

Received 5 April 2024, accepted 14 May 2024, date of publication 17 May 2024, date of current version 29 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402360

RESEARCH ARTICLE

A Semantic Context-Aware Automatic Quality Scoring Method for Machine Translation Based on Pretraining Language Model

FANGMIN TAN¹ AND HUAJU WANG² 

¹Huainan Vocational and Technical College, Huainan 232001, China

²School of Foreign Languages, Anhui University of Science and Technology, Huainan 232001, China

Corresponding author: Huaju Wang (2020600003@aust.edu.cn)

This work was supported in part by Anhui Provincial Quality Engineering Project, China, under Grant 2021kcszsfkc384 and Grant 2022jyxm1453; and in part by Huainan Vocational and Technical College Research Project, China, under Grant 2024HZSK001.


ABSTRACT Nowadays, machine translation has been a prevalent Internet application. But there still lacks mature intelligent algorithms to automatically evaluate quality of machine translation results. Considering the complexity inside machine intelligence-based semantic comprehension, we resort to pretraining language model (PLM) to deal with this challenge. Hence, this paper proposes a semantic context context-aware automatic quality scoring method for machine translation based on a specific PLM. The purpose of introducing the calculation of sentiment vectors in research is to consider emotional information in machine translation quality automatic scoring methods, in order to improve the accuracy and robustness of scoring. In particular, a novel PLM that combines multiple key features and tasks is established, which is utilized to make encoding towards largescale initial sentences and object sentences. It is finely tuned by integrating two typical pretraining structures. By applying the proposed PLM to complex semantic context and analysis tasks, we finally demonstrate its effectiveness through experiments on the News Crawl corpus and WMT dataset. The obtained results show that the proposal method has achieved significant improvements in various evaluation indicators, demonstrating its superiority in the quality evaluation of machine translation by perceiving semantic contexts. Through comparison experiments, efficiency of the proposal can be acknowledged.

INDEX TERMS Pretraining language model, semantic context, machine translation, quality evaluation, deep learning.

I. INTRODUCTION

With the continuous deepening of globalization and the rapid development of information technology [1], machine translation, as an important language communication tool, plays an increasingly important role in cross-cultural communication. However, although the development of deep learning technology has made significant progress in machine translation in recent years, there are still challenges in evaluating translation quality [2], [3]. Traditional manual evaluation methods are not only time-consuming and labor-intensive, but may also have subjective biases [4]. Therefore, there is an

urgent need for an automated evaluation method to accurately and efficiently evaluate the quality of machine translation. Machine translation quality assessment has always been a challenging task [5], [6]. Traditional manual evaluation methods are not only time-consuming and laborious, but also susceptible to subjective factors, making it difficult to achieve accurate and objective evaluation goals [7]. In order to overcome this problem, researchers have proposed various automatic quality evaluation methods for machine translation [8], among which the evaluation method based on perceptual semantic context has received much attention [9]. This method extracts semantic information from both the source and target languages, and models the contextual context to more accurately evaluate the quality of machine

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed .

translation [10]. However, existing scoring methods still have some limitations, such as insufficient understanding of semantic information and limited ability to handle complex contexts.

The aim of this study is to explore an automatic evaluation method for machine translation quality based on an integrated pre training model that perceives semantic contexts, in order to improve the accuracy and efficiency of machine translation evaluation. Semantic situational awareness refers to the ability to consider the context, background, and relevant information of a text or language when understanding it. In the field of machine translation, semantic situational awareness refers to the machine translation system's ability to accurately understand the meaning of the source language text and transform it into a target language text that matches the original text, while retaining the semantic context of the original text. This includes considering factors such as the relationships between words, phrases, sentences, and texts, contextual information, language habits, etc., to ensure that the translation results are not only accurate but also in line with human language communication habits and contexts.

Perceived semantic context plays a crucial role in machine translation, encompassing the semantic meaning of words, contextual information, and contextual relationships between texts. It is crucial for accurate understanding and translation of the source language text. Therefore, we will focus on how to comprehensively consider the factors of perceptual semantic context when evaluating the quality of machine translation, in order to improve the objectivity and accuracy of scoring. We first introduce the current research status and existing problems in the field of machine translation, and then elaborate in detail on the design and implementation of a machine translation quality automatic scoring method based on integrated pre training models, including key steps such as model architecture, feature extraction, and scoring calculation. Next, we validate the effectiveness and performance advantages of our proposed method through experiments, and compare and analyze it with traditional evaluation methods. The main contributions of the research are as following two points:

- 1) Research utilizes the semantic representation capabilities learned from large-scale corpora to map source language sentences and target language sentences into a shared semantic space, in order to better capture the semantic similarity between sentences.

- 2) Our method also considers the influence of context, and this study introduces the concept of perceptual semantic context, integrating human language perception and context understanding abilities into automatic scoring of machine translation, thereby improving the reliability of scoring results and providing new references for related research in the field of natural language processing.

With the rapid development of artificial intelligence, the research on automatic quality evaluation methods for machine translation will be of great significance. The research results of this article provide new ideas and

directions for further improving the quality automatic evaluation method of machine translation. In future work, we will continue to deepen our research, further expand the application of this method in other natural language processing tasks, and promote the continuous improvement of machine translation technology.

II. LITERATURE REVIEW

The research on automatic scoring methods of machine translation quality has attracted much attention. Especially with the development of deep learning technology, neural network-based methods have been widely used in this field. In my research, I focus on a method to evaluate the quality of machine translation for the perception of semantic situations, and realize the automatic scoring of translation quality with the help of an integrated pre-trained model. Under this research theme, I delve into the work of some relevant scholars. The research of Kahlon and Singh [11] has promoted the development of machine translation, which uses self-attention mechanisms to model long-distance dependencies and achieve impressive performance. The success of this model provides a solid foundation for subsequent research.

Zhou et al. [12] obtained rich language representations through large-scale unsupervised learning, and this model has made a breakthrough in understanding text semantics, providing a new way to improve the accuracy of machine translation quality evaluation. Kiros et al. [13] have proposed a method to evaluate the quality of machine translation using pre-trained language models. They used a pre-trained model to make semantic representations of the translated sentences, and then assessed the translation quality through similarity calculations. This method improves the accuracy of evaluation to some extent. Zhang et al. [14] explored an attention-mechanism-based approach to improve the effectiveness of machine translation quality evaluation. They propose an attention-based two-way matching network for comparing semantic similarities between source and target languages to more accurately assess translation quality.

Bai and An [15] proposed a method to integrate text information with other modal information (such as images, videos, etc.) for the quality evaluation of machine translation. By combining multiple information, translation quality can be evaluated more comprehensively, especially for translation tasks involving multimodal content. Shamsolmoali et al. [16] explored the application of transfer learning to quality evaluation of machine translation. By transferring knowledge between source and target languages, they improve the ability to evaluate accuracy in cross-language translation tasks, especially in resource-scarce language pairs. Nenkova et al. [17] proposed A method for quality evaluation of machine translation using knowledge graphs. By introducing semantic information in external knowledge base, text content can be better understood, thus improving the accuracy and robustness of translation quality evaluation.

Combining the research results of these scholars, I designed an integrated pre-trained model approach to

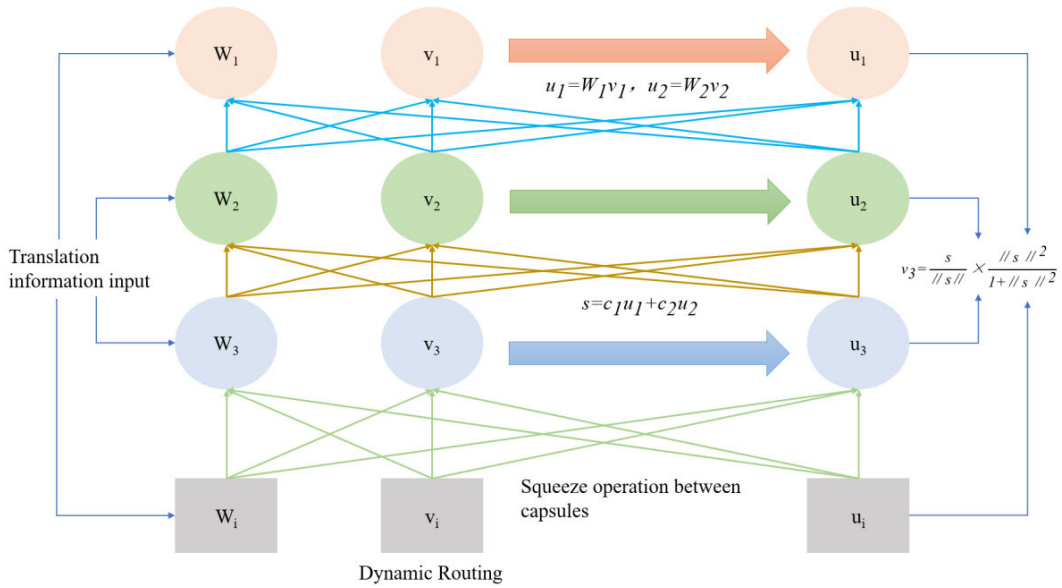


FIGURE 1. Emotional vector capsule network structure.

combine the semantic representations of multiple pre-trained models to capture semantic information about translation quality more comprehensively and accurately. Through the experimental verification on large-scale translation data, I demonstrate the superiority of this method in the task of quality evaluation of machine translation, and provide a new idea and method for improving the performance of machine translation system.

III. SEMANTIC AWARENESS IN MACHINE TRANSLATION A. EMOTIONAL VECTOR CAPSULE NETWORK

The Emotion-aware Capsule Network is a deep learning model based on the structure of the capsule network [18], designed to capture the emotional information of input text and integrate it into machine translation [19], [20]. In machine translation tasks, understanding the emotional information of the source language text is very important to accurately express the semantics and context [21]. Emotion vector capsule networks are expected to improve the performance and naturalness of machine translation systems by effectively learning and representing emotional content in text [22]. The network structure is shown in Figure 1.

The vector capsule network has a high degree of generalization ability for position and posture information, and the interaction between capsules is carried out through dynamic routing algorithms. In traditional neural networks, neurons output scalars, while the output of each capsule is a vector [23]. The squeezing operation formula between capsules is shown below.

$$u_1 = W_1v_1, u_2 = W_2v_2 \quad (1)$$

$$s = c_1u_1 + c_2u_2 \quad (2)$$

$$v_3 = \frac{s}{||s||} \times \frac{||s||^2}{1 + ||s||^2} \quad (3)$$

Among them, W_1 and W_2 are learnable parameter matrices, v_1 and v_2 are the vectors output by the low-level capsules, c_1 and c_2 are the weights of the low-level feature vectors, s is the weighted sum of u_1 and u_2 , and u_1 and u_2 are the vectors obtained by multiplying the learnable parameter W_i with the input vector v_i . The vector u_i encodes the relative positional relationships between the low-level features and the high-level features, which are contained in the learnable parameter matrix w_i . V_3 is the output vector calculated by the high-level capsule based on v_1 and v_2 . The modulus of vector v_3 represents the probability of the existence of higher-level features, and the direction of vector v_3 represents the pose information of the features [24]. The hyperparameters of the sentiment vector are set as shown in Table 1.

TABLE 1. Hyperparameter settings for sentiment vectors.

Hyperparameter	Value
optimizer	Adam
Learning rate	2e-6,8e-8
Learning decay rate	0,0.1
L2 regularization	0.01,1e-8
Number of learning rounds	5,6,8
Batch size	32,32,64
Gradient cropping	6,12,18

The next layer h is a convolutional capsule layer with multiple convolutional channels. The dimension of the capsule is M , and the size of the convolutional kernel is $(l-f+1) \times 1$. By increasing the size of the convolutional kernel, the perceived field of view of the model is expanded, and nonlinear squeezing is applied in the convolutional capsule layer h . The last layer of the model is the text capsule layer

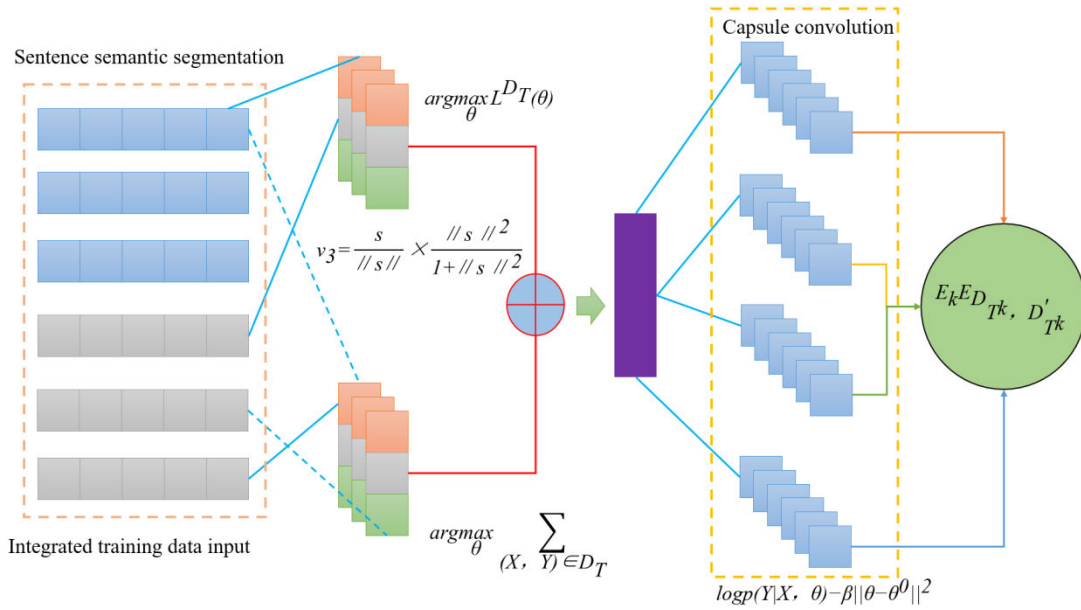


FIGURE 2. Specific language emotion parameter learning process.

v , and the information exchange between the convolutional capsule layer and the text capsule layer uses a static routing method.

B. PARAMETER LEARNING OF SPECIFIC LANGUAGE EMOTIONS

The parameter learning of specific language emotions involves the understanding and expression of emotions, and in machine translation, emotional information may have a significant impact on translation results [25]. For example, when translating from one language to another, if the source language contains emotionally rich vocabulary or sentences [26], the translation system needs to be able to accurately capture this emotional information and convert it into corresponding emotional expressions in the target language to maintain translation quality and emotional consistency.

Given any initial parameters θ^0 , the prior distribution of the corresponding neural machine translation model parameters satisfies an isotropic Gaussian distribution $\theta_i \sim N(\theta_i^0, \frac{1}{\beta})$, where $1/\beta$ Represents variance. According to the prior distribution, the learning process of a specific task is $Learn(D_T; \theta^0)$ is described as the logarithmic posterior probability that maximizes the model parameters of the given data D_T .

$$\begin{aligned} Learn(D_T; \theta^0) &= \text{arg max}_{\theta} L^{D_T}(\theta) \\ &= \text{arg max}_{\theta} \sum_{(X, Y) \in D_T} \log p(Y|X, \theta) - \beta \|\theta - \theta^0\|^2 \end{aligned} \quad (4)$$

In the above formula, the first term on the right side of the equation corresponds to the maximum likelihood criterion for

training neural machine translation models, and the second term is a constraint term, which prevents new learning of model parameters θ Stay away from initial parameters θ^0 , thereby alleviating overfitting problems in the absence of sufficient training data. Different languages may have different ways of expressing emotions, with some languages expressing emotions more directly, while others may be more implicit or euphemistic. Therefore, in machine translation, it is necessary to consider how to convert and adjust emotional expressions between different languages, in order to make the translation results more in line with the culture and expression habits of the target language.

By training specific high resource language pairs mentioned above, the specific high-level resource language referred to here refers to the parameter learning of specific language emotions, using data from the News Crawl corpus, the model can simulate translation scenarios for low resource tasks and find corresponding parameters.

$$\begin{aligned} L(\theta) &= E_k E_{D_{T^k}, D'_{T^k}} \left[\sum_{(X, Y) \in D'_{T^k}} \log p(Y|X; Learn(D_{T^k}; \theta)) \right] \end{aligned} \quad (5)$$

In the formula, k represents a meta learning scenario, where D_T and $D_{T'}$ follow a uniform distribution on dataset T . It can be seen that the SGD method is used to approximate the maximized meta learning objective function. For each meta learning scenario, randomly sample a high resource task T , and then sample two batches of training samples D_{T^k} and D'_{T^k} from the selected tasks. Use the first batch of training samples to train specific task model parameters, and the second batch of training samples to evaluate the model. At a learning rate of μ The process of updating specific task model

parameters is as follows.

$$\theta'_k = \text{Learn}(D_{T^k}; \theta) = \theta - \mu \nabla_{\theta} L^D \tau^k(\theta) \quad (6)$$

It is expected to obtain parameters from the first batch of training θ' k is used for the evaluation of the second batch dataset D'_{T^k} , and the calculated gradient ∇_{θ} is applied update meta parameters as meta gradients.

$$\theta \leftarrow \theta - \mu' \sum_k \nabla_{\theta} L^D \tau^k(\theta'_k) \quad (7)$$

In the above formula, μ' The meta learning rate. By using the above method, the learned initial parameters are in a relatively balanced state, not too close to the source task, and can quickly adapt to new tasks in a small number of learning steps. The specific language emotion parameter learning process is shown in Figure 2.

The parameter learning of specific language emotions plays an important role in semantic context perception in machine translation. By performing parameter learning in pretrained models and supervised learning or transfer learning, it is possible to better capture specific language emotional features in source language sentences and improve the effectiveness of automatic scoring methods for translation quality.

IV. AUTOMATIC SCORING FOR MACHINE TRANSLATION QUALITY BASED ON INTEGRATED PRETRAINING

A. PRETRAINING AUTOMATIC SCORING FRAMEWORK

The pretrained automatic scoring framework is a machine learning based method that evaluates the quality of machine translation by using pretrained models. The core idea of this framework is based on integrated pretraining models. Firstly, a large-scale bilingual corpus is used to pretrain a neural network model. The model obtains good semantic representation ability by learning the semantic relationship and translation features between the source and target languages. Then, during the evaluation phase, a pretrained model is used to encode the translated sentences to be evaluated and obtain their semantic representations.

The advantage of pretrained models is that they can extract knowledge from a large amount of annotated data and learn the general semantic representation of language. Due to the unsupervised training of the pretrained model, it can utilize a large number of unlabeled corpora for training, thus having significant advantages in terms of data. The results of the data integration training phase collected in this study are shown in Figure 3. In addition, pretrained models can greatly reduce the time and computational resources required to train machine translation models from scratch, thereby reducing the difficulty of developing new models. Pretrained machine translation models provide an effective way for semantic situational awareness. It can draw knowledge and experience from corpora, learn common semantic representations of source and target languages, and greatly improve the effectiveness of automatic scoring methods for translation quality. Pretrained machine translation models are also expected to

become the main direction of future machine translation research and application.

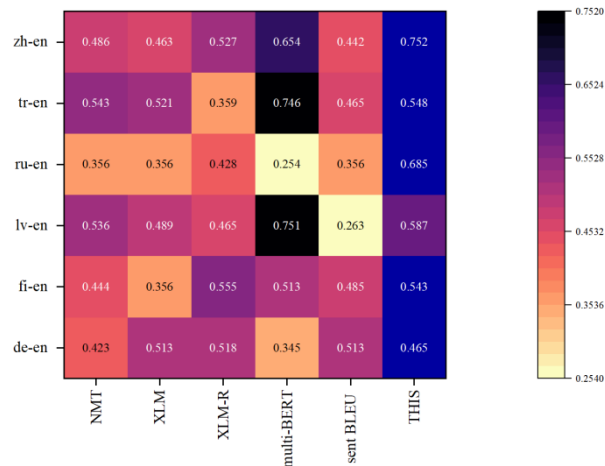


FIGURE 3. Integrated training phase achievements.

B. PRETRAINING DATA FUSION FOR AUTOMATIC TRANSLATION QUALITY EVALUATION

In the pretrained data fusion layer, we integrated the outputs of dependency syntactic parser, BERT encoder, and semantic role annotation parser to obtain context representations with syntactic and semantic awareness. Given $PR A = \{root_A, Pa 1, \dots, Pa m\}$. $PR B = \{root_B, Pb 1, \dots, Pb m\}$ and two vectors, which are the encoder outputs of the dependency syntactic parser model for two sentences, with lengths of $m + 1$ and $n + 1$, respectively. Due to each syntactic structure containing manually added root nodes, it is necessary to cut the encoder output of the dependency syntactic parser model to match the length of the original input sequence [28].

$$P_A = \{p_1^a, \dots, p_m^a\}, P_B = \{p_1^b, \dots, p_n^b\} \quad (8)$$

In the training of pretrained machine translation models, the fusion of pretrained data is also very important. This is because there are certain differences between different datasets in machine translation tasks. If only a single dataset is used for pretraining, this model may not perform well in some tasks. Therefore, a common solution is to fuse pretraining data between multiple datasets. By obtaining the cropped encoder output of the dependency syntax parsing model, we connect them based on their length:

$$P_{pair} = \{p_1^a, \dots, p_m^a, p_1^b, \dots, p_n^b\} \quad (9)$$

Given $B = \{o_1, \dots, o_l\}$ as the output of the BERT encoder, as BERT's output is at the sub word level. So we need to convert the final vector output by the dependency syntax parser to the sub word level, and finally obtain the syntactic aware context representation as $P = \{p_1, \dots, p_n\}$. The pre training model can capture rich semantic information through large-scale corpus learning, and fine tune on specific tasks to achieve better performance. In translation quality assessment,

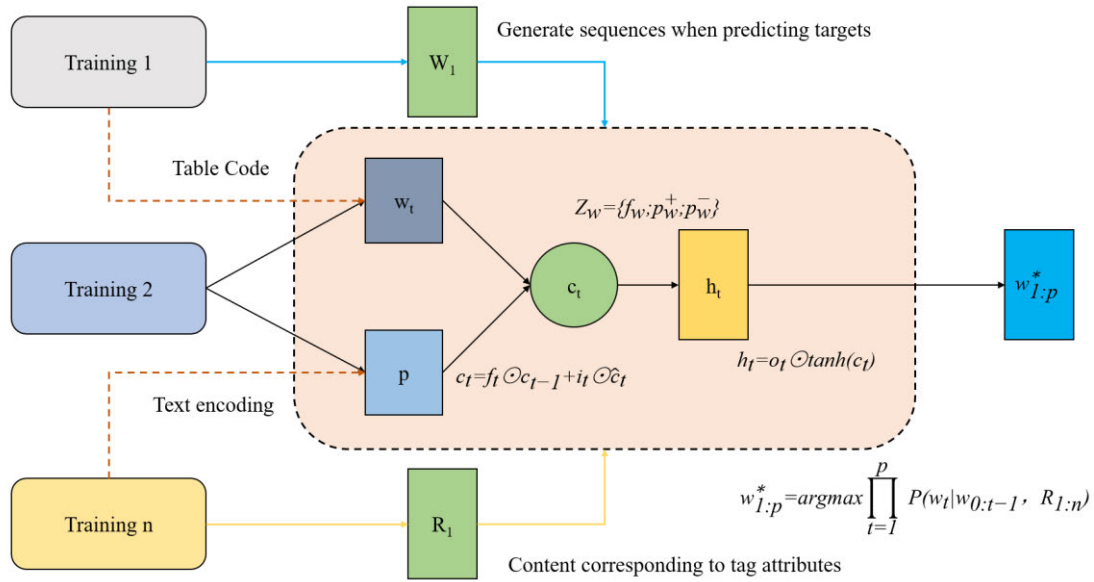


FIGURE 4. Integrated pretraining model running process.

the pre training model can more accurately assess the quality of translation by modeling the semantic relationship between the source language and the target language.

In the encoder section, this method utilizes bidirectional acquisition of deeper semantic features and uses pruning algorithms to obtain several candidate predicates and arguments from all options. The decoder section selects the highest score, which is calculated by predicate, argument, and predicate argument pairs. We maintained the same settings as the original model. The output of the last layer is represented as a hidden layer.

$$S = \{s_1, \dots, s_n\} = SRL\{x_1, \dots, x_n\} \quad (10)$$

In the formula, given an input sentence w_1, \dots, w_n , we obtain its word embedding vector $X_s = \{x_1, \dots, x_n\}$, where n is the length of the sentence. The final output of semantic role annotation is the hidden layer representation of semantics.

Given an input sentence, we first obtain its embedding vector representation, with the root node manually added. Then, the embedding vector is passed to obtain the hidden representation of the dependency syntax represented as P^R :

$$P^R = \{root, p_1, \dots, p_n\} = BiAffine(root_0, x_1, \dots, x_n) \quad (11)$$

In the pretraining stage, we can merge raw data from different datasets and adjust model parameters on these data to learn more universal language representations. In this way, the model can better address the challenges in specific tasks. Of course, the data fusion of pretrained models involves many related factors, such as data quality and data volume, which need to be analyzed in specific applications to customize the optimal strategy for actual situations. The integrated pretrained pseudocode is shown in Algorithm 1.

Algorithm 1 Integrated pretraining workflow

- 1: Input: The learnable parameter matrix W_i , the vector v_i output by the low-level capsule, the parameter variable i , the convolutional capsule layer h of the convolutional channel, the meta learning scene k , and the learning rate are μ .
- 2: Squeeze operation between capsules
- 3: **for all** $i = 1$ to n **do**
- 4: The effect of input on neurons eq-1
- 5: Calculate the fitness of each particle
- 6: **for** $W_i=1$: n
- 7: $u_1=W_1v_1, u_2=W_2v_2$
- 8: Calculated output vector
- 9: $s=c_1u_1+c_2u_2$
- 10: **if not** performing ideal enough in the task
- 11: Fusion of pretraining data
- 12: **else**
- 13: Final semantic role annotation
- 14: **end for**
- 15: **end for**

C. AUTOMATIC SCORING PARAMETER SETTING AND OPERATION PROCESS

Before evaluation, it is necessary to set some parameters for automatic scoring. The pretrained model parameters include the size of the hidden layer of the neural network, word vector dimension, learning rate, etc. The feature extraction parameters determine which features are extracted from the machine translation results, such as sentence length, word repetition rate, grammar error rate, etc. The score calculation parameters are used to calculate the final translation quality score, such as the weights and combination methods of different features.

A study is conducted on a given table T, which consists of n attribute value pair records $\{R_1, R_2, \dots, R_n\}$, each record R_i containing a word sequence $\{d_1, d_2, \dots, d_m\}$ and their

related domain attribute representations $\{Z_{d1}, Z_{d2}, \dots, Z_{dm}\}$. The output of the Z_{dm} model is to generate a text description S for table T , which consists of p characters $\{w_1, w_2, \dots, w_m\}$. We formalize the table to text generation process as a prediction of the probability model. When predicting the target, we generate a sequence $w^* 1:p$, which maximizes the probability $P(w_{1:p}|R_{1:n})$. The formula for generating $w^* 1:p$ is as follows.

$$w_{1:p}^* = \operatorname{argmax} \prod_{t=1}^p P(w_t | w_{0:t-1}, R_{1:n}) \quad (12)$$

The study utilized a machine translation model based on semantic context awareness for table to text generation tasks. Structured modeling is carried out to address the different hierarchical characteristics between tables and regular sequences, while improving the capture of various information in the table in terms of attention mechanism, in order to better generate descriptive text.

Attribute embedding marks the key points of each word in the corresponding content of an attribute by its corresponding attribute name and the position it appears in the table [29], [30]. We define the attribute corresponding to the word w as a triplet.

$$Z_w = \{f_w; p_w^+; p_w^-\} \quad (13)$$

Among them, $(p_w^+; p_w^-)$ represents the embedding representation of positional information, which respectively represents the position of word W in the record sequence from front to back and from back to front. The purpose of a table encoder is to use an LSTM encoder to encode each word d_j in the table along with its attribute embedding Z_{dj} into the hidden state representation h_j .

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \quad (14)$$

$$h_t = o_t \odot \tanh(c_t) \quad (15)$$

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{pmatrix} = \begin{pmatrix} \text{sigmoid} \\ \text{sigmoid} \\ \text{sigmoid} \\ \text{tanh} \end{pmatrix} W_{4n,2n}^c \begin{pmatrix} d_t \\ h_{t-1} \end{pmatrix} \quad (16)$$

where $i_t, f_t, o_t \in [0, 1]$ are the input gate, forget gate, and output gate, respectively, \hat{c}_t and c_t are the values of the suggested cell and the true cell at time t , and n is the size of the hidden layer. The running process of the integrated pretraining model is shown in Figure 4.

We embed semantic contextual information into the pre-trained model. Specifically, we introduce semantic contextual labels into the training data and embed these labels into the encoding layer of the pre-trained model, allowing the model to translate according to different semantic contexts. In this way, we can better adapt to different contexts and obtain more accurate translation results.

D. AUTOMATIC SCORING METHOD COMBINED WITH BERTSCORE

BERTSCORE is a scoring method based on the pre training model BERT (Bidirectional Encoder Representations from

Transformers) [31], which aims to measure translation quality by comparing the similarity between reference translation and machine translation. Compared with traditional indicators such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall Oriented Understudy for Gisting Evaluation), BERTSCORE pays more attention to the understanding of semantics and context, so it is more accurate in evaluating translation quality [32]. In my research, I took BERTSCORE as a part of the automatic scoring method, and combined the integration method of the pre training model. This integration method can include a combination of multiple different pre training models. The automatic scoring method combined with BERTSCORE is of great significance in the evaluation of machine translation quality in semantic context perception. It can not only improve the accuracy and reliability of scoring, but also better capture the semantic information of translated text, thus providing an important reference for the improvement and optimization of machine translation systems.

BERTSCORE can effectively alleviate the problem of semantic matching errors by introducing a context word embedding mechanism to calculate similarity. Secondly, it is aimed at addressing the issue of difficulty in capturing remote dependency relationships and punishing key sequence changes at the semantic level. For example, given a window of size 2, for a sentence ‘‘A because B’’ instead of ‘‘B because A’’, the word overlap based method will only slightly penalize the exchange problem of the causal clause, which is particularly evident in scenarios where A and B are both long and short phrases. BERTSCORE can effectively capture remote dependencies and word order information in sentences through the context word embedding mechanism. The calculation method of BERTSCORE can be divided into the following steps:

(1) For the given reference translation x and machine translation \hat{x} , Represent their word symbols by using a pretrained model BERT based on context embedding. This method can generate different vector representations for the same word in different statements based on the surrounding words that make up the target word in the context.

(2) Perform a soft similarity matching on the vector representations output by the pretrained model BERT for the reference translation and machine translation, rather than precise matching or heuristic matching similar to n-ary grammar matching. Calculate the reference translation word symbol x_i and the machine translation word symbol \hat{x}_j the cosine similarity of j is in the following form:

$$\cos(\theta) = \frac{x_i^T \hat{x}_j}{\|x_i\| \|\hat{x}_j\|} \quad (17)$$

(3) Based on the similarity matrix, a maximum similarity score is accumulated and normalized for the reference translation symbols and machine translation symbols. The accuracy, recall, and F1 values of BERTSCORE are obtained

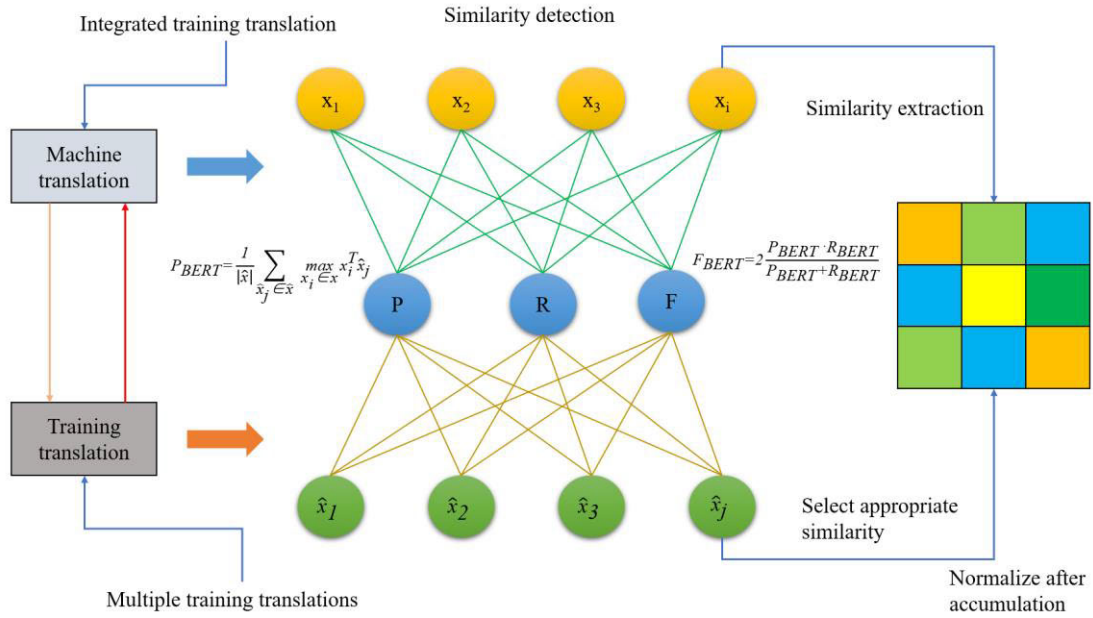


FIGURE 5. Schematic diagram of automatic evaluation calculation.

as follows:

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \quad (18)$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \quad (19)$$

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (20)$$

Previous studies on similarity matching have shown that rare words can better reflect sentence similarity than ordinary words. Therefore, the author also considers using the Inverse Document Frequency (IDF) function to assign different weights to different words, as shown in Figure 5.

Bilingual Evaluation Study (BLEU) is a commonly used evaluation metric for evaluating the quality of machine translation. Bilingual translation refers to the use of translation data between two languages for translation. Usually, we use source language sentences and corresponding target language sentences as translation data. Using this data, we can evaluate the quality of machine translation by calculating the BLEU (Bilingual Evaluation Understudy) score for translating source language sentences into target language sentences. The formula is as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (21)$$

BP is the penalty factor for being too short, which means the penalty coefficient based on when the translated sentence is shorter than the reference sentence. When translating long sentences, the BP penalty factor is not used, as N-grams already contain penalties for long sentences. Compared to traditional BLEU based evaluation methods, automatic evaluation methods based on bilingual translation

and BERTSCORE are more accurate and comprehensive in capturing semantic similarity and contextual information. This is of great significance for improving the translation quality and semantic accuracy of machine translation systems.

V. EXPERIMENTS AND ANALYSIS

During pretraining, multiple monolingual corpora are used, and self-supervised pretraining is achieved by randomly masking word elements in the input sequence. The study uses the News Crawl corpus and WMT dataset. The News Crawl corpus (also known as the News Commentary dataset) is a large-scale news article dataset that contains news content from multiple languages and topics. This dataset is commonly used for training machine translation because it contains rich texts from different fields and language pairs. Training with the News Crawl dataset can enhance the generalization ability of machine translation systems. WMT (Workshop on Machine Translation) is an important conference and evaluation task in the field of machine translation. The WMT dataset is part of the WMT evaluation task and contains parallel corpora of multiple language pairs. These datasets are typically created by expert manual translation or crowdsourced translation and are of high quality. The WMT dataset is widely used in the training, tuning, and evaluation of machine translation.

Both datasets provide rich corpus resources for machine translation researchers and developers, which can be used to build machine translation models, optimize models, and evaluate system performance. Meanwhile, due to the large size of the dataset, it can cover a variety of languages and

topics, which helps to improve the performance and accuracy of machine translation systems.

A. THE INFLUENCE OF NOISE LEVEL ON AUTOMATIC SCORING METHODS

The impact of noise on automatic scoring methods cannot be ignored. As the noise level increases, the quality of machine translation may decrease, which can lead to inaccuracies in automatic scoring methods. Due to the increase in noise level, the semantics of translation will be distorted, resulting in significant semantic differences between the evaluation results and the reference translation, thereby affecting the accuracy and stability of automatic scoring. The automatic scoring method based on integrated pretraining models can to some extent reduce the impact of noise on scoring. Pretrained models can learn rich semantic information from large-scale data and model the features of input sentences. Therefore, although noise levels may have a negative impact on the quality of machine translation, integrated pretrained models can improve the robustness and accuracy of scoring by considering the overall semantic context of sentences, as shown in Figure 6.

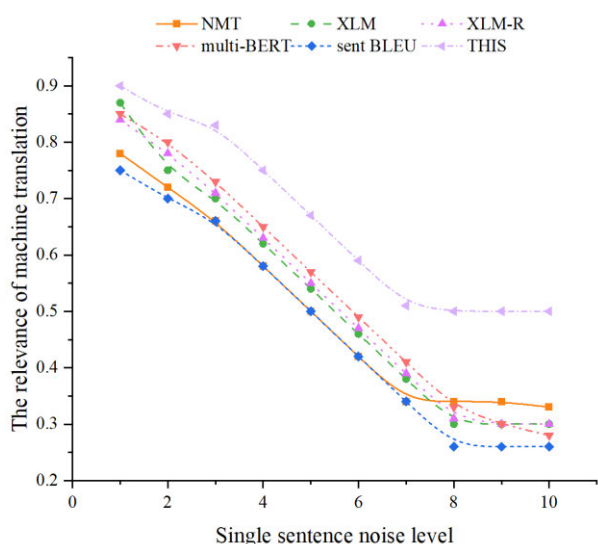


FIGURE 6. The influence of noise level on automatic scoring methods.

The results in Figure 6 show that the correlation of different machine translation models and their translation quality gradually decrease at different levels of correlation, which may reflect the performance changes of the models when dealing with translation tasks of different difficulty or complexity. The correlation of NMT model, XLM model, XLM-R model, multi BERT model, and send BLEU model gradually decreases from higher initial values to lower levels. This indicates a decrease in performance of these models when facing more challenging translation tasks. The method obtained from the study showed a more gradual decrease in correlation, and the final correlation value obtained was relatively stable. This means that the method can maintain

more stable performance when dealing with translation tasks of different difficulty levels.

B. COMPARISON OF ACCURACY OF METHODS WITH DIFFERENT NOISE LEVELS

In practical applications, translation quality assessment may face interference from various noise sources, such as spelling errors, grammar errors, ambiguity, etc. Therefore, when studying scoring methods, considering the impact of different noise levels on scoring accuracy can better simulate actual usage scenarios and improve the practicality and applicability of scoring methods. Comparing the accuracy of methods with different noise levels can help us understand the sensitivity of different scoring methods to different noise sources. This can further improve the accuracy, reliability, stability, and robustness of the scoring method, and enhance the reliability of the translation quality evaluation results. By comparing the accuracy of methods with different noise levels, the applicability and practicality of scoring methods can be evaluated, thereby expanding the scope of application of scoring methods. This comparison can also promote the improvement and optimization of scoring methods, improve the overall performance and efficiency of scoring methods, as shown in Figure 7 and Figure 8.

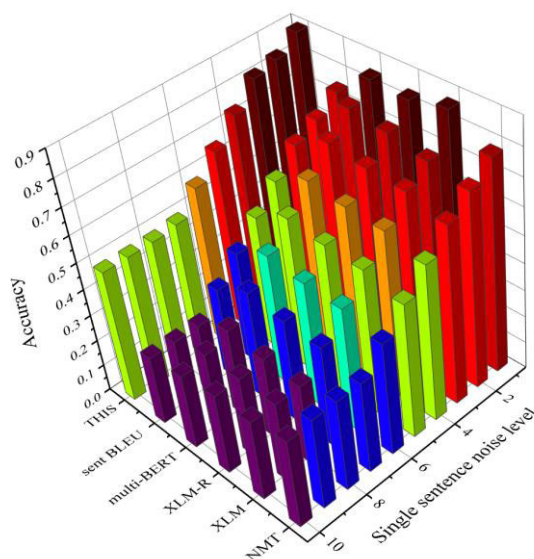


FIGURE 7. Comparison of accuracy of methods with different noise levels.

The results in Figures 7 and 8 show that the sentence level correlation of the NMT model is the lowest among all language pairs, and is below 0.5 in most language pairs. The performance of the XLM model has improved compared to the NMT model, but it is still relatively low overall, with no correlation score exceeding 0.6. The XLM-R model performs well in most language pairs, with correlation scores ranging from 0.4 to 0.6. In most language pairs, the correlation score of the multi BERT model is high, especially in lv en and tr en language pairs, with a score of 0.7 or above. The performance of the send BLEU model varies greatly in terms

of language pairs, with scores ranging from 0.3 to 0.7. The correlation score of this research model in most language pairs is between 0.4 and 0.7. Overall, different models exhibit certain differences in sentence level correlation across different language pairs. The perceptual semantic context machine translation ensemble pretrained model obtained from this study performs well in most language pairs. However, it should be noted that the performance of each model is also influenced by other factors, such as the quality and quantity of training data, as well as model tuning.

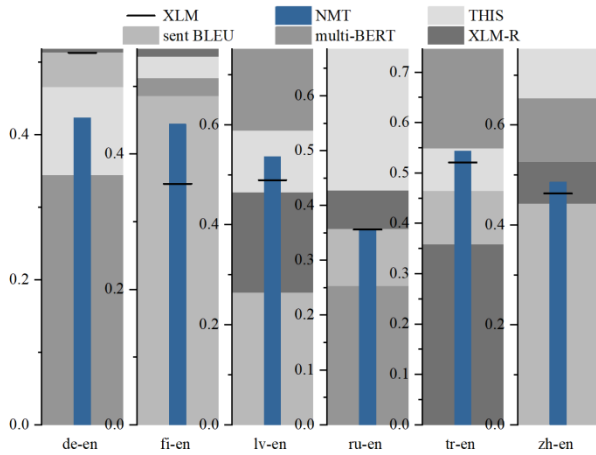


FIGURE 8. Comparison of section methods.

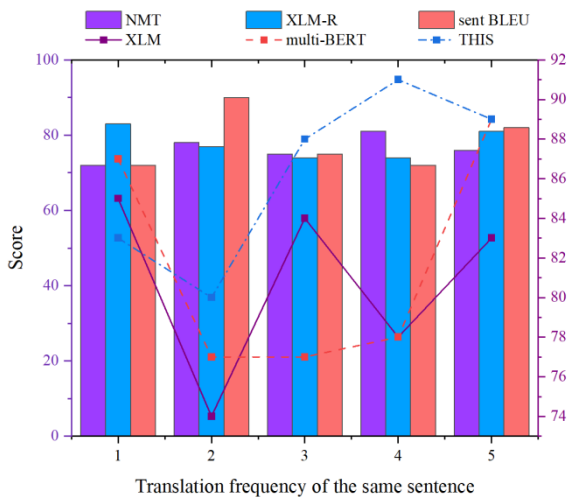


FIGURE 9. Comparison between the final score and the original meaning of the sentence.

C. COMPARISON BETWEEN THE FINAL SCORE AND THE ORIGINAL MEANING OF THE SENTENCE

By comparing the final score with the original meaning of the sentence, we can evaluate the effectiveness and accuracy of the automatic scoring method. If the final score is consistent with the comparison of the original meaning of the sentence, that is, the sentence with a higher score is closer in meaning to the original sentence, then we can consider

the automatic scoring method to be reliable and accurate. On the contrary, if there is a significant difference between the final score and the comparison of the original meaning of the sentence, we need to re-examine the effectiveness of the automatic scoring method and may take other improvement measures. By comparing with the original meaning of the sentence, we can verify the accuracy of the automatic scoring method in evaluating translation quality. If the score matches the original meaning of the statement, we have reason to believe in the reliability of the automatic scoring method. By comparing scores and the original meaning of sentences, we can understand which sentences are translated more accurately and which may have issues. This provides guidance and direction for improving translation quality, as shown in Figure 9.

Figure 9 shows that the machine translation score of the root NMT model is 85. This score represents the performance of the NMT model in machine translation tasks, with higher scores indicating better translation performance of the model. The machine translation score of XLM model is 74. Compared with the NMT model, the XLM model has a lower score, indicating that its translation performance may not be as good as the NMT model. The machine translation score of XLM-R model is 74. Compared with the NMT and XLM models, the XLM-R model also scores lower, which may indicate that its translation performance is not as good as the NMT model. The machine translation score of the multi BERT model is 77. Compared with the previous model, the score of the multi BERT model is lower, indicating that its translation performance may not be as good as the NMT model. The machine translation score of the send BLEU model is 90. This higher score may indicate that the model performs relatively well in machine translation tasks. The machine translation score of the pretrained model for perceptual semantic context machine translation ensemble obtained from this study is 80. Compared with the previous model, the pretrained model of perceptual semantic context machine translation ensemble obtained from this study generally has a slightly higher score and a smaller range of variation, making it more stable.

VI. CONCLUSION

A neural network model capable of perceiving semantic contexts was constructed by pretraining large-scale bilingual corpora. This model can learn the representation of semantic information and improve its performance through self-supervised learning. This step lays the foundation for the accuracy of subsequent scoring. In the scoring process, we adopted an integrated approach to combine the outputs of multiple pretrained models to obtain more comprehensive and accurate scoring results. This integration strategy can reduce the bias between individual models and improve overall robustness. Through comparative experiments with traditional natural language processing methods, the experimental results show that the method proposed in this study has significant advantages in automatic quality evaluation

tasks for machine translation. Our model can better capture semantic information, thereby obtaining more accurate results in the scoring process.

A perceptual semantic context scoring method based on integrated pretraining models has been proposed in this study, which has achieved significant results in automatic quality scoring tasks for machine translation. This study provides a more accurate and reliable evaluation method, which provides strong support for the improvement and application of machine translation. In the future, we will continue to conduct in-depth research to further enhance the application value of this method in the machine translation industry and expand its potential in other natural language processing tasks.

REFERENCES

- [1] Z. Guo, K. Yu, A. Jolfaei, G. Li, F. Ding, and A. Beheshti, "Mixed graph neural network-based fake news detection for sustainable vehicular social networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15486–15498, Dec. 2023.
- [2] I. Dunder, S. Seljan, and M. Pavlovski, "Automatic machine translation of poetry and a low-resource language pair," in *Proc. 43rd Int. Conf. Inf. Commun. Electron. Technol. (MIPRO)*, Sep. 2020, pp. 1034–1039.
- [3] F. M. Cabeceran, R. M. Costa-Jussà, and M. J. B. Acebal, "Linguistic-based evaluation criteria to identify statistical machine translation errors," in *Proc. 14th Annu. Conf. Eur. Assoc. Mach. Transl.*, 2010, pp. 167–173.
- [4] A. Zaretskaya, G. C. Pastor, and M. Seghiri, "Integration of machine translation in CAT tools: State of the art, evaluation and user attitudes," *Skase J. Transl. Interpretation*, vol. 8, no. 1, pp. 76–89, 2015.
- [5] E. Chatzikoumi, "How to evaluate machine translation: A review of automated and human metrics," *Natural Lang. Eng.*, vol. 26, no. 2, pp. 137–161, Mar. 2020.
- [6] M. S. Maučec and J. Brest, "Slavic languages in phrase-based statistical machine translation: A survey," *Artif. Intell. Rev.*, vol. 51, no. 1, pp. 77–117, Jan. 2019.
- [7] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks," *Int. J. Comput. Vis.*, vol. 128, no. 4, pp. 891–908, Apr. 2020.
- [8] J. Hutchins, "Machine translation: A concise history," *Comput. Aided Transl., Theory Pract.*, vol. 13, nos. 29–70, p. 11, 2007.
- [9] J. Moorkens, A. Toral, S. Castilho, and A. Way, "Translators' perceptions of literary post-editing using statistical and neural machine translation," *Transl. Spaces*, vol. 7, no. 2, pp. 240–262, Nov. 2018.
- [10] Y. Wang, F. Tian, and D. He, "Non-autoregressive machine translation with auxiliary regularization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 5377–5384.
- [11] N. K. Kahlon and W. Singh, "Machine translation from text to sign language: A systematic review," *Universal Access Inf. Soc.*, vol. 22, no. 1, pp. 1–35, Mar. 2023.
- [12] X. Zhou, Y. Li, and W. Liang, "CNN-RNN based intelligent recommendation for online medical pre-diagnosis support," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 3, pp. 912–921, May 2021.
- [13] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.
- [14] Z. Zhang, K. Chen, and R. Wang, "Neural machine translation with universal visual representation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–9.
- [15] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018.
- [16] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, and J. Yang, "Image synthesis with adversarial networks: a comprehensive survey and case studies," *Inf. Fusion*, vol. 72, pp. 126–146, Aug. 2021.
- [17] A. Nenkova, J. Chae, and A. Louis, "Structural features for predicting the linguistic quality of text: applications to machine translation, automatic summarization and human-authored text," in *Proc. Conf. Eur. Assoc. Comput. Linguistics*. Berlin, Germany: Springer, 2009, pp. 222–241.
- [18] C. Matuszek, D. Fox, and K. Koscher, "Following directions using statistical machine translation," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2010, pp. 251–258.
- [19] A. Alzu'bi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 20–54, Jul. 2015.
- [20] J. Zhang, X. Wang, H. Zhang, H. Sun, and X. Liu, "Retrieval-based neural source code summarization," in *Proc. IEEE/ACM 42nd Int. Conf. Softw. Eng. (ICSE)*, Oct. 2020, pp. 1385–1397.
- [21] J. Zhang, Y. Miao, and J. Yu, "A comprehensive survey on computational aesthetic evaluation of visual art images: Metrics and challenges," *IEEE Access*, vol. 9, pp. 77164–77187, 2021.
- [22] K. Bu, Y. Liu, and X. Ju, "Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning," *Knowl.-Based Syst.*, vol. 283, Jan. 2024, Art. no. 111148.
- [23] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating semantic models with (Genuine) similarity estimation," *Comput. Linguistics*, vol. 41, no. 4, pp. 665–695, Dec. 2015.
- [24] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4066–4079, Aug. 2018.
- [25] J. Hitschler, S. Schamoni, and S. Riezler, "Multimodal pivots for image caption translation," 2016, *arXiv:1601.03916*.
- [26] L. Zhou, J. Zhang, and C. Zong, "Synchronous bidirectional neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 91–105, Apr. 2019.
- [27] O. Caglayan, *Multimodal Machine Translation*. Le Mans, France: Le Mans Université, 2019.
- [28] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 6, p. 420, Nov. 2021.
- [29] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, Feb. 2016.
- [30] Y. Shen et al., "A deep learning-based data management scheme for intelligent control of wastewater treatment processes under resource-constrained IoT systems," *IEEE Internet Things J.*, early access, doi: 10.1109/JIOT.2024.3388043.
- [31] J. A. Evans and P. Aceves, "Machine translation: Mining text for social theory," *Annu. Rev. Sociology*, vol. 42, no. 1, pp. 21–50, Jul. 2016.
- [32] B. Wang, A. Wang, F. Chen, Y. Wang, and C.-C.-J. Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA Trans. Signal Inf. Process.*, vol. 8, no. 1, p. e19, 2019.



FANGMIN TAN received the bachelor's degree from Hunan University of Science and Technology, in 2003. She is currently an Associate Professor with Huainan Vocational and Technical College. Her research interests include applied linguistics, natural language processing, and text mining.



HUAJU WANG received the bachelor's degree from Anhui Normal University, in 2004, and the master's degree from Nanjing University, in 2011. Currently, she is the Vice Dean of the English Language Department, Anhui University of Science and Technology. Her research interests include semantic analysis, intelligent computing, and artificial intelligence.