Hindawi
Complexity
Volume 2020, Article ID 6657344, 11 pages
https://doi.org/10.1155/2020/6657344

WILEY | Hindawi

Research Article

Neural Network Machine Translation Method Based on Unsupervised Domain Adaptation

Rui Wang

School of Foreign Studies, Xi'an University of Finance and Economics, Xi'an 710006, China

Correspondence should be addressed to Rui Wang; wangrui23@xaufe.edu.cn

Received 20 October 2020; Revised 10 December 2020; Accepted 14 December 2020; Published 24 December 2020

Academic Editor: Wei Wang

Relying on large-scale parallel corpora, neural machine translation has achieved great success in certain language pairs. However, the acquisition of high-quality parallel corpus is one of the main difficulties in machine translation research. In order to solve this problem, this paper proposes unsupervised domain adaptive neural network machine translation. This method can be trained using only two unrelated monolingual corpora and obtain a good translation result. This article first measures the matching degree of translation rules by adding relevant subject information to the translation rules and dynamically calculating the similarity between each translation rule and the document to be translated during the decoding process. Secondly, through the joint training of multiple training tasks, the source language can learn useful semantic and structural information from the monolingual corpus of a third language that is not parallel to the current two languages during the process of translation into the target language. Experimental results show that better results can be obtained than traditional statistical machine translation.

1. Introduction

At present, with the gradual deepening of international exchanges, people's demand for language translation is increasing day by day [1, 2]. However, there are so many kinds of languages in the world, and the Internet has become the most convenient platform for obtaining information, and users have an increasingly urgent demand for online translation [3]. There are many kinds of languages on the Internet, each language has a lot of ambiguity, and the language is changing all the time, which puts higher requirements on translation services [4, 5]. In the prior art, in order to realize automatic machine translation, the currently commonly used techniques are methods based on the neural network [6, 7] and methods based on statistical machine translation [8, 9].

The former is neural machine translation (NMT). The latter is statistical machine translation (SMT). Iswarya and Radha [10] used an unsupervised method to achieve cross-language embedding and trained a good word-to-word model. On the basis of this work, Imankulova et al. [11] generated a pseudoparallel corpus for training through noise reduction and reverse translation and obtained good experimental results. Lee et al. [12] used a character-level decoder to improve the quality of morphologically rich language translation. Morente-Molinera et al. [13] selected granular information of words and characters in the encoder and used multiple attentions on the decoding side to make information of different granularities collaboratively help translation. Zhang et al. [14] modelled the similarity between language pairs in the same language family. Their encoder was composed of character-level one-way RNN and word-level two-way RNN and used a top-down hierarchical attention mechanism to obtain words first. Park et al. [15] proposed regularization of subwords, using a unary language model to generate multiple candidate subword sequences, enriching the input of the encoder to enhance the robustness of the translation system. Zhao et al. [16] introduced the representation of multigranularity BPE to obtain the semantic representation of vocabulary on average. Zhang et al. [17] believed that the encoder word vector layer, decoder word vector layer, and decoder output layer have different functions, so the choice of BPE granularity for different layers should also be different. Zhang et al. [18] used noise-

reducing autoencoders and adversarial training to map the two languages to the same implicit space and iteratively trained translation models in both directions. Wang et al. [19] first pretrained the word vectors and implemented an unsupervised translation model using autoencoder and reverse translation. Dabre et al. [20] believed that the previous unsupervised translation model uses a shared encoder to encode the semantic representation of different languages, which can easily lose the respective characteristics of different languages, thereby limiting translation performance. Therefore, they proposed that each language should use its own encoder for modelling and only share the weights of the last few layers of the encoder and the first few layers of the decoder.

However, the data-weighting method in the prior art assigns weights to those sentences based on their similarity to the corpus in the domain. The above existing technologies are inseparable from the serious problem of the annotated corpus, and the original training needs to be combined. However, the data-weighting method in the existing technologies assigns weights to those sentences according to the similarity with the corpus in the domain. The serious problems in the existing technology without annotation corpus need the original training corpus segmentation of several small elements leading to increase in the number of model parameters such as complicated operations, which makes the neural network performance of the machine translation is inefficient and cannot accurately obtain between the various areas of adaptive [21, 22]. In order to solve the above problems, this paper proposes an adaptive neural network machine translation in the unsupervised field. In this paper, the matching degree of translation rules is measured by adding relevant topic information to the translation rules and dynamically calculating the similarity between each translation rule and the document to be translated during the decoding process. Finally, through the joint training of multiple training tasks, the source language can learn useful semantic and structural information from the monolingual corpus of the third language which is not in line with the current two languages in the process of translation to the target language.

2. Machine Translation Related Technologies

2.1. Machine Translation Framework. At this stage, statistical machine translation is divided into a generative noise channel model [23, 24] and a discriminative log-linear model [25, 26]. We assume that the source sentence is s and the target sentence t.

2.1.1. Noise Channel Model. The noise channel is proposed based on the coding idea in information theory. In this model, the machine translation task is regarded as the information transmission process of the target sentence e being transformed into the source language s after passing through a noise channel. The process of searching for t that maximizes the translation probability $P(t|s)$ is as follows:

$$tp = \arg\max P(t|s). \tag{1}$$

According to the Bayesian principle, the above formula can be converted to

$$tp = \arg\max\left(\frac{P(t)P(s|t)}{P(s)}\right). \tag{2}$$

The translation model based on the noise channel cannot use more knowledge than the source sentence and target sentence in the translation process, and the importance of the language model and the translation model is fixed and cannot be adjusted according to the actual situation.

2.1.2. Log-Linear Model. The translation system based on the log-linear model decomposes the translation probability into a series of combinations of features:

$$P(t|s) = \frac{e^{\sum_{i=1}^{n}\alpha_i Q_i(t,s)}}{\sum e^{\sum_{i=1}^{n}\alpha_i Q_i(t,s)}}. \tag{3}$$

The translation system based on the log-linear model is very flexible, and some additional descriptive features can be added as needed, such as the number of words contained in translation candidates, the number of rules. Figure 1 shows the construction process of a translation system based on a logarithmic linear model. We can see that the machine translation system contains three parts of data: training data, development data, and test data. The language model is trained on large-scale monolingual training data. The translation system obtains bilingual word alignment information through machine learning methods on bilingual parallel training data and extracts translation rules and estimates their probabilities. The translation system adjusts the feature weights by minimizing error rate training on independently developed data. System performance evaluation is based on existing models and weights to translate test data and evaluate its performance.

2.2. Unsupervised Domain Adaptation. Effective feature extraction is a common basic element of various machine learning methods. As shown in Figure 2, suppose the current layer is a p-dimensional vector V and the previous layer is a q-dimensional vector V_f. First, construct a p-dimensional output layer O, and initialize the parameters of the two layers randomly. Given the input I, the hidden layer state H, and the output layer result O′, then, use the difference between O and O′ as the loss for backpropagation to update the parameters of the two layers. The single hidden layer neural network constructed in this way can be understood as a process of encoding the input I to obtain the hidden layer H and decoding the hidden layer G to obtain the input I. If $q < p$, the parameters obtained by such training can compress O while minimizing the coding loss. If $q \geq p$, then we need to add a regularization factor to the loss function for sparse coding or dimension upgrade.

In the research of domain adaptation, deep learning algorithms mainly learn the intermediate representations

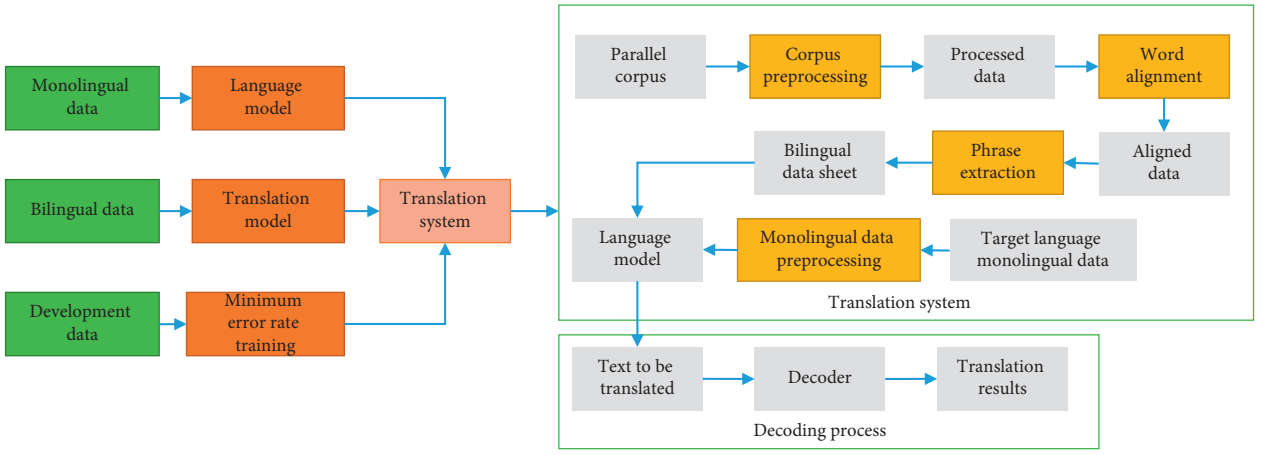

Figure 1: Machine translation system framework.

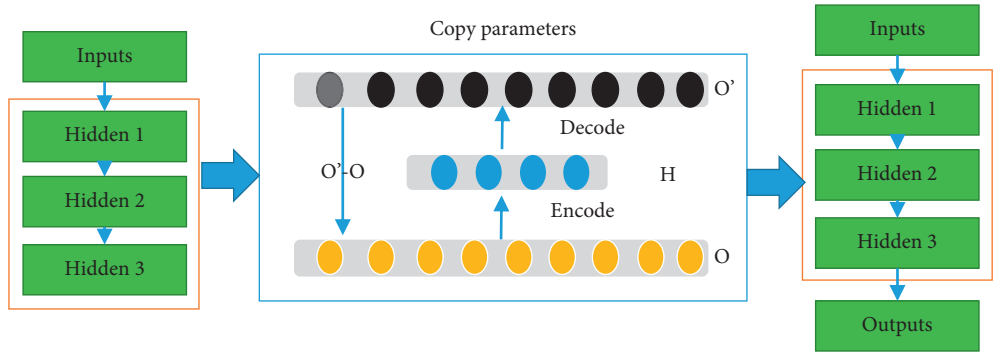

Figure 2: Unsupervised training based on automatic coding.

between input and output. The motivation behind these intermediate representations is that the results of these intermediate representations can bring better cross-domain machine learning performance. Since deep learning can carry out unsupervised training, massive open domain data can be used to learn the topic information representation of this domain. Deep learning is one of the fastest-growing fields in the field of machine learning in recent years. It has made breakthroughs in many natural language processing applications and is a direction worth trying.

3. Machine Translation Algorithm Based on Unsupervised Domain Adaptation

3.1. Sequence-Dependency Structure. This paper uses Transformer as the basic structure to create an encoder and a decoder for each language and share the three principles of parameter training of some layers. A training task is established between English, French, and German at the same time, and the training model is obtained. For example, when training English⟶French and English⟶German tasks, because French and German have similar language structures, useful semantic and structural information can be jointly learned from different target languages.

The main feature of Transformer is that it does not rely on RNN or CNN, but only uses the self-attention mechanism to achieve an end-to-end translation model. The self-attention mechanism is to perform attention calculation on each word in a sentence and all other words in the sentence. The purpose is to learn the dependencies within the sentence and capture the internal structure of the sentence. The structure diagram of Transformer's architecture is shown in Figure 3.

The encoder and decoder of Transformer are both multilayer network structures, and both the encoder and decoder contain M identical layers. In the encoder, each layer contains two sublayers, namely, the self-attention mechanism layer and the feedforward neural network layer. In each layer of the decoder, there are 3 sublayers. In addition to a mask self-attention mechanism layer and a feedforward neural network layer, there is also a multihead attention mechanism to the decoder output. The residual connection is used between the sublayers, and the method of residual connection can be expressed by the following formula:

$$S^i = S^{i-1} + Y_{S^l}\left(S^{i-1}\right). \tag{4}$$

Among them, S^i represents the output of the i-th sublayer and Y_{S^l} represents the function of the layer.

3.2. Bilingual Single-Task Model. In this paper, s and t are used to represent the set of sentences in the source language and the target language; M_s and M_t are, respectively, the

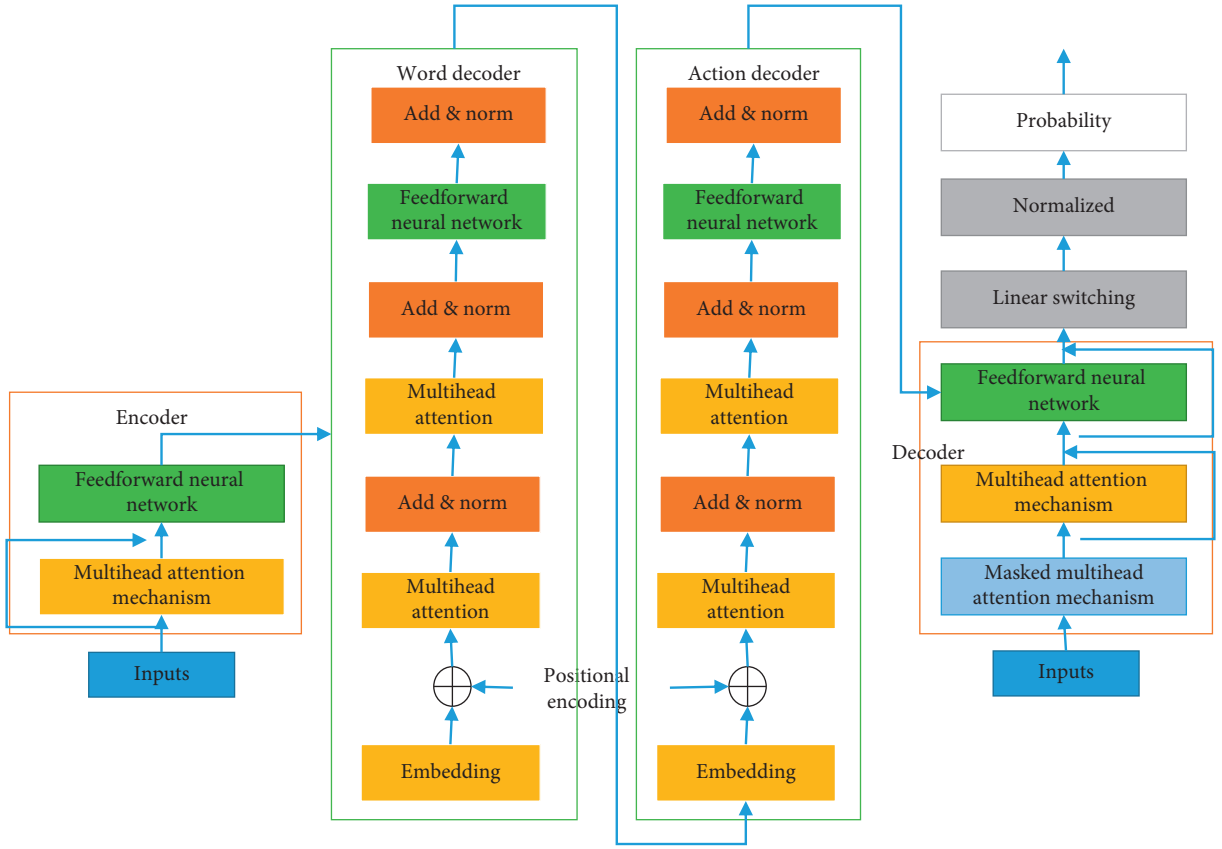

Figure 3: Transformer architecture.

language models trained by the monolingual of the source language and the target language; and $M_{s\longrightarrow t}$ and $M_{t\longrightarrow s}$ are used to represent the source language. The process of the single-task model is mainly composed of the following three steps to the predicted probability of the target language and the translation model from the target language to the source language.

(1) Initialization: The initialization of the model is roughly divided into two ways—the first method uses word2vec to train the word vectors of the two languages separately and then maps the word vectors of the two languages to the same latent space by learning a transformation matrix. In this way, a bilingual vocabulary with good accuracy can be obtained. The second method uses byte-pair encoding (BPE) of words as subword units. The advantage of this is that while reducing the size of the vocabulary, it eliminates the problem of "UNK" in the translation process. In addition, compared with the first method, the second method chooses to mix and scramble the two monolingual corpora to learn word vector features together. The source language and the target language can share the same vocabulary.

(2) Language model: In the bilingual single-task model, the noise reduction autoencoder of the language model minimizes the loss function as

$$\begin{aligned}\text{Loss}_f = &\left(-\log M_{s\longrightarrow s}\left(a|Kt\left(a\right)\right)\right)\\ &+\left(-\log M_{t\longrightarrow t}\left(b|Kt\left(b\right)\right)\right).\end{aligned} \tag{5}$$

Among them, $M_{s\longrightarrow s}$ indicates that the sentence s belongs to the expectation of the cross-entropy loss of S and K (a) indicates the sentence after adding noise a to the existing sentence s; the method is to exchange the positions of some words in the sentence or delete some words. Language model: the training process of is essentially taking the sentence K (b) added with noise as the source input sentence, and the initial sentence s as the target input sentence.

(3) Reverse translation: The process of reverse translation is a process of training pseudoparallel sentence

pairs as parallel sentence pairs. The training loss function is shown in formula (7).

$$\begin{aligned}\text{Loss}_b = &\left(-\log M_{s\longrightarrow t}\left(b|K't\left(b\right)\right)\right)\\ &+\left(-\log M_{t\longrightarrow s}\left(a|K't\left(a\right)\right)\right).\end{aligned} \tag{6}$$

The reverse translation process is to treat both (K' (b), b) and (K' (a), a) as parallel sentence pairs for training, transforming unsupervised problems into supervised ones. Repeating (2) and (3) is the complete bilingual single-task model training process.

3.3. Multilanguage and Multitask Model. The multilanguage multitask model is the model obtained by training multitask under the Transformer architecture. Assuming that there are currently 3 languages L1, L2, and L3 monolingual corpus, which are not parallel to each other before, the multitask model includes 6 training tasks, namely, L1⟶L2, L2⟶L1, L1⟶L3, L3⟶L1, L2⟶L3, and L3⟶L2. Inspired by the research of Yang et al. [27], in order to distinguish the semantic structure of each language while learning the useful structural information contained in the other language, this research establishes an encoder and a decoder for each language, but share some of the parameters of the layer. The optimization process of the parameter λ is shown in the following formula:

$$F\left(\lambda\right) = \arg\max\left(\sum_{m=1}^{M}\left(\frac{1}{U}\sum_{i=1}^{U}\log P\left(a_i^m|b_i^m;\lambda\right)\right)\right). \tag{7}$$

Among them, $M = \{1, 2, 3, 4, 5, 6\}$ is the index of the translation task; U is the number of sentence pairs; and a and b are the sentences in the source language and target language in the current translation task. Such parameter settings enable different language pairs to learn useful information in other languages.

In order to strengthen the role of shared latent space, this paper trains a generative adversarial network G to establish a three classification task between three encoders corresponding to three languages. Its role is to predict the category of the current coding language. Convert the cross-entropy loss shown in the following formula:

$$\begin{aligned}F_G\left(\lambda_G\right) = &-\text{ED}_{s'\in L}\left(\log P\left(L = L_1|\text{ED}_{s'}\right)\right)\\ &-\text{ED}_{s'\in L}\left(\log P\left(L = L_2|\text{ED}_{s'}\right)\right)\\ &-\text{ED}_{s'\in L}\left(\log P\left(L = L_3|\text{ED}_{s'}\right)\right).\end{aligned} \tag{8}$$

Among them, ED (s') represents the prediction result of the currently encoded sentence s' through the encoder of the L language, s' may come from the source language or the target language; λ_G is the parameter for generating the confrontation network G; and $\text{L} \in \{\text{L1, L2, L3}\}$.

3.4. Topic Similarity Model. The topic model is a statistical model used to discover abstract topics in the fields of machine learning and natural language processing. The topic similarity model measures the degree of similarity between the translation rules and the topic distribution of the language to be translated. In order to calculate the similarity between the translation rule and the language to be translated, we need to assign a distribution probability to the subject at the same time for the source language and target language of the translation rule. Use this probability to characterize the distribution relationship between the source language and target language of this rule on each topic.

If s is used to represent the source language part of the translation rule, t is used to represent the target language part of the translation rule, topic_s is used to represent the topic set of the source language, and topic_t is used to represent the topic set of the target language, then for any translation rule, there will be two distributions of rules to topics: P (topic_s|s) represents the topic distribution probability of the source language part of the translation rule in the source language, P (topic_t|t) represents the topic distribution of the target language part of the translation rule in the target language probability.

In the topic similarity model, you can choose the Hellinger Disaster (HD) to calculate the similarity of the topic between the translation rules and the document to be translated. Among them, the HD similarity evaluation method is a symmetric algorithm, which has been widely used to compare the similarity between two distributions. Assuming that the distribution P (topic|s) from the translation rule to the topic and the distribution P (topic|t) from the document to the topic is given, the formula for calculating the similarity between the two can be written as

$$\text{Dis}\left(P\left(\text{topic}|\right), P\left(\text{topic}|s\right)\right) = \sum_{i=1}^{n}\left(\sqrt{P\text{topic} = i|t)} - \sqrt{P\text{topic} = i|s)}\right)^2. \tag{9}$$

Obviously, by comparing all the translation candidates and the HD of the language to be translated, the similarity between the translation candidates and the translated language can be obtained. In information theory, the smaller HD distance represents the greater similarity, because our task is to find the translation with the greatest similarity between the selected language and the translated language as the final translation result. With the addition of the topic similarity model, our goal is to select the translation rules that are most similar to the translated language to realize the basis for adaptive translation using topic information.

3.5. Machine Translation Model and Process. In phrase-based statistical machine translation, the source language sentence $s = \{s_1, s_2, \ldots, s_n\}$ is translated using a logarithmic linear model. By comparing all translation candidates with the HD distance of the language to be translated, translation candidates can be obtained by the similarity between the translated language and the translated language, finding the translation with the greatest similarity between the selected and translated language after translation as the final translation result. The target translation with the greatest similarity $t = \{t_1, t_2, \ldots, t_n\}$:

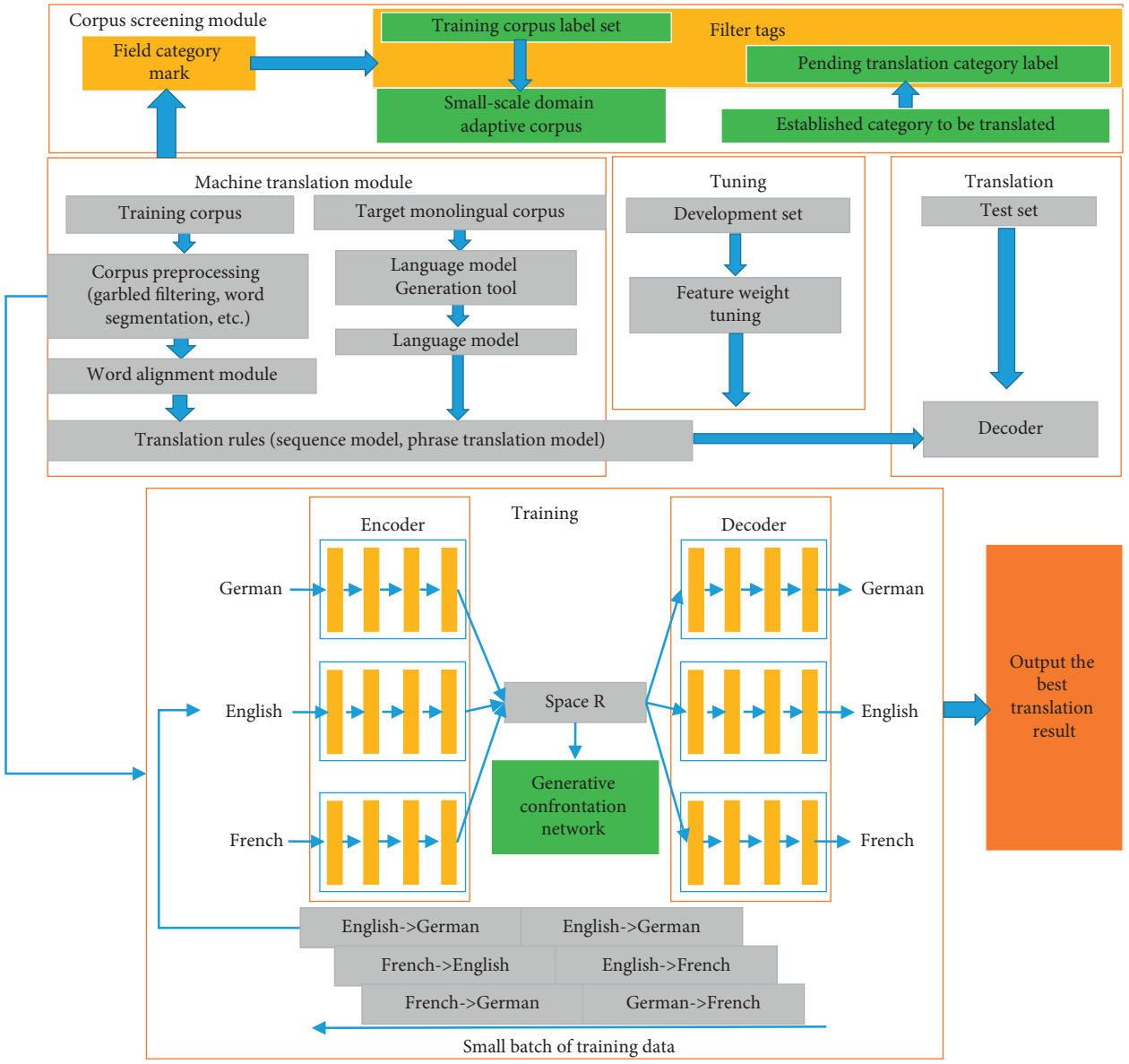

Figure 4: Flow chart of the machine translation algorithm in this article.

$$\text{Fin} = \arg\max \sum_{i=1}^{n} \lambda_n \text{Dis}(s,t). \tag{10}$$

Among them, Dis (s, t) is the characteristic function and λ_n is the characteristic weight.

The machine translation algorithm in this paper includes three stages of processing, training, tuning, and translation. As shown in Figure 4, it is necessary to prepare training data, monolingual target language corpus, development set, and test set.

The training data is bilingual translation corpus, mostly sentence alignment. After preprocessing and word alignment, various translation rules are obtained, including phrase translation table, ordering probability, and maximum entropy ordering parameters.

For the monolingual target language corpus, you can use the target language side of the training data, or you can add more monolingual data, mostly at the sentence level, to train the language model. In addition to the various translation rules and language models generated during the training process, the operation of the decoder also requires feature weights. The process of tuning is to select feature weights on the development set.

The development set is a collection of sentences in the source language, and each source language sentence has one or more reference translations in the target language. Tuning on the development set usually uses minimum error

training. It requires the decoder to continuously iterate the current feature parameters, automatically calculate and compare BLEU scores, and then change the weights to decode again until the upper limit of the number of iterations is reached or the translation system is stable. This is a problem of multidimensional parameter optimization. The decoder can implement the translation process by using the translation rules, language model, and feature weights obtained during the training process.

Use the test set to perform the translation and perform BLEU scoring to observe the translation effect of the translation system.

4. Results and Discussion

4.1. Experimental Setup. The experiment selects 10 million single sentences in English, German, and French from the WMT2007 to WMT2010 corpus. The experiment uses Adam as the optimizer, the deactivation rate (dropout) is set to 0.1, the dimension of the word is set to 512, the maximum sentence length is 175, and sentences with more than 175 words will be intercepted by the superlong part. The training step is 3.5×105 and the rest of the model parameters are set to the default parameters of the Transformer model. In the three-language multitask translation model, the vocabulary of the three languages is shared and the BPE operand is set to 85000. The fast text tool is used to train the cross-language word vector learning for the subworded training set.

In the evaluation of phrase-level translation, if a translation result candidate is the same as any one of the standard answers, we consider it to be correct. In the evaluation of sentence-level translation, the evaluation index of the translation result uses the case-insensitive 4-element BLEU value and uses the bootstrap resampling method to test the significance of the evaluation result.

4.2. Performance Comparison between Single-Task Model and Multitask Model. Figure 5 summarizes the translation performance of the single-task model and the multitask model on the test set. It can be seen from Figure 5 that the multitask model of this article has improved in the four translation tasks, but the effect of the improvement is quite different. In the two translation tasks of English⟶German and German⟶English, the test results showed that the BLEU value improved less. In the two translation tasks of German⟶French and French⟶German, the performance improved significantly, and the BLEU value increased by about 2.88 and 3.01 percentage points, but on the two translation tasks of English⟶French and French⟶English, the translation performance of the multitask model decreases.

In the multitask model of this article, a shared vocabulary is used for multiple languages, so it is particularly important to choose an appropriate vocabulary. In this regard, this paper has also done several experiments for comparative analysis. The experimental results are shown in Table 1. It can be seen from Table 1 that when the BPE operands are 85000 and 90000, the experimental results are relatively good, but the BLEU values of the two sets of BPE operands are not much different. And in the case of some language pairs, the BLE value of the BPE operand of 90000 is lower than the BLE value of the BPE operand of 85000. It is estimated that when the size of the vocabulary is further increased, the improvement of the experimental results is not significant. Therefore, in the final model of this article, the size of the BPE operand is selected to be 85000.

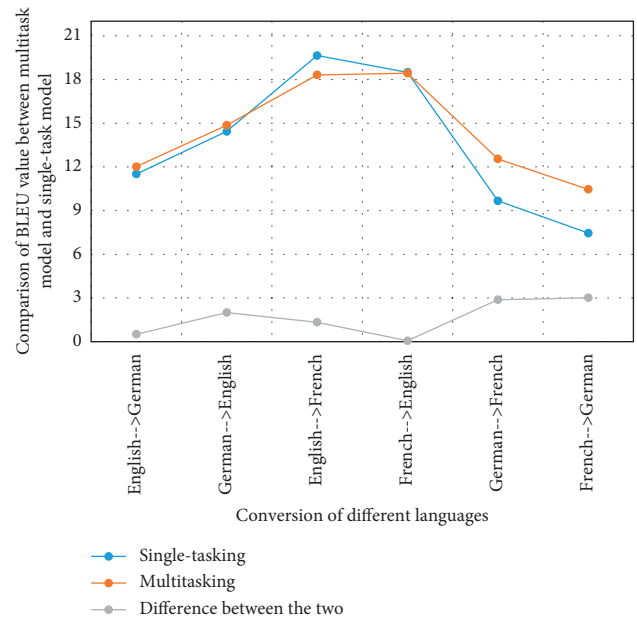

Figure 5: Performance comparison between multitask model and single-task model.

Table 1: Comparison of experimental results of different vocabulary sizes.

Conversion of different languages	BLEU value			
	60000	80000	85000	90000
English-->German	9.09	10.98	12.53	11.78
German-->English	13.99	14.31	14.97	15.01
German-->French	17.13	18.76	18.22	18.65
French-->German	16.12	17.88	18.52	18.05
English-->French	11.76	11.67	12.59	12.56
French-->English	9.88	10.22	10.67	11.08

In order to compare the training speed, this study counts the parameters that need training in the experiment. In the bilingual single-task translation task, the order of magnitude of the parameters is 1.3×10^8, while in the multilanguage and multitask translation model of this article, the total parameters are about 1.7×10^8. The number of parameters of the multilanguage translation model is only 1.3 times that of the bilingual translation system, which is much smaller than the sum of the parameters of the 6 tasks that are trained separately. Compared with the single-task model, the total training time of the multitask model is approximately reduced by half. In order to compare the translation performance and convergence speed of the two models more

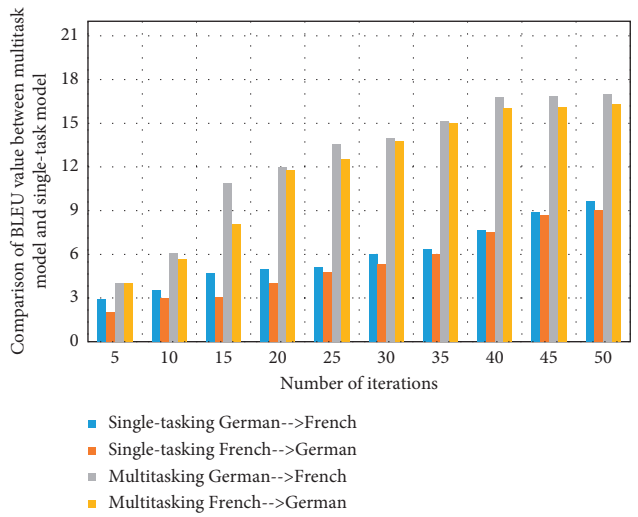

Figure 6: Comparison of the two models on the German <--> French translation task.

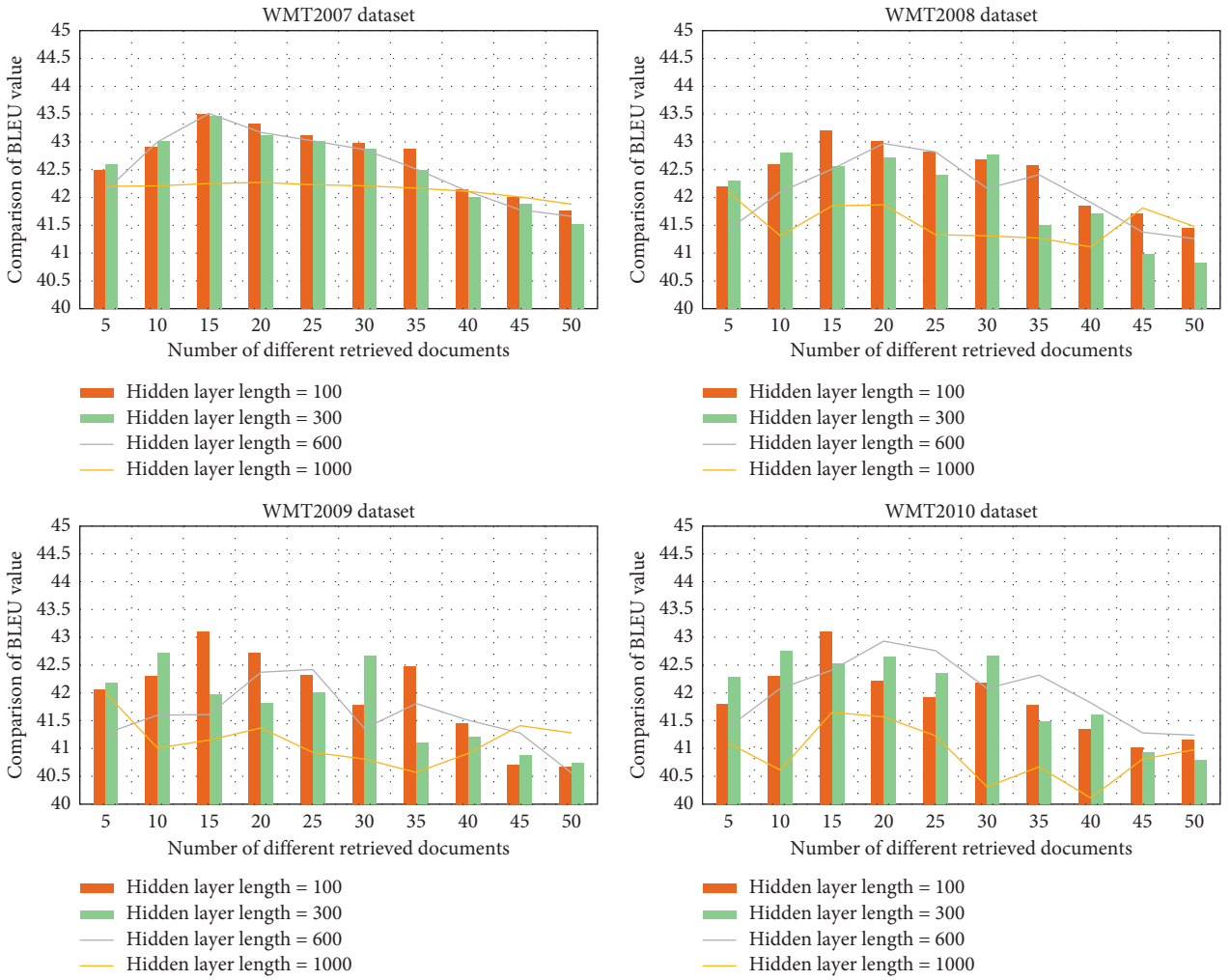

Figure 7: The impact of the number of retrieved documents and the length of the hidden layer on the accuracy of machine translation.

Table 2: Phrase-level translation accuracy.

	Single-tasking	Literature [15]	Literature [19]	Literature [20]	This paper
Top-1	0.621	0.812	0.789	0.822	0.819
Top-2	0.672	0.823	0.821	0.837	0.837
Top-3	0.721	0.856	0.867	0.856	0.8895
Top-4	0.739	0.889	0.892	0.891	0.919
Top-5	0.751	0.923	0.929	0.921	0.967

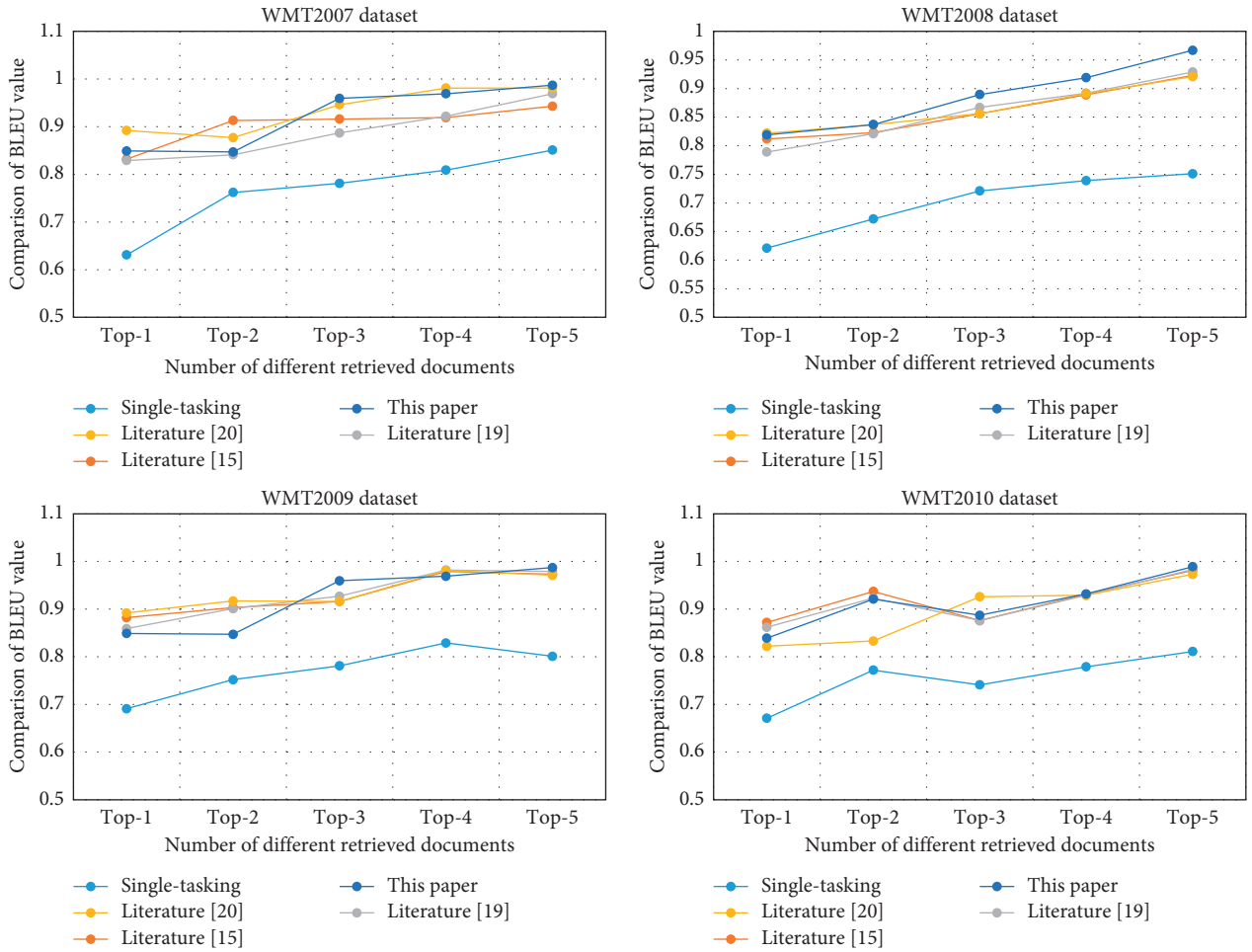

Figure 8: Sentence translation accuracy rate.

intuitively, on the German⟶French and French⟶German translation tasks where the translation effect has changed the most, a line graph is used to compare the bilingual single-task model and the multilingual proposed in this article; the effect of the multitask model is shown in Figure 6.

4.3. The Influence of the Number of Retrieved Documents and the Length of the Hidden Layer. We compared the effects of the number of retrieved documents and the length of the hidden layer on the accuracy of the translation model, and the results are shown in Figure 7. We found that for most results, the number of retrieved documents achieves the best translation accuracy when $N = 10$. This result confirms that the topic similarity of the information retrieval method is very helpful for determining topic information, and then it is helpful for choosing translation rules to an important role. However, in the experiment, when N is large, for example, when $N = 50$, the translation performance drops sharply. This is because as the number of retrieved documents further increases, topic-irrelevant documents will be introduced into the learning of the neural network. Irrelevant documents will bring topic-irrelevant real words, which will affect the performance of neural network learning.

Another important factor is the length L of the hidden layer vector in the neural network. In neural network learning, this parameter is usually adjusted by experience. In Figure 7, it can be seen that when L is small, the accuracy of the translation system is relatively high. In fact, in the

case of $L \leq 600$, the difference in translation performance is very small. However, when $L = 1000$, translation accuracy is worse than the other cases. The main reason is that the number of parameters in the neural network is so large that it cannot be learned well. We know that when $L = 1000$, there are a total of 100000×1000 parameters between the linear and nonlinear layers of the network. The current training data size is not enough to support this kind of network parameter level training, so the model is likely to fall into the local optimal and unacceptable topic representation information.

4.4. Phrase- and Sentence-Level Translation Performance. In the phrase-level translation process, Table 2 shows the accuracy rates of the top 5 phrase translation result candidates. It can be seen from the experimental results that our method and the methods proposed in literature [15], literature [19], and literature [20] are significantly better than the single-task translation model, which proves that our method is obtaining the latest translation. There is a great advantage in knowledge.

In the sentence-level translation evaluation, we tested the translation quality of different types of text and compared it with other algorithms. The experimental results are shown in Figure 8. Although the translation method in this article does not use any pretrained model, its translation results are comparable to traditional machine translation results based on massive training data. This shows that the translation knowledge obtained by the algorithm in this paper is very efficient.

5. Conclusion

Each has its own characteristics and flexible forms, making automatic language processing, including machine translation between languages, a difficult problem to be solved. At the same time, how to provide users with high-quality translation services has become a difficult problem to solve. Therefore, this article measures the matching degree of the translation rules by adding relevant subject information to the translation rules and dynamically calculating the similarity between each translation rule and the document to be translated during the decoding process. Then, through the joint training of multiple training tasks, the source language can learn useful semantic and structural information from the monolingual corpus of a third language that is not parallel to the current two languages during the process of translation into the target language. Finally, simulation experiments prove the effectiveness of the proposed algorithm. Experiments show that the algorithm used in this paper is significantly better than the comparison algorithm method, and only using part of the training data can achieve a better translation effect than the original training data, which improves the translation performance while reducing the translation system training and decoding costs.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no known conflicts of financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Scientific Research Program Funded by Shaanxi Provincial Education Bureau: Xi'an Tour Text Translation Strategy Research in Terms of Prototype and Model Theory (Program no. 18JK0298).

References

[1] D. Guo, W. Zhou, A. Li et al., "Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation," *IEEE Transactions on Image Processing*, vol. 29, pp. 1575–1590, 2019.

[2] R. Shadiev, A. Sun, and Y. M. Huang, "A study of the facilitation of cross-cultural understanding and intercultural sensitivity using speech-enabled language translation technology," *British Journal of Educational Technology*, vol. 50, no. 3, pp. 1415–1433, 2019.

[3] O. Kernberg, "Treinta métodos para destruir la creatividad de los candidatos a psicoanalistas," *Revista de Psicoanálisis*, vol. 85, pp. 47–62, 2019.

[4] I. R. Beiler and J. Dewilde, "Translation as translingual writing practice in English as an additional language," *The Modern Language Journal*, vol. 104, no. 3, pp. 533–549, 2020.

[5] H. C. Ouertani, L. Tatwany, and L. Tatwany, "Augmented reality based mobile application for real-time Arabic language translation," *Communications in Science and Technology*, vol. 4, no. 1, pp. 30–37, 2019.

[6] S. Jha, A. Dey, R. Kumar, and V. Kumar-Solanki, "A novel approach on visual question answering by parameter prediction using faster region based convolutional neural network," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 5, pp. 30–37, 2019.

[7] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting transformer to end-to-end spoken language translation," in *Proceedings of the INTERSPEECH 2019. International Speech Communication Association (ISCA)*, pp. 1133–1137, Graz, Austria, September 2019.

[8] S. K. Mahata, D. Das, and S. Bandyopadhyay, "Mtil2017: machine translation using recurrent neural network on statistical machine translation," *Journal of Intelligent Systems*, vol. 28, no. 3, pp. 447–453, 2019.

[9] Y. Xia, "Research on statistical machine translation model based on deep neural network," *Computing*, vol. 102, no. 3, pp. 643–661, 2020.

[10] P. Iswarya and V. Radha, "Adapting hybrid machine translation techniques for cross-language text retrieval system," *Journal of Engineering Science and Technology*, vol. 12, no. 3, pp. 648–666, 2017.

[11] A. Imankulova, T. Sato, and M. Komachi, "Filtered pseudo-parallel corpus improves low-resource neural machine translation," *ACM Transactions on Asian and Low-Resource*

Language Information Processing (TALLIP), vol. 19, no. 2, pp. 1–16, 2019.

[12] J. Lee, K. Cho, and T. Hofmann, "Fully character-level neural machine translation without explicit segmentation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 365–378, 2017.

[13] J. A. Morente-Molinera, G. Kou, C. Pang, F. J. Cabrerizo, and E. Herrera-Viedma, "An automatic procedure to create fuzzy ontologies from users' opinions using sentiment analysis procedures and multi-granular fuzzy linguistic modelling methods," *Information Sciences*, vol. 476, pp. 222–238, 2019.

[14] B. Zhang, D. Xiong, J. Su, and H. Duan, "A context-aware recurrent encoder for neural machine translation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2424–2432, 2017.

[15] C. Park, Y. Yang, K. Park, and H. Lim, "Decoding strategies for improving low-resource machine translation," *Electronics*, vol. 9, no. 10, p. 1562, 2020.

[16] L. Zhao, A. Zhang, Y. Liu, and H. Fei, "Encoding multi-granularity structural information for joint Chinese word segmentation and POS tagging," *Pattern Recognition Letters*, vol. 138, pp. 163–169, 2020.

[17] Z. Zhang, H. Zhao, K. Ling et al., "Effective subword segmentation for text comprehension," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1664–1674, 2019.

[18] Y. Zhang, R. Barzilay, and T. Jaakkola, "Aspect-augmented adversarial networks for domain adaptation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 515–528, 2017.

[19] J. Wang, Z. Fu, M. Niu, P. Zhang, and Q. Zhang, "Multi-feedback pairwise ranking via adversarial training for recommender," *Chinese Journal of Electronics*, vol. 29, no. 4, pp. 615–622, 2020.

[20] R. Dabre, C. Chu, and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–38, 2020.

[21] L. H. Baniata, S. Park, and S.-B. Park, "A multitask-based neural machine translation model with part-of-speech tags integration for Arabic dialects," *Applied Sciences*, vol. 8, no. 12, p. 2502, 2018.

[22] M. R. Costa-jussà, "From feature to paradigm: deep learning in machine translation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 947–974, 2018.

[23] N. Pourdamghani and K. Knight, "Neighbors helping the poor: improving low-resource machine translation using related languages," *Machine Translation*, vol. 33, no. 3, pp. 239–258, 2019.

[24] Y. Liu, H. Li, and M. Wang, "Single image dehazing via large sky region segmentation and multiscale opening dark channel model," *IEEE Access*, vol. 5, pp. 8890–8903, 2017.

[25] Y. Liu, C. M. Vong, and P. K. Wong, "Extreme learning machine for huge hypotheses re-ranking in statistical machine translation," *Cognitive Computation*, vol. 9, no. 2, pp. 285–294, 2017.

[26] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, and J. Xie, "A hierarchy-to-sequence attentional neural machine translation model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 623–632, 2018.

[27] Z. Yang, W. Chen, F. Wang et al., "Unsupervised neural machine translation with weight sharing," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 46–55, Melbourne, Australia, July 2018.