Fusion of Improved BERT and BiLSTM Models to Improve English-Chinese Terminology Translation

Zhongwei Mei[1]
[1]School of Foreign Languages, Luoyang Normal University, Luoyang 471934, Henan, China
Corresponding author: Zhongwei Mei (e-mail: lynumzw@163.com)

ABSTRACT To address weak domain adaptability and insufficient modeling of long-range contextual dependencies in specialized English-Chinese terminology translation, this paper proposes a neural translation model integrating improved BERT and BiLSTM. The model first pre-trains BERT on the public WMT corpus and domain-specific corpora, and introduces term boundary markers to enhance input representation, improving recognition of specialized terminology and domain adaptability. A gated-attention BiLSTM encoder is then designed to strengthen local contextual awareness and sequence-structure modeling, especially around term boundaries, thereby complementing BERT's global representation and capturing the semantic context of terms more accurately. A feature-fusion module integrates multi-layer representations from BERT and BiLSTM, and an attention-driven decoder generates the final translation. Experimental results show that the proposed model outperforms advanced baselines across key metrics. Terminology translation accuracy reaches 88.1%, domain consistency reaches 92.0%, term-boundary F1-score reaches 87.6%, and context sensitivity reaches 86.5%, verifying the effectiveness of the method in improving the accuracy and robustness of specialized terminology translation.

INDEX TERMS BERT, BiLSTM, terminology translation, feature fusion, domain adaptation

I. INTRODUCTION

As the internationalization of professional fields accelerates, high-quality terminology translation is becoming increasingly important in professional scenarios such as scientific and technological literature, medical reports, legal documents, patent documents, and international standards, especially concerning specialized terms in machinery specifications and fiber composition labeling for global trade. Existing neural machine translation models generally have problems such as weak generalization ability, semantic drift, and domain mismatch when dealing with professional terminology [1]. General-purpose models often struggle to balance terminology accuracy with contextual consistency. Although the translation is fluent, it lacks professionalism. This seriously restricts the credibility and deployment effect of professional translation systems in practical applications, and a targeted modeling method is urgently needed to solve this problem.

This paper focuses on the translation of English-Chinese professional terms, and the research objects cover polysemous and highly professional terms in three high-value fields: medicine, information technology, and law. Naveen and Trojovský confirmed that errors in cross-domain terminology translation are mainly due to insufficient context windows or lack of domain knowledge [2]. There are also studies that systematically sorted out the semantic differences between different AI (Artificial Intelligence) models in medical terminology translation [3], and that the same words often require different translations in different contexts [4,5]. Tang found that the legal term "consideration" is translated very differently in contract law and common law [6]. Guerberof Arenas et al. pointed out that term processing leads to longer fixation time and more return gazes [7]. Studies have consistently shown that effectively modeling domain adaptability and context dependence is the key to improving the quality of terminology translation.

In recent years, many researchers have tried to explore term translation from two paths: pre-trained language models and sequence tagging models [8,9]. Although the former performs well in general translation tasks, it lacks explicit modeling of term boundaries. Although the latter can identify term units, it is difficult to capture global semantics and cross-sentence dependencies. However, these methods have their own limitations in professional scenarios. Arnaud et al directly fine-tuned mBART (Multilingual Bidirectional and Auto-Regressive Transformers) on medical corpus, but did not process term boundaries, resulting in text being disassembled [10]. Other works first used BiLSTM-CRF to identify terms and then called a universal translator, resulting in a semantic gap between the recognition and

generation stages. The introduction of an external term dictionary to enhance input, but without joint optimization with the neural encoder, resulted in limited gains [11,12]. Rahman et al continued to pre-train RoBERTa (Robustly optimized BERT approach) on the corpus and designed a context-aware decoding mechanism to distinguish subtle differences in words [13].

This paper proposes a model for English-Chinese terminology translation that integrates an improved BERT and BiLSTM; by pre-training BERT on WMT and corpora from the medical, legal, and IT fields, along with an added corpus of standards and technical specifications, it internalizes the semantics of specialized terminology. It introduces [TERM_START]/[TERM_END] boundary markers and combines them with a gated attention BiLSTM encoder to dynamically model terminology structure and long-range contextual dependencies. A multi-granularity feature fusion module is designed to integrate global sentence-level semantics with local terminology representations. The Bahdanau attention LSTM (Long Short-Term Memory) decoder is initialized to achieve terminology-sensitive translation. Experiments show that this method outperforms baselines in terms of terminology accuracy, boundary F1, and context sensitivity, providing a reliable solution for highly specialized machine translation.

II. ALGORITHM DESIGN

A. IMPROVEMENT OF BERT BASED ON DOMAIN ADAPTATION

Based on the BERT-base-Chinese initialization model [14], this paper performs further pre-training on corpora from three high-value fields: science and technology, medicine, and law, to construct a domain-adaptive BERT. The overall model architecture is shown in Figure 1.

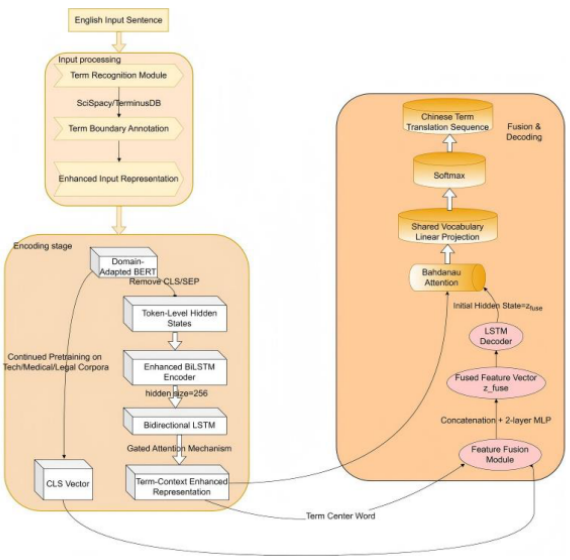

FIGURE 1. Overall model architecture

The system identifies terms in English sentences and inserts [TERM_START]/[TERM_END] tokens to construct augmented input. Contextual word vectors are extracted using BERT and combined with a gated attention BiLSTM to generate enhanced term-context encodings. The [CLS] vector is fused with the term center word features to form a zfuse, which initializes the hidden state of the LSTM decoder. During decoding, Bahdanau attention dynamically focuses on term information. Finally, a shared vocabulary and Softmax are used to generate accurate and fluent Chinese translations, achieving end-to-end term awareness and precise translation.

1) Pre-Training Target Model Architecture

This paper builds a domain-adaptive BERT based on BertForMaskedLM from the Hugging Face Transformers library. BERT-base-Chinese is used as the initialization model, and the model parameters are only fine-tuned. The pre-training task continues with the standard Masked Language Modeling (MLM) [15]. For any input sequence, 15% of the tokens are randomly selected to form a mask set $M \subseteq \{1, \ldots, T\}$ and perturb them according to the following distribution:

$$\tilde{x}_t = \begin{cases} [\text{MASK}], & \text{with probability } 0.8, \\ x_{\text{rand}} \sim \text{Uniform}(V), & \text{with probability } 0.1, \\ x_t, & \text{with probability } 0.1, \end{cases} \quad \forall t \in M \tag{1}$$

DA-BERT (Domain-Adapted BERT) maps the perturbation sequence $\tilde{x}$ into a context-aware hidden state sequence through an $L = 12$-layer Transformer encoder:

$$\begin{aligned} H &= \text{Transformer}_\Theta(\tilde{x}) \\ &= [h_1, h_2, \ldots, h_T] \in \mathbb{R}^{T \times d}, \quad d = 768 \end{aligned} \tag{2}$$

Each layer contains a multi-head self-attention mechanism:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \ldots, head_h) W^O \tag{3}$$

$$head_i = \text{softmax}\left(\frac{Q W_i^Q (K W_i^K)^\top}{\sqrt{d_k}} \right) V W_i^V \tag{4}$$

The MLM loss function is defined as the negative log-likelihood of the masked word [16]:

$$\begin{aligned} L_{MLM}(\Theta) &= -\frac{1}{|M|} \sum_{t \in M} \log P_\Theta(x_t | \tilde{x}) \\ &= -\frac{1}{|M|} \sum_{t \in M} \log \frac{\exp(e_{x_t}^\top h_t)}{\sum_{v \in V} \exp(e_v^\top h_t)} \end{aligned} \tag{5}$$

$e_v^\top \in \mathbb{R}^d$ is the embedding vector of word v, using input-output weight sharing strategy.

2) Optimization Strategy and Training Configuration

The model optimization adopts the AdamW algorithm [17], and its parameter update rules are as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla_\Theta L_{MLM}(\Theta_t) \tag{6}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla_\Theta L_{MLM}(\Theta_t))^2 \tag{7}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{8}$$

$$\Theta_{t+1} = \Theta_t - \eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} + \lambda\Theta_t\right) \tag{9}$$

The hyperparameters are set as learning rate $\eta = 2\times10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay coefficient $\lambda = 0.01$. The training adopts the dynamic batching strategy [18], and the length of the dynamic combination of samples is used to maximize the GPU memory utilization. The average effective batch size is 32, and a total of 20 epochs are trained.

3) Term Mask Recovery

500 are randomly extracted from the WMT test set, and all subwords of the 500 professional terms $T = \{\tau_j\}_{j=1}^{500}$ terms are replaced with [MASK] in the original context. Their recovery ability is evaluated after inputting them into the model. It is assumed that the term τ_j contains m_j subwords after word segmentation:

$$P_\Theta(\tau_j|s_j) = \prod_{k=1}^{m_j} \frac{\exp(e_{w_k^{(j)}}^\top h_{t_k})}{\sum_{v\in V}\exp(e_v^\top h_{t_k})} \tag{10}$$

Term Recovery Accuracy (TRA) is the proportion of top-5 predictions that contain correct terms:

$$TRA = \frac{1}{|T|}\sum_{j=1}^{|T|} I\left[\tau_j \in \text{Top-5}\left(\arg\max_{c\in V^{m_j}} \log P_\Theta(c|\tilde{s}_j)\right)\right] \tag{11}$$

After training is completed, the output context-sensitive word vector is directly fed into the subsequent improved BiLSTM encoder to provide high-quality, domain-adapted semantic representation for term boundary identification and long-distance dependency modeling.

B. TERM-AWARE INPUT REPRESENTATION

External tools (SciSpacy, TerminusDB) are used solely for annotating term boundaries in the training and evaluation data. The model itself is trained end-to-end, with term markers provided as input without further external intervention during inference. The term boundaries of the input English sentence $s = [w_1, w_2, \ldots, w_n]$ are identified. Specialized tools are used for different fields. SciSpacy [19] is used in the medical and biological fields, and the term extraction module of TerminusDB [20] is used in the legal and information technology fields. Both output a term span set $T = \{(s_j, e_j)\}_{j=1}^m$. The sentence is input into the WordPiece segmenter of BERT-base-Chinese for subword segmentation to obtain a subword sequence. For each term span $[s_j, e_j]$, the subword index interval $[p_j, q_j]$ it covers is determined:

$$\begin{aligned} p_j &= \min\{k|\text{char start}(x_k) \geq s_j\}, \\ q_j &= \max\{k|\text{char end}(x_k) \leq e_j\} \end{aligned} \tag{12}$$

If a term spans multiple subwords, a special token [TERM_START] is inserted before the first subword, and [TERM_END] is inserted after the last subword [21]. To maintain semantic compatibility, the embeddings for [TERM_START] and [TERM_END] are initialized using the average of relevant pre-trained token embeddings and are fine-tuned during training. After the above processing, the original subword sequence is expanded into an enhanced sequence. Then, word embedding, position embedding, and term type embedding are constructed. The final input representation is the element-by-element addition of the three:

$$z_i = e_i^{word} + e_i^{pos} + e_i^{type}, \quad i = 0, 1, \ldots, \tilde{T} \tag{13}$$

This representation is fed into the DA-BERT encoder, which outputs a context vector that explicitly encodes term boundary signals. Figure 2 shows an example of a term-aware input representation.

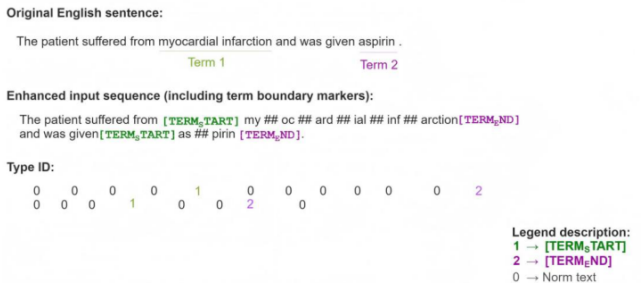

FIGURE 2. Example of term-aware input representation

Figure 2 illustrates the construction process of the "term-aware input representation". After term recognition, the original English sentence is segmented using BERT, and [TERM_START] and [TERM_END] tokens are inserted into the BERT tokenization results to explicitly define term boundaries. A type ID is annotated below each token, which is used by the embedding layer to distinguish semantic roles. This representation, which serves as the DA-BERT input, enables the model to accurately perceive term scope and provide structured semantic clues for subsequent translation.

C. IMPROVEMENTS TO BILSTM WITH GATED ATTENTION

This paper designs a bidirectional LSTM encoder with gated attention enhancement. It is assumed that the hidden state

sequence output by DA-BERT after encoding the input sequence is:

$$H^{bert} = [h_1^{bert}, h_2^{bert}, \ldots, h_{L+2}^{bert}] \in \mathbb{R}^{(L+2)\times 768} \quad (14)$$

In Formula 14, L is the number of original subwords, and h_1^{bert}, h_{L+2}^{bert} are [CLS] and [SEP] vectors respectively. After removing these two special tokens, the valid word sequence is:

$$X = [x_1, x_2, \ldots, x_L] = [h_1^{bert}, \ldots, h_L^{bert}] \in \mathbb{R}^{L\times 768} \quad (15)$$

X is input into the bidirectional LSTM layer, and the forward LSTM is recursive by time step:

$$\overrightarrow{h}_t = LSTM_{fwd}(x_t, \overrightarrow{h}_{t-1}), \quad t = 1, 2, \ldots, L \quad (16)$$

The backward LSTM reverse recursion is:

$$\overleftarrow{h}_t = LSTM_{bwd}(x_t, \overleftarrow{h}_{t+1}), \quad t = L, L-1, \ldots, 1 \quad (17)$$

The hidden state dimension is set to $d_{hid} = 256$. The bidirectional outputs are concatenated to get the BiLSTM representation:

$$h_t^{bi} = [\overrightarrow{h}_t; \overleftarrow{h}_t] \in \mathbb{R}^{512}, \quad t = 1, \ldots, L \quad (18)$$

At the same time, a gated attention mechanism is introduced to dynamically enhance the context related to terms [22]. The attention weight is calculated:

$$e_t = v^\top \tanh(W_a h_t^{bi} + b_a), \quad \alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{L} \exp(e_k)} \quad (19)$$

In Formula 19, $W_a \in \mathbb{R}^{d_a\times 512}$, $v \in \mathbb{R}^{d_a}$, $d_a = 128$, $b_a \in \mathbb{R}^{d_a}$ are learnable parameters. Then, the gating unit is designed to adjust the degree of fusion between the original BiLSTM state and the attention weighted context, and the gating vector is defined as:

$$g_t = \sigma(W_g[h_t^{bi}; c]), \quad c = \sum_{k=1}^{L} \alpha_k h_k^{bi} \quad (20)$$

In Formula 20, σ is the sigmoid function; $W_g \in \mathbb{R}^{512\times 1024}$; c is the global context vector. The final enhanced hidden state is:

$$\tilde{h}_t = g_t \odot h_t^{bi} + (1 - g_t) \odot c \quad (21)$$

$\odot$ represents the element-by-element product. This gating mechanism allows the model to adaptively inject global semantics while preserving local temporal information, especially strengthening contextual consistency near the term position. Figure 3 shows the structure of the gated attention mechanism:

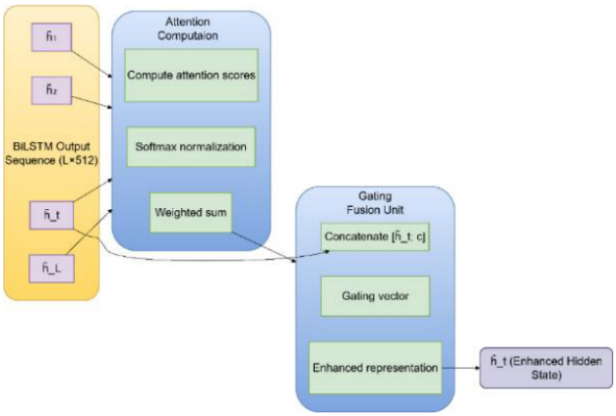

FIGURE 3. Gated attention mechanism structure diagram

Figure 3 shows the internal computational flow of the gated attention mechanism. The gated unit dynamically adjusts the fusion ratio of the local state $\tilde{h}_t$ and the global context c through the sigmoid gate vector g_t, introduces the term position mask, and determines the set of subword indexes covered by the term [23]. During the training process, a slight gradient suppression is applied to the enhanced hidden state of nonterm positions, so that the parameter update is more focused on modeling the term-related context. The final output sequence $\tilde{H} = [\tilde{h}_1, \ldots, \tilde{h}_L] \in \mathbb{R}^{L\times 512}$ is used as the source representation for the subsequent feature fusion module. By effectively capturing long-distance dependencies through BiLSTM and combining it with the gated attention mechanism, the semantic drift problem caused by insufficient context modeling in the translation of professional terms in traditional sequence models is alleviated [24].While Transformerbased models excel at capturing long-range dependencies, BiLSTM remains effective in modeling local syntactic structures and sequential dependencies. The gated attention mechanism further enables dynamic focus on term-related contexts, mitigating semantic drift caused by inadequate local modeling.

D. BERT-BILSTM MULTI-GRANULARITY FEATURE FUSION

1) Global Semantic Vector Extraction

The hidden state corresponding to the [CLS] tag is directly extracted from the output sequence of the DA-BERT encoder and is denoted as $h_{[CLS]} \in \mathbb{R}^{768}$. This vector has been optimized into a semantic summary of the entire sentence during the MLM pre-training and domain training process, and can effectively encode the domain theme, language style, and overall semantic tendency [25].

2) Contextual Positioning of Term Center

The term boundary markers constructed in this paper backtrack to determine the start and end index interval $[p, q]$ of a term in a subword sequence. To accurately capture the core

semantics of a term, the term center subword position c is defined as:

$$c = \left\lfloor \frac{p+q}{2} \right\rfloor \tag{22}$$

Subsequently, the hidden state $\tilde{h}_c \in \mathbb{R}^{512}$ of the corresponding position is extracted from the output improved BiLSTM enhanced sequence $\tilde{H} = [\tilde{h}_1, \ldots, \tilde{h}_L]$ as the local context-sensitive representation of the term.

3) Multi-Granularity Feature Concatenation and Nonlinear Mapping

The global semantic vector $h_{[CLS]}$ is concatenated with the local term representation $\tilde{h}_c$ along the feature dimension to form a joint feature [26]:

$$z_{concat} = [h_{[CLS]}; \tilde{h}_c] \in \mathbb{R}^{1280} \tag{23}$$

z_{concat} is input into a two-layer multilayer perceptron (MLP) for nonlinear transformation:

$$z_1 = \mathrm{ReLU}(W_1 z_{concat} + b_1), \quad W_1 \in \mathbb{R}^{512\times1280}, b_1 \in \mathbb{R}^{512} \tag{24}$$

$$z_{fuse} = \mathrm{ReLU}(W_2 z_1 + b_2), \quad W_2 \in \mathbb{R}^{512\times512}, b_2 \in \mathbb{R}^{512} \tag{25}$$

The ReLU activation function in Formulas 24 and 25 introduces nonlinear expression capabilities, enabling the fused feature $z_{fuse} \in \mathbb{R}^{512}$ to effectively alleviate the vanishing gradient problem [27], enabling the network to learn the complex coupling pattern between global and local features. The resulting fused vector z_{fuse} is used as the initial hidden state $h_0^{dec} = z_{fuse}$ of the decoder LSTM and as the initial value of the query vector in the Bahdanau attention mechanism [28]. This avoids terminology misinterpretation or domain drift caused by initial state ambiguity [29].

E. TERM TRANSLATION DECODING BASED ON ATTENTION MECHANISM

1) Decoder Structure and Initialization

This system uses a single-layer LSTM as the basic decoding unit and introduces the Bahdanau attention mechanism to achieve dynamic alignment between the source and target. At each decoding step t, the decoder hidden state h_{t-1}^{dec} and the source BiLSTM enhanced sequence representation $\tilde{H} = [\tilde{h}_1, \ldots, \tilde{h}_L] \in \mathbb{R}^{L\times512}$ jointly calculate the attention weight [30]:

$$e_{t,i} = v_a^\top \tanh(W_a[h_{t-1}^{dec}; \tilde{h}_i]), \quad i = 1, \ldots, L \tag{26}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{L}\exp(e_{t,j})}, \quad c_t = \sum_{i=1}^{L}\alpha_{t,i}\tilde{h}_i \tag{27}$$

$W_a \in \mathbb{R}^{512\times1024}$ and $V_a \in \mathbb{R}^{512}$ are learnable parameters. The context vector c_t is concatenated and input into LSTM:

$$h_t^{dec} = LSTM([y_{t-1}; c_t], h_{t-1}^{dec}) \tag{28}$$

2) Vocabulary Mapping and Output Generation

The output layer uses the Chinese vocabulary shared with the input, and generates the word probability distribution [31] through linear projection and Softmax:

$$o_t = W_o[h_t^{dec}; c_t] + b_o, \quad P(y_t|y_{<t}, x) = \mathrm{Softmax}(o_t) \tag{29}$$

In Formula 29, $W_o = E_{zh}^\top$, $E_{zh} \in \mathbb{R}^{|V_{zh}|\times512}$ are the Chinese word embedding matrices, which realize weight sharing to reduce parameters and improve generalization ability.

3) Training and Inference Strategies

The training phase adopts a teacher-forcing strategy [32], takes the true target sequence as input, and minimizes the cross-entropy loss [33]:

$$L_{dec} = -\sum_{t=1}^{T_y}\log P(y_t^{gold}|y_{<t}^{gold}, x) \tag{30}$$

In Formula 30, T_y is the length of the Chinese translation. The inference stage adopts the beam search strategy [34,35], and the beam width is set to 5, which improves the accuracy of term translation while ensuring the fluency of the translation.

III. EXPERIMENT AND VERIFICATION

A. EXPERIMENTAL DESIGN

1) Dataset Construction

The experiment uses data from the WMT Chinese-English machine translation training set, a dataset derived from the WMT 2021 News Translation Task. This dataset, comprised of data from ParaCrawl, News-Commentary, UN Parallel Corpus, WikiMatrix, and CCMT, totals 25 million bilingual sentence pairs. The selected data covers three major professional fields: medicine, information technology (IT), and law. Each example includes an English term, its context, the corresponding Chinese term, and its translation. The data is divided into training, validation, and test sets in an 8:1:1 ratio, with each field separated to prevent cross-domain data leakage. The data classification is shown in Table 1.

TABLE 1. WMT data sample classification

Domain	Train	Validation	Test	Total
Medical	9600	1200	1200	12000
IT	8000	1000	1000	10000
Legal	6400	800	800	8000

To prevent data leakage, terms in the test set do not appear in the DA-BERT pre-training corpus. This ensures that the

Term Recovery Accuracy evaluation reflects the model's true generalization ability.

2) Training and Optimization Settings

In order to carry out corpus training, the training environment of the corpus required by the model in this paper is shown in Table 2.

TABLE 2. Training environment

Serial Number	Experimental configuration	Content
1	Hardware	4 × NVIDIA V100 GPU
2	Random seed	42
3	Deep learning framework	PyTorch + Hugging Face Transformers
4	Optimizer	AdamW
5	Learning rate	2×10^{-5}
6	Batch size	32
7	Maximum number of training rounds	20

This model, along with the Transformer-base, BERT-Seq2Seq (Sequence-to-Sequence), BiLSTM-CRF (Conditional Random Field), and mBART-large baselines, uses the AdamW optimizer (learning rate = 2×10^{-5}, batch size = 32, max epochs 20). Early stopping (patience = 3) is monitored based on validation set term accuracy. The BERT-base-Chinese vocabulary is used, and the Chinese output embedding weights are shared. The experiments are run on 4×V100 GPUs with a fixed random seed of 42 to ensure reproducibility.

B. TERMINOLOGY TRANSLATION ACCURACY

During the testing phase, term boundaries are identified for each input English sentence. By backtracking the subword interval covered by the [TERM_START] and [TERM_END] tokens on the source side, the precise span $[p, q]$ of the term in the source sentence is determined. Using the alignment weights output by the gated attention BiLSTM encoder, this span is mapped to the corresponding character position in the target Chinese translation, extracting the model-generated term translation fragment. TA (Term Accuracy) is calculated separately for the three fields of medicine, IT, and law, and then the weighted average is used to obtain the overall TA.

$$\begin{aligned} TA_{domain} &= \frac{N_{correct}^{domain}}{N_{total}^{domain}} \times 100\%, \\ TA_{overall} &= \frac{\sum_{domain} N_{correct}^{domain}}{\sum_{domain} N_{total}^{domain}} \times 100\% \end{aligned} \tag{31}$$

$N_{total}^{Medical} = 1200$, $N_{total}^{IT} = 1000$, and $N_{total}^{legal} = 800$. The obtained values are compared with the above baseline model, as shown in Figure 4:

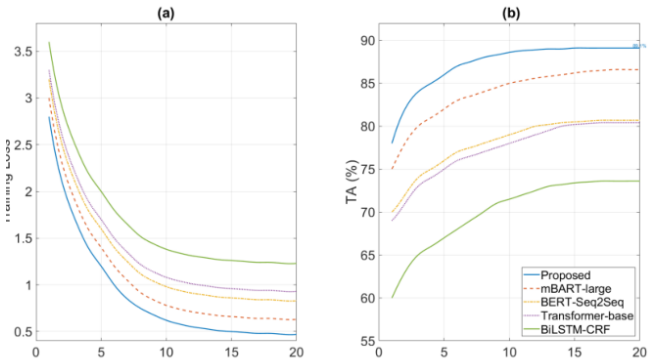

FIGURE 4. Comparison of the proposed model with the baseline model: (a) Training loss curve; (b) Term translation accuracy curve

Figure 4 consists of (a) the training loss curve and (b) the validation term translation accuracy curve. The horizontal axis represents the training round, and the vertical axis represents the training loss value and the TA percentage. The training loss of the proposed model is significantly lower than that of other models. Figure (b) shows that its TA far exceeds the next best performance of mBART-large and Transformer-base.

C. DOMAIN CONSISTENCY SCORE

RoBERTa-large is selected as the backbone model for domain classification and fine-tuned on corpora from the three domains of science, medicine, and law. The Chinese translations generated by all models on the WMT test set are standardized. Then, the pre-processed translations are fed into the fine-tuned RoBERTa-large classifier to obtain its predicted domain labels. The domain consistency score is defined as the classification accuracy:

$$DCS = \frac{N_{consistent}}{N_{total}} \times 100\% \tag{32}$$

The DCS (Domain Consistency Score) is calculated for each domain to analyze the model's domain-preserving capabilities in different domains, allowing to determine whether domain semantic drift is caused by mistranslated terminology or linguistic shift. All models are evaluated using the same RoBERTa classifier to avoid bias introduced by classifier differences. For baselines that do not explicitly model domain information, translations are more susceptible to domain confusion due to the influence of general corpus priors.

D. BLEU SCORE

BLEU-2 and BLEU-3 are somewhat limited in their use and are not as effective as BLEU-4 in indicating the linguistic fluency and structural rationality of a translation. However, they can serve as auxiliary indicators to enhance the depth of analysis. This design uses the Python nltk.translate.bleu_score module to calculate the BLEU

(Bilingual Evaluation Understudy) score. SmoothingFunction().method4 is enabled. This method smoothes short translations or high-order n-gram matches when they are zero, preventing BLEU value distortion caused by zero scores. The BLEU score is calculated for each translation, and then the arithmetic mean is taken for the entire test set to provide the final BLEU score.

E. TERM BOUNDARY F1 VALUE

Each Chinese translation in the WMT test set is annotated character-by-character using the BIO (Beginning–Inside–Outside) scheme. The first character of a term is marked with a "B"; subsequent characters are marked with an "I"; non-term characters are uniformly marked with an "O". For example, if "neuron" is a term in the translation "Neuronal Cells", the annotation sequence is [B, I, I, O, O]. The Python open-source library seqeval is used to perform a character-by-character comparison of the predicted label sequence with the reference label sequence. Precision, recall, and F1 scores are calculated using a standard sequence annotation evaluation protocol. The term boundary F1 score is calculated for each of the three sub-domains of medicine, IT, and law, and the overall weighted average result is reported.

F. CONTEXT SENSITIVITY

128 frequently polysemous terms are manually selected from the WMT test set, each paired with two domain contexts, for a total of 256 examples. A term's translation is considered correct only when it is completely consistent character-level with the reference translation, ensuring that the CSI (Context Sensitivity Index) only measures the model's ability to accurately understand contextual semantics. The CSI is calculated as follows:

$$CSI = \frac{N_{correct}}{N_{total}} \times 100\% \tag{33}$$

$N_{total} = 256$ is the total number of test samples, and $N_{correct}$ is the number of correct translations. All models generate translations using beam search on the same subset.

IV. RESULTS AND DISCUSSION

A. DOMAIN ADAPTATION AND TERM-AWARE INPUT TO IMPROVE TRANSLATION ACCURACY

The experiment uses different models to calculate the translation accuracy of terms in the fields of medicine, information technology, and law, and the results are shown in Figure 5.

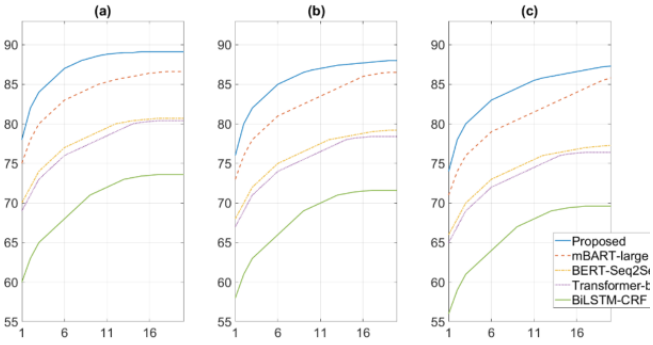

FIGURE 5. TA comparison chart: (a) TA comparison chart in the medical field; (b) TA comparison chart in the information technology field; (c) TA comparison chart in the legal field

Figure 5 shows the term translation accuracy (TA) of the proposed model and four baselines across three domains: medicine, information technology, and law. The proposed model significantly outperforms all domains and converges smoothly, improving from 78% in the first round to 89.1% in medicine, 88.0% in IT, and 87.3% in law. These results underscore the efficiency and stability of our method, largely attributable to domain-adaptive pre-training and explicit term boundary awareness. The former, DA-BERT, is trained on specialized corpora such as medicine and law, giving the model strong domain semantic priors from the outset and shortening the transition from general to specialized semantics. The latter, through the introduction of markers, accurately locates the complete span of multi-word terms, preventing fragmented translation.

B. DEEP DOMAIN KNOWLEDGE INJECTION TO ENSURE TRANSLATION DOMAIN CONSISTENCY

The domain consistency score is a key indicator for measuring whether the query is "semantically correct", as shown in Figure 6.

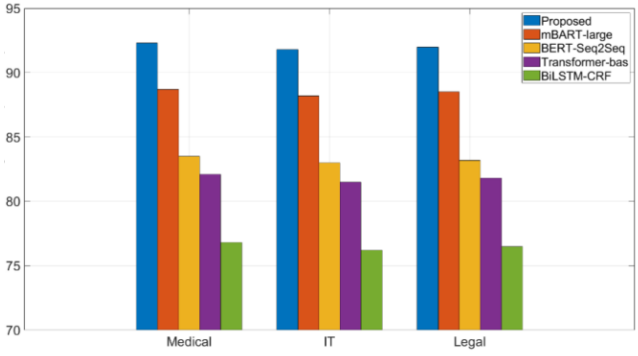

FIGURE 6. DCS value statistics

The data shows that the system's designed model leads in all fields, achieving 92.3% in medicine, 91.7% in IT,

and 92.0% in law. The suboptimal model, mBART-large, scores slightly worse, while traditional models such as Transformer-base and BiLSTM-CRF lag significantly behind.

The feature fusion module nonlinearly integrates the global [CLS] vector of the BERT with the term center representation extracted by the BiLSTM, and uses this to initialize the decoder hidden state. This allows the generated translation to be anchored in the correct domain context, thereby improving DCS. A sampling analysis reveals that the main types of errors in the baseline model include term generalization, mixed use of domain terms, and semantic drift caused by missing context. By introducing term boundary markers and a gated attention mechanism, the proposed model can better preserve domain-specific context and reduce such errors. The proposed model achieves both formal accuracy and semantic attribution in the translation of specialized terminology, thereby improving the domain robustness of terminology translation.

C. OVERALL TRANSLATION QUALITY

Figure 7 compares the BLEU-1 to BLEU-4 scores of the proposed model against four baseline models in three professional domains: medicine, information technology, and law. As the n-gram order increases, the BLEU scores of each model decrease. The proposed model leads across all orders and domains, demonstrating its superiority in generating fluent, accurate, and well-structured translations.

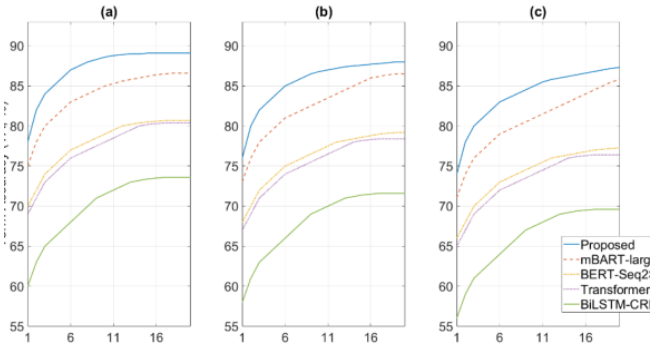

FIGURE 7. BLEU score comparison chart: (a) Comparison chart in the medical field; (b) Comparison chart in the field of information technology; (c) Comparison chart under legal field

The high BLEU-1 score indicates that the proposed model is more accurate in vocabulary selection. Its leading BLEU-2 and BLEU-3 scores reflect the rationality of its word order and phrase structure. The fusion feature zfuse, which jointly anchors global domain semantics and local term centers, maintains its leading BLEU-4 score, indicating that the model can generate complete semantic units that conform to professional language styles. The decoder focuses on the correct professional context from the outset, thereby maintaining terminology accuracy while taking into account the overall coherence of the translation.

D. GATED ATTENTION AND FEATURE FUSION TO EFFECTIVELY ENHANCE BOUNDARY RECOGNITION AND CONTEXT UNDERSTANDING

The term boundary F1 score and context sensitivity measure the model's ability to recognize term structure and respond to contextual differences, respectively. Figure 8 shows the F1 score and context sensitivity values for various fields.

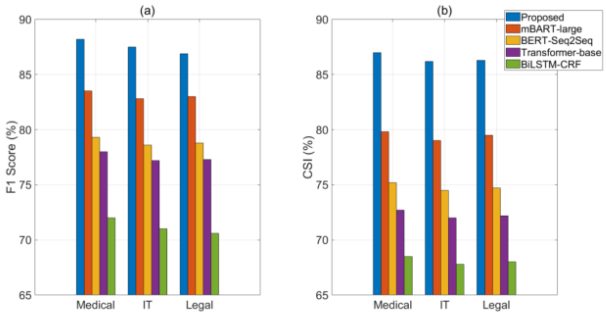

FIGURE 8. Comparison of F1 value and context sensitivity: (a) F1 comparison chart; (b) CSI comparison chart

Figure 8 shows the cross-domain comparison results of the proposed model and four baseline models in terms of term boundary recognition ability and context sensitivity. The proposed model achieves 88.2 %, 87.6%, and 87.0% F1 and 87 %, 86.2%, and 86.3% CSI respectively in terms of term boundary recognition ability and context sensitivity in different fields, surpassing the suboptimal model. This performance breakthrough stems from the deep integration of structure perception and dynamic context modeling. Explicit boundary markers provide the model with prior signals of term structure, while the gated attention BiLSTM builds on this by dynamically adjusting the fusion ratio of local states to global semantics based on context. The feature fusion module further nonlinearly integrates BERT's global semantic summary with BiLSTM's local term representation to generate a highly discriminative initialization state, guiding the decoder to accurately generate domain-adapted translations. The four mechanisms of domain-adaptive pretraining, term structure perception, gated context modeling, and multi-granularity feature fusion work together to construct an end-to-end translation framework that can both "recognize term structure" and "understand term context", alleviating the problems of domain mismatch and semantic drift in the translation of specialized terminology.

V. CONCLUSIONS

This paper proposes an English-Chinese terminology translation model that integrates an improved BERT and BiLSTM architecture, specifically designed to enhance the accuracy of technical terms found in engineering and material science documentation. Through domain-adaptive pretraining, explicit term boundary marking, a gated attention BiLSTM encoder, and multi-granularity feature fusion, this

model systematically addresses the issues of weak domain adaptation and insufficient context modeling. Evaluation results confirm the superiority of our model over mainstream baselines in terminology accuracy, consistency, and robustness. It achieves explicit modeling of terminology structure and context-sensitive generation, and bridges the semantic gap between recognition and translation through end-to-end optimization. This model provides an effective solution for the practical application of machine translation in highly specialized fields such as medicine, law, IT, and the complex technical documentation used in manufacturing and quality control.

AUTHOR CONTRIBUTIONS

Zhongwei Mei designed, collected and analyzed the data, and drafted the manuscript. Zhongwei Mei conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Zhongwei Mei 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 was supported by, This article is a preliminary result of the 2024 Culture Research Special Project of Henan Province Culture Revitalization Program — “Research on Henan Translators from the Perspective of Civilization Mutual Learning (2024XWH167)”.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Associate Professor Zhen Wang for his invaluable assistance. His significant contributions were pivotal to the completion of related experiments, data organization, and literature review.

REFERENCES

[1] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, “Progress in machine translation,” Engineering, vol. 18, pp. 143-153, 2022, doi: 10.1016/j.eng.2021.03.023.

[2] P. Naveen and P. Trojovský, “Overview and challenges of machine translation for contextually appropriate translations,” Iscience, vol. 27, no. 10, 110856, 2024, doi: 10.1016/j.isci.2024.110878.

[3] R. Al-Jarf, “Translation of Arabic Folk Medical Terms with Om and Abu by AI: A Comparison of Microsoft Copilot and DeepSeek,” Journal of Medical and Health Studies, vol. 6, no. 4, pp. 45-58, 2025, doi: 10.32996/jmhs.2025.6.4.8.

[4] X. E. Zhang, “A study of cultural context in Chinese-English translation,” Region-Educational Research and Reviews, vol. 3, no. 2, pp. 11-14, 2021, doi: 10.32629/rerr.v3i2.303.

[5] S. A. Mohamed, A. A. Elsayed, Y. F. Hassan, and M. A. Abdou, “Neural machine translation: past, present, and future,” Neural Computing and Applications, vol. 33, no. 23, pp. 15919-15931. doi: 10.1007/s00521-021-06268-0, 2021.

[6] G. Tang, “Translation of the proper nouns in legal English,” Theory and Practice in Language Studies, vol. 11, no. 6, pp. 711-716, 2021, doi: 10.17507/tpls.1106.15.

[7] Guerberof ArenasA, J. Moorkens, and S. O’Brien, “The impact of translation modality on user experience: an eye-tracking study of the Microsoft Word user interface,” Machine Translation, vol. 35, no. 2, pp. 205-237, 2021, doi: 10.1007/s10590-021-09267-z.

[8] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, “Pre-trained language models and their applications,” Engineering, vol. 25, pp. 51-65, 2023, doi: 10.1016/j.eng.2022.04.024.

[9] B. Min, H. Ross, E. Sulem, et al., “Recent advances in natural language processing via large pre-trained language models: A survey,” ACM Computing Surveys, vol. 56, no. 2, pp. 1-40, 2023, doi: 10.1145/3605943.

[10] É. Arnaud, M. Elbattah, M. Gignon, and D. Gilles, “Learning Embeddings from Free-text Triage Notes using Pretrained Transformer Models,” HEALTHINF, vol. 5, pp. 835-841, 2022, doi: 10.5220/0011012800003123.

[11] Y. Zhu, “A knowledge graph and BiLSTM-CRF-enabled intelligent adaptive learning model and its potential application,” Alexandria Engineering Journal, vol. 91, pp. 305-320, 2024, doi: 10.1016/j.aej.2024.02.011.

[12] S. Arslan, “Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text,” Neural Computing and Applications, vol. 36, no. 15, pp. 8371-8382, 2024, doi: 10.1007/s00521-024-09532-1.

[13] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, “RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis,” IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 9, pp. 1-12, 2025, doi: 10.1109/TETCI.2025.3572150.

[14] Y. Cui and M. Liang, “Automated scoring of translations with BERT models: Chinese and English language case study,” Applied Sciences, vol. 14, no. 5, pp. 1925-1925, 2024, doi: 10.3390/app14051925.

[15] S. Doddapaneni, G. Ramesh, M. Khapra, A. Kunchukuttan, and P. Kumar, “A primer on pretrained multilingual language models,” ACM Computing Surveys, vol. 57, no. 9, pp. 1-39, 2025, doi: 10.1145/3727339.

[16] M. Jiralerspong, J. Bose, I. Gemp, C. Qin, Y. Bachrach, and G. Gidel, “Feature likelihood divergence: evaluating the generalization of generative models using samples,” Advances in Neural Information Processing Systems, vol. 36, pp. 33095-33119, 2023, doi: 10.52202/075280-1436.

[17] P. Zhou, X. Xie, Z. Lin, and S. Yan, “Towards understanding convergence and generalization of AdamW,” IEEE transactions on pattern analysis and machine intelligence, vol. 46, no. 9, pp. 6486-6493, 2024, doi: 10.1109/TPAMI.2024.3382294.

[18] Z. Ren and Z. Zhou, “Dynamic batch learning in high-dimensional sparse linear contextual bandits,” Management Science, vol. 70, no. 2, pp. 1315-1342, 2024, doi: 10.1287/mnsc.2023.4895.

[19] A. Jolly, V. Pandey, I. Singh, and N. Sharma, “Exploring biomedical named entity recognition via SciSpaCy and BioBERT models,” The Open Biomedical Engineering Journal, vol. 18, no. 1, pp. 1-10, 2024, doi: 10.2174/0118741207289680240510045617.

[20] B. Ranjgar, A. Sadeghi-Niaraki, M. Shakeri, F. Rahimi, and S. Choi, “Cultural heritage information retrieval: past, present, and future trends,” IEEE Access, vol. 12, pp. 42992-43026, 2024, doi: 10.1109/ACCESS.2024.3374769.

[21] C. Si, Z. Zhang, Y. Chen, F. Qi, X. Wang, Z. Liu, Y. Wang, and et.al, “Sub-character tokenization for Chinese pretrained language models,” Transactions of the Association for Computational Linguistics, vol. 11, pp. 469-487, 2023, doi: 10.1162/tacl_a_00560.

[22] A. Hernández and J. M. Amigó, “Attention mechanisms and their applications to complex systems,” Entropy, vol. 23, no. 3, pp. 118-147, 2021, doi: 10.3390/e23030283.

[23] S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Mariano, J. T. Chiu, et al., “Simple and effective masked diffusion language models,” Advances in Neural Information Processing Systems, vol. 37, pp. 130136-130184, 2024, doi: 10.52202/079017-4135.

[24] H. Zhang, K. Chen, X. Bai, Y. Xiang, and M. Zhang, “Paying more attention to source context: Mitigating unfaithful translations from large language model,” arXiv preprint arXiv:2406.07036, 2024.

[25] L. Yang, H. Chen, Z. Li, X. Ding, and X. D. Wu, “Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling,” IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 7, pp. 3091-3110, 2024, doi: 10.1109/TKDE.2024.3360454.

[26] M. Abulaish, M. Fazil, and M. J. Zaki, “Domain-specific keyword extraction using joint modeling of local and global contextual semantics,” ACM

Transactions on Knowledge Discovery from Data (TKDD), vol. 16, no. 4, pp. 1-30, 2022, doi: 10.1145/3494560.

[27] S. Kılıçarslan, K. Adem, and M. Çelik, "An overview of the activation functions used in deep learning algorithms," Journal of New Results in Science, vol. 10, no. 3, pp. 75-88, 2021, doi: 10.54187/jnrs.1011739.

[28] S. Abinaya, M. Deepak, and A. S. Alphonse, "Enhanced Image Captioning Using Bahdanau Attention Mechanism and Heuristic Beam Search Algorithm," IEEE Access, vol. 12, pp. 78945-78956, 2024, doi: 10.1109/ACCESS.2024.3431091.

[29] I. R. Visan, "Errors and Difficulties in Translating Maritime Terminology," Translation Studies: Theory and Practice, vol. 3, no. 1, pp. 66-84, 2023, doi: 10.46991/TSTP/2023.3.1.066.

[30] X. Wen and W. Li, "Time series prediction based on LSTM-attention-LSTM model," IEEE access, vol. 11, pp. 48322-48331, 2023, doi: 10.1109/ACCESS.2023.3276628.

[31] S. Lv, S. Lu, R. Wang, L. Yin, Z. Yin, S. A.AIQahtani, et al., "Enhancing Chinese Dialogue Generation with Word– Phrase Fusion Embedding and Sparse SoftMax Optimization," Systems, vol. 12, no. 12, pp. 516-516, 2024, doi: 10.3390/systems12120516.

[32] Y. Hao, Y. Liu, and L. Mou, "Teacher forcing recovers reward functions for text generation," Advances in Neural Information Processing Systems, vol. 35, pp. 12594-12607, 2022, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/51ae7d9db3423ae96cd6afeb01529819-Abstract-Conference.html.

[33] J. Lu and S. Steinerberger, "Neural collapse under cross-entropy loss," Applied and Computational Harmonic Analysis, vol. 59, pp. 224-241, 2022, doi: 10.1016/j.acha.2021.12.011.

[34] J. Choo, D. Kwon Y, J. Kim, J. Jae, A. Hottung, K. Tierney, et al., "Simulation-guided beam search for neural combinatorial optimization," Advances in Neural Information Processing Systems, vol. 35, pp. 8760-8772, 2022, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/39b9b60f0d149eabd1fff2d7c7d5afc4-Abstract-Conference.html.

[35] M. R. Islam, A. A. Lima, S. C. Das, M. F. Mridha, A. R. Prodeep, and Y. Watanobe, "A comprehensive survey on the process, methods, evaluation, and challenges of feature selection," IEEE Access, vol. 10, pp. 99595-99632, 2022, doi: 10.1109/ACCESS.2022.3205618.