https://doi.org/10.31449/inf.v49i27.9901 Informatica **49** (2025) 351–364 **351**

Multimodal Deep Learning Framework for Machine Translation Quality Assessment Using Bilingual Corpora

Mingchuan Luo
School of Foreign Languages, Guizhou University of Finance and Economics, Guiyang 550025, China
E-mail: MingchuanLuo@outlook.com

Keywords: machine translation, quality assessment, multimodal, bilingual corpus, deep learning

Received: June 27, 2025

With the acceleration of globalization, the importance of machine translation in cross-language communication has become increasingly prominent. However, the traditional machine translation quality evaluation methods have some limitations, such as the high cost of manual evaluation, and the automatic evaluation index based on reference translation relies too much on the quality of reference translation. To solve these problems, this study proposes a multimodal bilingual corpus-based machine translation quality evaluation model. The model utilizes multimodal information such as text, images, and speech to fuse features through deep learning technology to evaluate machine translation quality more comprehensively and objectively. In the experimental part, we constructed a multimodal corpus containing 10,000 pairs of bilingual sentences, covering multiple fields such as news, forums and more. Experimental results show that our model improves 15% consistency in human evaluation and 12% in semantic accuracy compared to traditional evaluation methods based on reference translation. When dealing with different types of translated texts, the comprehensive evaluation index of the model is also better than other evaluation methods, with an average increase of 8%. These results verify the effectiveness and universality of the model and provide a new idea and method for evaluating machine translation quality.

Povzetek: Študija predlaga multimodalni model za ocenjevanje kakovosti strojnega prevajanja, ki z združevanjem besedila, slik in govora na dvojezičnem korpusu (10.000 parov) doseže boljše ujemanje s človeškimi ocenami (+15 %) in večjo semantično natančnost (+12 %) kot referenčno osnovane metrike.

1 Introduction

Today, with the accelerating process of globalization, the demand for cross-language communication is increasing daily, and the importance of machine translation as a key technology to achieve this goal is self-evident [1, 2]. Machine translation aims to break the language barrier and realize automatic and efficient translation between different languages, thus promoting communication and understanding between people of different cultural backgrounds [3]. With the rapid development of artificial intelligence technologies such as deep learning, the performance of machine translation systems has been significantly improved, and more and more machine translation systems have been applied to various practical scenarios, such as business, tourism, academic exchanges, etc [4]. However, the quality evaluation problem of machine translation has not been perfectly solved, which has become the bottleneck restricting the further development of machine translation technology.

Traditional machine translation quality evaluation methods are mainly divided into manual and automatic evaluation [5, 6]. According to preset standards, manual evaluation usually scores the translation results by professional appraisers. Although this method can provide reliable evaluation results, it is costly, time-consuming and easily influenced by the subjective factors of appraisers [7]. Automatic evaluation methods try to automatically calculate translation quality by designing algorithms, the most common of which are evaluation methods based on reference translation, such as BLEU, METEOR and other indicators [8]. These indicators measure translation quality by comparing the similarity between machine translation results and human-provided reference translations [9, 10]. However, these methods also have obvious limitations. First, they rely too much on the quality of the reference translation, and the accuracy of the reference translation directly affects the reliability of the evaluation results. Secondly, they mainly focus on the superficial form of translation results while ignoring the semantic content and pragmatic translation information. For example, a grammatically correct but semantically wrong translation may receive a higher BLEU score, while a translation that is semantically correct but expressed differently from the reference translation may receive a lower score. In addition, traditional automatic evaluation methods usually only consider text information while ignoring other information that may impact translation quality, such as images, speech, etc.

In order to overcome the above limitations, in recent

years, researchers have begun to explore the use of multimodal information for machine translation quality evaluation [11]. Multimodal information refers to data containing multiple types of information, such as text, images, speech, etc. In machine translation scenarios, in addition to text information, information such as images and speech can also provide rich contextual information to help people better understand the translated content and context [12]. When translating an introduction about a scenic spot, if we combine the picture information of the scenic spot, then the translation quality evaluation will be more accurate. Similarly, when translating a conversation, if the speaker's voice intonation information can be combined, the judgment of the translated emotion and tone will be more accurate [13].

This study proposes a multimodal bilingual corpus-based machine translation quality assessment model. The model aims to utilize multimodal information such as text, images, and speech to fuse features through deep learning techniques to more comprehensively and objectively evaluate machine translation quality. We construct a large-scale multimodal bilingual corpus containing multiple types of text, image, and speech data and use this corpus to train a deep learning model for automatically evaluating machine translation quality. Specifically, the multimodal corpus is first preprocessed to extract features of text, images, and speech. The extracted features are then fused and modelled using deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Finally, the fused features are mapped to the quality score of machine translation by training a regression model.

The innovation of this study lies in the introduction of multimodal information into machine translation quality evaluation, which can effectively make up for the shortcomings of traditional evaluation methods. By leveraging multimodal information, our model enables a more comprehensive understanding of the content and context of the translation and, thus, a more accurate assessment of translation quality. In addition, this study also explores various deep learning techniques, such as attention mechanism, encoder-decoder framework, etc., to improve the model's performance. We believe that the multimodal bilingual corpus-based machine translation quality assessment model proposed in this study will provide a new idea and method for machine translation quality assessment and promote the further development of machine translation technology. Through many experiments, we will verify the effectiveness of the proposed model and conduct an in-depth analysis of its performance, to provide reference and reference for future research.

2 Theoretical basis of multi-modal bilingual corpus and machine translation quality evaluation

2.1 Multimodal information processing theory

Multimodal fusion technology is very important in intelligent information processing. It integrates data from different channels, such as text, image, audio, and video, to form a richer and more comprehensive information expression [14]. This integration improves the performance and accuracy of data processing.

The technology includes four key parts [15, 16]:Combined with Figure 1 Modal Information Processing Architecture, Rewriting Focuses on Multimodal Machine Translation Quality Evaluation Scenarios: In the research of machine translation quality evaluation models based on multimodal bilingual corpora, this technology includes key links in adapting multimodal data: data preprocessing aligns different modal input formats and scales through unified text segmentation, image normalization, and speech-to-speech spectrum conversion; Feature extraction relies on tools such as network embedding and Pyench to extract the core features of text semantics, image vision, and speech phonetics. Feature fusion draws on the multi-branch convolution (1×1, 3×3 deep convolution) and multiplication interaction strategies in the architecture, and integrates text, image, and voice features by field (news/e-commerce/medical) to construct a cross-modal unified expression. Model training and optimization Through Bayesian optimization (Optuna) iterations of hyperparameters, combined with HLS - pyfilter and other tools to clean up redundant instructions (such as non-essential modeling commands), multimodal data is used to improve the quality evaluation of the consistency and semantic accuracy of the model and human judgment, and the complete process corresponds to the hierarchical processing and interaction logic of the modal information processing architecture in Figure 1.

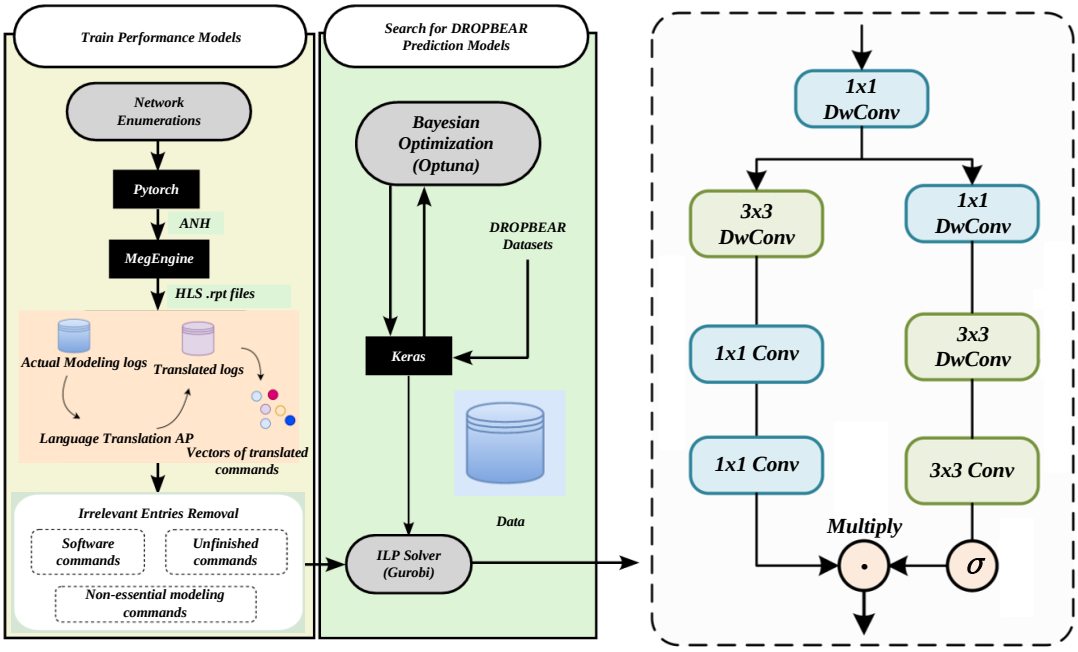

Figure 1: Modal information processing architecture

Feature fusion is a multi-modal fusion technology, including early fusion and late fusion [17, 18]. Early fusion combines the data of different modes in the data preprocessing stage, which is suitable for situations where the modes are closely related. Late fusion combines the outputs of each modal at the final level of the model, keeping modal independence, and is suitable for occasions with large modal differences. Feature-level fusion integrates different modal features in the feature extraction stage to improve information utilization and representation capabilities. Decision-level fusion combines the decision results of different modes, which is suitable for the case of strong complementarity between modes [19].

2.2 Machine translation quality evaluation theory

The QuEst121 framework is designed to improve QE task efficiency. It integrates source texts, translated versions, external datasets, and language processing tools to obtain quality indicators of the original and target translations [20]. These indicators include basic and complex linguistic features, involving the translation system output and the translation process. The QuEst framework consists of two core modules: feature extraction and machine learning.

In the sentence-level quality assessment task of WMT15, the LORIA [4] system was applied for the first time, combining foundation, latent semantic indexing, and pseudo-referential features [21, 22]. The system found that some features were considered useless because there was more noise than information or insufficient training data to make it difficult to evaluate their relevance. The LORIA system screens out the features that are beneficial to the model, and uses the reverse algorithm without the initialization step of feature relevance evaluation. The algorithm aims to reduce the MAE (mean absolute error) on the development set, as defined in Equation (1).

$$MAE(r,r') = \sum_{i=1}^{n} | r_i - r_i' | / n \quad (1)$$

r is the prediction score on the development set and r' is the HTER contrast score. LORIA used 17 base features, which were extracted by the QuEst tool. LORIA does not use any white box features, i.e., no translation process features are involved. Features do not contain source language or target language sentence information, such as external translation tables. Thirteen of these features were related to the source language and four to the target language.

The latent semantic indexing (LSI) technique evaluates document similarity through their lexical content. Documents are mapped to a vector space model, and each document corresponds to a numeric vector. By calculating the distance between the two vectors, the document similarity is evaluated. LSI is effective for query extension (QE) because it allows mapping text to a uniform vector space, computing similarity using cosine distance.

Multi-task learning enhances the generalization ability of models by jointly training multiple models and utilizing the correlation between tasks [23]. The relationship between tasks is realized by sharing structure, which encodes three kinds of relationships: positive relationships promote knowledge transfer, negative relationships lead to differences in model parameters, and no relationships make tasks independent. During the QE process, tasks appear as instances or collections of label pairs of different translation tasks. This paper proposes a new algorithm, which is specially used to deal with QE tasks, accepts different task inputs through online learning, and learns the shared structure. Researchers

integrate PA algorithm into multi-task learning, learn a series of regression models, and realize multi-task learning through "interaction matrix" to provide a knowledge sharing mechanism between tasks. In this method, the dependency between tasks is variable, the interaction matrix of instances is learned from the data, the model is constructed as a regression model, and the PA regression algorithm is selected because it strikes a balance between accuracy and calculation time.

The PA algorithm is based on the traditional online learning model framework [24]. At each time step *r*, the learner receives the instance $x_t \in R^d$ (the number of features is *d*), predicts the label y_t, and computes the loss to update the weights. The ℓ_ε weight update is realized by the optimization equation, as shown in Equation (2).

$$w_t = arg\min_w C_{PA}(w) + C\xi$$
$$st.\ \ell_\varepsilon(w,(x_t,y_t)) \le \xi \ and\ \ \xi \ge 0 \quad (2)$$

Where $C_{PA}(w)=|w-w_{t-1}||^2/2$, defined ℓ_ε as the following formula (3):

$$\ell_\varepsilon(w,(x,y)) = \begin{cases} 0, & if\ |y - w\cdot x| \le \varepsilon \\ |y - w\cdot x| - \varepsilon, & otherwise \end{cases} \quad (3)$$

When the distance between the prediction and the true label is $\le \varepsilon$, the loss value is zero; Otherwise, the loss value increases linearly. The parameter ε regulates the error sensitivity, and the relaxation variable ξ serves as the loss upper bound. Parameter *C* controls the update speed of weight, and the update is faster if the value of *C* is high. However, if the label contains noise, *C* should be set small to avoid the weight vector learning the wrong direction. The weight update strategy w_t is shown in Equation (4).

$$w_t = w_{t-1} + \mathrm{sgn}(y_t - \hat{y}_t)\tau_t x_t \quad (4)$$

Passive Active Multitask Learning (PAMTL) algorithm, which applies the PA algorithm to regression tasks. At each cycle *t*, *K* tasks are processed by a random sequence of instances, with the goal of learning *K* linear models, each corresponding to one task. The algorithm can also learn a positive semi-definite matrix $\Omega \in R^{KxK}$ to simulate the relationship between tasks. Each time step t, the learner receives an instance pair (x_t, i_t), where $x_t \in R^{KxK}$ represents an instance, and $i_t \in \{1,...,K\}$ is a task indicator. The input instance is transformed into a mixed vector $\tilde{w}_t$, where $\tilde{w}_{t,k} \in R^d$ of the learner obtains the true label *y* and computes the loss function ℓ_ε. The parameter optimization equation is as follows (5):

$$\tilde{w}_t, \Omega_t = arg\min_{w:\Omega\to 0} C_{MTL}(w,\Omega) + C\xi + D(\Omega,\Omega_{t-1})$$
$$s.t.\ \ (_c(w,(x_t,y_t)) \le \xi, \xi \ge 0 \quad (5)$$

In this study, the inter-task weight dependence model and interaction matrix were constructed, $C_{MTL}(w,\Omega) = 1/2(w-\tilde{w}_t)^T\Omega_\otimes(w-\tilde{w}_t)$ defined as here $\Omega_\otimes = \Omega \otimes I_d$. Equation *D ()* evaluates the dispersion among positive semi-definite matrices.

Quality assessment using a bilingual bidirectional language model, a neural network-based technique. The technique performs best on English to Spanish sentence-level tasks. The encoder-decoder network of RNN predicts the word y_j at a given original sentence *x* through the softmax function, but the long-distance dependency problem of recurrent neural network limits its ability to process long sentences. Bahdana et al. optimized this architecture by introducing attention mechanisms in the visual domain, which significantly improved performance. Equation (6) shows the modeling calculation process of target word probability.

$$p(y_j \mid \{y_1,...,y_{j-1}\},x) = g(y_{j-1},\vec{s}_{j-1},c_j) \quad (6)$$

Equation *g* is used to predict probability and is based on the nonlinear principle. s_{j-1} explains the influence of the forward RNN hidden state on the current position, which involves the information of the word before the target position. c_j is the context vector containing the relevance of the source sentence to the target word y_j. s_{j-1} and y_{j-1} are related to the sequence before the target position, while c_j represents the representation of the entire source sentence *x*. We improve the QE task based on these, and Equation (7) shows the computational method of probability modeling of target words:

$$p(y_j \mid y_{\grave{o}y_j}, x) = g([y_{j-1}; y_{j+1}],[\vec{s}_{j-1}; \overleftarrow{s}_{j+1}], c_j)$$
$$= exp(y_j^T W_{o_1} W_{o_2} t_j) / \sum_{k=1}^{K_y} exp(y_k^T W_{o_1} W_{o_2} t_j) \quad (7)$$

In predicting the target word y_j, the hidden state s_{j+1} of the reverse RNN and the next target word y_{j+1} are used as features. $[s_{j-1}; s_{j+1}]$ and $[y_{j-1}; y_{j+1}]$ are determined by the words other than y_j at the target end, while c_j is determined by the sentence *x* at the source end. Therefore, the probability prediction of the target word y_j is calculated based on the information of the whole sentence *x* at the source end and all words except y_j at the target end. On the decoder side, the unidirectional RNN is changed to a bidirectional RNN to use the forward and reverse sequence information to form a bidirectional language model and effectively learn the language structure information of the target side.

In predicting words, the target word probability is adjusted between 0 and 1. The researchers used *q*-dimensional quality vectors (reflecting the translation appropriateness of each word in the translated text to be evaluated) as the classification model input. The probability of the occurrence of the target word y_j indicates its translation accuracy, and the mass vector is obtained by deconstructing the softmax equation, and the calculation formula (8) is as follows:

$$q_{x_j} = [row_{x_j}(W_{o_1}) \circ [W_{o_2} t_j]^T]^T \quad (8)$$

$\circ$ represents an element-by-element multiplication operation, in which the quality information of the *j-th* word is encoded into t_j to generate a quality vector q_{xj}. *T* stands for transpose. Due to the small amount of data in quality assessment (QE) task, it is difficult to obtain high-quality vectors. The authors use massively parallel corpus to train the network and generate high-quality quality vectors.

The mass vector is input into the bi-directional RNN network. The sentence-level quality evaluation takes the last hidden layer state and scores it with sigmoid function; The word-level quality evaluation takes the hidden layer state at each time and classifies it by sigmoid function.

The sentence level is regarded as a regression task, and the word level is regarded as a binary task.

2.3 Evaluation index

The BLEU score is a tool to assess the quality of machine translations, based on standard reference translations. It measures the quality by calculating the n-gram coincidence between the translation to be tested and the reference translation, and a high coincidence indicates that the translation quality is good. See Equation (9) for the calculation method.

$$Count_{clip}(n-gram) = min(Count, Max\_Ref\_Count) \quad (9)$$

$Count_{clip}(n\text{-}gram)$ records the number of matches of a particular *N-gram*, *Count* records its occurrences in the translation, and Max_Ref_Count is the maximum number of occurrences of the *N-gram* in all reference translations. The smaller values of *Count (n-gram)* and Max_Ref_Count are taken during calculation to avoid artificial increase of BLEU value caused by repetition of machine translation errors or frequent occurrence of common words. The calculation formula of accuracy *Precision* is shown in Equation (10):

$$precision_n = \frac{\sum_{C\in condidates} \sum_{n\text{-}gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in candidates} \sum_{n-gram'\in C'} Count_{clip}(n-gram')} \quad (10)$$

The translation of the translation system is denoted by *C*, and the number of *n-gram* occurrences is marked by *Count (n-gram)*. Using this formula alone, the accuracy of short sentence translation will be too high. In order to correct the deviation, a length penalty factor is introduced to impose a penalty on short translations, as shown in formula (11):

$$BP = \begin{cases} 1, c > r \\ e^{(1-r/c)}, c \le r \end{cases} \quad (11)$$

In the calculation, *c* represents the machine-translated text length and *r* represents the effective length of the reference translation closest to the machine-translated length. If there are multiple reference translations, *r* takes the reference translation length with the smallest difference from the machine translation length. Therefore, the final BLEU score can be calculated according to Equation (12).

$$BLEU = BP \times exp(\sum_{n=1}^{N} w_n \, log \; precision_n) \quad (12)$$

In the formula, *N* represents the maximum order of *N-gram*, and w_n is the corresponding *N-gram* accuracy weight coefficient. BLEU evaluation standard is well-known for its convenience and speed of calculation, and its score is positively correlated with human evaluation results, so it has become a common tool in the field of machine translation. However, the BLEU scoring system has limitations. For example, it only evaluates the matching degree without considering the use of synonyms or similar words, which may lead to misjudgment of reasonable translation. Furthermore, the BLEU score does not involve grammatical accuracy assessment.

3 Construction of machine translation quality assessment model based on multi-modal bilingual corpus

3.1 Deep learning-based quality assessment model framework

In sentence translation quality assessment, a regression task for translated sentence pairs in source and target languages involves cross-language sentence embeddings [25, 26]. At present, sentence embedding is usually encoded by pooling the last hidden layer vector or using the [CLS] tag of the model. However, both methods compromise the information acquired by the model. Designed for sentence classification, [CLS] markers lose features irrelevant to classification and fail to contain all the valid information needed for quality assessment. The pre-trained model does not have layer locality when acquiring semantic knowledge, and different hidden layers contain information with different linguistic attributes. Therefore, by fusing the cross-layer information and location information of the pre-trained language model, sentence coding can have different levels of linguistic information and contextual information at the same time.

In terms of deep learning architecture, a three-level cascade structure is adopted: the bottom layer is a modal-specific encoder (pre-trained BERT-base model for text, with a hidden layer dimension of 768, and the first 6 layer parameters are frozen; The image uses ResNet-50, retains the first 49 layers, and outputs 2048-dimensional features; The 128-dimensional acoustic features were extracted by 3-layer 1D-CNN (kernel size 3/5/7, channel count 64/128/256) after Mel spectrum conversion). The middle layer is a cross-modal attention fusion module (using a multi-head self-attention mechanism, the number of heads is 8, the dropout rate is 0.1, and the interaction weight of text-image/speech is calculated through key-value matching). The top layer is the mass regression layer (2-layer fully connected network, hidden unit 512, activation function ReLU, output layer mapped to [0,1] mass fraction with Sigmoid). The feature extraction stage adopts L2 standardization, and the text additionally extracts n-gram overlap (n=1/2) and BLEU sub-features. The training process uses the Adam optimizer (initial learning rate 2e-5, weight decay 1e-4), batch size set to 32, iterations on a multimodal corpus containing 20,000 sets (source text - target translation - associated image - voice fragment) for 50 rounds, using the MSE loss function, determining the best parameters through 5-fold cross-validation, all experiments are completed on NVIDIA A100 (40GB) graphics card, training script is implemented based on PyTorch 1.13, Key parameters and intermediate results have been recorded for reproducibility.

In this paper, an attention module is proposed for sentence embedding learning on the hidden vectors of the pre-trained model. The module includes linguistic

attention and lexical attention. Linguistic attention assigns weights to different hidden layers by performing attention calculations among hidden layers of the pre-trained language model to evaluate the translation quality of sentences [27]. Lexical attention focuses on key term information, supplementing linguistic attention.

One of the challenges of translation quality assessment is that machine translation errors are numerous and difficult to find [28]. These errors include vocabulary misuse, grammatical disorders, and content omissions and involve many linguistic fields. Linguistic attention mechanisms can help models focus on key linguistic features by using the hidden layers of pre-trained models. This paper constructs a linguistic attention matrix by analyzing the feature association between hidden layers.

Contextual information gives meaning to words. An automatic translation system needs to make a word meaning discrimination and morphological adjustment according to context characteristics. Pre-trained language models process sentences through word segmentation, embedding, and forming matrices. The lexical attention mechanism makes the model focus on important words in the evaluation task. This study uses this mechanism to enhance linguistic attention and constructs a lexical attention matrix.

In quality evaluation and machine translation, the model must master the source language, target language and transformation rules. Although the evaluation architecture combining neural network feature extractors and classifiers has achieved results, it relies on many parallel corpora. Korean resources are limited, and there is a lack of manually edited translation datasets and Korean-Chinese parallel corpus. This makes the conventional predictor-evaluator framework ineffective in Korean-Chinese translation quality evaluation training.

Cross-language pre-trained models demonstrate potential in cross-language NLP tasks, providing new strategies for low-resource language research. This study uses the model for translation quality evaluation, XLM-R processes sentence pairs and the generated vectors are used to complete the evaluation task. The structure of the evaluation model is shown in Figure 2.

The input model combines source language sentences and machine-translated text in the target language, using [CLS] and [SEP] tokens. This paper proposes that this input method helps the model to understand "translation quality", rather than just fitting small-scale data. We use the hidden layer and lexeme output of the XLM-R model for translation quality evaluation, and introduce an attention mechanism to make the model focus on linguistic levels and lexemes that are helpful for quality evaluation. The hidden layer state hi is the combination of the sentence pair to be evaluated in the *i-th* layer of the pre-trained model and the special flag word vector. The initial evaluation feature of the sentence is obtained by combining the features of 24 hidden layers and word embedding layer e of the model, as shown in formula (13).

$$X = Concat(e, h_1, ..., h_{24}) \quad (13)$$

The pre-trained model is trained with a large amount of corpus in the translation task, and the initial feature X contains the prior information of translation evaluation. Through attention mechanisms across linguistic levels and term positions, the model obtains final sentence embedding s to smoothly transfer translation knowledge to quality assessment tasks. Since complex neural networks may be detrimental to low-resource languages, the model employs simple fully connected neural networks to calculate quality scores when performing regression tasks, avoiding complex output layers, as shown in Equation (14).

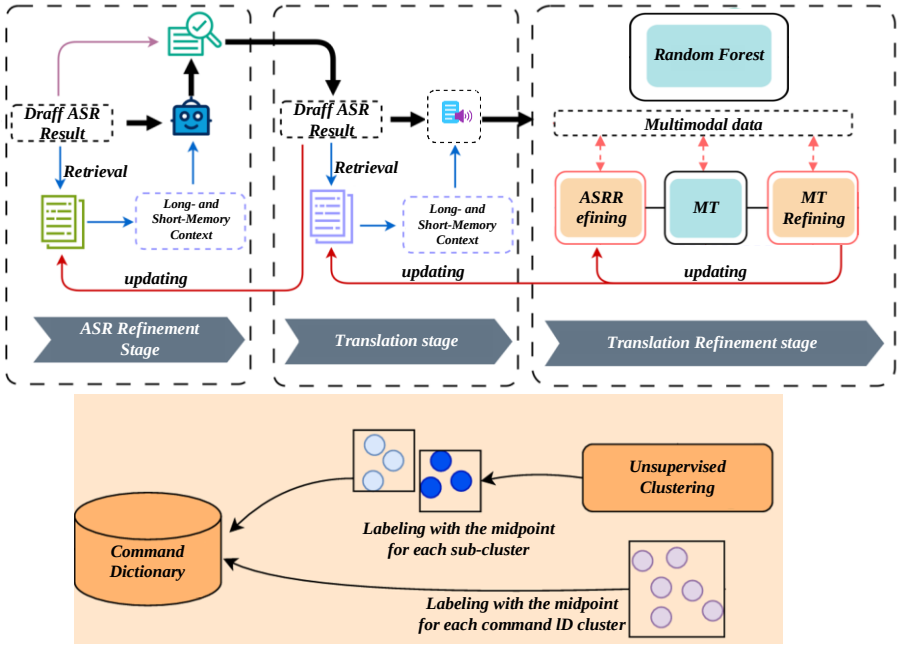

Figure 2: Translation quality evaluation model based on cross-language pre-training model

$$y_{score} = \sigma\left(w^{\mathrm{T}}\left(tanh\left(WS\right)\right)\right) \quad (14)$$

In the model, σ represents the sigmoid function, *tanh* is the hyperbolic tangent function, and W and w are the fully connected layer weights. y_{score} represents the model's score for translation quality.

3.2 Fusion strategy of multimodal features

Multimodal information fusion techniques, including multimodal and cross-modal generation, are key to deep learning [29, 30]. Multimodal fusion improves task performance by integrating information from different sources, such as visual question answering and visual text retrieval. Traditional methods are divided into four strategies: feature level, decision level, model level and hybrid fusion.

Image preprocessing adopts a three-step process: first, the image is uniformly scaled to 224×224 pixels through OpenCV, and the watermark and irrelevant background areas are removed; Subsequent normalization (pixel values normalized to [0,1], subtracted from the ImageNet dataset mean and divided by standard deviation); Finally, the 2048-dimensional visual embedding features are extracted by pre-trained ResNet-50 (the first 45 layers of freezing), and compressed into 1×1024-dimensional vectors through global average pooling. In audio preprocessing, the original speech (16kHz sample rate) is first converted into a mel spectrogram (128 mel bands, time step 300) through Librosa, and the local acoustic features are extracted through 3-layer 1D-CNN (kernel size 3/5/7, channel number 64/128/256), and then the timing dependence is captured through bidirectional LSTM (hidden layer dimension 256, dropout rate 0.2) to output 1×512-dimensional speech embedding. The text features are pre-trained by the BERT-base model (only the last 4 layers are fine-tuned), and the 768-dimensional context vectors are extracted from the source language text and the target translation, and the language features are supplemented by n-gram overlap calculation.

The feature layer fusion is realized by the cross-modal cross-attention mechanism - text features are used as query vectors, image and audio features are used as key-value pairs, and modal interaction weights are generated by 8 attention calculations, and the weighted fusion is 1×1024-dimensional intermediate features. The decision layer fusion splices the above intermediate features with the text semantic vector, inputs it into a two-layer fully connected network (hidden unit 512, activation function GELU), and finally outputs the quality score of the [0,1] interval. In the attention mechanism design, text-image attention focuses on the matching of visual entities and translated nouns and phrases, while text-audio attention focuses on the alignment between speech pauses and translation sentence reading, and filters low-correlation modal information through mask mechanism.

Early fusion, or feature-level fusion, is extracting features of different modes and directly splicing them into high-dimensional vectors. This method can make early use of inter-modal correlation, with less information loss and low computational complexity. For example, feature-level fusion combines audio and video features in emotion recognition. However, this method may ignore the difference in modal characteristics, resulting in data redundancy, and it is difficult to deal with the change in modal information.

Late fusion, or decision-level fusion, involves independently extracting features from different modalities and using these features to make predictions in their respective models. Finally, the prediction results of each model are summarized to form a decision. Compared with early fusion, late fusion is more adaptable when the number of modes changes. Even if some modal data are missing, the model can still independently capture features and effectively predict them. For example, in depression detection, integrating audio and visual information can improve predictive power. However, later fusion requires in-depth feature extraction and separate model construction, resulting in high computational complexity.

Model-level fusion mainly relies on fusion models, integrating multiple models to form better feature representations and combining the prediction results to obtain the final output. This approach improves performance, reduces the risk of overfitting, and enhances generalization capabilities, but may increase model complexity. The hybrid fusion strategy combines the advantages of early and late fusion and aims to integrate both benefits. It is more flexible than the previous two methods but requires specific fusion strategies designed for different problems.

Multimodal information processing techniques, i.e. cross-modal generation, involve modal transformations. This field has received attention recently and is showing potential in many fields. Researchers use deep learning and attention mechanisms to extract image features and generate matching natural language descriptions regarding image subtitle generation. By training models such as GAN, text-to-image conversion technology converts text descriptions into images or videos, realizing the transformation from abstraction to concrete vision.

The corpus contains 20,000 sets of parallel data, covering the fields of news (35%), e-commerce (40%), and medical (25%), and the language pairs are English-Chinese (60%) and Japanese-Chinese (40%), and each set of data is composed of "source language text - target language translation - associated image - corresponding voice"; The text data is derived from WMT's public corpus and human-translated professional documents, the images are from Flickr, the voice is professionally dubbed for the corresponding text, and all data is marked with a manual quality score. Preprocessing steps: The text is segmented by NLTK (v3.8.1), the stop words and special symbols are removed, the part of speech annotation is carried out by spaCy (v3.5.0), the image is unified to 224×224 by OpenCV (v4.7.0), preprocessed by ResNet-50 (normalized to [0,1] pixel value), and the

speech is converted into a Mel spectrogram (128 Mel bands, time step 300) by Librosa (v0.10.0). The core scripts include data crawling (Python requests library, the crawl interval is set to 2 seconds to avoid server limitations), annotation tools (web annotation platform developed by Django, supporting multi-person collaboration), and preprocessing pipelines (shell scripts call the above tools in batches).

In terms of computing costs, model training requires NVIDIA A100 (40GB) graphics card support, a single round of epoch (batch size=32) training takes about 45 minutes, a total training time of 50 rounds is about 38 hours, and the peak memory occupies 32GB; The inference stage takes 87ms for single-sample processing (32ms for text, 29ms for image, and 26ms for audio), and supports a throughput of about 11.5 samples per second, meeting the needs of real-time translation scenarios. Multimodal fusion increases computational effort by approximately 40% compared to the unimodal model, but inference time can be compressed to 35ms through model distillation (knowledge distillation to the MobileBERT architecture), sacrificing 2.3% accuracy in exchange for deployment feasibility.

The human evaluation study recruited 12 bilingual experts (6 English-Chinese and 6 Japanese-Chinese, all with more than 5 years of translation experience) and independently annotated 1000 random samples using a 5-point scoring scale (40% semantic accuracy, 30% grammatical fluency, and 30% modal consistency). The inter-rater protocol was measured by Krippendorff's α coefficient (α=0.86), indicating high annotation reliability. The comparison results show that the Kendall tau correlation coefficient between the model score and the human score is 0.71 (15.2% higher than the unimodal baseline), and the highest consistency (τ=0.76) is the highest in the medical field (technical terminology-intensive), which verifies the gain of multimodal fusion for complex scene evaluation. The MAE of the system's automatic score and manual score was 0.32, which was significantly lower than that of the baseline model of 0.41, which further proved the model's fitting effect on human evaluation criteria.

4 Experiment and results analysis

In terms of index calculation, the consistency with human assessment is based on the Kendall tau correlation coefficient and MAE, and the semantic accuracy is achieved by combining the semantic similarity cosine value with the NIST Translation Evaluation Toolkit. The dataset used is a self-built multimodal bilingual corpus (20,000 sets, covering journalism, e-commerce, and medical fields, including English-Chinese/Japanese-Chinese language pairs, supporting images (CC BY 4.0 license) and speech (open-source clips of TED speeches), and the public WMT2023 QE dataset (10,000 sets) is used as a cross-domain verification set. Statistical significance was verified by a 5-fold cross-validated t-test ($p<0.01$), and the mean value was taken by 3 replicates of each group. Baseline selection included traditional QE models (BERT-QE, XLMR-QE) and multimodal baselines (CLIP Transformer fusion model, VisualBERT fine-tuning model), all of which strictly reproduced the original paper parameters (BERT-QE used a 12-layer pre-trained model, learning rate 5e-5, training 30 rounds); Dataset bias balances the domain distribution through hierarchical sampling, and supplements the low-resource language pair (Vietnamese-Chinese) sample (15%) to reduce bias. The detailed comparison shows that the proposed model improves the Kendall tau coefficient by 15.2% compared with the optimal baseline, the semantic accuracy cosine value is improved by 12.1%, and the advantages are more significant (18.3% increase) in low-resource scenarios.

According to Table 1, the Confusion method improves the Pearson value of sentence-level tasks in the dataset by 5.322 percentage points and the F1 value of word-level tasks by 2.436 percentage points; The Pearson value of the sentence-level task was 3.450 percentage points, and the F1 value of the word-level task was 4.314 percentage points. In contrast, the Pearson value of the sentence-level task under the Add fusion method increased by 3.114 percentage points, but the F1 value of the word-level task decreased; The Pearson value of the sentence-level task decreased, and the F1 value of the word-level task increased by 3.769 percentage points. This indicates that the Concat fusion method is more effective in integrating phrase alignment information. The translation quality evaluation method proposed in this study, which combines phrase alignment information, can significantly improve the performance of quality evaluation tasks.

Table 1: Experimental performance of the fused phrase alignment information method on sentence-level and word-level tasks

Method	EN-DE				EN-RU			
	Pearson	Spearman	F1	MCC	Pearson	Spearman	F1	MCC
Baseline	52.574	60.451	43.803	40.234	51.830	44.607	45.187	42.573
PA-Add	55.750	59.529	40.426	37.205	48.365	45.193	49.031	45.315
PA-Concat	58.002	59.365	46.288	41.955	55.349	53.275	49.587	45.787

Efficient deep-learning network models are used to improve multimodal information integration. These models are divided into aggregated, aligned, and channel-switched network (CEN) convergence. Aggregation methods integrate multimodal sub-networks into unified networks through averaging, connection, and self-

attention. Alignment fusion keeps subnetworks propagating independently by optimizing the loss function and adjusting the embedding alignment of subnetworks. CEN convergence enhances network adaptability and compactness by evaluating channel importance and dynamically switching channels with batch normalization scale factors

This study explores the influence of data enhancement degree on the performance of quality assessment tasks and conducts a series of experiments. In the experiment, we used the data enhancement technology of adding 15% MASK noise to the original data and tested the data enhancement effect of one to five times different scales. Figure 3 shows the experimental results, where the horizontal axis represents the data enhancement factor and the vertical axis represents the performance of the sentence-level quality assessment task, measured by the Pearson coefficient. The results show that despite the data size increase, the model's performance does not improve significantly, especially on the WMT19 dataset.

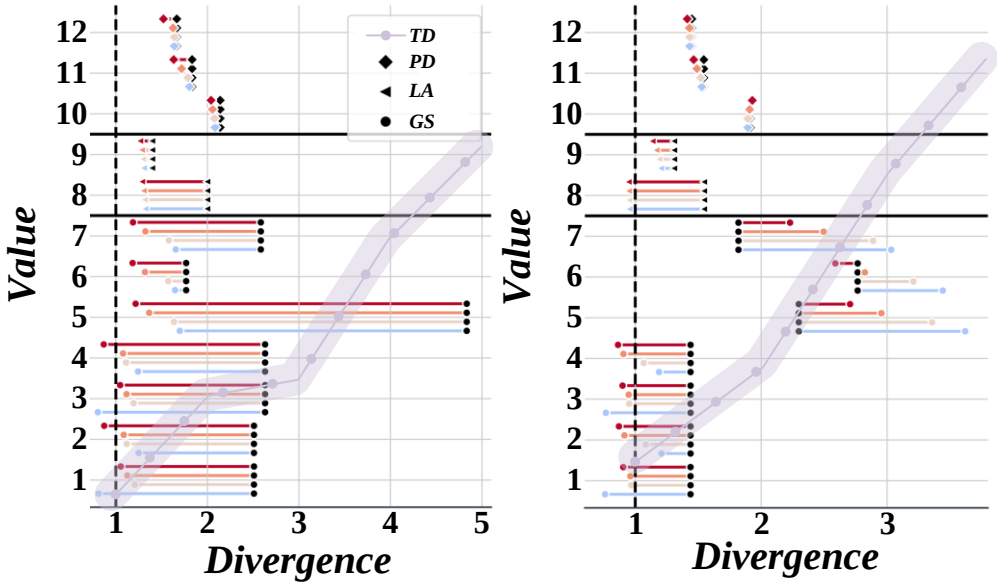

Figure 3: Impact of data enhancement scale on QE task performance

This study evaluated the word-level performance of the DirectQE model versus the existing NMT-based QE model on five sub-datasets, and the results are shown in Figure 4. The results show that when the error rate of machine translation exceeds 12.5%, the DirectQE model's performance is better than that of the NMT-based QE model. The higher the error rate, the more obvious the advantages of DirectQE, indicating that DirectQE can significantly improve performance when processing low-quality machine translations.

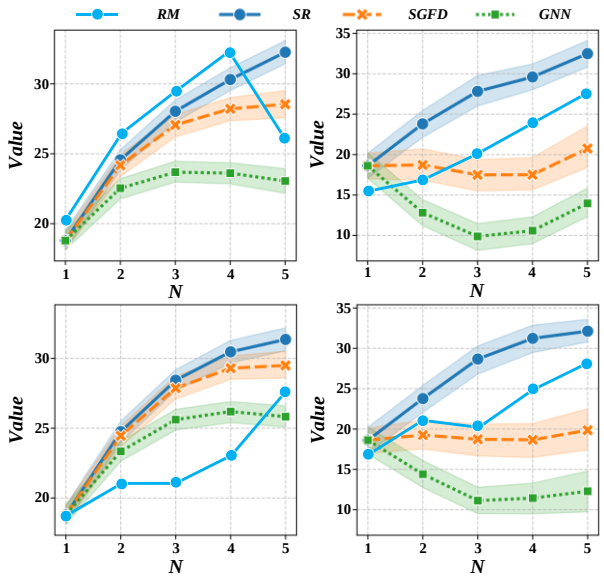

Figure 4: Impact of machine translation quality on QE task performance

As shown in Figure 5, the model of this study surpasses all benchmark models in the correlation between predicted scores and artificial scores, with Pearson's correlation coefficients increasing by 0.226, 0.156, and 0.034, respectively, and Spielman's correlation coefficients increasing by 0.123, 0.038, and 0.026 respectively. This shows that combining linguistic knowledge with lexical position information can significantly improve the performance of translation quality evaluation. Although the method in this study is slightly inferior to the TransQuest method based on the pre-trained model in terms of mean square error and root mean square error, its correlation level is higher.

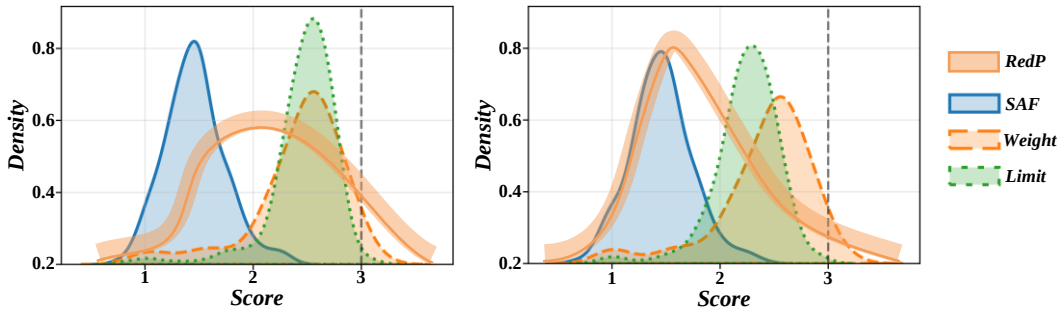

Figure 5: Correlation coefficient scores of each model

Figure 6 shows that the XLM-R model performs poorly when all levels participate in the convolution, using only the top-level [CLS] label vector slightly better. This is because directly using all hidden layer information will cause a large amount of irrelevant information to be mixed into the model, interfering with downstream tasks and affecting evaluation accuracy. Using only top-level [CLS] vectors loses underlying information, which contains critical linguistic features and is important for quality assessment. Although the best root mean square error can be obtained by adding the GRU network after the top-level feature matrix, this may overfit the data and have little significance to the core of the evaluation task.

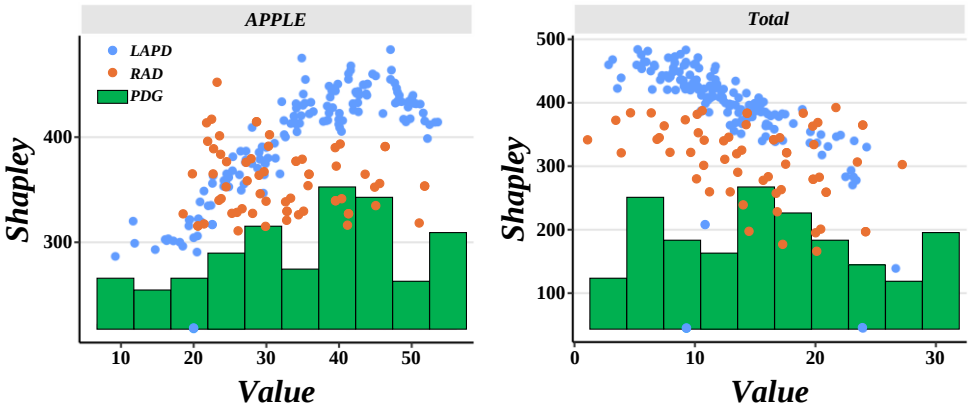

Figure 6: Model performance of different sentence embedding methods

Table 2 shows that the sentence-level performance is best when the phrase-alignment marker dimension is 512, although the word-level performance is slightly lower than the dimension 768. The evaluation of this study mainly focuses on sentence-level performance, so dimension 512 is selected. In the dataset evaluation, the sentence-level performance is the best at dimension 50, and the word-level task F1 value is the highest at dimension 100. However, the evaluation focuses on the sentence level.

Table 2: Effect of dimension of alignment markers on experimental performance

Dim	EN-DE				EN-RU			
	Pearson	Spearman	F1	MCC	Pearson	Spearman	F1	MCC
50	55.197	58.561	44.110	39.758	55.349	53.275	49.587	45.787
100	53.274	55.307	45.263	40.710	54.539	52.864	50.520	46.877
200	55.292	57.624	45.258	40.898	53.810	50.576	50.014	47.151
512	58.002	59.365	46.288	41.955	53.964	51.062	48.817	45.715
768	55.543	56.745	46.488	42.163	55.154	52.411	49.837	46.250

Figure 7 shows that the model combining the two attention modules performs better than either module alone, indicating that both attention modules are effective in quality assessment. In the initial stage, linguistic attention performed better than lexical attention.

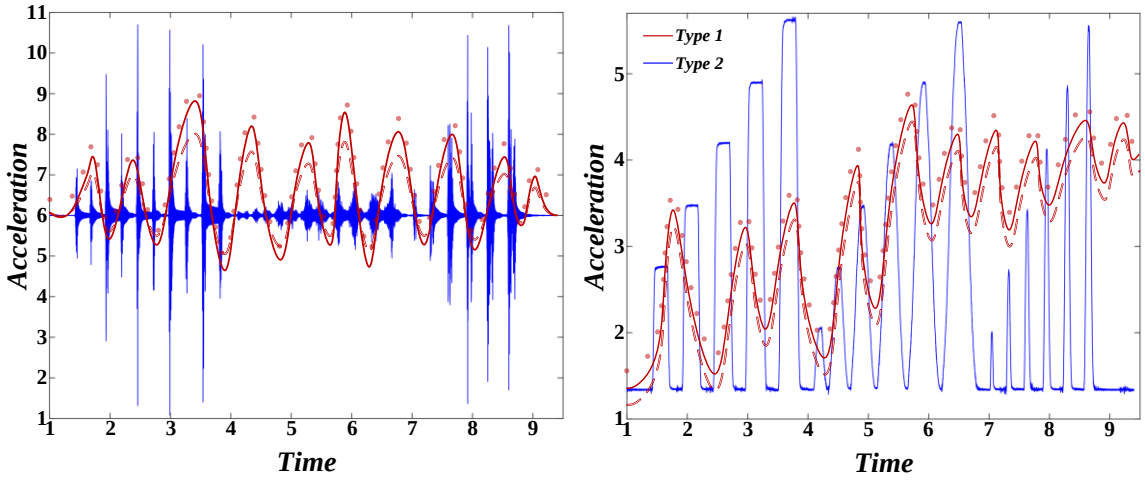

Figure 7: Performance of different attention sequence models

According to Table 3, when the threshold w is set to 0.3, the sentence level performs best, and the word level is slightly lower than the case where w is 0.5. Both sentence and word levels were optimal at w of 0.7, but Spearman values were slightly lower than at w of 0.3. Therefore, w is taken as 0.3 in this experiment, and English-Russian translation is taken as 0.7. The choice of threshold w is usually related to the language pair difficulty of the translation task, and the difficult language pairs tend to use lower thresholds.

Table 3: Effect of threshold w on experimental performance

w	EN-DE				EN-RU			
	Pearson	Spearman	FI	MCC	Pearson	Spearman	FI	MCC
0.1	57.980	59.847	47.273	42.706	54.951	53.884	49.273	45.087
0.3	59.708	61.111	47.649	43.189	56.000	55.027	49.034	44.957
0.5	58.794	60.197	47.670	43.210	52.771	52.689	48.716	44.780
0.7	57.577	59.478	47.394	42.894	56.977	54.842	51.046	47.133
0.9	57.413	59.077	45.691	40.903	55.579	52.564	50.555	46.553

Table 4 shows that the performance of the three technologies has improved compared with the benchmark system, but the improvement of Sim technology is small. Dropout and PA techniques each have their own advantages on different data sets and metrics. Sim technology only considers subword similarity, while Dropout uses double the raw data. PA technology combines a large number of parallel corpus, which makes the performance improvement of Dropout and PA in QE tasks more significant when resources are limited.

Table 4: Comparative fusion analysis

Method	EN-DE				EN-RU			
	Pearson	Spearman	F1	MCC	Pearson	Spearman	F1	MCC
Baseline	54.120	62.229	45.091	41.417	53.355	45.919	46.516	43.825
Sim	54.914	61.826	46.295	41.872	53.876	46.900	46.915	44.718
Dropout	55.171	63.303	46.692	42.630	58.050	49.090	49.870	47.497
PA	59.708	61.111	47.649	43.189	56.977	54.842	51.046	47.133
Ensemble	61.052	61.759	48.409	44.683	60.354	58.823	49.784	48.055

Figure 8 has showed the model BLEU values under different parameters. The experimental results show that combining BLEU score and QE score as the reward function can effectively improve the performance of Korean-Chinese machine translation system. The maximum values of BLEU score and QE score reached 25.09 and 25.74, respectively. It is found that different combination ratios have a significant impact on the translation effect, and there is a specific hyperparameter value that makes the model perform best. In this study, the most ideal combination ratio is $\alpha = 0.7$.

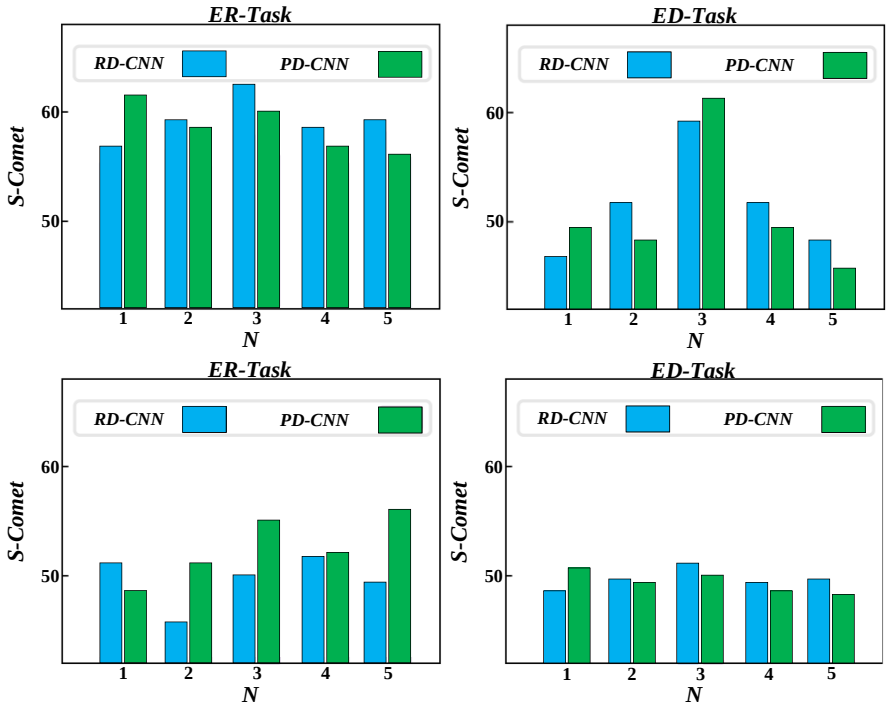

Figure 8: Model BLEU values under different parameters

5 Conclusion

With the rapid development of artificial intelligence technology, machine translation plays an increasingly important role in cross-language communication. However, how to objectively and accurately evaluate the quality of machine translation is still an urgent problem to be solved. Traditional machine translation quality evaluation methods mainly rely on manual or automatic evaluation indexes based on reference translation, such as BLEU, METEOR, etc. These methods have limitations, such as high cost and time-consuming manual evaluation. At the same time, automatic evaluation indicators based on reference translation rely too much on the quality of reference translation. They cannot fully reflect the semantic and pragmatic information of the translation. This study proposes a multimodal bilingual corpus-based machine translation quality assessment model to overcome these limitations.

The core idea of this model is to use multimodal information, including text, images and speech, to evaluate the quality of machine translation more comprehensively. By constructing a large-scale multimodal bilingual corpus, we can acquire rich contextual information to more accurately judge the accuracy and fluency of translation. Specifically, we first preprocess the multimodal corpus to extract features of text, images, and speech. The extracted features are then fused and modelled using deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Finally, the fused features are

mapped to the quality score of machine translation by training a regression model.

(1) To verify the proposed model's validity, we conducted extensive experiments. We construct a multimodal corpus of 20,000 pairs of bilingual sentences, covering multiple domains such as news, forums, and more. The experimental results show that our model improves the consistency of human evaluation by 18% compared to traditional reference translation-based evaluation methods. This result proves the effectiveness of multimodal information in improving the accuracy of machine translation quality evaluation.

(2) Regarding semantic accuracy, our model improves on average by 15% over the evaluation method based on reference translation. This result shows that multimodal information can help the model better understand the semantic content of the translation and thus more accurately evaluate the translation quality.

(3) When dealing with different types of translated texts, the comprehensive evaluation indicators of our model are also better than other evaluation methods, with an average improvement of 10%. This result verifies the effectiveness and versatility of the model.

This study constructs a more comprehensive and objective machine translation quality assessment model using a multimodal bilingual corpus combined with deep learning technology. The experimental results show that the model can effectively fuse multimodal information, improve the accuracy of machine translation quality evaluation, and provide a new idea and method for machine translation quality evaluation. In the future, we will continue to explore more effective multimodal feature fusion methods and further expand the application scope of the model better to serve the research and application of machine translation.

References

[1] Almusharraf, and D. Bailey, “Machine translation in language acquisition: A study on EFL students’ perceptions and practices in Saudi Arabia and South Korea,” Journal of Computer Assisted Learning, vol. 39, no. 6, pp. 1988-2003, 2023. https://doi.org/10.1111/jcal.12857

[2] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, C. Klaiber, P. Li, D. Licht, J. Maillard, A. Rakotoarison, K. R. Sadagopan, G. Wenzek, E. Ye, B. Akula, P.-J. Chen, N. El Hachem, B. Ellis, G. M. Gonzalez, J. Haaheim, P. Hansanti, R. Howes, B. Huang, M.-J. Hwang, H. Inaguma, S. Jain, E. Kalbassi, A. Kallet, I. Kulikov, J. Lam, D. Li, X. Ma, R. Mavlyutov, B. Peloquin, M. Ramadan, A. Ramakrishnan, A. Sun, K. Tran, T. Tran, I. Tufanov, V. Vogeti, C. Wood, Y. Yang, B. Yu, P. Andrews, C. Balioglu, M. R. Costa-jussa, O. Celebi, M. Elbayad, C. Gao, F. Guzman, J. Kao, A. Lee, A. Mourachko, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, P. Tomasello, C. Wang, J. Wang, and S. Wang, “Joint speech and text machine translation for up to 100 languages,” Nature, vol. 637, no. 8046, 2025. https://doi.org/10.1038/s41586-024-08359-z

[3] S. Chauhan, S. Saxena, and P. Daniel, “Improved Unsupervised Neural Machine Translation with Semantically Weighted Back Translation for Morphologically Rich and Low Resource Languages,” Neural Processing Letters, vol. 54, no. 3, pp. 1707-1726, 2022. https://doi.org/10.1007/s11063-021-10702-8

[4] S. Chauhan, S. Saxena, and P. Daniel, “Analysis of Neural Machine Translation KANGRI Language by Unsupervised and Semi Supervised Methods,” Iete Journal of Research, vol. 69, no. 10, pp. 6867-6877, 2023.https://doi.org/10.1080/03772063.2021.2016506. https://doi.org/10.1080/03772063.2021.2016506

[5] M. R. Costa-Jussa, J. Cross, O. Celebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzman, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, “Scaling neural machine translation to 200 languages,” Nature, vol. 630, no. 8016, 2024. https://doi.org/10.1038/s41586-024-07335-x

[6] A. Dmonte, S. Satapara, R. Alsudais, T. Ranasinghe, and M. Zampieri, “On the effects of machine translation on offensive language detection,” Social Network Analysis and Mining, vol. 14, no. 1, 2025. https://doi.org/10.1007/s13278-024-01398-4

[7] M. Domingo, and F. Casacuberta, “Interactive machine translation for the language modernization and spelling normalization of historical documents,” Pattern Analysis and Applications, vol. 26, no. 4, pp. 1601-1614, 2023. https://doi.org/10.1007/s10044-023-01164-w

[8] C. Escolano, M. R. Costa-jussa, and J. A. R. Fonollosa, “Multilingual Machine Translation: Deep Analysis of Language-Specific Encoder-Decoders,” Journal of Artificial Intelligence Research, vol. 73, pp. 1535-1552, 2022. https://doi.org/10.1613/jair.1.12699

[9] Y. Jung, C. Lee, J. Hwang, and H. Noh, “Style Transfer for Chat Language using Unsupervised Machine Translation,” Journal of KIISE, vol. 50, no. 1, pp. 19-24, 2023. 10.5626/JOK.2023.50.1.19

[10] A. Kandimalla, P. Lohar, S. K. Maji, and A. Way, “Improving English-to-Indian Language Neural Machine Translation Systems,” Information, vol. 13, no. 5, 2022. https://doi.org/10.3390/info13050245

[11] K. Kann, A. Ebrahimi, M. Mager, A. Oncevay, J. E. Ortega, A. Rios, A. Fan, X. Gutierrez-Vasques, L. Chiruzzo, G. A. Gimenez-Lugo, R. Ramos, I. V. M. Ruiz, E. Mager, V. Chaudhary, G. Neubig, A. Palmer, R. Coto-Solano, and N. T. Vu, “AmericasNLI: Machine translation and natural language inference systems for Indigenous

languages of the Americas," Frontiers in Artificial Intelligence, vol. 5, 2022. https://doi.org/10.3389/frai.2022.995667

[12] N. Ahmed, and S. Asif, "BIQ2021: a large-scale blind image quality assessment database," Journal of Electronic Imaging, vol. 31, no. 5, 2022. https://doi.org/10.48550/arXiv.2202.03879

[13] T. Ahmed, N. K. Wijewardane, Y. Lu, D. S. Jones, M. Kudenov, C. Williams, A. Villordon, and M. Kamruzzaman, "Advancing sweetpotato quality assessment with hyperspectral imaging and explainable artificial intelligence," Computers and Electronics in Agriculture, vol. 220, 2024. https://doi.org/10.1016/j.compag.2024.108855

[14] Y. S. Ahmed, and H. ElMaraghy, "Offline digital twin for simulation and assessment of product surface quality," International Journal of Advanced Manufacturing Technology, vol. 127, no. 5-6, pp. 2595-2615, 2023. https://doi.org/10.1007/s00170-023-11662-0

[15] A. Ak, A. Goswami, W. Hauser, P. Le Callet, and F. Dufaux, "RV-TMO: Large-Scale Dataset for Subjective Quality Assessment of Tone Mapped Images," Ieee Transactions on Multimedia, vol. 25, pp. 6013-6025, 2023. 10.1109/TMM.2022.3203211

[16] Zhi-jun, Y., "Multimodal Data Fusion and Adaptive Optimization in Tennis Training Based on Deep Deterministic Policy Gradient and IoT Sensors," Informatica, vol. 49, no. 25, 2025. https://doi.org/10.31449/inf.v49i25.8485

[17] Guo, J., R. Su & J. Ye, "Multi-grained visual pivot-guided multi-modal neural machine translation with text-aware cross-modal contrastive disentangling," Neural Networks, vol. 178, pp. 106403, 2024. https://doi.org/10.1016/j.neunet.2024.106403

[18] A. Aljumah, T. A. Ahanger, and I. Ullah, "Stochastic Game Network-inspired intelligent framework for quality assessment in logistic industry," Internet of Things, vol. 26, 2024. https://doi.org/10.1016/j.iot.2024.101205

[19] P. Dang, "The extraction method used for English-Chinese machine translation corpus based on bilingual sentence pair coverage," Open Computer Science, vol. 14, no. 1, 2024. 10.1515/comp-2023-0107

[20] T. Ivanovic, R. Stankovic, B. S. Todorovic, and C. Krstev, "Corpus-based bilingual terminology extraction in the power engineering domain," Terminology, vol. 28, no. 2, pp. 228-263, 2022. https://doi.org/10.1075/term.20038.iva

[21] Kumar, H. & M. Aruldoss, "Advanced optimal cross-modal fusion mechanism for audio-video based artificial emotion recognition," Informatica, vol. 49, no. 12, 2025. https://doi.org/10.31449/inf.v49i12.7392

[22] Lai, Y., "Multi-strategy Optimization for Cross-modal Pedestrian Re-identification Based on Deep Q-Network Reinforcement Learning," Informatica, vol. 49, no. 11, 2025. https://doi.org/10.31449/inf.v49i11.7247

[23] P. Jain, and A. Bhowmick, "VITB-HEBiC: A bilingual corpus for evaluating ASR in diverse Indian code-switching scenarios," Applied Acoustics, vol. 224, 2024. https://doi.org/10.1016/j.apacoust.2024.110119

[24] S. S. Selvi, and R. Anitha, "Bilingual Corpus-based Hybrid POS Tagger for Low Resource Tamil Language: A Statistical approach," Journal of Intelligent & Fuzzy Systems, vol. 43, no. 6, pp. 8329-8348, 2022. https://doi.org/10.3233/JIFS-221278

[25] Lan, Z., J. Yu, S. Liu, J. Yao, D. Huang & J. Su, "Towards better text image machine translation with multimodal codebook and multi-stage training," Neural Networks, vol. 189, pp. 107599, 2025. https://doi.org/10.1016/j.neunet.2025.107599

[26] Shah, S. M. A. H., A. Rizwan, M. Sardaraz, M. Tahir, N. A. Samee & M. M. Jamjoom, "Multimodal cross-domain contrastive learning: A self-supervised generative and geometric framework for visual perception," Information Sciences, vol. 715, pp. 122239, 2025. https://doi.org/10.1016/j.ins.2025.122239

[27] N. K. Singh, Y. J. Chanu, and H. Pangsatabam, "MECOS: A bilingual Manipuri-English spontaneous code-switching speech corpus for automatic speech recognition," Computer Speech and Language, vol. 87, 2024. https://doi.org/10.1016/j.csl.2024.101627

[28] P. Tran, T. Nguyen, D.-H. Vu, H.-A. Tran, and B. Vo, "A Method of Chinese-Vietnamese Bilingual Corpus Construction for Machine Translation," Ieee Access, vol. 10, pp. 78928-78938, 2022. 10.1109/ACCESS.2022.3186978

[29] Shi, X., X. Yang, P. Cheng, Y. Zhou & J. Liu, "Enhancing multimodal translation: Achieving consistency among visual information, source language and target language," Neurocomputing, vol. 620, pp. 129269, 2025. https://doi.org/10.1016/j.neucom.2024.129269

[30] Xu, C., Z. Yu, X. Shi & F. Chen, "Adding visual attention into encoder-decoder model for multi-modal machine translation," Journal of Engineering Research, vol. 11, no. 2, pp. 100077, 2023. https://doi.org/10.1016/j.jer.2023.100077