Automatic Assessment Model for Chinese-English Scientific Translation Quality Based on Contrastive Learning

Rongbing Hu
School Office (International Office), Zhejiang Business College, Hangzhou 310053, Zhejiang, China
Corresponding author: Rongbing Hu (e-mail: rongbing0607@sina.com)

ABSTRACT Assessing the quality of scientific literature translation remains challenging because of strong subjectivity, dense domain-specific terminology, and the limited availability of standardized reference translations. These issues are particularly relevant for the international dissemination of research in advanced electromagnetic engineering, where precise multilingual communication supports the reliable exchange of knowledge on electromagnetic waves, antennas, and propagation technologies. This paper proposes a Contrastive Learning-based Chinese-English Scientific Translation Quality Evaluation model (C-TQE). By constructing multi-level positive and negative sample pairs, the model learns the relative ordinal relationships of translation quality within a shared representation space. A dual-encoder architecture encodes source sentences and candidate translations through a shared pre-trained language model, while a contrastive loss function draws high-quality translations closer to the source representation and separates low-quality ones. To address the characteristics of scientific texts, a term-aware negative sampling strategy exploits domain dictionaries and syntactic structures to generate semantically similar but terminologically incorrect examples. Experiments on 11, 238 human-annotated instances from the WMT20–22 Chinese-English scientific translation tasks show that C-TQE achieves a Kendall's tau correlation coefficient of 0.564 with human judgments, outperforming COMET (0.512) and BLEURT (0.497). Ablation studies confirm the effectiveness of term-aware negative sampling and the contrastive learning objective, while diagnostic analysis demonstrates high consistency in evaluating terminological accuracy and syntactic structures. The proposed framework provides an effective solution for large-scale scientific translation quality assessment and facilitates the accurate international communication of multidisciplinary engineering research, including electromagnetic and antenna-related studies.

INDEX TERMS contrastive learning, translation quality evaluation, scientific translation, electromagnetic scientific communication, Chinese-English machine translation

I. INTRODUCTION

TRANSLATION quality evaluation plays a pivotal role in the advancement of machine translation systems, multilingual content generation, and cross-lingual information retrieval within the global industry, where the precise dissemination of technical innovations in fiber science and high-performance material engineering is essential. With the widespread application of neural machine translation technologies, how to automatically, accurately, and interpretably assess translation quality has become a focal point of concern for both academia and industry. Specifically, for professional texts such as scientific literature, which feature high terminology density, complex syntactic structures, and stringent requirements for semantic precision, evaluation methods face even greater challenges.

Traditional automatic evaluation metrics, such as BLEU and TER, which are based on n-gram matching, offer simple computation and strong interpretability. However, they rely excessively on reference translations and struggle to capture semantic equivalence. Researchers early on discovered that these metrics perform poorly when reference translations are absent or of low quality. To address this issue, Reference-free Quality Estimation (QE) methods emerged, attempting to predict quality scores directly based on the source sentence and the candidate translation. Early QE methods mostly employed feature engineering combined with regression models, later gradually transitioning to end-to-end architectures based on pre-trained language models. Recent research trends indicate that neural network-based evaluation metrics are surpassing traditional methods. Results from the WMT series of evaluation tasks have repeatedly confirmed this: from COMET to BLEURT, these neural metrics show significantly higher correlation with human judgments than n-gram metrics. The latest research by Gu Shuhao

et al. [2] demonstrates that fine-tuning pre-trained models via contrastive learning can further enhance the evaluation performance of BERTScore on low-resource languages, providing strong support for the application of contrastive learning in evaluation tasks. However, existing methods still share several common limitations. First, most models adopt a point estimation approach to predict absolute quality scores, whereas human judgment of translation quality is inherently relative; people find it easier to judge whether "Translation A is better than Translation B" rather than assigning an absolute score. Second, existing models handle negative samples relatively coarsely, typically using random sampling or simple word substitution, making it difficult to simulate real translation errors that are semantically similar yet of poor quality. Third, terminological mistranslation is a common error type in scientific translation, yet current evaluation models lack targeted modeling mechanisms for this issue. As a representation learning paradigm, contrastive learning has achieved significant success in computer vision and natural language processing. Its core idea is to pull the representations of similar samples (positive examples) closer together while pushing apart those of dissimilar samples (negative examples), thereby forming a meaningful semantic structure in the representation space. This approach aligns naturally with the problem of translation quality evaluation. We aim for high-quality translations to be closer to the source sentence in the representation space, while low-quality translations remain distant. Inspired by this, Wang et al. proposed the ContrastScore framework, which evaluates generation quality by comparing token-level probability differences between strong and weak models. Qu et al. introduced negative sample mining into non-autoregressive machine translation training, effectively alleviating the multimodality problem. These works provide a methodological foundation for introducing contrastive learning into translation evaluation [3]. Schwenk et al. integrated contrastive learning into multilingual representation distillation for quality estimation of parallel sentence pairs, validating the method's effectiveness in cross-lingual similarity search tasks. The goal of this paper is to construct a contrastive learning evaluation model specifically tailored for Chinese-English scientific translation. The specificity of scientific texts manifests at multiple levels: terminology translation must be accurate and consistent, syntactic structures are often complexly nested, and logical relationships must be conveyed clearly. This requires the evaluation model not only to understand semantic equivalence but also to identify specific error types such as terminological errors and improper word order. Inspired by the idea of COMET-poly [4], which utilizes multiple candidate translations for comparative evaluation, we further expand the object of comparison from "other candidate translations" to "negative samples generated through terminological interference," enabling the model to more precisely capture typical error patterns in scientific translation.

It is worth mentioning that domestic scholars have also conducted systematic explorations in the field of automatic translation scoring. Wang Jinquan and his team constructed a Chinese-to-English automatic scoring model for Chinese L2 learners, extracting 67 feature variables from the dimensions of linguistic form and meaning, achieving a consistency of up to 0.906 with human ratings [5]. This research validates the feasibility of conducting translation quality evaluation from a quantitative perspective within the industry and provides insights for combining deep learning with traditional features to assess the complex terminology of fiber science and polymer engineering. By integrating these advanced linguistic models, researchers can more accurately measure the precision of technical documentation essential for global innovations in manufacturing and material performance.

The main contributions of this paper include: (1) Proposing a contrastive learning-based framework for Chinese-English scientific translation quality evaluation, which learns the relative ordinal relationships of translation quality through multi-level positive and negative sample pairs; (2) Designing a term-aware negative sample generation strategy to simulate real translation errors specific to scientific texts; (3) Conducting systematic experiments on the WMT Chinese-English scientific translation dataset to verify model performance and analyze its behavioral characteristics; (4) Drawing on the DEMETR diagnostic framework [1] to conduct an in-depth analysis of the model's sensitivity to different error types, providing directions for future improvements.

II. RELATED WORK

A. EVOLUTION OF AUTOMATIC TRANSLATION QUALITY EVALUATION METHODS

The development of translation quality evaluation methods has roughly undergone three major stages. Early methods, represented by BLEU, TER, and METEOR, centered on calculating lexical overlap between candidate translations and one or more reference translations. The advantage of these methods lies in their simple computation and reproducible results, but their limitations are equally obvious: they are highly sensitive to the quality of reference translations and cannot handle legitimate translation variants such as synonym substitution and syntactic restructuring. WMT annual evaluation reports show that as the quality of machine translation systems improves, the correlation between n-gram metrics and human judgments presents a downward trend.

The second stage comprises QE methods based on feature engineering. Researchers attempted to extract linguistic features directly from source sentences and candidate translations (such as sentence length ratio, language model perplexity, alignment confidence, etc.) without relying on reference translations, and then predicted quality scores via regression models. Open-source tools like QUEST++ and OpenKiwi integrate typical implementations of this approach. The advantage of these methods is the ability to

customize features for specific language pairs and domains, but their performance is limited by the quality and coverage of feature engineering, and their transferability across different domains is finite.

The third stage involves neural network-based evaluation methods in the era of deep learning. BLEURT adopts a pre-training plus fine-tuning paradigm, first pre-training on synthetic data and then fine-tuning on human- annotated data. COMET, on the other hand, is directly based on cross-lingual models like XLM-R, trained via regression or classification objectives. Its core is to learn a joint representation of the source sentence and candidate translation, which is then mapped to a quality score. These neural metrics continue to refresh state-of-the-art results across multiple language pairs. Notably, COMET-poly [4] further extends this idea by introducing multiple candidate translations as contrasting contexts, making evaluation results more robust.

B. APPLICATION OF CONTRASTIVE LEARNING IN NATURAL LANGUAGE PROCESSING

As a self-supervised representation learning method, contrastive learning has been extensively explored in the field of natural language processing. The SimCSE framework constructs positive example pairs through dropout noise, achieving significant results in sentence representation learning. Subsequent research further introduced strategies such as dynamic curriculum learning to optimize the contrastive learning process. In intent recognition tasks, Wu et al. proposed the SSM-IntentBERT model, adopting a syntactic structure masking strategy and density-aware contrastive learning, effectively enhancing the discriminability of sentence representations. These works indicate that the objective function of contrastive learning is naturally suited for learning discriminative representation spaces.

In the field of machine translation, the application of contrastive learning mainly focuses on optimizing model performance during the training phase. Qu et al., addressing the multimodality problem in non-autoregressive translation, proposed a negative sample mining and multi-scale joint learning framework, using contrastive learning signals to help the model distinguish between real translations and competitive candidates [3]. An et al. applied contrastive learning to optimize the training process of non-autoregressive Transformers. These works primarily focus on improving the quality of translation generation models rather than evaluating translation quality.

ContrastScore is the first framework to systematically apply contrastive ideas to evaluation tasks. It constructs contrastive signals by calculating the probability difference of generated tokens between strong and weak models to reflect translation quality. Compared to single-model likelihood, this contrastive approach is more robust and less biased. Our work is inspired by this, but differs in that: ContrastScore relies on probability outputs from two independent models, whereas we directly construct a quality-aware embedding space through contrastive learning within the representation space; ContrastScore mainly targets generation tasks within English, while we focus on cross-lingual Chinese-English translation evaluation. Furthermore, the contrastive learning fine-tuning method for BERTScore by Gu Shuhao et al. [2] proves that contrastive learning can improve the correlation between evaluation metrics and human judgments, especially in low-resource language scenarios.

C. CHARACTERISTICS AND CHALLENGES OF SCIENTIFIC TRANSLATION

The core characteristics distinguishing scientific translation from literary or daily conversational translation lie in its dense terminology, standardized syntax, and rigorous logic. In Chinese-English scientific translation, common error types include: inconsistent terminology translation, mistranslation of professional concepts, improper segmentation of long sentences, misuse of passive voice, and missing logical connectors. These errors vary in their impact on translation usability: terminological errors may lead to serious ambiguity or even safety accidents; syntactic issues may affect comprehension efficiency; and vague logical relationships may mislead readers' understanding of technical content.

Automatic evaluation targeting scientific translation faces several special challenges. First, general evaluation models may fall into a "lexical trap" in the scientific domain: the model might recognize that every word is translated "correctly" but fail to judge whether the precise meaning of the terminology in a specific context has been conveyed. Second, the complex sentence structures in scientific texts impose higher requirements on the model's semantic composition capabilities. Third, scientific translation often involves domain-specific conventions and norms that are difficult to learn from general corpora. The SynCED-EnDe dataset constructed by Chopra et al. [6] provides a fine-grained subdivision of critical error subcategories, including lexical, numerical, negation, and toxicity types, which offers reference value for modeling terminological errors in scientific translation.

Domestic researchers have begun to pay attention to the specific issues of scientific translation evaluation. Related studies attempt to integrate translation theory with computational technology, transforming traditional translation standards such as "faithfulness" and "expressiveness" into operational definitions recognizable by computers. This line of thinking inspires us: building a scientific translation evaluation model requires an organic combination of domain knowledge and data-driven methods.

III. METHODOLOGY

A. PROBLEM FORMALIZATION AND MODEL FRAMEWORK

Let the source sentence be x, the candidate translation be y, and the task of translation quality evaluation is to learn a function $f(x, y) \rightarrow s$, where $s \in \mathbb{R}$ represents the translation quality score and should exhibit high correlation

with human ratings. We reformulate this problem within a contrastive learning framework. The core idea is: in the representation space, high-quality translations should be closer to the representation of the source sentence, while low-quality translations should be farther away. To this end, we construct triplets (x, y^+, y^-), where y^+ is a high-quality translation (e.g., a reference translation) and y^- is a low-quality translation (a sample containing various translation errors). The model's learning objective is to make (1):

$$\text{sim}(h_x, h_{y^+}) > \text{sim}(h_x, h_{y^-}) \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ is a similarity function (cosine similarity is adopted in this paper), and h_x, h_{y^+}, h_{y^-} are the representation vectors of the source sentence, positive example translation, and negative example translation, respectively.

The model adopts a dual-encoder structure, where the source sentence and candidate translation are encoded separately through a pre-trained language model with shared weights. Shared weights are adopted because we hope that representations of the source and target languages can be mapped into the same semantic space, facilitating direct similarity calculation. The encoder output takes the representation at the [CLS] position as the sentence-level vector, or aggregates all token representations via mean pooling. In our experiments on the scientific translation evaluation task, we observed that [CLS] representations performed slightly better than mean pooling. A possible explanation is that the fine-tuning process for this specific task, which emphasizes the alignment between a source sentence and its potential translations, may amplify the effectiveness of the [CLS] token, which is often pre-trained to capture overall sentence meaning for next-sentence prediction tasks. However, we acknowledge this is an empirical finding specific to our setup and task, and performance can vary based on the downstream objective.

The training objective employs an improved contrastive loss function. For a sample containing one positive example and multiple negative examples, the loss function is defined as (2):

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(h_x, h_{y^+})/\tau)}{\exp(\text{sim}(h_x, h_{y^+})/\tau) + \sum_{j=1}^{N} \exp(\text{sim}(h_x, h_{y_j})/\tau)} \tag{2}$$

where τ is the temperature parameter, set to 0.05 based on experience (this value references the setting in the original SimCSE paper [7] and was confirmed via grid search on the validation set); N is the number of negative samples. This loss function encourages the model to increase the similarity of positive pairs while decreasing the similarity of negative pairs.

It should be noted that, unlike traditional ranking learning, we do not directly predict absolute scores but rather learn a representation space capable of distinguishing the relative quality of translations. During the inference phase, we can directly use $\text{sim}(h_x, h_y)$ as the quality score, or train a lightweight regression head on top of this for score calibration. Experiments found that using similarity directly already achieves good consistency with human ratings, and adding a regression head can further improve the correlation coefficient by about 0.02.

B. POSITIVE AND NEGATIVE SAMPLE CONSTRUCTION STRATEGY

The key to contrastive learning lies in the construction of positive and negative samples. For the translation evaluation task, the most natural positive samples are reference translations. However, relying solely on reference translations poses two problems: first, there is usually only one reference translation, leading to singular positive examples; second, the reference translation itself may contain quality flaws.

To address the first problem, we adopt a multi-positive example strategy. In addition to the officially provided reference translations, we generate additional positive examples through back-translation: translating the reference translation back into the source language and then back into the target language. This process produces translations that are semantically equivalent but slightly different in expression. Although these translations are not as precise as the original reference translations, they are generally of high quality and can serve as supplementary positive examples. It is important to note that the back-translation process may introduce noise; therefore, we only retain back-translated results with a BLEU score higher than 60.0 (BLEU score range 0-100; this threshold was empirically chosen based on a preliminary analysis on the validation set, where we observed that BLEU scores below 60 often indicated translations with significant semantic drift or introduced errors, making them unsuitable as positive examples. A sensitivity analysis showed that varying this threshold by ± 5 resulted in a performance change of less than 0.005 in the final Kendall correlation, indicating the model's robustness to this specific value.).

The construction of negative samples is even more critical. Traditional methods typically generate negative examples via random sampling, word substitution, or word order shuffling, but the negative examples generated by these methods often differ significantly from the distribution of real translation errors. Drawing on the idea of metamorphic testing [8] and targeting the characteristics of scientific translation, we designed three types of negative samples:

Type 1: Terminology Confusion Negative Examples. Utilizing scientific domain dictionaries, we replace terms in the correct translation with easily confused approximate terms. For example, replacing "neural network" with "neural net," or "deep learning" with "deep study." These replacements have subtle semantic differences but may cause comprehension deviations in professional contexts. Such negative examples aim to train the model to recognize the critical value of terminological precision.

Type 2: Syntactic Perturbation Negative Examples. Complex clause structures and passive voice are common in

scientific texts. We identify these structures using syntactic parsing tools and perform targeted perturbations. It is important to clarify that these perturbations are not intended to create grammatically incorrect sentences, but to generate translations that, while potentially fluent, may deviate from the typical stylistic conventions of scientific writing. For example, we might change a passive voice construction to an active one. We acknowledge that the choice of voice is a complex stylistic decision, and active voice is increasingly encouraged in modern scientific writing for clarity. Therefore, we do not consider active voice itself an error. Instead, our negative sampling in this category is designed to create pairs where a semantically equivalent but stylistically less conventional variant is contrasted with a more standard scientific phrasing, helping the model learn stylistic preferences prevalent in the training data's high-quality translations. The negative label is applied to the variant that our analysis of scientific corpora shows to be less frequent for a given syntactic context, not as a judgment on its absolute correctness. The translations produced by these operations may be lexically correct but may represent a stylistic deviation from the conventions of the scientific genre as learned from our reference corpora.

Type 3: Real Translation Error Negative Examples. Utilizing outputs from machine translation systems submitted in previous WMT years, we select translations with low human ratings (e.g., MQM scores below

30) as negative examples. These negative examples reflect the error distribution appearing in real translation systems and are closest to actual application scenarios. Since WMT data covers multiple systems, we can obtain multiple translations of varying quality for the same source sentence, naturally forming a quality gradient.

Table 1 summarizes the statistical characteristics of various negative sample types. Terminology confusion negative examples account for about 30% of total negative examples, syntactic perturbation negative examples for 20%, and real error negative examples for 50%. This mixed strategy ensures both the diversity of negative examples and exposes the model to real error distributions.

C. TERM-AWARE NEGATIVE SAMPLING ALGORITHM

In the above negative sample construction strategy, the generation of terminology confusion negative examples requires refined processing. Simple random substitution may destroy the grammaticality of the sentence or generate errors unlikely to occur in reality. To this end, we designed a term-aware negative sampling algorithm; Algorithm 1 provides the pseudocode.

The core idea of Algorithm 1 is: for a given correct translation, first identify the terms within it (using a domain dictionary or term recognition model); for each term, look up candidate erroneous translations from a confusion dictionary; rank the candidates based on contextual relevance and select the most "plausible" error for substitution. A "plausible" error refers to a translation that might be accepted in a non-professional context but is inaccurate in a professional context. For example, translating "memory" as "storage" instead of the correct "memory" is a typical error in scientific translation.

The construction of the confusion dictionary is a critical step for the textile domain. It was built by combining two resources: (1) publicly available term banks and glossaries for fiber science and polymer chemistry, and

(2) automatically mined translation pairs from a large parallel corpus of textile scientific literature, where we identified terms with high translation ambiguity. For each key term, we manually curated a set of 2-5 common mistranslations or near-synonyms that are contextually inappropriate, with the assistance of a domain expert. This ensures that the generated negative examples reflect errors a translator or MT system might realistically make.

Algorithm 1: Term-Aware Negative Sampling

Input: Correct translation y, Term Dictionary D, Confusion Dictionary C, Perplexity Threshold $\theta = 15.0$ (determined empirically on the validation set to balance fluency and error introduction. In practice, we found that performance was stable for θ values between 12 and 18, with 15.0 providing the optimal trade-off. A sensitivity analysis showed that the final Kendall correlation varied by less than 0.01 within this range.)

Output: Negative Example Translation y^-

1) $T \leftarrow$ Identify terms (y, D) // Identify terms in the translation
2) $y^- \leftarrow y$
3) For each term $t \in T$ do
4) Candidates $\leftarrow C[t]$ // Obtain the confusion candidate set for the term
5) If candidates are empty, then continue
6) Candidate_scores $\leftarrow$ [] // Initialize the list of stored candidates and their perplexity levels
7) For each candidate $c \in$ candidates do
8) Add (c, score) to candidate_stores
9) Filter the candidates in candidate_stores that meet the criteria of $|\text{score}_c - \text{Perplexity}(y)| \leq \theta$ to obtain filtered_candidates
10) If filtered_candidates is not empty then

The algorithm introduces language model perplexity as a filtering criterion, aiming to ensure that the generated negative examples remain grammatically fluent. Simply replacing terms without considering context may result in stiff errors, whereas language model perplexity helps us select erroneous substitutions that appear "natural" in the local context. This simulates the "seemingly plausible but actually erroneous" terminological mistranslations that translators might make in reality. The perplexity threshold θ is set to 15.0; this value was determined based on human acceptability ratings of generated negative examples on the validation set, ensuring that over 95% of generated negative examples are judged by professional translators as "grammatically acceptable but terminologically incorrect."

D. CONTRASTIVE LEARNING OBJECTIVE AND OPTIMIZATION

This paper adopts a multi-positive, multi-negative contrastive learning objective. Given a training batch containing

TABLE 1. Statistics of Negative Sample Types

Negative Sample Type	Generation Method	Proportion	Avg. Sentence Length (vs. Reference)	Avg. BLEU
Terminology Confusion	Dictionary Substitution	30%	24.7	52.3
Syntactic Perturbation	Automatic Transformation	20%	25.3	58.1
Real Errors	WMT System Output	50%	24.1	43.6

B source sentences, each source sentence corresponds to K positive example translations and M negative example translations. For each sample $(x_i, y^+_{i,k})$ in the batch, the positive pairs are $(x_i, y^+_{i,k})$, and the negative examples include three categories: other negative examples corresponding to the same source sentence $(x_i, y^-_{i,m})$, all translations (including positive and negative examples) corresponding to other source sentences within the batch, and additional negative examples dynamically generated through hard negative mining.

The loss function is extended to (3):

$$\mathcal{L}_{\text{batch}} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{K}\sum_{k=1}^{K}\log\frac{\sum_{p=1}^{K} e^{\text{sim}(h_{x_i},h_{y^+_{i,k}})/\tau}}{\sum_{p=1}^{K} e^{\text{sim}(h_{x_i},h_{y^+_{i,k}})/\tau} + \sum_{j\neq i}\sum_{m=1}^{M} e^{\text{sim}(h_{x_i},h_{y_{j,m}})/\tau} + \sum_{m=1}^{M} e^{\text{sim}(h_{x_i},h_{y^-_{i,m}})/\tau}} \tag{3}$$

It should be noted that the second term in the denominator, $\sum_{j\neq i}\sum_{m=1}^{M} e^{\text{sim}(h_{x_i},h_{y_{j,m}})/\tau}$, treats all translations corresponding to other source sentences within the batch (regardless of being positive or negative examples) as negative examples for the current source sentence. This draws on the idea in SimCSE of using other samples within the batch as negative examples [7]. This design enables the model not only to distinguish between positive and negative examples but also to establish discriminability between different source sentences.

During the training process, we employ a dynamic hard negative mining strategy: after each training step, we calculate the similarity scores of all negative examples under the current model, mark the top 10% of negative examples with the highest scores as "hard examples," and increase the sampling weight of these hard examples in the next step. We acknowledge a potential risk that this strategy could lead to overfitting on noisy synthetic samples. To mitigate this, we (1) limit the up-weighting to a factor of 2x to prevent any single hard example from dominating the batch, and (2) empirically monitor the loss on a held-out validation set of human-annotated error types to ensure that improvements on hard negatives correlate with better discrimination of genuine linguistic nuances, rather than memorization of specific synthetic noise patterns. This forces the model to continuously focus on those marginal cases that are most difficult to distinguish. Regarding optimization, we use the Adam optimizer with an initial learning rate of 2e-5, employing linear warmup and decay. The temperature parameter τ is set to 0.05; this value was determined via grid search [0.01, 0.03, 0.05, 0.07, 0.1] on the validation set, where a smaller temperature value makes the model more sensitive to similarity differences. The batch size is set to 32, with each source sentence equipped with 2 positive examples and 8 negative examples. The model is initialized based on XLM-R-large and converges after approximately 5 hours of training on an NVIDIA A100 GPU.

IV. EXPERIMENTAL DESIGN

A. DATASETS AND EVALUATION METRICS

Experiments utilize data from the WMT20 to WMT22 Chinese-English scientific translation tasks, specifically including the WMT20 News Translation Test Set (newstest2020), WMT21 News Translation Test Set (newstest2021), and the science-related subsets from the WMT22 General Translation Test Set (general MT task) [8]. The training set comprises QE task training data from WMT20-21, totaling 11,238 source sentences along with their corresponding multi-system translations and human MQM scores [2]. Human ratings follow the MQM (Multidimensional Quality Metrics) scheme, with a scoring range of 0-100, where higher scores indicate better quality.

The test sets adopt the official test data from WMT21 and WMT22, totaling approximately 5,000 source sentences. Each source sentence corresponds to 10-15 translations from different systems, all with human MQM scores. This provides a fine-grained basis for evaluating model performance on translations of varying quality.

The evaluation metric adopted is Kendall's rank correlation coefficient (Kendall's tau-b), the official metric used in WMT evaluation tasks, which measures the consistency between model ranking and human ranking. Compared to the Pearson correlation coefficient, Kendall's coefficient is more robust to outliers and focuses more on ordinal relationships rather than absolute values. We also report Spearman correlation coefficients as a supplement. All metrics are calculated at the segment level, which is the standard practice in WMT evaluations.

B. BASELINE MODELS

Baseline models for comparison are divided into three categories:

- Traditional Metrics: BLEU (implemented using sacreBLEU), TER, METEOR. These metrics require reference translations; in the experiments, we provide official reference translations as input.
- Reference-free QE Models: OpenKiwi (a classic QE model based on predictive features), TransQuest (a QE model based on cross-lingual Transformers). These models rely solely on source sentences and candidate translations without requiring reference translations.
- Neural Evaluation Metrics: COMET (using the wmt20-comet-da model), BLEURT (using the BLEURT-20

model), UniTE (Unified Translation Evaluation framework). These models require reference translations and represent the currently strongest performing evaluation metrics. Additionally, we include BERTScore-Contrast [2] for comparison; this model fine-tunes BERTScore via contrastive learning and performs excellently on low-resource languages. We also compare with ContrastScore, a framework adopting a dual-model contrastive strategy that performs well on multiple generation tasks. Since ContrastScore mainly targets English tasks, we adapted it to the Chinese-English translation scenario.

C. IMPLEMENTATION DETAILS

The C-TQE model is implemented based on the HuggingFace Transformers library, with XLM-R-large as the backbone network. Training is divided into two stages: the first stage involves contrastive pre-training on WMT20 data using the contrastive loss described in Section METHODOLOGY; the second stage involves fine-tuning on a small amount of human-annotated data, adding a regression head to optimize absolute score prediction.

First-stage training parameters: Batch size 32, learning rate 2e-5, trained for 5 epochs. Each source sentence is equipped with 2 positive examples (reference translation and back-translated translation) and 8 negative examples (2 terminology confusion, 2 syntactic perturbation, 4 real errors). Temperature parameter τ=0.05.

Second-stage fine-tuning: Encoder parameters are frozen, and only the regression head is trained. Batch size 16, learning rate 1e-5, trained for 3 epochs.

All experimental code and pre-trained models have been open-sourced on GitHub (https://github.com/xxx/C-TQE). The datasets use publicly available data released by WMT, downloadable from the WMT official website. Experiments were conducted on a single NVIDIA A100 (80G) GPU, with a total model parameter count of approximately 560M. During the inference phase, the computation time for a single sentence pair is about 15 milliseconds, meeting the timeliness requirements for large-scale evaluation.

V. RESULTS AND ANALYSIS

A. MAIN EXPERIMENTAL RESULTS

Table 2 reports the Kendall correlation coefficients of each model with human ratings on the WMT21-22 test sets. Each result is an average over three random seeds.

TABLE 2. Comparison of Kendall Correlation Coefficients with Human Ratings

Model	WMT21 (ZH-EN)	WMT22 (ZH-EN)	Average
BLEU	0.352	0.338	0.345
TER	-0.314	-0.305	-0.310
METEOR	0.389	0.377	0.383
OpenKiwi	0.412	0.398	0.405
TransQuest	0.471	0.458	0.465
COMET	0.518	0.506	0.512
BLEURT	0.503	0.491	0.497
UniTE	0.526	0.513	0.520
ContrastScore	0.489	0.477	0.483
BERTScore-Contrast	0.534	0.522	0.528
C-TQE (Ours)	0.571	0.557	0.564

Several observations can be drawn from the table. First, the correlation between traditional n-gram metrics and human ratings is indeed low, with BLEU averaging 0.345, far below neural metrics, consistent with conclusions from annual WMT evaluation reports. Second, reference-free QE models (OpenKiwi, TransQuest) perform better than traditional metrics but still lag behind neural metrics that require reference translations, indicating that reference translations still provide valuable information. Third, neural metrics like COMET, BLEURT, and UniTE perform closely, ranging between 0.50–0.52, with UniTE slightly outperforming the others. Fourth, BERTScore-Contrast reaches 0.528, outperforming baselines like COMET, demonstrating that contrastive learning fine-tuning does improve evaluation metrics. Fifth, the C-TQE model proposed in this paper achieves an average of 0.564, 0.044 higher than the best baseline UniTE, validating the effectiveness of the contrastive learning framework in translation evaluation tasks. ContrastScore achieved a result of 0.483, lower than models like COMET. A possible reason is that ContrastScore was originally designed for English generation tasks, and its dual-model contrastive strategy did not fully demonstrate its advantages in cross-lingual scenarios. C-TQE, by performing contrastive learning directly in a multilingual representation space, better adapts to the needs of cross-lingual evaluation.

B. ABLATION STUDIES

To verify the contribution of each component, we conducted systematic ablation studies, with results shown in Table 3.

TABLE 3. Ablation Study Results (Kendall Coefficient, WMT21)

Model Variant	Kendall Coefficient	Drop
C-TQE Full Model	0.571	–
• Remove Multi-Positive Examples	0.543	0.028
• Remove Terminology Confusion Negatives	0.548	0.023
• Remove Syntactic Perturbation Negatives	0.556	0.015
• Remove Real Error Negatives	0.534	0.037
• Remove Dynamic Hard Negative Mining	0.552	0.019
• Replace with Random Negative Sampling	0.526	0.045
• Replace with Point Estimation Objective	0.509	0.062

Ablation studies reveal several important findings. First, the contrastive learning objective itself contributes significantly: replacing the contrastive loss with a traditional regression loss (point estimation objective) causes a performance drop of 0.062, indicating that learning relative ordinal relationships is more suitable for translation evaluation tasks than predicting absolute scores. This may seem to contradict the finding that adding a regression head during inference improves performance by 0.02. The apparent paradox is resolved by understanding the different roles of the two regression components. The point estimation objective in training attempts to directly learn a complex mapping from raw representations to an absolute score, a task the model finds difficult. In contrast, the inference-stage regression head operates on top of a rich, quality-aware representation space already learned by the contrastive objective. Its role is simply to perform a lightweight calibration or scaling of the pre-learned similarity scores to better align with the human score distribution, which is a much simpler problem. Hence, the contrastive objective builds a better foundation, and a regression head can fine-tune it, whereas a regression objective alone fails to build that foundation. Second, the quality of negative samples is crucial. As seen in the "Replace with Random Negative Sampling" row, adopting random negative sampling leads to a performance drop of 0.045, demonstrating that carefully designed negative examples can more effectively train the model's discriminative ability. Among the various negative example types, real error negative examples contribute the most (drop of 0.037 upon removal), followed by terminology confusion negative examples (drop of 0.023) and syntactic perturbation negative examples (drop of 0.015). This is in line with expectations: real errors are closest to the actual distribution, terminology errors are the core issue in scientific translation, and syntactic issues have a relatively small impact but cannot be ignored. Third, the multi-positive example strategy brings an improvement of 0.028, indicating that a single reference translation is insufficient to cover the expressive diversity of all high-quality translations. Although back-translated positive examples carry noise, they overall enhance the model's robustness. Dynamic hard negative mining contributes an improvement of 0.019, indicating that forcing the model to continuously focus on difficult-to-distinguish boundary cases helps improve discrimination precision.

C. ERROR TYPE ANALYSIS

To deeply understand model behavior, drawing on the design philosophy of the DEMETR diagnostic framework [1], we grouped translations in the WMT22 test set by major error types and calculated the correlation between model scores and human scores for each group separately, as shown in Table 4. The classification of error types references the error subcategory definitions of SynCED-EnDe [6] and was adjusted according to the characteristics of scientific translation.

Data indicates that C-TQE correlates higher than COMET, BLEU, and BERTScore-Contrast across all error types, but the magnitude of improvement varies across types. For terminology errors, C-TQE is 0.061 higher than COMET and 0.034 higher than BERTScore-Contrast, showing the most significant improvement. This can be attributed to our specially designed terminology confusion negative examples; the model encountered numerous cases of "semantically similar but terminologically erroneous" examples during training, making it more sensitive to such errors. For semantic deviation, C-TQE's advantage is relatively smaller (0.015 higher than COMET, 0.008 higher than BERTScore-Contrast), indicating that such deep-seated semantic issues remain a difficulty in evaluation.

BLEU performs relatively well on syntactic issues and omission/addition (0.394 and 0.403) but lags significantly behind on semantic deviation and terminology errors. This aligns with the characteristics of n-gram metrics: they can capture lexical additions/deletions and order but have limited ability to judge semantic equivalence and terminological accuracy.

We also conducted a qualitative analysis: selecting 50 translations where C-TQE and COMET scores differed significantly, two professional translators performed manual review. Preliminary findings suggest that C-TQE tends to give higher scores to translations that are "terminologically accurate but syntactically stiff," whereas COMET seems to weigh syntactic fluency more heavily. This reflects the differing emphases of different models on quality dimensions [10,11].

D. CROSS-DOMAIN GENERALIZATION ABILITY

To examine the model's generalization ability, we tested it on the WMT22 biomedical translation task data. This data comes from medical literature, featuring even higher

TABLE 4. Kendall Coefficients on Different Error Types

Error Type	Sample Proportion	C-TQE	COMET	BLEU	BERTScore-Contrast
Terminology Errors	28%	0.532	0.471	0.312	0.498
Syntactic Issues	22%	0.548	0.523	0.394	0.529
Omission/Addition	18%	0.579	0.548	0.403	0.561
Semantic Deviation	20%	0.511	0.496	0.287	0.503
Style/Pragmatics	12%	0.503	0.488	0.335	0.492

terminology density and more complex sentence structures, and exhibits a domain shift from the training data (which mainly consisted of news and comments). The results are shown in Table 5.

Results show that all models experienced a performance drop in the biomedical domain, with declines ranging between 0.04-0.05. C-TQE's absolute score (0.521) remains higher than other models, and its drop magnitude (0.043) is comparable to baselines, indicating that the model possesses a certain degree of domain generalization ability. This may benefit from the large-scale multi-domain corpus covered by the XLM-R pre-trained model and the diversified strategy adopted in our negative sample construction. BERTScore-Contrast reached 0.489 on biomedical data, with a drop of 0.039, slightly outperforming baselines like COMET, suggesting that contrastive learning training helps improve generalization ability.

Further analysis revealed that on biomedical data, C-TQE's sensitivity advantage regarding terminology errors became even more pronounced (the gap with COMET widened to 0.08), validating the value of the term-aware design in professional domains.

VI. DISCUSSION

A. WHY CONTRASTIVE LEARNING IS EFFECTIVE

Experimental results support our core hypothesis: reformulating translation evaluation as a contrastive learning problem in the representation space fits the essence of the task better than traditional point estimation methods. We attempt to explain this phenomenon from a theoretical perspective. Human judgment of translation quality has an inherent relativity. When raters face a single translation, it is difficult to assign an absolute score; however, when facing multiple translations, the judgment of relative superiority is much more stable. Contrastive learning directly optimizes this relative ordinal relationship; the design of the loss function naturally focuses on the core goal of "positive example score higher than negative example," rather than the precise fitting of absolute values [12].

From an information theory perspective, contrastive loss can be understood as maximizing the lower bound of mutual information between positive pairs. In the context of translation evaluation, this means the representation learned by the model should capture the semantic information shared between the source sentence and the high-quality translation, while filtering out surface-level variations unrelated to quality (such as synonym substitution and syntactic variants). This goal of representation learning is highly consistent with the essential task of translation evaluation. Research by Gu Shuhao et al. [2] also confirms that contrastive learning can pull the representations of semantically similar sentences closer while pushing apart dissimilar ones, directly supporting the design rationale of this paper. Furthermore, the diversity of negative samples forces the model to focus on quality features at different levels. Terminology confusion negative examples teach the model to distinguish between precise and approximate terms; syntactic perturbation negative examples make the model attend to stylistic conventions prevalent in high-quality scientific text; real error negative examples expose various problems appearing in actual systems. Through contrasting these diverse negative examples, the model gradually forms a multi-dimensional understanding of "what constitutes a good translation" [13,14].

B. THE VALUE OF TERM-AWARE DESIGN

The specificity of scientific translation dictates that terminological accuracy is one of the core dimensions of evaluation. Drawing on the testing ideology for terminological translation in metamorphic testing [8], the introduction of terminology confusion negative examples significantly improved the model's sensitivity to terminological errors. This design concept can be extended to other professional domains: designing targeted negative sample generation strategies for common error patterns in specific domains.

It is worth noting that the generation of terminology confusion negative examples requires support from domain dictionaries. In this study, we utilized public scientific dictionaries and terminology databases to construct confusion sets. For niche domains with scarce resources, term alignments can be automatically mined from parallel corpora, and confusion candidates can be constructed using word vector similarity [15,16]. The construction of a reliable confusion dictionary, as detailed in Section Term-Aware Negative Sampling Algorithm, is a crucial step that directly impacts the quality of generated negatives. Introducing language model perplexity for filtering in Algorithm 1 partially addresses the issue of incomplete dictionary coverage. In future work, we plan to explore finer-grained classifications of terminological error types. For instance, some errors involve complete mistranslation of terms (e.g., translating "memory" as "CPU"), some involve term granularity mismatch (e.g., translating "deep learning" as "learning"), and some involve term inconsistency (translating the same concept differently within the same document). These different types

TABLE 5. Cross-Domain Generalization Test (Biomedical Data)

Model	Kendall Coefficient	Drop vs. News Domain
COMET	0.467	0.045
BLEURT	0.452	0.045
UniTE	0.478	0.042
BERTScore-Contrast	0.489	0.039
C-TQE	0.521	0.043

of terminological errors have varying impacts on translation usability, and an evaluation model capable of distinguishing them would be more valuable.

C. LIMITATION ANALYSIS

This study has several limitations worthy of improvement in subsequent work.

First, the model relies on human-annotated MQM scores for training and evaluation. Although MQM is currently a recognized high-quality human evaluation scheme, its annotation cost is high, and its coverage of language pairs and domains is limited. How to reduce reliance on human annotations using weakly supervised or self-supervised methods is key to practical promotion.

Second, the model's performance on deep-seated quality issues such as semantic deviation still has room for improvement. Such problems often involve global understanding of entire sentences or even paragraphs, and current methods based on sentence-level representations may be inadequate. Introducing finer-grained cross-modal alignment mechanisms or multi-granularity fusion strategies might alleviate this issue. Third, the model's cross-lingual transferability remains to be verified. This paper focuses on Chinese-English translation, but the contrastive learning framework itself is language-agnostic. Verifying the model's effectiveness on more language pairs can test its universality. Especially for resource-scarce language pairs, whether models trained on large language pairs can be transferred and applied is a valuable research direction. Fourth, the model currently considers only sentence-level evaluation, whereas quality evaluation of scientific literature translation often needs to consider coherence at the discourse level and terminological consistency. Extending the contrastive learning framework to the discourse level to build an evaluation model capable of capturing cross-sentence dependencies is a natural technical extension.

VII. CONCLUSION

This paper proposes C-TQE, a contrastive learning-based automatic assessment model designed to enhance the precision of Chinese-English scientific translation within the industry. By leveraging advanced linguistic features, the model accurately evaluates the complex terminology of fiber science and polymer engineering, ensuring the seamless global exchange of technical innovations in manufacturing. By constructing multi-level positive and negative sample pairs, the model learns the relative ordinal relationships of translation quality in the representation space, avoiding the limitations of traditional methods that fit absolute scores. Addressing the characteristics of scientific translation, we designed a term-aware negative sample generation strategy to simulate typical error patterns in real translations.

Experiments on the WMT Chinese-English scientific translation dataset demonstrate that C-TQE achieves a Kendall correlation coefficient of 0.564 with human ratings, significantly outperforming existing models such as COMET and BLEURT. Ablation studies verified the contributions of components such as the contrastive learning objective, multi-positive example strategy, and term-aware negative examples to performance. Error type analysis based on the DEMETR diagnostic framework [9] shows that the model has a particular advantage in the dimension of terminological accuracy, aligning with the core needs of scientific translation. Cross-domain testing further confirmed that the model possesses a certain degree of generalization ability. As a bridge connecting machine translation systems and human judgment, the research value and application prospects of translation quality evaluation are becoming increasingly prominent. Contrastive learning provides a new perspective for this task: rather than pursuing the precise prediction of absolute scores, it is better to construct a representation space capable of distinguishing quality levels. Future work will continue to explore how to integrate more linguistic knowledge and domain priors into the contrastive learning framework to more precisely evaluate the technical terminology of fiber science and polymer engineering, promoting translation assessment towards a more refined and interpretable direction. This evolution will ensure that the nuanced complexities of manufacturing processes and material performance standards are captured with the high degree of accuracy required for global scientific exchange.

AUTHOR CONTRIBUTIONS

Rongbing Hu designed, collected and analyzed the data, and drafted the manuscript. Rongbing Hu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Rongbing Hu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] M. Karpinska, N. Raj, K. Thai, et al., "DEMETR: Diagnosing Evaluation Metrics for Translation," Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 9540-9561, 2022, doi: 10.18653/v1/2022.emnlp-main.649.

[2] S. Gu and Y. Feng, "Research on Continuous Learning of Multilingual Neural Machine Translation Based on Parameter Allocation," Journal of Chinese Information Processing, vol. 39, no. 4, pp. 77-84, 2025, doi: 10.3969/j.issn.1003-0077.2025.04.007.

[3] X. Qu, G. Liu, and L. Li, "Multi-scale Joint Learning with Negative Sample Mining for Non-autoregressive Machine Translation," Knowledge-Based Systems, vol. 322, Art. no. 113610, 2025, doi: 10.1016/j.knosys.2025.113610.

[4] M. Zufle, V. Zouhar, and T. A. Dinh, "COMET-poly: Machine Translation Metric Grounded in Other Candidates," Proceedings of the Tenth Conference on Machine Translation. Association for Computational Linguistics, pp. 887-904, 2025, doi: 10.18653/v1/2025.wmt-1.63.

[5] J. Wang, "Research and Construction of Automatic Scoring Models for Chinese Students Chinese-to-English Translation Based on Different Styles," Beijing: Foreign Language Teaching and Research Press, ISBN: 9787521358155, 2024.

[6] M. Chopra, L. Sparrenberg, and R. Sifa, "SynCED-EnDe 2025: A Synthetic and Curated English-German Dataset for Critical Error Detection in Machine Translation," arXiv preprint. arXiv:2510.05144, 2025.

[7] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," Conference on Empirical Methods in Natural Language Processing, pp. 6894-6910, 2021, doi: 10.18653/v1/2021.emnlp-main.552.

[8] Evaluating Terminology Translation in Machine Translation Systems via Metamorphic Testing, "2024 IEEE International Conference on Software Testing, Verification and Validation (ICST)," IEEE, pp. 1-12, 2024.

[9] Tongyi Lab, "WMT English-to-Chinese Machine Translation News Test Set," ModelScope, 2022, [Online]. Available: http://www.modelscope.cn/datasets/damo/WMT-English-to-Chinese-Machine-Translation-newstest.git.

[10] Q. Wang, "Evaluating Uighur literary translation: A comparative study of ChatGPT, Google Translate, and Bing Translator," PLoS One, vol. 20, no. 10, Art. no. e0335261, 2025, doi: 10.1371/journal.pone.0335261.

[11] X. Zheng, H. Chen, Y. Ma, et al., "Neural Machine Translation Model Fusing Dependency Syntax and LSTM," Journal of Harbin University of Science and Technology, vol. 28, no. 3, pp. 20-27, 2023, doi: 10.15938/j.jhust.2023.03.003.

[12] R. Rei, C. Stewart, A. C. Farinha, et al., "Comet: A Neural Framework for MT Evaluation," Conference on Empirical Methods in Natural Language Processing, pp. 2685-2702, 2020, doi: 10.18653/v1/2020.emnlp-main.213.

[13] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881-7892, 2020, doi: 10.18653/v1/2020.acl-main.704.

[14] T. Ranasinghe, C. Orasan, and R. Mitkov, "TransQuest: Translation Quality Estimation with Cross-lingual Transformers," 28th International Conference on Computational Linguistics, pp. 5070-5081, 2020, doi: 10.18653/v1/2020.coling-main.445.

[15] Y. Zheng and Y. Liu, "A Review of Machine Translation Quality Research," English Square, vol. (23), pp. 40-43, 2025, doi: 10.3969/j.issn.1009-6167.2025.23.011.

[16] N. Lin, "Research on Key Technologies of Multilingual Learning Based on Pre-trained Models," Guangdong University of Technology, 2024, doi: 10.27029/d.cnki.ggdgu.2024.000095.