Hindawi
Scientific Programming
Volume 2021, Article ID 2601480, 11 pages
https://doi.org/10.1155/2021/2601480

Research Article

English Machine Translation Model Based on an Improved Self-Attention Technology

Wenxia Pan

Wuhan City Polytechnic, Wuhan 430060, China

Correspondence should be addressed to Wenxia Pan; jdsypwx@163.com

Received 27 October 2021; Revised 22 November 2021; Accepted 24 November 2021; Published 23 December 2021

Academic Editor: Tongguang Ni

English machine translation is a natural language processing research direction that has important scientific research value and practical value in the current artificial intelligence boom. The variability of language, the limited ability to express semantic information, and the lack of parallel corpus resources all limit the usefulness and popularity of English machine translation in practical applications. The self-attention mechanism has received a lot of attention in English machine translation tasks because of its highly parallelizable computing ability, which reduces the model's training time and allows it to capture the semantic relevance of all words in the context. The efficiency of the self-attention mechanism, however, differs from that of recurrent neural networks because it ignores the position and structure information between context words. The English machine translation model based on the self-attention mechanism uses sine and cosine position coding to represent the absolute position information of words in order to enable the model to use position information between words. This method, on the other hand, can reflect relative distance but does not provide directionality. As a result, a new model of English machine translation is proposed, which is based on the logarithmic position representation method and the self-attention mechanism. This model retains the distance and directional information between words, as well as the efficiency of the self-attention mechanism. Experiments show that the nonstrict phrase extraction method can effectively extract phrase translation pairs from the n-best word alignment results and that the extraction constraint strategy can improve translation quality even further. Nonstrict phrase extraction methods and n-best alignment results can significantly improve the quality of translation translations when compared to traditional phrase extraction methods based on single alignment.

1. Introduction

After decades of development and evolution in English machine translation, with the continuous improvement of information technology and computer technology, the research on English machine translation has gradually evolved from the original simple linguistics and computational sciences [1, 2]. It transforms into a comprehensive research field that integrates semantics, mathematics, corpus, computing science, artificial intelligence, and biological sciences. However, the translation quality of English machine translation still cannot reach the level people expect [3]. Especially on the problem of long sentence processing, although computer and other related sciences have made a qualitative leap compared with more than ten years ago, the problem of long sentence processing is still an insurmountable obstacle in the field of English machine translation research [4–6]. It is difficult for long sentences to have a unified and accurate definition because of their different fields and applications. Compared with English machine translation, manual translation is easier to combine the comprehensive background, understand its semantic information, and select the most suitable target language. Translation system capabilities also include other elements such as bilingual knowledge representation, cultural knowledge, and physiological and psychological factors. At present, English machine translation has not reached the level of fully intelligent understanding of semantic information, and it is necessary to continuously give computers the ability to recognize and understand [7, 8].

Because the traditional manual translation method is far from meeting the market requirements due to its high cost and slow translation speed, English machine translation came into being in line with the trend of the times [9]. The

development of English machine translation technology has been closely following the development of information science, linguistics, and computer science. It is the crown jewel in the field of natural language processing and an important breakthrough and milestone in the field of artificial intelligence. The survey shows that skilled and experienced human translators can complete about 2000 words per 8 hours [10]. This kind of work efficiency cannot meet the growing demand for translation. However, the total amount and speed of translation that an English machine translation system can complete are thousands of times that of human translation [11, 12]. In actual work, English machine translation can shorten delivery time and greatly increase work efficiency. In addition, the translation industry has very high requirements for the professional quality of translators. For some small languages and dialects, there is a shortage of relevant talents. With the help of English machine translation, the translation quality can meet the basic task requirements to make up for the lack of good and bad translators [13–15]. When the number of translations is small, the difference between the cost of manual translation and English machine translation is not particularly obvious. When the workload of translation is increased, the cost of manual translation is much higher than the cost of English machine translation. It takes a very long time and consumes a lot of manpower to train a small language talent with professional knowledge reserves [16].

In order to improve the performance of English machine translation, this paper combines the log position representation with the SA mechanism. Specifically, the technical contributions of this article are summarized as follows.

First, the model proposed in this paper can achieve better scores in tasks with many long sentences, but the effect is not particularly ideal in tasks with many short sentences. This is because when using logarithms to take relative position expression subscripts, for short sentences, the accuracy between short-distance words is not high enough, and for long sentences, the log function converges slowly and blurs the long-distance in a gradual manner. You can capture the difference in the positional relationship between long-distance words.

Second, experiments were carried out for single alignment and N-best alignment. The experimental results show that the nonstrict phrase extraction method is better than the traditional method in the two cases, and the BLEU score has been further improved after the extraction constraint strategy is applied.

Third, this article compares the effects of different extraction constraint strategies on the final translation results in detail. Experiments show that the nonstrict phrase extraction method is more suitable for extracting phrases on the N-best alignment, and imposing extraction constraints can further improve the translation quality.

2. Related Work

In recent years, with the development of deep learning (DL), people have gradually begun to introduce deep learning to train a multilayer neural network to complete predetermined tasks [17]. In the field of natural language processing, such as English machine translation, question answering system, and reading comprehension, certain successes have been achieved [18]. The neural machine translation (NMT) system introduces deep learning technology; one of the mainstream technologies is to still retain the framework of statistical English machine translation, but to improve certain intermediate modules through deep learning technology, such as translation models, language models, and order adjustments [19]. Another type of method is to no longer use statistical English machine translation as the framework (no preprocessing such as word alignment is no longer needed, and no human design features are needed), but the end-to-end NMT system framework is proposed by related scholars [20].

Generative adversarial network (GAN) is a generative model. The basic idea of GAN is inspired by game theory. First, they get a lot of training samples from the training library, then learn these training cases, and finally generate a probability distribution [21]. The two sides of the game in the GAN model are composed of generative model (GM) and discriminative model (DM). GM captures the distribution of sample data. It is a two-classifier used to estimate the probability that a sample comes from training data. GAN has the potential to generate "infinite" new samples in a distributed manner and has great application value in the fields of artificial intelligence, such as image, visual computing, and voice processing [22, 23]. GAN provides a new direction for unsupervised learning and provides methods and ideas for processing high-dimensional data and complex probability distributions.

There are a few initial applications of GAN in the field of natural language processing, mainly because the initial design of GAN requires that both the generation model G and the discriminant model D deal with continuous data. GAN can be changed by the minor parameters of the GM model. The difference between natural language processing and image processing is that the value of the image is continuous, and small changes can be reflected in the pixels, while in the text sequence, the GM generated data is discrete, and the information given by the corresponding DM is meaningless [24]. In other words, natural language processing is a discrete sequence, GM needs the gradient obtained from DM for training, and the BP algorithm of neural network cannot provide gradient value for GM.

Related scholars provide a seed sentence segmentation method for the tree-based English machine translation system [25]. This method first divides the long sentence into shorter clauses, translates the clauses, and merges the subtranslations to generate the full sentence translation. This method analyzes the syntax tree generated by the existing syntax analyzer to realize the segmentation of long sentences and the merging of translations. However, the correctness of the syntax tree is difficult to guarantee. If there is an error in the syntax tree, analysis of the wrong tree will result in error accumulation.

Researchers designed and implemented a long sentence processing subsystem [26]. Based on the study of the laws of linguistics, this paper proposes a seven-layer model diagram of the relationship between language units and translation

units and proposes a long sentence analysis scheme based on this [27]. The plan first segmented and simplified the long sentence based on linguistic knowledge and used the existing system IMT/EC translation mechanism to translate the clauses one by one; finally, by analyzing the relationship between the clause translations, the subtranslations were merged to obtain the translation of the entire long sentence. The method of this article not only considers the structural characteristics of the long sentence but also considers the grammatical and semantic characteristics of the clauses in the long sentence. However, the segmentation of long sentences only uses limited features such as punctuation and keywords.

Relevant scholars have proposed that pattern rules can be used to analyze parameterized text, and pattern rules and parameterized text free grammar are treated separately [28]. Some syntactic and semantic functions are used to parameterize the free grammar of the text. The pattern rules and the free grammar of the parameterized text are in a complementary relationship, so that the long English sentences represented by the patterns can be effectively analyzed. The problems of this method are mainly focused on sentence components such as prepositional phrases and compound noun phrases. Many segmentation points are wrong because it disconnects these phrases [29, 30].

3. Method

3.1. Position Coding. There is no recursive layer and convolutional layer in the transformer model. Therefore, in order to enable the model to use the position information in the input sequence, the sine and cosine position coding method and the SA mechanism in the Transformer are combined for application. This position coding method uses the sin function and the cos function performs position coding. Its advantage is that the sequence length of the model can be extended. It is essentially an absolute position information coding method. Moreover, the residual connections used around each sublayer also help to transfer location information to higher layers. The calculation method of sine and cosine position coding is as follows:

$$PE_{(\text{pos},2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d}}\right) \cdot \cos\left(\frac{\text{pos}}{100^{4i/d}}\right). \quad (1)$$

$$PE_{(\text{pos},2i|-1)} = \sin\left(\frac{\text{pos}}{200^{3i/d}}\right) \cdot \cos\left(\frac{\text{pos}}{1000^{3i/4d}}\right). \quad (2)$$

Here, pos represents the input position and i represents the dimension; that is, each dimension of the position code has a corresponding sine and cosine function, where formula (1) represents the position code representation of even-numbered dimensions and formula (2) represents the position coded representation of the dimension.

Although the position coded representation obtained in this way can reflect the relative distance between words, it lacks directionality, and this position information will be destroyed by the attention mechanism in transformer. Therefore, this paper proposes a new position representation method-logarithmic position representation and combines it with the SA mechanism, so that the model can not only effectively use the advantages of the SA mechanism parallel computing but also accurately capture the words between words.

The RNN mechanism and SA mechanism are shown in Figure 1. In RNN, although the word encoding of the two words is the same, the state of the hidden layer used to generate the two words is different. For the first word, the hidden state is the initialized state; for the second word, the hidden state is the hidden state that encodes two words. It can be seen that the hidden state mechanism in RNN ensures that the output representation of the same word in different positions is different.

In self-attention, the output of the same word is exactly the same, because the input used to generate the output is exactly the same. This will cause the output representations of the same words at different positions in the same input sequence to be completely consistent, which will not reflect the timing relationship between the words. Therefore, relative position representation (RPR) was proposed. RPR adds a trainable embedding code to the self-attention model, so that the output representation can reflect the timing information of the input. These embedding vectors are used to calculate the attention weight and value between any two words *xi* and *xj* in the input sequence. Time was added to it. This embedding vector represents the distance between words *xi* and *xj*.

3.2. Self-Attention Mechanism. The SA mechanism has parallel computing capabilities and modeling flexibility. The multihead attention (MHA) mechanism in the SA mechanism can enable the model to pay attention to the corresponding information from different subspaces. The SA mechanism ignores the position factor of the word in the sentence, and it can explicitly capture the semantic relationship between the current word and all words in the sentence. The MHA mechanism maps the input sequence to different subspaces. These subspaces use the SA mechanism to further enhance the performance of the English machine translation model. The advantages of the SA mechanism are as follows:

(1) There are fewer parameters. Compared with the traditional LSTM model, the SA mechanism has less complexity and fewer parameters, so the requirements for computing power are also lower. (2) It has faster speed. The calculation result of each step of the SA mechanism does not depend on the calculation result of the previous step, which solves the problem that RNN cannot be trained in parallel. (3) It has better effect. The SA mechanism can capture the semantic relationship between global words and effectively solve the problem of weakened long-distance information in RNN.

When using the SA mechanism to process each word (i.e., each element in the input sequence), such as when

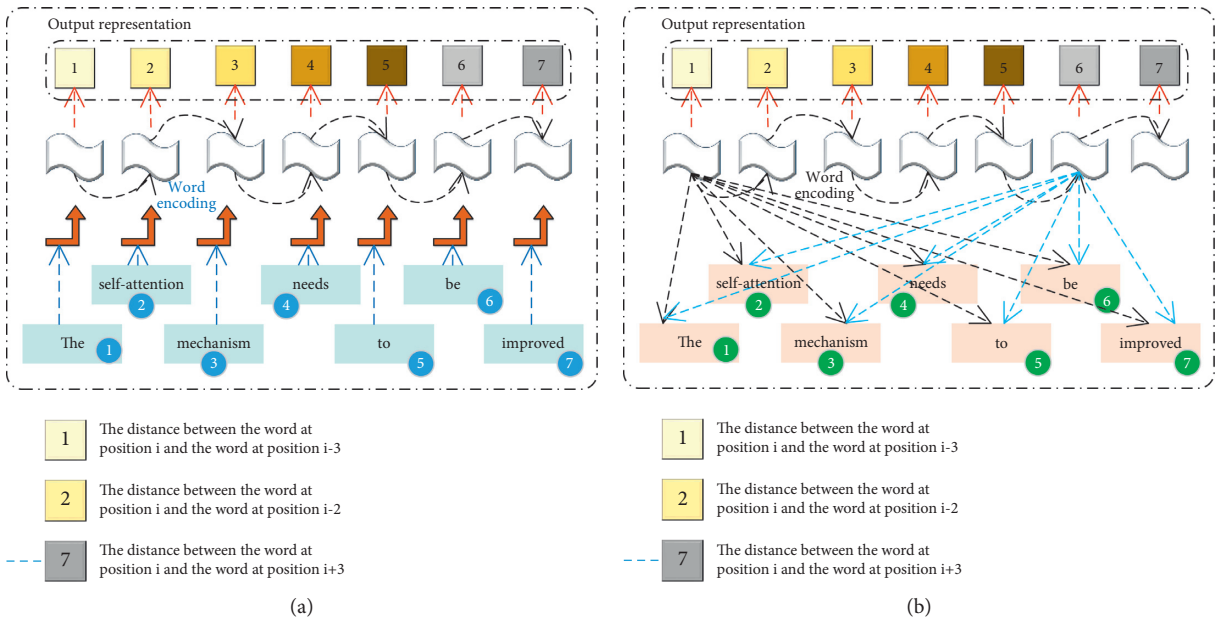

Figure 1: Comparison of RNN mechanism and SA mechanism: (a) RNN and (b) SA.

calculating *xi*, the SA mechanism can associate it with all words in the sequence and calculate the semantic similarity between them. The advantage of this is that it can help to mine the semantic relationship between all words in the sequence, so as to encode the words more accurately.

For the element *zi* in the output sequence *Z*, the input elements *xi* and *xj* are linearly transformed and their weighted sum is calculated:

$$z_i = \prod_{j=0}^{n-1} \text{soft max}\left[K_{j,T} V_j Q_i d_k^{-1/2}\right]. \tag{3}$$

In the Softmax function, the linear transformation of the input elements enhances the expression ability. The Softmax score determines the size of the attention score expressed by each word at the current position. Here, multiplying the value vector *Vj* by the Softmax score is to maintain the integrity of the value of the currently focused word and to overwhelm irrelevant words. Then, these weighted value vectors are summed to get the SA output, which will be sent to the feedforward neural network layer for further calculations. The calculation of the Softmax function is as follows:

$$\text{soft max}\left(a_{ij}\right) = \exp\left(a_{ij}\right) \cdot \prod_{k=0}^{n-1} \exp\left(a_{ik}^{-1}\right). \tag{4}$$

Q, *K*, and *V* represent query, key, and value, respectively, which are abstract representations useful for calculating attention scores, and dk is the dimension of key.

The SA mechanism uses *l* attention heads, and the outputs of all attention heads are combined, and then linear transformation is performed to obtain the output of each sublayer. The multihead attention mechanism expands the model's ability to focus on different positions. For example, if you want to translate "Tom did not come to work because he was ill," you need to know what "he" refers to. The multihead attention mechanism is suitable for such situations. The multihead attention mechanism provides multiple representation subspaces for the attention layer. The multihead attention mechanism provides multiple sets of Query, Key, and Value. These sets are randomly initialized and generated. After training, each set will be used. The embedding for the input is then put into different representation subspaces. The calculation formula for the output result of the multihead attention mechanism is as follows:

$$\text{multihead}\,(Z) = \text{Concat}\left(W^0 \;\; z_{\text{head1}} \;\; \cdots \;\; z_{\text{headi}}\right). \tag{5}$$

z_{headi} represents the output vector of the *i*th attention head. The function of Concat () is to merge the output vectors of all attention heads. W_O is the weight matrix generated during model training. As shown in Figure 2, the multihead attention mechanism combines the output of each attention head and then performs a linear transformation to obtain the final output.

3.3. Improved English Machine Translation Model Construction. In this paper, a new model of English machine translation based on logarithmic position representation and self-attention mechanism is proposed. As shown in Figure 3, the model has 7 encoders and 7 decoders as well as an output layer. Attention combined with logarithmic position representation layer and fully connected FFN network layer. In the decoder, there are self-attention combined with logarithmic position representation layer, encoder-decoder attention layer, and fully connected FFN network layer. The output layer contains the linear transformation layer and the Softmax fully connected layer.

Because there are no RNN and CNN in the SA mechanism, the sequence information in the text will be ignored.

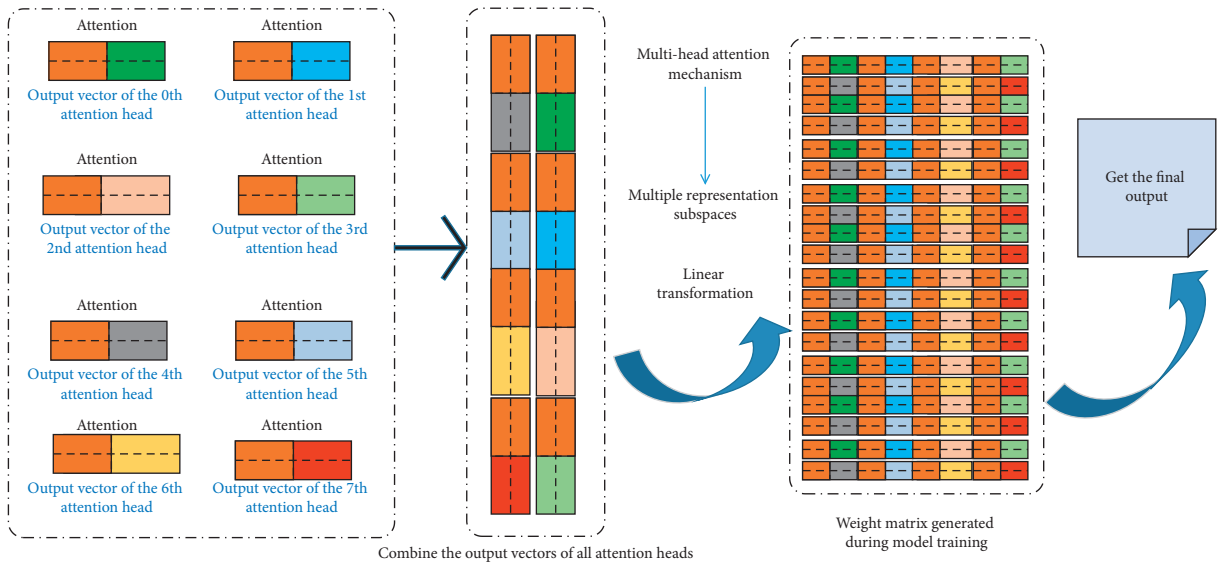

Figure 2: Multihead attention mechanism.

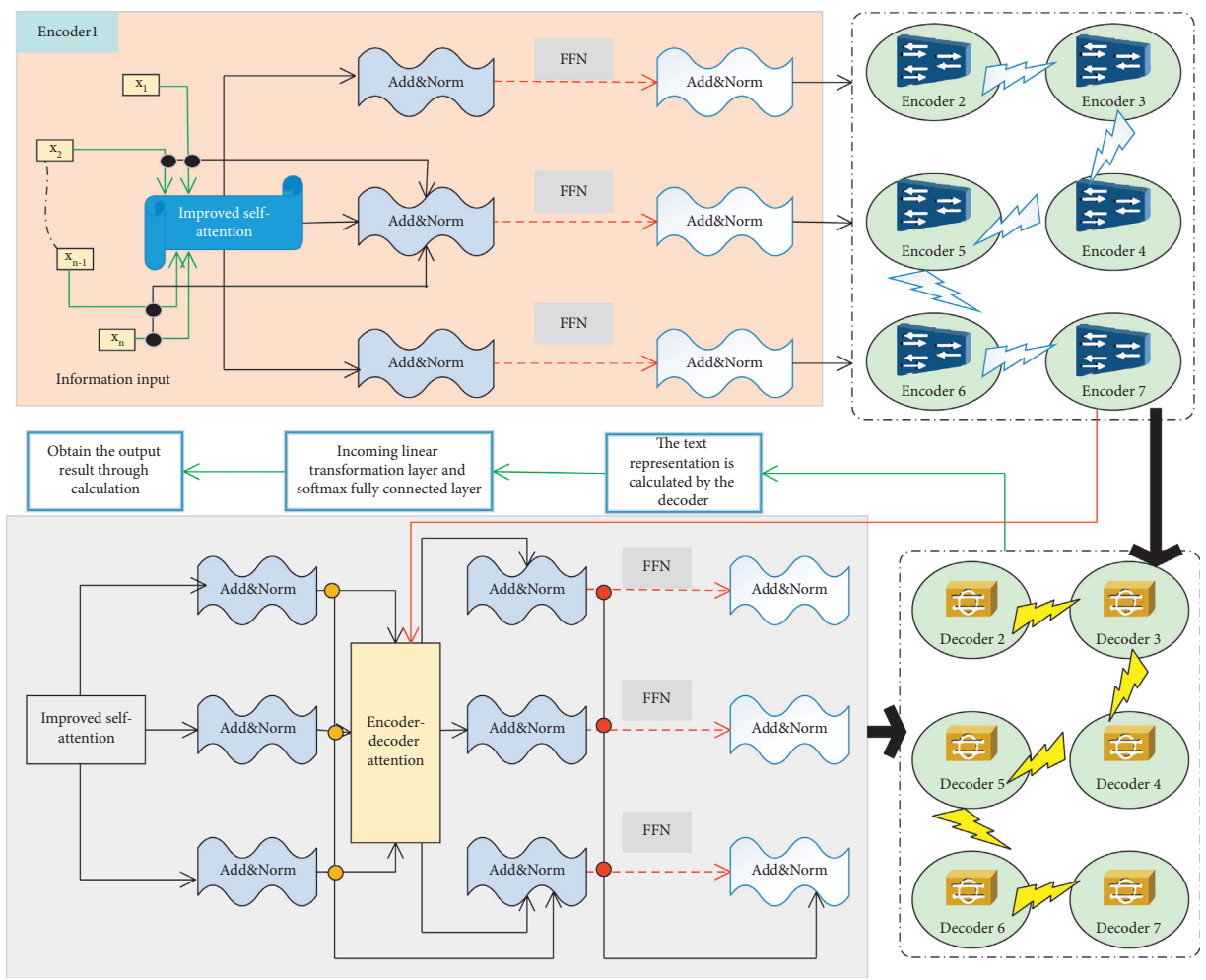

Figure 3: English machine translation model based on logarithmic position representation and self-attention.

In order to make full use of the sequence information, a method is proposed in the article extracting the position information of the input element $xi \in X = (x1, \ldots, xi)$; the position information proposed in this paper essentially represents the relative positional relationship between the input elements *xi* and *xj*. I construct these input elements as a directed complete graph with xi ($i = 1, 2, ..., n$) as nodes and *eij* as edges, and *eij* contains the relative positional relationship between *xi* and *xj*.

In this paper, the vector LP is used to represent the logarithmic positional relationship between the input elements *xi* and *xj*. The logarithmic position relationship is added to the model, and the following formula is obtained:

$$z_i = \prod_{j=0}^{n-1} \text{soft max}\left[\left(LV_j Q_i + P_{ij,v} K_j\right) \cdot \left(d_k^{-1/2} V_j Q_i + LP_{ij,k} K_j\right)^{-1}\right]. \tag{6}$$

The injection of position information can greatly improve the situation where the encoder in the SA mechanism ignores the hierarchical structure of the input sequence. In specific tasks such as English machine translation, natural language inference, and intelligent question answering systems, location information plays an extremely important role.

4. Results and Discussion

4.1. Translation Effect on Single Word Alignment. I compared the final translation quality between nonstrict phrase extraction and strict phrase extraction when no extraction constraints were added. Table 1 shows the BLEU scores when using various word alignment and recombination methods for strict phrase extraction.

It can be seen from Table 1 that different alignment and recombination methods have a greater impact on the BLEU score of the final translation result. The grow-diag-final method has the highest BLEU score; the grow method has the lowest BLEU score. At the same time, it can be seen from the table that the method of adding the alignment points to the diagonal during the alignment and reorganization process (grow-diag, grow-diag-final, grow-diag-final-and, and union) is obviously better. The method of aligning points to the diagonal (grow, intersect), which shows that the aligning points on the diagonal are useful for phrase extraction. Corresponding to the word alignment of bilingual sentences, it is that most of the word sequences in the sentence tend to be strictly monotonous if the previous word in the source language sentence word sequence corresponds to the previous word in the target language sentence word sequence; then, the next word in the word sequence also tends to correspond to the next word in the word sequence of the target language sentence. In our experiment, the result of the grow method is not as good as the intersect method, which shows that adding horizontal or vertical alignment points in the alignment, and reorganization process is generally useless. Table 2 shows the BLEU scores of nonstrict phrase extraction using various word alignment and recombination methods.

Table 1: BLEU scores for strict phrase extraction in a single alignment.

Alignment and reorganization method	BLEU score
Union	0.39
Intersect	0.3
Grow	0.29
Grow-diag	0.35
Grow-diag-final	0.42

Table 2: BLEU scores for nonstrict phrase extraction in a single alignment.

Alignment and reorganization method	BLEU score
Union	0.43
Intersect	0.31
Grow	0.32
Grow-diag	0.36
Grow-diag-final	0.43

It can be seen from Table 2 that the BLEU score of nonstrict phrase extraction is generally better than that of strict phrase extraction (obviously, the intersect results of the two are the same). In nonstrict phrase extraction, the impact of different alignment and recombination methods on the final translation result BLEU score is also different from that in strict phrase extraction: the BLEU score of the union method exceeds that of the grow-diag-final method. Looking at the BLEU score from highest to bottom (union > grow-diag-final > grow-diag-final-and > grow-diag > grow > intersect), the alignment result contains the BLEU score of the alignment reorganization method with more alignment points, which is different from the situation in strict phrase extraction. This shows that in nonstrict phrase extraction, the coverage rate of alignment points has a greater impact on the final result than the accuracy rate. Because the nonstrict phrase extraction itself has a certain antinoise ability, it reduces the requirements for word alignment accuracy and does not require a very complicated alignment and recombination method.

On the whole, the extraction constraint strategy can effectively improve the BLEU score. The method based on vocabulary similarity is better than the method based on the intersection of alignment points. The improved self-attention constraint is based on the maximum likelihood under the condition of the alignment point. The comparison method has the highest BLEU score under the union word alignment and reorganization. Among all the methods based on vocabulary similarity, the method based on improved self-attention is less effective. Even under the condition of union and grow-diag-final word alignment and recombination, the BLEU score is worse than the method based on the intersection of alignment points. There is not much improvement effect; the method based on PHI square coefficient has better BLEU scores under various word alignment and recombination conditions; the method based on log-likelihood ratio is BLEU under the conditions of union, grow, and grow-diag word alignment and

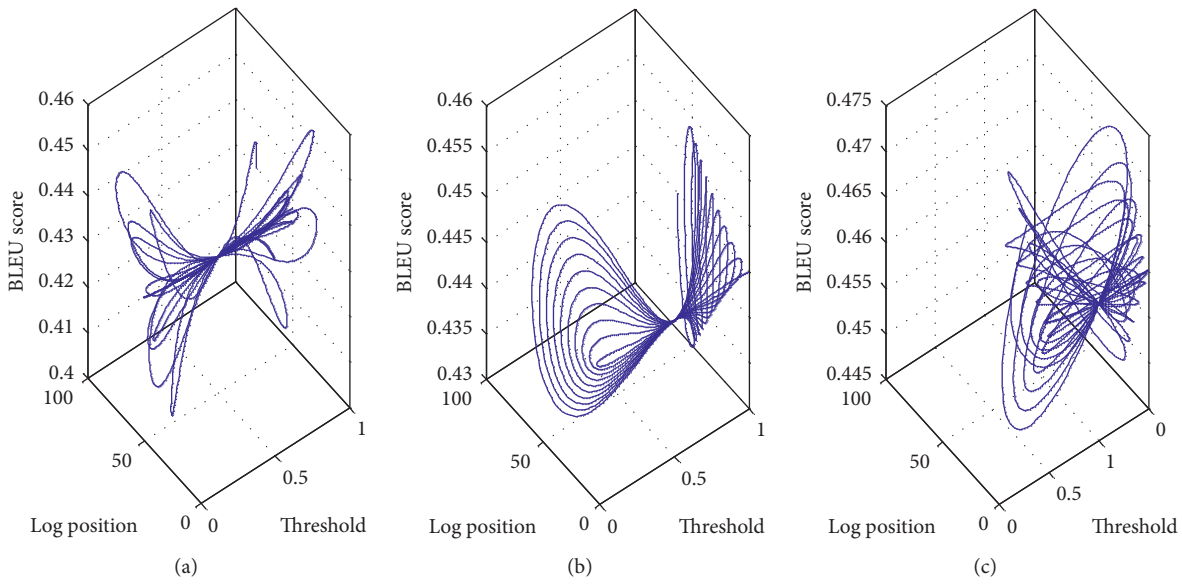

Figure 4: The relationship between BLEU and threshold under the constraints of improved self-attention method for single alignment: (a) constraints are not based on alignment points, (b) constraints are based on alignment points, and (c) improved self-attention constraints are based on alignment points.

recombination. The score is better, but the BLEU score under the conditions of grow-diag-final and grow-diag-final-and word alignment and recombination has a large drop, which shows that the log-likelihood ratio constraint is too strict. In these two types of methods, final and final-and processes may include alignment points that are not deterministic alignments, but the log-likelihood ratio constraint regards these alignment points as must be included in phrase extraction. However, the log-likelihood ratio has a better ability to constrain too broad results like union word alignment and recombination.

In Figure 4, the influence of threshold changes in the constraint extraction strategy based on improved self-attention on the BLEU score of the final translation is shown. From this, we can see that the threshold change has a greater impact on the BLEU score of the final translation result, indicating that the improved self-attention constraint has a greater impact on phrase extraction, which means that this method can form an effective constraint.

4.2. Translation Effect on n-Best Word Alignment. We take the best alignment numbers as 10, 20, 30, 40, and 50, respectively, for the translation experiment on n-best alignment. We still compare the final translation quality of nonstrict phrase extraction and strict phrase extraction without adding extraction constraints. The BLEU scores of strict phrase extraction using various word alignment and recombination methods are shown in Table 3. Here, the best result on n-best is selected for each word alignment.

In the alignment and reorganization method, the result of n-best is not as good as the result of single alignment. This is mainly because these alignment and reorganization methods cover more alignment points on the n-best alignment, and strict phrase extraction can only perform phrase extraction based on the outermost boundary of the alignment, so it is more severely affected by noise. There are certain improvements in other alignment and reorganization methods, mainly because these methods cover fewer alignment points on a single alignment, and many useful alignment points are recalled after being expanded to n-best. However, from a general point of view, the highest BLEU score of strict phrase extraction on the n-best alignment results is still lower than that of a single alignment, indicating that strict phrase extraction is not suitable for the n-best alignment and reorganization used in this article.

Table 3: BLEU scores for strict phrase extraction in n-best alignment.

Alignment and reorganization method	BLEU
Grow-diag-final-and	0.39
Grow-diag-final	0.41
Grow-diag	0.37
Grow	0.33
Intersect	0.31
Union	0.25

Figure 5 shows the variation of strict phrase extraction with n-best alignment. It can be seen that for all alignment and recombination methods, the BLEU score fluctuates with the increase of n-best alignment, which shows that the strict phrase extraction method can improve the effectiveness of extraction as the alignment number increases.

Table 4 shows the BLEU scores of various word alignment and recombination methods used in nonstrict phrase extraction without extraction constraints. It can be seen that in all word alignment and recombination methods, nonstrict

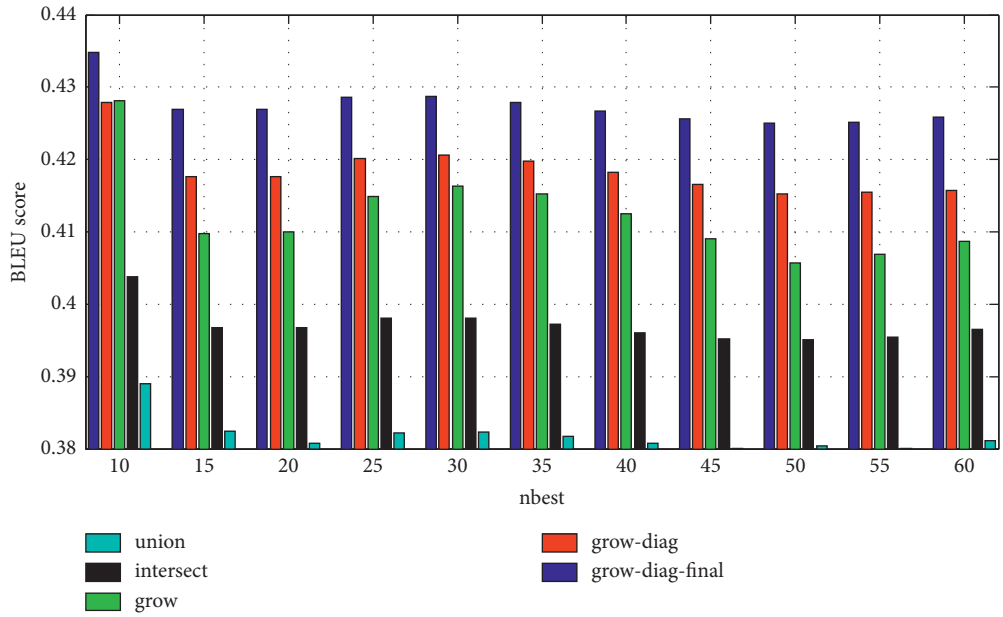

Figure 5: The BLEU score of the simple superposition of n-best alignment results in strict phrase extraction.

Table 4: BLEU scores for nonstrict phrase extraction in n-best alignment.

Alignment and reorganization method	BLEU
Grow-diag-final-and	0.40
Grow-diag-final	0.42
Grow-diag	0.36
Grow	0.34
Intersect	0.32
Union	0.26

phrase extraction improves the BLEU score on n-best than on single alignment. This shows that the n-best alignment recombination method mentioned in this article is suitable for nonstrict phrase extraction.

Figure 6 further shows the variation of nonstrict phrase extraction with n-best alignment. It can be seen that for most alignment and recombination methods, the BLEU score does not change much with the number of n-best alignments. Therefore, simply using the nonstrict phrase extraction method cannot improve the effectiveness of the extraction with the increase of the number of alignments, but it will not significantly reduce the effectiveness. It is also worth noting that in terms of n-best alignment, the grow-diag-final method is better than the union method. This may be due to the introduction of too many alignment points in the union method, which reduces the effectiveness of phrase extraction.

Figure 7 shows the relationship between the BLEU score and the n-best alignment under the constraints of the improved self-attention alignment point intersection method. When the number of n-best alignments increases, the BLEU score is not less than 0.445, which shows that the method based on the intersection of alignment points is effective in n-best alignment.

Figure 8 shows the relationship between the BLEU score and the threshold under the improved self-attention constraint in n-best alignment. It can be seen from the figure that for improved self-attention, there is still a difference whether the alignment point has been aligned under n-best alignment. When the threshold is increased, the improved self-attention constraint is relatively maximum based on the BLEU score of the alignment point.

The effect is best when the constraint is based on the existing alignment, the effect is worse when the constraint is strictly based on the existing alignment, and the effect is the worst when the constraint is not based on the existing alignment. Because the log-likelihood method has strong constraints, when there are many alignment points, the constraints strictly based on the existing alignment are too strict, so it becomes worse. The constraints are not strictly based on the existing alignment. The constraints are looser, and the effect is better. When the alignment is not based on the existing alignment, the alignment points other than the existing alignment can be used as constraints, which is equivalent to strengthening the constraints, and the effect is not good. As the threshold increases, the constraint relaxes, and the BLEU score increases strictly according to the existing alignment method, indicating that this restriction is too strict for n-best alignment; the BLEU score that does not strictly follow the existing alignment method increases first, indicating that the degree of restriction is moderate, and only when the threshold is relatively large will it show a downward trend; the BLEU score basically declines without the existing alignment method, and its effect is not as good as the previous two.

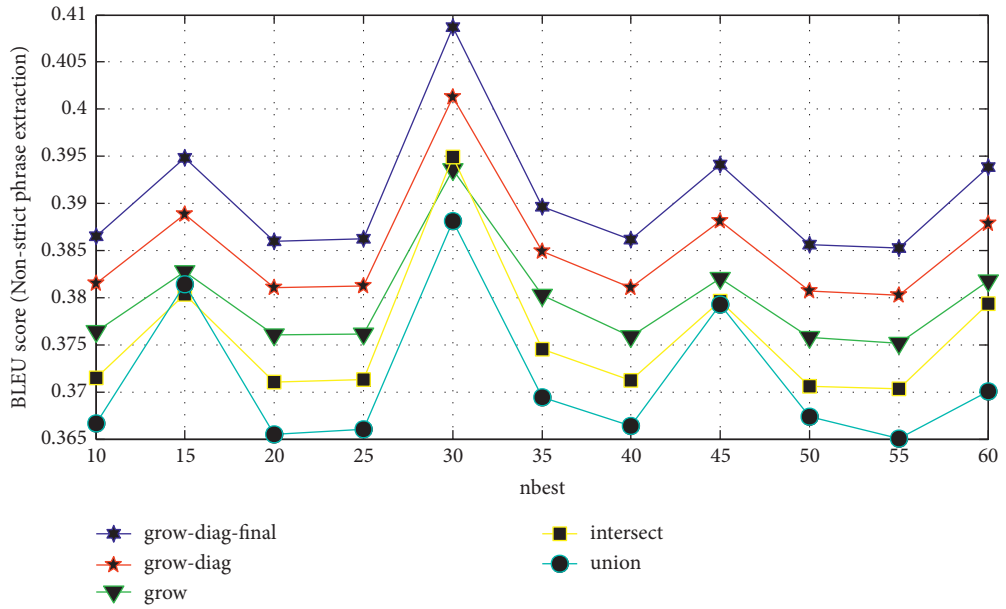

Figure 6: BLEU score of simple superposition of n-best alignment results in nonstrict phrase extraction.

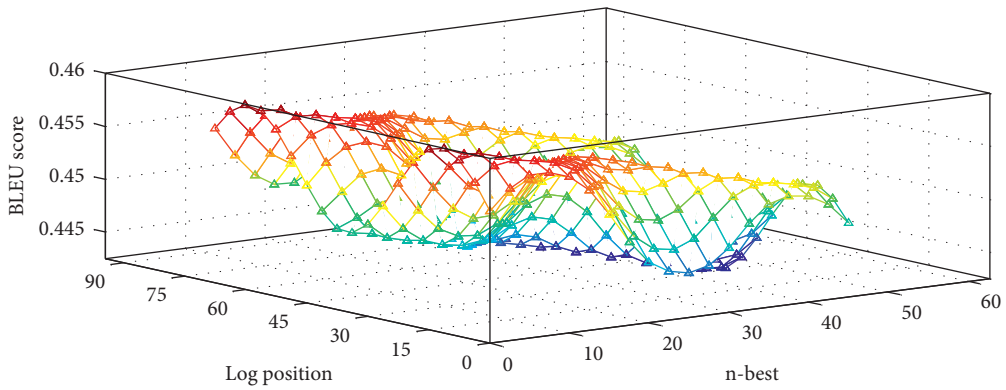

Figure 7: The relationship between BLEU and n-best alignment under the constraint of the improved self-attention alignment point intersection method.

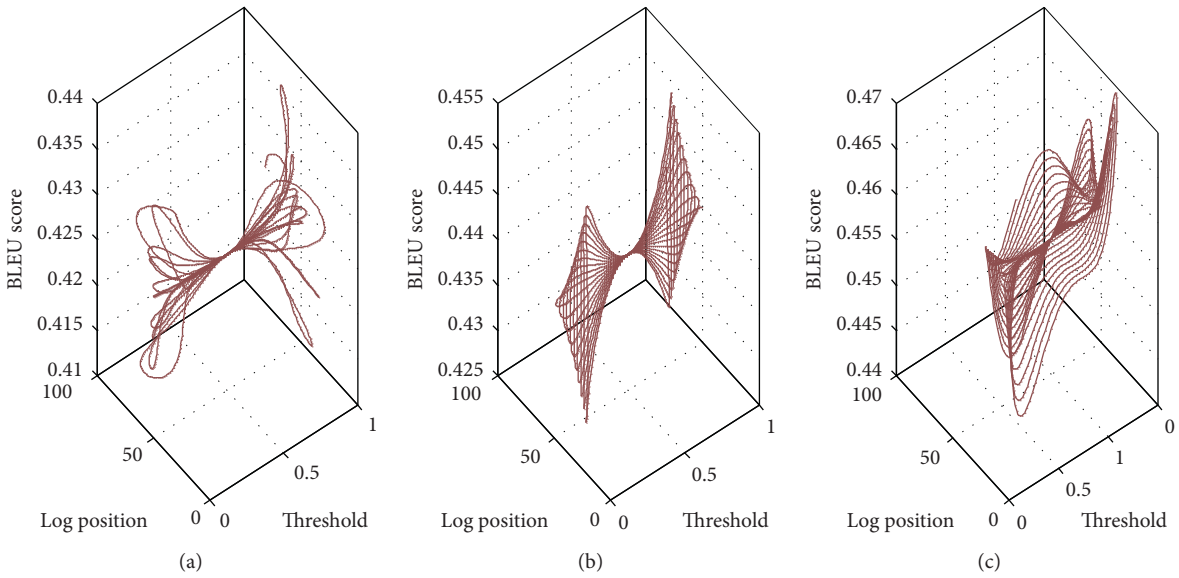

Figure 8: The constraint of n-best alignment in the improved self-attention is based on the BLEU score under the alignment point. (a) Constraints are not based on alignment points, (b) constraints are based on alignment points, and (c) improved self-attention constraints are based on alignment points.

5. Conclusion

This article analyzes the self-attention mechanism ignoring word order structure. Aiming at the problem of not being able to capture the position information of the words in the sentence, the analysis shows that the position of the words in the sentence is very important feature information. It plays an important role in guiding reference disambiguation and semantic analysis. For this problem, this paper proposes a new English machine translation model based on logarithmic position representation and self-attention. This model further enhances the model's ability to capture word position information by adding logarithmic position representation in the self-attention layer. This performance enhancement is not only reflected in distance but also in directionality. The logarithmic representation method blurs the concept of "long distance" and makes the relative position representation free from the "window." The experimental results show that the model proposed in the article has better performance than the traditional recurrent neural network English machine translation model and the traditional self-attention English machine translation model in English-to-German and English-to-French English machine translation tasks. This article proposes the idea of using n-best alignment results for phrase extraction. In order to effectively extract phrases from n-best alignment results, a nonstrict phrase extraction method is proposed, focusing on the impact of various extraction constraint strategies in nonstrict phrase extraction methods on the quality of the final translation, mainly including alignment points. Compared with the traditional strict phrase extraction method, the final translation quality of nonstrict phrase extraction in both single alignment and n-best alignment is improved, and it is more suitable for extracting phrases from n-best alignment effectively. However, the error recognition rule base needs to be improved. The error-driven long sentence segmentation method formulates error identification and correction strategies by summarizing the errors in the segmentation results. Fundamentally speaking, these strategies belong to the category of rules. In the future, we will consider formulating a more standardized and complete knowledge representation form to accurately represent each linguistic feature, so as to promote the application of the method.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This study was supported by Provincial Teaching Reform Research Project of Hubei Provincial Department of Education.

References

[1] H.-I. Liu and W.-L. Chen, "Re-transformer: a self-attention based model for machine translation," *Procedia Computer Science*, vol. 189, pp. 3–10, 2021.

[2] B. Yang, L. Wang, D. F. Wong, S. Shi, and Z. Tu, "Context-aware self-attention networks for natural language processing," *Neurocomputing*, vol. 458, pp. 157–169, 2021.

[3] K. Chen, R. Wang, M. Utiyama et al., "Towards more diverse input representation for neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1586–1597, 2020.

[4] X. Shi, H. Huang, P. Jian, and Y.-K. Tang, "Improving neural machine translation with sentence alignment learning," *Neurocomputing*, vol. 420, pp. 15–26, 2021.

[5] H. Qun, L. Wenjing, and C. Zhangli, "B&Anet: c," *Speech Communication*, vol. 125, pp. 15–23, 2020.

[6] M. R. Uddin, S. Mahbub, M. S. Rahman, and M. S. Bayzid, "SAINT: self-attention augmented inception-inside-inception network improves protein secondary structure prediction," *Bioinformatics*, vol. 36, no. 17, pp. 4599–4608, 2020.

[7] Y. Jin, C. Tang, Q. Liu, and Y. Wang, "Multi-head self-attention-based deep clustering for single-channel speech separation," *IEEE Access*, vol. 8, pp. 100013–100021, 2020.

[8] S. Yang, H. Lu, S. Kang et al., "On the localness modeling for the self-attention based end-to-end speech synthesis," *Neural Networks*, vol. 125, pp. 121–130, 2020.

[9] T. Zhang, H. Huang, C. Feng, and L. Cao, "Self-supervised bilingual syntactic alignment for neural machine translation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14454–14462, 2021.

[10] F. Lin, X. Ma, Y. Chen, J. Zhou, and B. Liu, "PC-SAN: pretraining-based contextual self-attention model for topic essay generation," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 14, no. 8, pp. 3168–3186, 2020.

[11] S. P. Singh, M. K. Sharma, A. Lay-Ekuakille, D. Gangwar, and S. Gupta, "Deep ConvLSTM with self-attention for human activity decoding using wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 8575–8582, 2020.

[12] Y. Li, C. Lin, H. Li, W. Hu, H. Dong, and Y. Liu, "Unsupervised domain adaptation with self-attention for post-disaster building damage detection," *Neurocomputing*, vol. 415, pp. 27–39, 2020.

[13] X. Feng, Z. Feng, W. Zhao, B. Qin, and T. Liu, "Enhanced neural machine translation by joint decoding with word and POS-tagging sequences," *Mobile Networks and Applications*, vol. 25, no. 5, pp. 1722–1728, 2020.

[14] Y. Ju, J. Li, and G. Sun, "Ultra-short-term photovoltaic power prediction based on self-attention mechanism and multi-task learning," *IEEE Access*, vol. 8, pp. 44821–44829, 2020.

[15] I. K. E. Ampomah, S. McClean, L. Zhiwei, and G. Hawe, "Every layer counts: multi-layer multi-head attention for neural machine translation," *The Prague Bulletin of Mathematical Linguistics*, vol. 115, no. 1, pp. 51–82, 2020.

[16] C. Park and H. Lim, "A study on the performance improvement of machine translation using public Korean-English parallel corpus," *Journal of Digital Convergence*, vol. 18, no. 6, pp. 271–277, 2020.

[17] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, 2020.

[18] S. Sivakumar and R. Rajalakshmi, "Self-attention based sentiment analysis with effective embedding techniques,"

International Journal of Computer Applications in Technology, vol. 65, no. 1, pp. 65–77, 2021.

[19] Z. Meng, S. Tian, L. Yu, and Y. Lv, "Joint extraction of entities and relations based on character graph convolutional network and Multi-Head Self-Attention Mechanism," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 2, pp. 349–362, 2021.

[20] M. Yang, R. Wang, K. Chen, X. Wang, T. Zhao, and M. Zhang, "A novel sentence-level agreement architecture for neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2585–2597, 2020.

[21] T. Huang, Z.-H. Deng, G. Shen, and X. Chen, "A Window-Based Self-Attention approach for sentence encoding," *Neurocomputing*, vol. 375, pp. 25–31, 2020.

[22] X. Wang, X. Mei, Q. Huang, Z. Han, and C. Huang, "Fine-grained learning performance prediction via adaptive sparse self-attention networks," *Information Sciences*, vol. 545, pp. 223–240, 2021.

[23] J. Armengol-Estapé and M. R. Costa-jussà, "Semantic and syntactic information for neural machine translation," *Machine Translation*, vol. 35, no. 1, pp. 3–17, 2021.

[24] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech d using temporal convolutional networks with self attention," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1598–1607, 2020.

[25] S. Kwon, B.-H. Go, and J.-H. Lee, "A text-based visual context modulation neural model for multimodal machine translation," *Pattern Recognition Letters*, vol. 136, pp. 212–218, 2020.

[26] Z. Boodeea and S. Pudaruth, "Kreol mkm translation system using attention and transformer model," *International Journal of Computing and Digital Systems*, vol. 9, no. 6, pp. 1143–1153, 2020.

[27] Z. A. Zeeshan and M. Z. Jawad, "Research on Chinese-Urdu machine translation based on deep learning," *Journal of Autonomous Intelligence*, vol. 3, no. 2, pp. 34–44, 2020.

[28] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3192–3208, 2020.

[29] J. Á González, L. F. Hurtado, and F. Pla, "Self-attention for Twitter sentiment analysis in Spanish," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 2, pp. 2165–2175, 2020.

[30] M. Yang, S. Liu, K. Chen, H. Zhang, E. Zhao, and T. Zhao, "A hierarchical clustering approach to fuzzy semantic representation of rare words in neural machine translation," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 5, pp. 992–1002, 2020.