Deliberation Network-based Feature Fusion For Chinese- English Neural Machine Translation

Renshui Fan

School of Foreign Languages, Zhengzhou University of Science and Technology, Zhengzhou 450064 China

Corresponding author. E-mail: zhyongsfw@163.com

Received: April 08, 2025; Accepted: May 27, 2025

Neural machine translation (NMT) has achieved remarkable results in sentence-level translation, but the text problems of sentence-level translation, such as consistency and reference, are solved by using context information. Different from the previous methods using source context modeling, this paper proposes a novel Chinese-English neural machine translation that integrates target context information based on deliberation network. Specifically, with the help of the deliberating network, this paper makes a second translation of the source end of the text. The first translation is based on sentence level translation, and the second translation refers to the first translation of the whole text. The integration of domain knowledge into the translation model improves the effect of the translation model. The experimental results show that compared with the baseline model, the BLEU values of the Chinese-English and English-Chinese models are increased by 1.28 and 2.08.

Keywords: Chinese-English neural machine translation; Target context information; Deliberation network; Domain knowledge

http://dx.doi.org/10.6180/jase.202603_29(3).0003

1. Introduction

In recent years, neural machine translation has become a mainstream method in the field of machine translation, and its translation effect on multiple language pairs has reached the current optimal [1–3]. However, neural machine translation models usually take sentences as the translation unit and only model the current sentence, without considering the context information of the text. As a result, when the text is used as input, the dependency between the sentences in the text cannot be utilized, and the generated translation is prone to misturn, poor cohesion and other problems due to the absence of the text context. Recently, more and more relevant works have proposed various models to capture contextual information conducive to the translation of the current sentence from the source text other than the current sentence [4, 5], hoping to alleviate the problems of reference, translation consistency, translation coherence and other problems in the text translation through contextual information. Different from these efforts, this paper will explore how to effectively use the target text information (i.e., automatic translation) to improve the quality of text translation [6].

Text machine translation using the target-end (automatic) context usually involves two stages. In the first stage, the target translation at sentence level is obtained. In the second stage, by modeling part or all of the text translation of the first stage, the second translation is carried out. The models of the two phases can be independent of each other, or they can be related, such as sharing the encoder end. Zhang et al. [7] used the target context of part of the text to obtain the (sentence-level) translation of each sentence in the text in advance, and then repaired the text translation or regarded the text translation as similar to the source context. Chen et al. [8] needed to prepare local target context for each sentence in the chapter (such

as the first three sentences), and the model still took the sentence as the input unit. This would lead to a lot of computational redundancy in the model and increase the computational burden. In order to improve the cohesion between translated sentences, related reward functions were set up according to the text translation in reference [9]. In addition, references [10–12] used the method of caching vocabularies/sentences to cache the previous translation or the output state of the previous translation to try to maintain the consistency of the translation in the subsequent translation process. Different from these related works, this paper models the whole target context, obtains useful information conducive to the current sentence, carries out secondary translation, and obtains the final text level translation.

In general, text context modeling only uses the source information, while ignoring the target information. Empirically, the text context modeling method on the target end can solve the problems in text translation more directly [13, 14]. However, in the machine translation task, the input of the test process only contains the source text, so the sentence level translation of the text to be translated needs to be obtained first, and then the target text generated by the decoding is used to obtain the final text level translation [15]. Choo et al. [16] proposed the Grid Beam Search method to limit the translation of terms to ensure that terms and other words would appear in the translation. Ahammad et al. [17] improved the translation effect by combining the training data to solve the problem of sparse words in the training data. By using domain data set to fine-tune the model, San et al. [18] achieved a good improvement in the translation effect of the model in the domain. Siu et al. [19] integrated the terms in the conference scene into the translation model by means of splicing and fusion, and improved the translation quality of the model in the conference scene. Bala et al. [20] proved the effectiveness of the concatenation fusion method and improved the translation quality of models in the field of news by using the term base.

For the second stage, this paper takes the whole text as the input unit to translate the sentences in the text synchronously. Specifically, the proposed refinement network [21] consists of an encoder and two decoders, wherein the encoder is used to encode the sentences in the source text. The first decoder is used to output sentence-level translation. The input of the second decoder includes the output of the corresponding sentence on the encoding side, and the translation of the entire text of the first decoder. In order to make the model lightweight, this article shares parameters for the common parts of the two decoders. The experimental results of Chinese-English text translation show that the proposed method can improve the translation performance. It is commendable that with the improvement of the performance of the benchmark model, the proposed method is more effective because the first stage can output high-quality translations.

2. Materials and methods

2.1. Deliberation Network

Xia et al. [22] proposed deliberation networks for sequence generation. The deliberation network has a two-stage decoder, in which the first stage decoder is used to decode and generate the original sequence, and the second stage decoder revises the original statement through the elaboration process, so that it can produce a better sequence by observing future words from the original statement in the first stage. The network can make use of the global information by checking the contents before and after the sequence decoding process through a deliberation process.

2.2. Sentence level neural machine translation

Machine translation can be seen as the transformation of one sequence to another. In neural machine translation, the sequence to sequence transformation process can be implemented by the encoder-decoder framework. The encoder encodes the source end sequence [23], and the decoder decodes it into the sequence corresponding to the source end, which is generally called the target end. The optimization goal of sentence-level neural machine translation is to maximize the log-likelihood probability of the objective function on a given sentence-level training set $S_1 = \left\{\left\langle x^k, y^k \right\rangle\right\}_{k=1}^{K_1}$, as shown in Eq. (1).

$$L(\theta) = \max_{\theta} \frac{1}{K_1} \sum_{k=1}^{K_1} \log\left(P_\theta\left(y^k \mid x^k\right)\right) \tag{1}$$

Where $\left\langle x^k, y^k \right\rangle$ is the k -th parallel sentence pair and K_1 is the number of parallel sentence pairs in the training data. θ is the sentence level neural machine translation parameter that needs to be trained.

2.3. Text level neural machine translation

Similarly, in text level neural machine translation, where the input and output are one text, given the text level training set $S_2 = \left\{\left\langle X^k, Y^k \right\rangle\right\}_{k=1}^{K_2}$, the optimization objective function is shown in Eq. (2).

$$L(\Theta) = \max_{\Theta} \frac{1}{K_2} \sum_{k=1}^{K_2} \log\left(P_\Theta\left(Y^k \mid X^k\right)\right) \tag{2}$$

Where $\left\langle X^k, Y^k \right\rangle$ represents the k -th parallel chapter pair. K_2 is the number of parallel text pairs in the training data. Because text level neural machine translation parameters need to be trained. Specifically, the probability function of a parallel text can be further represented as shown in Eq. (3).

$$P\left(Y^k \mid X^k\right) = \prod_{i=1}^{L} P_\Theta\left(Y_i^k \mid X_i^k, D_i\right) \tag{3}$$

Where $\left\langle X_i^k, Y_i^k \right\rangle$ represents the i-th parallel sentence pair in the k-th chapter. L represents the number of sentences in the parallel text. D_i represents the text context of the i-th sentence. Specifically, the sentence-level probability function of the above formula is shown in Eq. (4).

$$P\left(Y^k \mid X^k\right) = \prod_{i=1}^{L}\prod_{j=1}^{m} P_\Theta\left(y_{i,j}^k \mid x_i^k, y_{i<j}^k, D_i\right) \tag{4}$$

Where m represents the sentence length of the target end. $y_{i,j}^k$ represents the j -th word in the i -th sentence of the target end in the k -th chapter.

In particular, this paper explores text machine translation that integrates full-text at the target end. Therefore, Eq. (4) can be further refined as:

$$P\left(Y^k \mid X^k\right) = \prod_{i=1}^{L}\prod_{j=1}^{m} P_\Theta\left(y_{i,j}^k \mid x_i^k, y_{i<j}^k, \widehat{Y}_{i\neq j}^k\right) \tag{5}$$

Where $\widehat{Y}_{i\neq j}^k$ represents the context of the text other than the i -th sentence in the k -th text. In training stage, $Y_{i\neq j}^k = \widehat{Y}_{i\neq j}^k$ represents the actual text context. In testing stage, $\widehat{Y}_{i\neq j}^k = \widetilde{Y}_{i\neq j}^k$ indicates the first decoding of the automatically generated text context.

2.4. Proposed Chinese-English neural machine translation

The purpose of this paper is to improve the performance of model translation by effectively modeling the context information of the target text and integrating the context information into the current sentence. The proposed model structure in this paper is shown in Fig. 1, which is implemented based on the Transformer neural machine translation model. The proposed model consists of encoder, decoder 1 and decoder 2.

Encoder. Consistent with standard sentence-level Transformer encoders, the same encoding layers are superimposed. Each encoding layer consists of a multi-head self-attention sub-layer and a fully connected feed-forward network complex sub-layer.

For a source text $X = (X_1, \cdots, X_L)$ containing L sentences, the hidden state of the encoder is $S = (S_1, \cdots, S_L) \in R^{L\times n\times d}$. Where $S_i = (s_{i,1}, \cdots, s_{i,n}) \in R^{n\times d}$ represents the hidden state of the i -th sentence X_i in the chapter. $s_{i,j}$ indicates the hidden state of the j -th word of the i-th sentence in the text. n indicates the sentence length at the source end. d indicates the hidden status dimension.

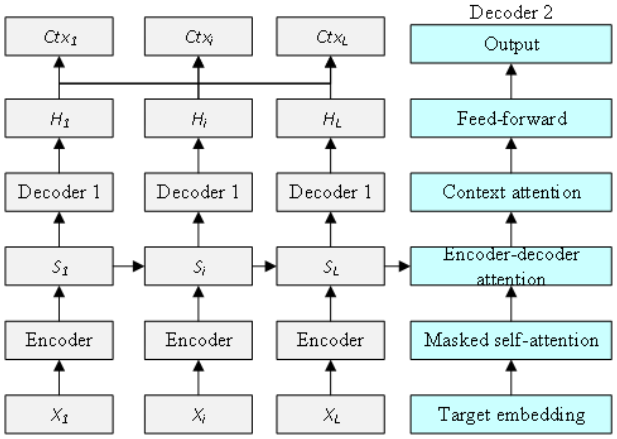

Fig. 1. Proposed network structure.

Decoder 1. That is, the decoding end of the first translation, which is consistent with the standard sentence-level Transformer decoder, and the decoding layer of decoder 1 consists of multi-head self-attention sub-layer, source-end context attention sub-layer and a fully connected feed-forward network complex sub-layer. Here, the encoder-decoder attention sub-layer is used to capture the information of the source end sentence.

Similarly, given the target text $Y = (Y_1, \cdots, Y_L)$, record the hidden state of decoder 1 output as $H = (H_1, \cdots, H_L) \in R^{L\times m\times d}$. Where $H_i = (h_{i,1}, \cdots, h_{i,n}) \in R^{m\times d}$ represents the hidden state of the i -th sentence Y_i in the text. $h_{i,j}$ represents the hidden state of the j -th word in the i-th sentence of the text. m represents the sentence length of the target end.

Decoder 2. That is, the decoding side of the second translation, whose structure is similar to that of decoder 1. The difference is that each layer of decoder 2 adds a target end context attention sub-layer between the encoder-decoder attention sub-layer and the fully connected feed-forward network complex sub-layer to fuse the target end context information.

In this paper, a dynamic weighting method is adopted to integrate the whole text information into a fixed length range, so as to solve the problem of inconsistent context length.

In this paper, a dynamic weighting method is adopted to integrate the whole text information into a fixed length range, so as to solve the problem of inconsistent context

length.

Sentence level representation. For the i-th sentence at the target end, the sentence level representation vector U_i of the sentence is obtained by sentence weighting [24], as shown in Eqs. (6) and (7).

$$\alpha_i = \text{softmax}\left(W^2 \tanh\left(W^1 H_i^T\right)\right) \tag{6}$$

$$U_i = \sum_{j=1}^{m} \alpha_{i,j} h_{i,j} \tag{7}$$

Where, $W^1 \in R^{d_1 \times d}$ and $W^2 \in R^{d_1}$ are trainable parameters. d_1 is the hidden layer dimension in the sentence weighted representation process. $U_i \in R^d$ is the sentence-level vector representation of the i -th sentence in the chapter. And so on, the sentence level vector of the entire chapter at the target end is represented by $U = (U_1, \cdots, U_L) \in R^{L \times d}$. In a similar way, it can obtain the sentence-level vector representation $V = (V_1, \cdots, V_L) \in R^{L \times d}$.

Context-weighted representation. For the i-th sentence, the target end context does not include the target end sentence vector of the current sentence itself, expressed as $H_{\neq i} = (H_1, \cdots, H_{i-1}, H_{i+1}, \cdots, H_L) \in R^{(L-1) \times m \times d}$. The corresponding sentence level at the source end is represented as $V_{\neq i} = (V_1, \cdots, V_{i-1}, V_{i+1}, \cdots, V_L) \in R^{(L-1) \times d}$, and the weight of the current sentence on the context is calculated, as shown in Eqs. (8) and (9).

$$\lambda_i = \text{softmax}\left(V_i V_{\neq 1}^T\right) \tag{8}$$

$$C_i = \sum_{j=1}^{L-1} \lambda_{i,j} \times H_{\neq i,j} \tag{9}$$

Where V_i is the sentence-level representation of the source end of the i-th sentence, and the context-weighted representation of the i-th sentence in the chapter is C_i.

Integrating the target text information into decoder 2. In this paper, a target-context attention sub-layer is added between the encoder-decoder attention sub-layer and the feed-forward neural network sub-layer of each layer of decoder 2. For the i-th sentence of the text, recording O_i, it represents the output of the encoder-decoder attention sub-layer. The new target context attention sub-layer uses a multi-head attention model to fuse the target text information C_i, as shown in equation (10).

$$T_i = \text{multihead}\ (O_i, C_i, C_i) \tag{10}$$

Training process and decoding process. Neural machine translation training parameters need a lot of parallel corpus, text level parallel corpus is very rare, but sentence level parallel corpus is relatively easy to obtain in some common languages. Similarly, this paper adopts a two-stage training method. The first stage trains the parameters required for sentence-level neural machine translation, as shown in Figure 1 for encoder and decoder 1. In the second stage, the parameters of the sentence-level model are fixed first, and then the parameters required by the text translation model are trained. In order to ensure that the effect brought by the model comes from the improvement of text information and verify the validity of the model proposed in this paper under large-scale sentence-level parallel corpus, the parameters of the same layer in decoder 1 and decoder 2 are shared, and only the parameters of the context information fusion layer are trained.

Different from the training process, in the decoding process, because there is no real text context, it is necessary to use the sentence-level model to automatically generate the trusted target end, and then use the automatically generated target end as the context for the second stage decoding.

3. Results and discussion

The Chinese-English parallel corpus extracted from LDC is used in the Chinese-English translation task [25], which includes sentence-level and texture-level corpus. The sentence-level corpus has about 2 million parallel sentence pairs, and the texture-level corpus has more than 60000 parallel sentence pairs. It contains about 800000 parallel sentences. On average, each chapter has 22 sentences. At the same time, the NIST 2006 dataset is used as the development set, and NIST2002, NIST2003, NIST2004, NIST2005, and NIST2008 are used as the test set. The development set and the test set consist of 100 and 580 chapters, including 1664 and 5833 sentences, respectively. In addition, the TED Chinese-English parallel discourse corpus is also used. There are more than 10000 parallel discourse pairs in the texture-level corpus, which contains about 200000 parallel sentence pairs in total, with an average of 15 sentences per chapter. At the same time, dev2010 data set is used as the development set, and tst2010, tst2011, tst2012, and tst2013 are used as the test set. The development and test sets consist of 60 and 365 chapters, including 887 and 5473 sentences, respectively. The sentence-level parallel corpus is still LDC parallel sentence pairs. When preprocessing the experimental data, the Jieba word segmentation tool is used to segment the Chinese sentences and the Moses script is used to preprocess the English sentences. Then, the processed Chinese and English data are sublexicalized by BPE tool respectively, and the operands on both Chinese and English data sets are 30000.

The text level parallel corpus training set for the English-

Table 1. Results of Chinese-English translation experiment.

Model	NIST02	NIST03	NIST04	NIST05	NIST08	All
Baseline1	42.66	44.11	42.51	40.90	31.88	40.23
Proposed	43.20	45.13	43.33	42.28	32.89	41.19
Baseline2	49.35	49.50	49.27	49.49	39.34	47.33
Proposed	50.87	51.16	50.93	50.62	41.22	49.00

Table 2. Results of the Chinese-English translation experiment in TED.

Model	dev2010	Tst201013
Baseline1	12.59	18.18
Proposed	12.83	18.79
Baseline2	14.36	20.51
Proposed	14.83	21.16

Table 3. Results of English-German translation experiment.

Model	news-test2015	news-test2016
Baseline1	20.33	22.20
Proposed	20.63	22.70
Baseline2	28.13	32.82
Proposed	29.16	33.91

German translation task is News Commentary v11 corpus, and the sentence level parallel corpus is derived from the WMT14 English-German translation task [26]. The text-level parallel corpus contains about 10000 parallel text pairs, with a total of about 200000 sentence pairs. The WMT14 English and German corpus contains 4 million sentence pairs. The development set is news-test2015, and the test set is news-test2016, which consist of 112 and 184 chapters, containing 2200 and 2999 sentences, respectively. Similar to the Chinese-English translation task, Moses script is used to preprocess English and German sentences respectively. Then, the processed data is sublexicalized.

The experimental model uses the Transformer model in OpenNMT and adds the functions required by the model in this paper on this basis. The encoder and decoder are set to six layers, the number of heads in the multi-head attention mechanism is 8 , and the input and output dimensions of the feed-forward neural network layer are 512 . The hidden layer dimension is 2048, using the Adam optimization function as the optimizer. Where β_1 is 0.9 , β_2 is 0.998 , and the Dropout probability is set to 0.1 . The learning rate is set to 1.0 and 0.5 in the first and second phases, respectively. During the translation of decoder 2, the Beam size is set to 5 . At the same time, in order to speed up the decoding speed in the first stage, the Beam size is set to 2 during the translation process of decoder 1 .

Based on two scenarios, namely whether the translation model is pre-trained with or without large-scale parallel sentences, two benchmark systems are defined in this paper. Baseline 1 only uses text level parallel corpus for sentence level translation. Baseline 2 uses large-scale parallel sentences to pre-train the model on the corpus, and then continues to train the model based on the text level parallel corpus (i.e. model fine-tuning). It is important to note that benchmark system 2 includes the data used in benchmark system 1.

This article uses the multi-bleu.perl test script and reports BLEU scores as test metrics. In addition, a self-sampling method is used to test the significance of the improvement of the performance BLEU value.

Table 1 shows the results of the Chinese-English translation experiment. As can be seen from Table 1, the model method proposed in this paper can significantly improve the performance of the two benchmark systems. In particular, when only the text parallel corpus is included, adding the text information can improve the test set by 0.96 BLEU compared with the sentence-level model Baseline1. When the model uses more sentence-level corpus, even though the performance of sentence-level model Baseline2 is greatly improved compared with Baseline1, the proposed method improves the test set by 1.67 BLEU, which is higher than the improvement based on Baseline1. The reason is that with the improvement of translation quality in the first stage, the proposed method can capture higher-quality text context information at the target end. Table 2 shows the results of the TED Chinese-English translation experiment. It can be seen from Table 2 that, compared with the sentence-level model Baseline1, adding text information can increase 0.61 BLEU in the test set. However, after adding LDC sentence-level corpus, compared with Baseline2, the effect is not significantly improved, possibly because there are certain differences between sentence-level corpus and discourse corpus.

Table 3 shows the results of English-German translation experiment. As can be seen from Table 3, the model in this paper still has a good performance. Compared with the sentence-level model Baseline1, adding the text information can improve the test set by 0.50 BLEU. When the model uses more sentence-level corpus, the proposed method improves the test set by 1.09 BLEU.

Table 4. Comparison of translation performance (BLEU) and parameter with target related work.

Model	parameter	NIST02	NIST03	NIST04	NIST05	NIST08	All
Baseline2	49	49.35	49.50	49.27	49.49	39.34	47.33
Proposed	57	50.87	51.16	50.93	50.62	41.22	49.00
Rachman Baseline2	49	48.88	48.60	48.92	49.34	40.72	47.19
Rachman	84	48.98	49.12	49.37	49.06	41.19	47.56
Zan Baseline2	60	48.64	47.55	47.80	48.35	38.32	45.98
Zan	72	48.98	48.06	47.92	48.54	38.39	46.38

In this section, we will take Chinese-English translation task as an example to conduct experimental analysis of the proposed text neural machine translation model integrating target context. At the same time, since the second scenario, which uses large-scale parallel sentence pairs to pre-train the model, achieves better translation results than the first scenario, Baseline2 will be used as the benchmark system in this section.

Rachman et al. [27] used the local target text information to assist the translation of the current sentence. Consistent with the experimental setting, this paper uses the source code and the corpus of this paper to run the reference system (Rachman (Baseline2)) and the text translation system (Rachman) which uses the first 3 sentences of both the source and target end of the current sentence as the text context. Table 4 is the experimental result. It can be seen from Table 4 that the performance of the benchmark system used in this paper is slightly better than that of the Rachman benchmark system on the test set. In the scenario where large-scale parallel sentence pairs are used, the Rachman method achieves only 0.37 BLEU improvement, even when both source and target context information are used. Compared with the Rachman method, the proposed model can significantly improve the translation performance. In addition, the method proposed by Zan et al. [28] was tested on the Transformer benchmark system (recorded Zan(Baseline2), Zan), and the proposed method by Zan achieved 0.40 BLEU. Table 4 compares the parameters of the model. On the one hand, the parameters of the benchmark system in this paper are similar to those of the Rachman benchmark system; On the other hand, because the parameters of the common sub-layer in the two decoders are shared, the proposed model introduces only about 16% of the parameters compared to the benchmark system, far fewer than the parameters in the Rachman model.

Li et al. [29] used part of the source-end text context and defined two encoders to encode the current sentence and context respectively. In this paper, the method is reproduced on the benchmark system (noted as Li). Table 5 compares the experimental results, and it can be seen from Table 5 that the method of using the target context in this paper is slightly better. The proposed method by Li achieves 1.49 BLEU improvement by using the sources-end method, and this paper achieves 1.67 BLEU improvement by using the target-end method. In addition, this article compares the method of using the target side context with the method of using the source side context (recorded as Proposed+s). As can be seen from Table 5, the target-end context method used in this paper increases the value of BLEU by 0.47 compared with the source context method used in this paper.

A. Effect of different decoding layers in decoder 1 on model performance.

Because the Transformer decoder contains different layers, each layer may contain different information. In order to explore the output influence of different decoding layers in decoder 1 on the model, Table 6 analyzes the experimental results using the different layers output in decoder 1 at the target end as the text context. That is, the H_j in Eq. (9) comes from the output of different layers in decoder 1. As can be seen from Table 6, which layer of encoder information is used has little impact on the performance of the final model. Relatively speaking, the best performance is achieved by modeling the last layer of output in decoder 1 as the text context.

B. Pronoun translation performance

In order to verify the effect of Pronoun translation after adding text context information, this paper analyzes the accuracy of pronoun translation [30]. Table 7 shows the pronoun translation accuracy of this model on the Chinese-English dataset test set. As can be seen from the results in Table 7, the target-end context can effectively improve the performance of reference translation.

4. Conclusions

Different from previous text translation methods that model the context of the source text, this paper proposes a text translation model that integrates the information of the target text. Specifically, with the help of the deliberation network, the source end of the text is translated twice. The first translation is based on sentence-level translation, and the second translation refers to the first translation of the whole

Table 5. Comparison of translation performance (BLEU) and parameter with source-side related work.

Model	parameter	NIST02	NIST03	NIST04	NIST05	NIST08	All
Baseline2	49	49.35	49.50	49.27	49.49	39.34	47.33
Proposed	57	50.87	51.16	50.93	50.62	41.22	49.00
Li	76	50.81	50.92	50.47	51.10	40.62	48.82
Proposed + s	57	50.32	50.61	50.04	50.73	40.79	48.53

Table 6. The output influence of encoding layers with different contexts in decoder 1 on the model.

Model	NIST02	NIST03	NIST04	NIST05	NIST08	All
Baseline2	49.35	49.50	49.27	49.49	39.34	47.33
Proposed+t1	50.72	50.49	50.64	50.53	41.55	48.82
Proposed+t3	50.81	50.37	50.54	50.66	41.77	48.88
Proposed+t5	50.84	51.05	50.64	50.57	41.52	48.93
Proposed+t6	50.87	51.16	50.93	50.62	41.22	49.00

Table 7. Pronoun translation accuracy on Chinese-English dataset.

Model	Accuracy/%
Baseline2	56.84
Proposed	58.25

text. In order to reduce the size of the newly introduced parameters in the model, the model shares the common sub-layer of the two decoders. Experimental results based on LDC Chinese-English text dataset and WMT English-German text dataset show that the proposed method can significantly improve translation performance under the condition of introducing fewer parameters. At the same time, with the improvement of the quality of the first translation (sentence level translation), the method in this paper is more effective.

References

[1] M. Al-Barham, I. Afyouni, K. Almubarak, A. Turky, I. A. T. Hashem, A. B. Nassif, I. Shahin, and A. Elnagar, (2025) *"Unlocking language boundaries: AraCLIP-transforming Arabic language and image understanding through cross-lingual models"* **Engineering Applications of Artificial Intelligence *151***: 110577. DOI: 10.1016/j.engappai.2025.110577.

[2] A. L. Tonja, O. Kolesnikova, A. Gelbukh, and G. Sidorov, (2023) *"Low-resource neural machine translation improvement using source-side monolingual data"* **Applied Sciences *13***(2): 1201. DOI: 10.3390/app13021201.

[3] T. Z. Shah, M. Imran, and S. M. Ismail, (2024) *"A diachronic study determining syntactic and semantic features of Urdu-English neural machine translation"* **Heliyon *10***(1): DOI: 10.1016/j.heliyon.2023.e22883.

[4] Y. Sun, S. Yin, H. Li, L. Teng, and S. Karim, (2019) *"GPOGC: Gaussian pigeon-oriented graph clustering algorithm for social networks cluster"* **IEEE Access *7***: 99254–99262. DOI: 10.1109/ACCESS.2019.2926816.

[5] M. A. Faheem, K. T. Wassif, H. Bayomi, and S. M. Abdou, (2024) *"Improving neural machine translation for low resource languages through non-parallel corpora: a case study of Egyptian dialect to modern standard Arabic translation"* **Scientific Reports *14***(1): 2265. DOI: 10.1038/s41598-023-51090-4.

[6] D. A. Sulistyo, D. D. Prasetya, F. A. Ahda, and A. P. Wibawa, (2025) *"Pivoted Low Resource Multilingual Translation with NER Optimization"* **ACM Journal of Data and Information Quality**: DOI: 10.1145/3727876.

[7] B. Zhang, I. Titov, B. Haddow, and R. Sennrich. "Beyond sentence-level end-to-end speech translation: Context helps". In: *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics. 2021, 2566–2578. DOI: 10.18653/v1/2021.acl-long.200.

[8] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, (2019) *"Neural machine translation with sentence-level topic context"* **IEEE/ACM Transactions on Audio, Speech, and Language Processing *27***(12): 1970–1984. DOI: 10.1109/TASLP.2019.2937190.

[9] M. Yang, R. Wang, K. Chen, X. Wang, T. Zhao, and M. Zhang, (2020) *"A novel sentence-level agreement architecture for neural machine translation"* **IEEE/ACM Transactions on Audio, Speech, and Language Processing *28***: 2585–2597. DOI: 10.1109/TASLP.2020.3021347.

[10] J. Smith, C. Quirk, and K. Toutanova. "Extracting parallel sentences from comparable corpora using document level alignment". In: *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*. 2010, 403–411. DOI: 10.3115/1690339.1690350.

[11] H. Yun, Y. Hwang, and K. Jung. "Improving context-aware neural machine translation using self-attentive sentence embedding". In: *Proceedings of the AAAI conference on artificial intelligence*. ***34***. 05. 2020, 9498–9506. DOI: 10.1609/aaai.v34i05.6494.

[12] J. Yu, L. Zhao, et al., (2021) *"A novel deep CNN method based on aesthetic rule for user preferential images recommendation"* **Journal of Applied Science and Engineering *24***(1): 49–55. DOI: 10.6180/jase.202102_24(1).0006.

[13] S. Zhu, S. Li, and D. Xiong, (2024) *"VisTFC: Vision-guided target-side future context learning for neural machine translation"* **Expert Systems with Applications *249***: 123411. DOI: 10.1016/j.eswa.2024.123411.

[14] X. Zhu, Q. Ruan, S. Qian, and M. Zhang, (2025) *"A hybrid model based on transformer and Mamba for enhanced sequence modeling"* **Scientific Reports *15***(1): 11428. DOI: 10.1038/s41598-025-87574-8.

[15] M. Bamoki, S. H. Wady, and S. Badawi, (2025) *"Holy Quran Kurdish Sorani translation dataset for language modelling"* **Data in Brief *60***: 111533. DOI: 10.1016/j.dib.2025.111533.

[16] J. Choo, Y.-D. Kwon, J. Kim, J. Jae, A. Hottung, K. Tierney, and Y. Gwon, (2022) *"Simulation-guided beam search for neural combinatorial optimization"* **Advances in Neural Information Processing Systems *35***: 8760–8772. DOI: 10.48550/arXiv.2207.06190.

[17] S. H. Ahammad, R. R. Kalangi, S. Nagendram, S. Inthiyaz, P. P. Priya, O. S. Faragallah, A. Mohammad, M. M. Eid, and A. N. Z. Rashed, (2024) *"Improved neural machine translation using Natural Language Processing (NLP)"* **Multimedia Tools and Applications *83***(13): 39335–39348. DOI: 10.1007/s11042-023-17207-7.

[18] M. E. San, S. Usanavasin, Y. K. Thu, and M. Okumura, (2024) *"A Study for Enhancing Low-resource Thai-Myanmar-English Neural Machine Translation"* **ACM Transactions on Asian and Low-Resource Language Information Processing *23***(4): 1–24. DOI: 10.1145/3645111.

[19] S. C. Siu. "Revolutionising translation with AI: Unravelling neural machine translation and generative pre-trained large language models". In: *New advances in translation technology: Applications and pedagogy*. Springer, 2024, 29–54. DOI: 10.1007/978-981-97-2958-6_3.

[20] S. Bala Das, D. Panda, T. Kumar Mishra, B. Kr. Patra, and A. Ekbal, (2024) *"Multilingual neural machine translation for indic to indic languages"* **ACM Transactions on Asian and Low-Resource Language Information Processing *23***(5): 1–32. DOI: 10.1145/3652026.

[21] G. Lin, A. Milan, C. Shen, and I. Reid. "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, 1925–1934. DOI: 10.1109/CVPR.2017.549.

[22] Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, and T.-Y. Liu, (2017) *"Deliberation networks: Sequence generation beyond one-pass decoding"* **Advances in neural information processing systems *30***:

[23] H. Li, Z. Li, X. Wang, M. Ibrar, and X. Zhu, (2024) *"Multi-keyword Ciphertext Sorting Search Based on Conformation Graph Convolution Model and Transformer Network in English Education"* **International Journal of Network Security *26***(4): 555–564. DOI: 10.6633/IJNS.202407_26(4).03.

[24] J. Tiedemann, M. Aulamo, D. Bakshandaeva, M. Boggia, S.-A. Grönroos, T. Nieminen, A. Raganato, Y. Scherrer, R. Vázquez, and S. Virpioja, (2024) *"Democratizing neural machine translation with OPUS-MT"* **Language Resources and Evaluation *58***(2): 713–755. DOI: 10.1007/s10579-023-09704-w.

[25] Z.-M. Gao, (2011) *"Exploring the effects and use of a Chinese–English parallel concordancer"* **Computer Assisted Language Learning *24***(3): 255–275. DOI: 10.1080/09588221.2010.540469.

[26] A.-M. De Cesare, A. Albom, D. Cimmino, and M. L. Spagnolo, (2020) *"Domain adverbials in the news: A corpus-based contrastive study of English, German, French, Italian and Spanish"* **Languages in Contrast *20***(1): 31–57. DOI: 10.1075/lic.17005.dec.

[27] F. H. Rachman, M. W. M. A. Syauqi, N. Ifada, S. Wahyuni, et al., (2025) *"Transformer Hyperparameter Tuning for Madurese-Indonesian Machine Translation"* **Engineering, Technology & Applied Science Research *15***(2): 22216–22225. DOI: 10.48084/etasr.9851.

[28] Z. Hongying, A. Javed, M. Abdullah, J. Rashid, and M. Faheem, (2025) *"Large Language Models With Contrastive Decoding Algorithm for Hallucination Mitigation in Low-Resource Languages"* **CAAI Transactions on Intelligence Technology**: e70004. DOI: 10.1049/cit2.70004.

[29] B. Li, Y. Weng, F. Xia, and H. Deng, (2024) *"Towards better Chinese-centric neural machine translation for low-resource languages"* **Computer Speech & Language *84***: 101566. DOI: 10.1016/j.csl.2023.101566.

[30] X. Meng, X. Wang, S. Yin, and H. Li, (2023) *"Few-shot image classification algorithm based on attention mechanism and weight fusion"* **Journal of Engineering and Applied Science *70***(1): 14. DOI: 10.1186/s44147-023-00186-9.