-aware neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle: ACL, 2020:3512-3518.

[7] SU J S, CHEN J X, LU Z Y, et al. A survey of document-level neural machine translation[J]. Technology Intelligence Engineering, 2020, 1(5):4-14.

[8] KUMAR S, BYRNE W. Minimum Bayes-risk decoding for statistical machine translation[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Massachusetts: Association for Computational Linguistics, 2004:169-176.

[9] MA W Y, MCKEOWN K. System combination for machine translation through paraphrasing[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015:1053-1058.

[10] FREITAG M, HUCK M, NEY H. Jane: open source machine translation system combination[C]//Proceedings of the Demonstrations at the 14th Conference of the 14th European Chapter of the Association for Computational Lin-

guistics. Gothenburg: Association for Computational Linguistics, 2014:29-32.
[11] ZOPH B, KNIGHT K. Multi-source neural translation[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016:30-34.
[12] FIRAT O, SANKARAN B, Al-ONAIZAN Y, et al. Zero resource translation with multi-lingual neural machine translation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016:268-277.
[13] GARMASH E, MONZ C. Ensemble learning for multi-source neural machine translation[C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee. Osaka: Association for Computational Linguistics, 2016:1409-1418.
[14] ZHOU L, HU W, ZHANG J, et al. Neural system combination for machine translation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017:378-384.
[15] TAN M, YIN M, DUAN X. System combination method for neural machine translation[J]. Journal of Xiamen University (Natural Science), 2019, 7(4):600-607. (In Chinese)
[16] ZHOU L, ZHANG J, KANG X, et al. Deep neural network based machine translation system combination[J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2020, 19(5):1-19.
[17] HUANG X C, ZHANG J C, TAN Z X, et al. Modeling voting for system combination in machine translation[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: IJCAI, 2020:3694-3701.
[18] SENNRICH R, BIRCH A, CURREY A, et al. The University of Edinburgh's neural MT systems for WMT17[C]//Proceedings of the 2nd Conference on Research on System Combination of Machine Translation 15Machine Translation. Copenhagen: Association for Computational Linguistics, 2017:389-399.
[19] WANG Y, CHENG S, JIANG L, et al. Sogou neural machine translation systems for WMT17[C]//Proceedings of the 2nd Conference on Machine Translation. Copenhagen: Association for Computational Linguistics, 2017: 410-415.
[20] SENNRICH R, HADDOW B, BIRCH A. Edinburgh neural machine translation systems for WMT 16[C]//Proceedings of the ist Conference on Machine Translation. Berlin: Association for Computational Linguistics, 2016: 371-376.
[21] WU Z Y, HOU H Y, BAI T G, et al. The research on multi-model fusion for Mongolian-Chinese neural machine translation based on CSGAN[J]. Journal of Jiangxi Normal University (Natural Science Edition), 2020, 44(2):153-159. (In Chinese)
[22] WANG Y J, LI X, YANG Y T, et al. Hybrid system combination framework for Uyghur-Chinese machine translation[J]. Information, 2021, 12(3):1-19.
[23] ZONG Q Q, LI M X. Research on neural machine translation based on re-decoding[J]. Journal of Chinese Information Processing, 2021, 35(6):39-46.
[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach: Association for Computational Linguistics, 2017: 5998-6008.
[25] ZHANG J, LUAN H, SUN M, et al. Improving the transformer translation model with document-level Context[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018:533-542.
[26] OTT M, EDUNOV S, BAEVSKI A, et al. Fairseq: a fast, extensible toolkit for sequence modeling[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Montréal: Association for Computational Linguistics, 2019:48-53.
[27] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Pennsylvania: Association for Computational Linguistics, 2002:311-318.
[28] WANG Kun. Urheen: A Chinese/English lexical analysis toolkit[EB/OL]. [2021-9-27]. http://www.nlpr.ia.ac.cn/cip/software.
[29] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016:1715-1725.
[30] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. (2014-11-22)[2021-9-27]. https://arxiv.org/pdf/1412.6980.pdf.

LIU Wenbin, born in 1995. He received the M. S. degree in Information Science in Institute of Scientific and Technical Information of China (ISTIC) in 2022, and the B. S. degree in Information Management and Information System of Wuhan University of Science and Technology in 2018. His research interests include natural language processing and data mining.