Eng Int Syst (2026) 1: 109–118

Engineering
Intelligent Systems

College English Translation Technology Based on Multi-Source Information Corpus Acquisition and Data Fusion

Wentao Meng[1], Lei Yu[2]*, and Yunyun Zhu[3]

[1] *Department of Basic Education, Beihai Campus, Guilin University of Electronic Technology, Beihai 536000, China*
[2] *School of Sports and Arts, Harbin Sport University, Harbin 230000, China*
[3] *Department of Economics and Management, Beihai Campus, Guilin University of Electronic Technology, Beihai 536000, China*

Translation scenarios for college English are moving in interdisciplinary and cross-disciplinary directions, meaning that traditional methods that rely on limited textbook materials are unable to meet the diverse needs of students. To solve the problems of insufficient coverage of multi-source corpora and limited adaptability of translation models, a corpus acquisition framework integrating knowledge discovery and domain discrimination was developed, and a translation model based on data fusion was designed. By using a web crawling system to capture bilingual discourse level corpora in new fields, and combining sentence segmentation strategies and dynamic programming to optimize alignment accuracy, high-quality parallel corpora were generated. At the same time, a bidirectional encoder representation combined with WordPiece model was proposed to enhance sequence annotation performance by integrating part-of-speech and syntactic dependency features. The experiment outcomes showed that the proposed model had an accuracy of 0.92 and a processing time of only 12 seconds. For the translation, the proposed model had an accuracy close to 1.0 after 900 iterations, a false positive rate reduced to 0.05, and a translation time of 16 seconds, significantly better than traditional models. The results indicate that multi-source corpus acquisition and data fusion techniques can effectively enhance the processing capability of translation systems for complex contexts, providing high-precision solutions for interdisciplinary English translation.

Keywords: Multi-source information; Corpus; Data fusion; Bidirectional encoder; WordPiece

1. INTRODUCTION

The development of interdisciplinary education has extended the demand for college English translation from general scenarios to professional fields. However, traditional methods rely on textbook corpora with significant limitations in terms of terminology coverage and genre diversity. The current translation technology needs to address two major challenges: the dynamic acquisition and alignment of multi-source corpora, and the precise capture of semantics in complex contexts. Although existing neural machine translation models perform well in general fields, they are susceptible to data singularity constraints when dealing with specialized texts, leading to mistranslations of terminology, and style deviations. In recent years, research has attempted to optimize translation performance through domain adaptation and data augmentation, fine-tuning cross-domain models by applying methods such as transfer learning, mixed monolingual/bilingual corpus training, etc. [1, 2]. However, these methods still have shortcomings in terms of corpus acquisition efficiency and the lack of a mechanism to update the knowledge base, making it difficult to adapt to the fast iteration characteristics of college English textbooks. In addition, traditional model fusion strategies often suffer from performance

**Corresponding author e-mail: yulei198307252025@163.com

degradation due to parameter conflicts, giving rise to an urgent need for more efficient fusion frameworks [3]. In response to these issues, a two-stage corpus acquisition model of "knowledge discovery domain discrimination" is proposed, which dynamically captures new domain data through a crawler system, combines dependency syntax analysis and dynamic programming optimization alignment accuracy, and constructs an extensible parallel corpus. At the level of translation technology, the Prefix-BART framework is designed to optimize prefixes and integrate multi-source features, enhancing the model's ability to capture professional terminology and syntactic structures. The research aims to provide an efficient translation system that can adapt to interdisciplinary needs, reduce semantic biases in professional text processing, and promote the practical application of machine translation in educational settings. The innovation of the research lies in the proposal of a corpus incremental update mechanism based on lifelong learning, and optimizing the sequence annotation method by integrating part-of-speech and syntactic features.

2. RELATED WORKS

In recent years, deep learning technology has made significant progress in terms of improving the accuracy and fluency of machine translation systems, especially neural network-based translation methods. To improve translation efficiency, Xiang et al. proposed a bidirectional long short-term memory translation model that can extract specific vocabulary from a text corpus and construct a translation graph to achieve topological translation. The research results showed that this model demonstrated superiority in testing, with a translation accuracy of 85.1%, effectively verifying the feasibility and effectiveness of the model in improving translation quality [4]. To improve the quality and efficiency of low resource language machine translation, Liu et al. developed a translation optimization algorithm grounded on biological evolution principles. This algorithm aimed to optimize translation results by simulating the process of biological evolution. The research results indicated that this method significantly improved translation accuracy and enhanced customer satisfaction, providing a new solution for low resource language translation [5]. To improve the accuracy of computer-aided translation in the field of traditional Chinese medicine translation, Xiumin et al. used artificial intelligence to develop a language feature optimization method and constructed a corresponding Chinese medicine corpus. The research results indicated that this method improved the translation efficiency of traditional Chinese medicine consultations and provided strong support for international exchanges in the field of traditional Chinese medicine [6]. To raise the effectiveness of classroom machine translation both domestically and internationally, Lee proposed a multimodal translation task optimization method for translation tasks that are difficult to emotionally judge. The research results indicated that this method significantly improved the acceptance and efficiency of machine translation among teachers and students, providing a more practical and efficient solution for classroom translation [7].

To achieve instance-aware, image-to-image translation, Kim et al. proposed a Transformer-based network architecture. This architecture utilized the self-attention module of Transformers to consider contextual information and achieves multimodal translation with styled code through adaptive instance normalization. The research results indicated that this method achieved efficient image translation across multiple domains with less content loss [8]. Guo et al. found that the encoder-decoder translation framework is prone to information dimensionality reduction and loss. Therefore, researchers proposed a dual generative adversarial network with cross-domain mapping. The experiment findings indicated that this method could effectively translate word length and sentence length sequences of neural activity into speech [9]. Mao et al. found that the problem of continuous translation between images has not been solved. To address this issue, they proposed an effective signature attribute vector that can perform continuous translation on different mapping paths in different domains. The results indicated that this method indeed produced higher quality continuous translation results [10]. Huang et al. found that text ambiguity and irrelevant information in images may lead to translation errors of some keywords in zero resource machine translation. To overcome this, the researchers introduced knowledge entities and used Transformer for assisted learning. The results indicated that this method focused on images and achieved state-of-the-art BLEU scoring in the field of zero resource machine translation [11].

In summary, in the field of college English translation technology, many research teams have conducted in-depth research on multi-source information corpus acquisition and data fusion, and have achieved significant results. Overall, research on college English translation methods is constantly evolving to become more intelligent and precise by integrating various technologies such as natural language processing, deep learning, and knowledge graphs. This study combines knowledge discovery and domain discrimination to obtain corpus, and introduces data fusion technology to optimize translation models, exploring how to use multi-source information to improve the accuracy and generalization ability of college English translation in order to better adapt to complex and changing translation needs.

3. METHODS

3.1 Multi-Source Information Corpus Acquisition in College English Translation

In machine translation research, corpus is key to achieving high-quality translation models. Traditional college English translation often relies on limited textbook texts, teaching aids, or publicly available bilingual corpora [12, 13]. However, with the increasing interdisciplinary, cross disciplinary, and cross-cultural communication, the translation application scenarios of college English have significantly expanded. Learners from different majors or fields face more diverse text types in the translation process, and professional terminology, expression methods, and genres are also vastly different. Therefore, it is crucial to obtain a multi-source information

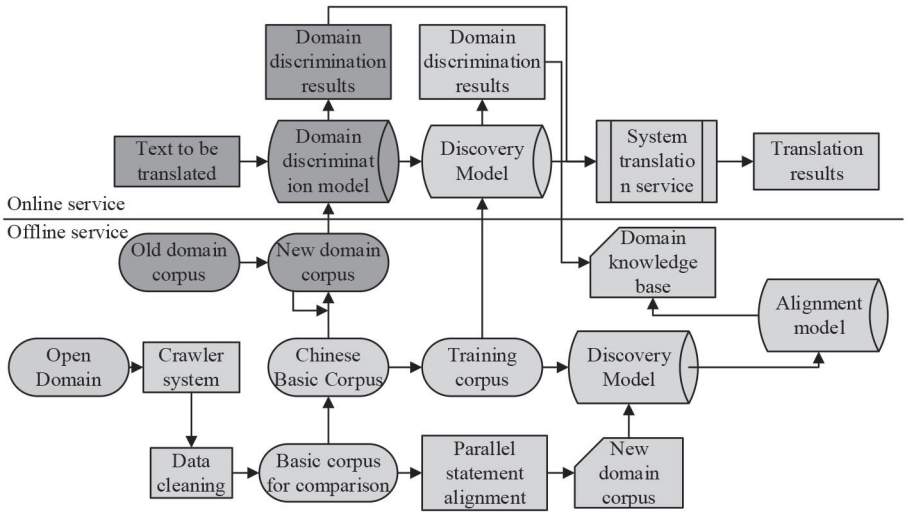

Figure 1 A corpus acquisition model based on knowledge discovery and domain discrimination.

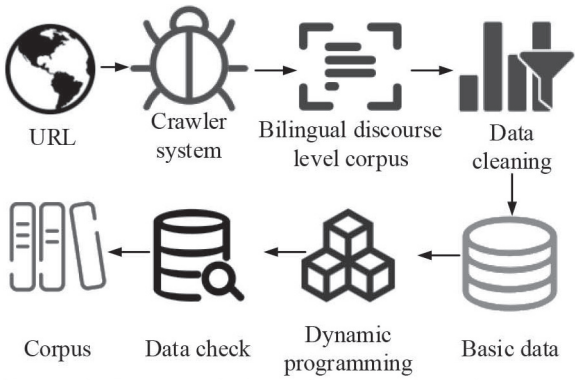

Figure 2 Process of obtaining new domain corpus.

corpus so as to fully cover various professional backgrounds and language styles. A corpus acquisition model based on knowledge discovery and domain discrimination has been proposed in this study. Its structure is shown in Figure 1.

As shown in Figure 1, the online service module processes the text to be translated input by the user. Firstly, the domain discrimination model is used for domain classification to obtain the domain discrimination result, which is then passed on to the source language sentence knowledge discovery model. This model queries the domain knowledge base to obtain domain category reference information, and combines it with the translation service of the lifelong learning translation system to generate the final translation result. The offline task involves domain discrimination and knowledge discovery training of source language sentences. Firstly, the system trains the domain discrimination model using classification corpora from both old and new domains [14–16]. The language data of the new field is obtained by means of a crawler system, and after data cleaning, it forms the basic Chinese corpus of the new field. Then, parallel sentences are aligned with the existing bilingual comparable basic corpus to generate a parallel corpus of the new field. Subsequently, the parallel corpus is used to train a source language sentence knowledge discovery model, which further learns and participates in the optimization of a lifelong learning translation system. At the same time, the system establishes a knowledge alignment model for bilingual parallel corpora, uses this model for knowledge discovery, stores the obtained information in the domain knowledge base, and feeds it back to the translation system.

Due to the changes in university textbooks and the complexity of their content, the model must be able to continuously acquire new domain corpora and, therefore, also needs to obtain different parallel corpora that can be used for translation from the corpus. Figure 2 illustrates how the new domain corpus is obtained.

As shown in Figure 2, firstly, the crawler system retrieves relevant data from the new domain corpus website, obtains bilingual discourse level corpus, and then cleans the data to remove invalid information, formatting errors, or low-quality data, ensuring the quality of the corpus. After data cleaning, the obtained bilinguals enter the next processing stage, where a sentence splitting strategy is adopted and optimized using vector representation and dynamic programming methods to improve the accuracy and usability of data matching [17]. After this step, the system checks and filters the data, removing low-quality or noisy data to ensure the quality and alignment accuracy of the corpus. Finally, the filtered data is stored in a new domain parallel corpus, providing high-quality training data for subsequent machine translation, knowledge

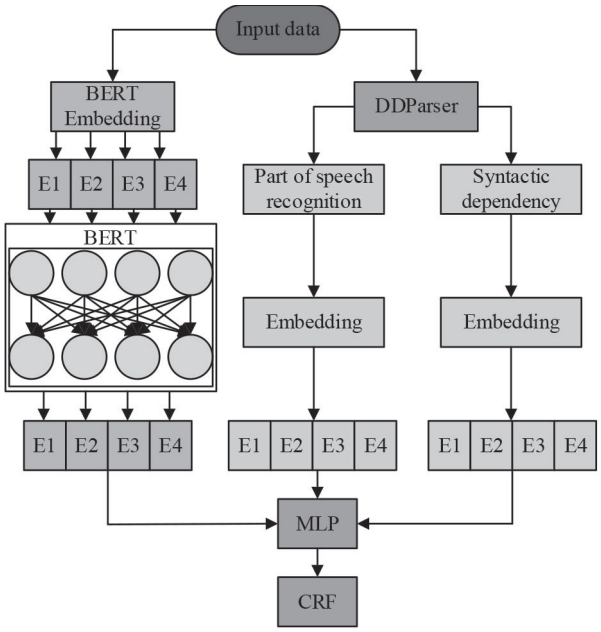

Figure 3 BERT-WordPiece model structure diagram.

discovery, and domain discrimination. In the optimization of sequence annotation tasks, a method combining Bidirectional Encoder Representation from Transformers (BERT) with WordPiece is studied, and its structure is shown in Figure 3.

As shown in Figure 3, the input text first undergoes a BERT word-embedding process to generate the basic word vector representation, which is expressed as equation (1).

$$H = \text{BERT}(X) \tag{1}$$

In equation (1), X represents the input text sequence, and H is the word vector representation generated by BERT. At the same time, the text is subjected to syntactic dependency analysis using Dependency Parser (DDParser) to obtain syntactic dependency structures, and Part-of-Speech Tagging (POS) modules are used to obtain part-of-speech tags [18]. These pieces of information are processed separately through part-of-speech embedding and syntactic dependency embedding to generate part of speech feature vectors and syntactic dependency feature vectors, whose expressions are shown in equation (2).

$$\begin{cases} E_{pos} = \text{Embedding}(POS) \\ E_{dep} = \text{Embedding}\ (Dependency) \end{cases} \tag{2}$$

In equation (2), E_{pos} represents the part of speech feature vector, and E_{dep} represents the syntactic dependency feature vector. Subsequently, these feature vectors from different sources are concatenated to form a comprehensive representation, which is expressed as equation (3).

$$E_{\text{concat}} = [H; E_{\text{pos}}; E_{\text{dep}}] \tag{3}$$

In equation (3), [;] represents the vector concatenation operation. The comprehensive representation is then input into a Multi-Layer Perceptron (MLP) for feature extraction, and the hidden state vector is calculated, as shown in equation (4).

$$E_{\text{mlp}} = \text{MLP}(E_{\text{concat}}) \tag{4}$$

In equation (4), E_{mlp} represents the implicit state vector. Finally, the implicit state vector is passed to the Conditional Random Field (CRF) for sequence annotation prediction, and the score function for sequence annotation is calculated, as shown in equation (5).

$$S(X, Y) = \sum_{i=1}^{n} \psi(E_{\text{mlp},i}, Y_i) + \sum_{i=1}^{n-1} \phi(Y_i, Y_{i+1}) \tag{5}$$

In equation (5), $\psi(E_{\text{mlp},i}, Y_i)$ represents the annotation score of the current word, and $\psi(E_{\text{mlp},i}, Y_i)$ represents the transfer score between labels. Finally, the optimal sequence label is found through the Viterbi decoding algorithm, and its expression is shown in equation (6).

$$Y^* = \arg\max_{Y} S(X, Y) \tag{6}$$

In equation (6), Y^* represents the optimal sequence label. By integrating BERT semantic features, part of speech information, and syntactic dependency information, the ability to understand text has been enhanced, and the performance of sequence annotation tasks has been improved. The overall architecture comprises the complete process from input text to final annotated output, where BERT is responsible for capturing contextual semantics, DDParser provides syntactic structure support, MLP extracts deep features, and CRF further optimizes sequence prediction to ensure the coherence and accuracy of output results.

3.2 University English Translation Technology Based on Data Fusion

After obtaining a multi-source information corpus, the translation technology of data fusion is used to translate college English. In the development of intelligent translation technology, data fusion has become an important means of improving

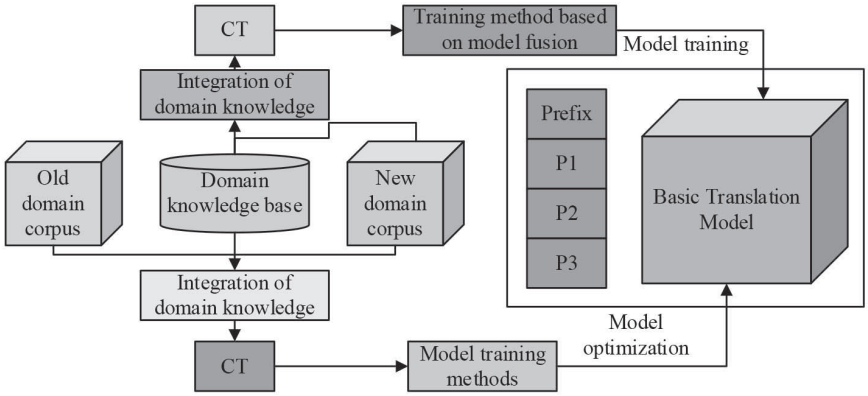

Figure 4 Machine translation model based on data fusion and model fusion.

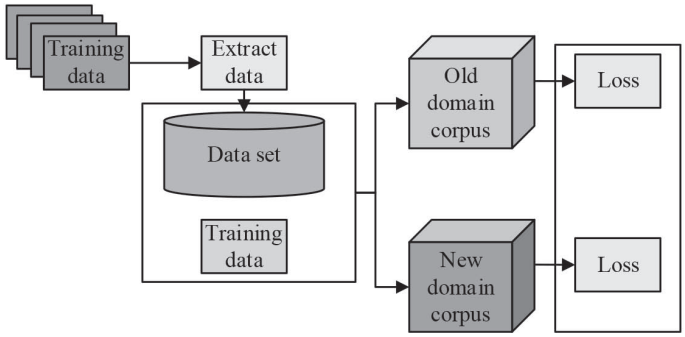

Figure 5 Process flow of translation model training method based on instance fusion.

the performance of translation systems. Traditional statistical machine translation and neural machine translation methods are often limited by the constraints of a single data source when dealing with complex contexts and polysemous words, making it difficult to accurately capture contextual semantics [19]. Data fusion technology integrates multiple language resources, including parallel corpora, monolingual corpora, POS, syntactic dependencies, and semantic knowledge bases, to improve the model's understanding of language features and optimize translation quality. Therefore, a machine translation model grounded on data fusion and model fusion is proposed. The model structure is depicted in Figure 4.

As shown in Figure 4, in this process, the input translation task is first integrated into the knowledge base by means of two domain knowledge fusion methods. Next, the model is optimized through a model training step using training data from different domains. This training data includes parallel corpora from different domains and has been adaptively adjusted by applying model fusion strategies [20]. Subsequently, a machine translation model for the new field is generated through model-based tuning. This model combines previous models with the knowledge features of the current field. For the final evaluation of the model, the translation system outputs the final translation result based on the prefix and trained translation parameters. In this model, due to the complexity and diversity of college English, there may be a problem of decreased translation performance. Therefore, a translation model training method based on instance fusion is proposed. Its training process is shown in Figure 5.

As shown in Figure 5, firstly, the old training data is collected and processed through the dataset extraction module, and combined with the knowledge base. Then, this old data is preprocessed by the dataset to extract useful features, forming training data that is input into the translation model. The new training data is also processed similarly and input into the translation model to generate translation outputs. During the training process, both domain models generate translation results and calculate the loss function. The loss function of the old translation model is obtained with equation (7).

$$Loss_{\text{old}} = \sum_{i=1}^{n} Loss(Y_{\text{old},i}, \hat{Y}_{\text{old},i}) \tag{7}$$

In equation (7), $Y_{\text{old},i}$ means the true translation of the ith word, $\hat{Y}_{\text{old},i}$ means the translation generated by the model, and the loss function evaluates the performance of the model by calculating the difference between the output and the target. The loss function of the new translation model is shown in equation (8).

$$Loss_{\text{new}} = \sum_{m}^{i=1} \text{Loss}(Y_{\text{new},i}, \hat{Y}_{\text{new},i}) \tag{8}$$

In equation (8), $Y_{\text{new},i}$ denotes the real translation of the ith word and $\hat{Y}_{\text{new},i}$ denotes the translation generated by the model. The loss function evaluates the performance of the model by calculating the difference between the output and the target. Finally, the two losses are combined to obtain the total loss function of the model, as denoted in equation (9).

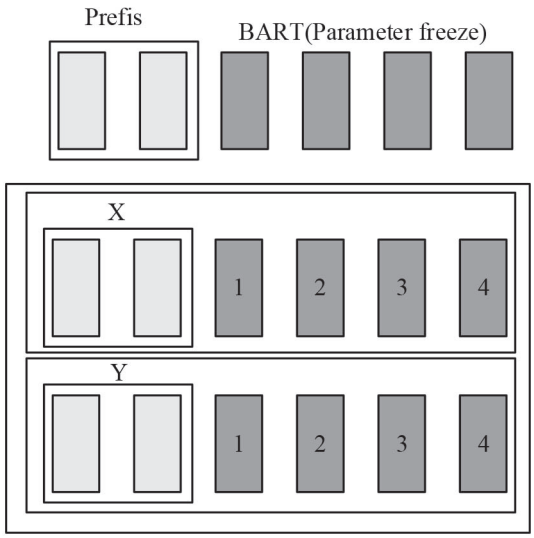

Figure 6 Translation model structure based on Prefix-BART.

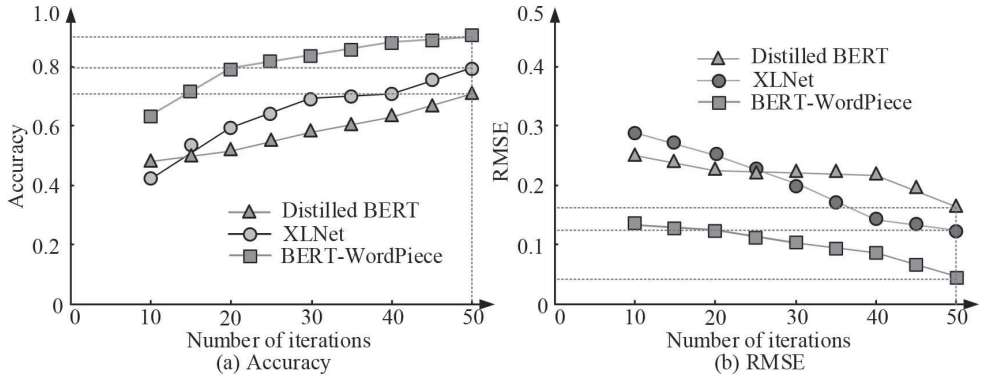

Figure 7 Comparison of accuracy and RMSE of various models.

$$Loss_t = Loss_{\text{old}} + Loss_{\text{new}} \tag{9}$$

In equation (9), $Loss_t$ represents the total loss function value, $Loss_{\text{old}}$ represents the loss function of the old translation model, and $Loss_{\text{new}}$ represents the loss function of the new translation model. The research uses Bidirectional and Auto-Regression Transformers (BART) models as the main models for machine translation. Their fusion form is shown in Figure 6.

As shown in Figure 6, the characteristics of the Prefix-tuning method are combined with the BART model for optimization. In this framework, the input text is first processed through multiple prefixes, which provide domain-specific contextual information for subsequent models. These prefixes are input into the BART model for further processing, where the BART model serves as a generative model that utilizes these prefixes to generate sequences related to translation tasks. Prefix-tuning improves the model's adaptability and efficiency in translation tasks by learning prefixes related to a specific task, making BART more focused on that task-specific information during processing. Within the model, the hidden state represents the encoding results of the input at different levels, while the positional index indicates the relative position of each word in the sentence. When the prefix is fed into the model along with the input text, the sentence representation is obtained through positional indexing and word vector indexing. During the actual translation process, the input sequence and target sequence are fed into the model, which generates corresponding outputs.

4. RESULTS

4.1 The Performance of Multi-Source Information Corpus Acquisition Model for College English Translation

The central processor used in the experimental hardware configuration was Intel Core i5-8750H, the graphics processor was NVIDIA Geforce GTX2080Ti, the video memory was 8GB, the memory was 16GB, and the operating system was Windows 10 system. The dataset comprised the publicly available TOEFL dataset, which consists of actual TOEFL test questions covering listening comprehension, reading comprehension, writing tasks, and speaking tasks. Each section was accompanied by professional ratings and annotations. The study selected Distilled BERT and Generalized Autoregressive Pretraining for Language Understanding (XLNet) as comparative models. The experiment results are shown in Figure 7.

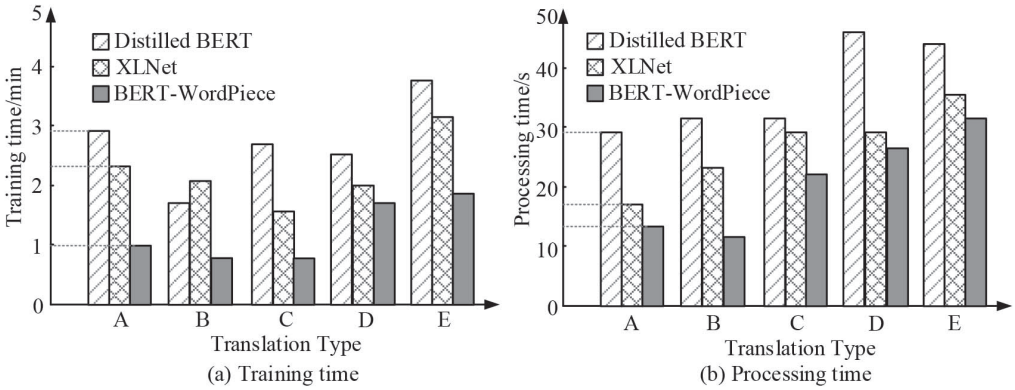

Figure 8 Efficiency analysis of each model.

Table 1 Comprehensive performance analysis of various models.

Model	Processing Time (s)	Accuracy	RMSE	Precision
Distilled BERT	22	0.75	0.2	0.8
XLNet	31	0.85	0.35	0.85
BERT-WordPiece	12	0.92	0.25	0.9
Model	Recall	F1-Score	AUC	Training Accuracy
Distilled BERT	0.7	0.741	0.82	0.8
XLNet	0.8	0.825	0.88	0.85
BERT-WordPiece	0.85	0.875	0.91	0.95

Figure 7(a) showcases the variation of accuracy of different algorithms with increasing iteration times. Figure 7(b) showcases the variation of root mean square error (RMSE) for different algorithms. In Figure 7(a), with the increase of iteration times, the accuracy of all three algorithms showed an upward trend. Among them, BERT-WordPiece had the most stable and high accuracy growth, reaching around 0.92 after 50 iterations, far higher than the other two algorithms. The accuracy of Distilled BERT was relatively low at 0.71, and although the growth rate was fast in the initial iterations, it tended to be flat. The accuracy of XLNet was slightly lower than BERT-WordPiece, but higher than Distilled BERT, at approximately 0.81. Figure 7(b) shows that with the increase of iteration times, the RMSE of all algorithms gradually decreases, indicating that the prediction error of the model was decreasing. The RMSE of Distilled BERT showed the most significant decrease, dropping to 0.15 after 50 iterations, demonstrating relatively stable performance. The RMSE value of BERT-WordPiece was relatively low, remaining at 0.05, while the RMSE of XLNet was close to 0.12. The experimental results showed that the proposed model had excellent performance. The results of the efficiency analysis are shown in Figure 8.

Figures 8(a) and 8(b) show the performance of three algorithms in terms of training time and processing time, respectively. In Figure 8(a), the training time of BERT-WordPiece was generally low, especially in translation types A and C, where the training time was significantly shorter than the other two algorithms, about 1 minute. The training time of BERT-WordPiece also performed well in translation type D, lower than XLNet, about 2 minutes, while the training time of Distilled BERT was generally longer, at a higher level in all translation types. From Figure 8(b), the processing time of BERT-WordPiece was generally short, especially in translation types A and B, with a processing time of nearly 10 seconds. However, the processing time of Distilled BERT and XLNet was generally higher than that of BERT-WordPiece, especially in translation type E, where the processing time of Distilled BERT reached 48 seconds. The experiment outcomes showed that the proposed method had excellent performance. The results for the comprehensive performance analysis of each model are given in Table 1.

According to Table 1, BERT-WordPiece performed the best in terms of processing time, with a processing time of 12 seconds, significantly lower than XLNet's 31 seconds and Distilled BERT's 22 seconds. In terms of accuracy, the performance of BERT-WordPiece was outstanding, reaching 0.92, far higher than the other two models. The accuracy of Distilled BERT was relatively low, at 0.75, while XLNet had an accuracy of 0.85, indicating that BERT-WordPiece had higher accuracy in handling complex tasks. In terms of RMSE, Distilled BERT had the lowest error of about 0.2, indicating good prediction accuracy, while XLNet had a higher error of 0.35, indicating poor prediction accuracy. In terms of accuracy and recall, BERT-WordPiece also performed well, with an accuracy of 0.9 and a recall of 0.85, demonstrating its strong ability to correctly predict and recognize actual positive examples. The experimental results showed that BERT-WordPiece performed well in all metrics, especially leading in accuracy, recall, and training accuracy.

4.2 The Performance of College English Translation Model Based on Data Fusion

To investigate the performance of the Prefix-BART model, BART and ProphetNet were used as comparative models. The results are shown in Figure 9.

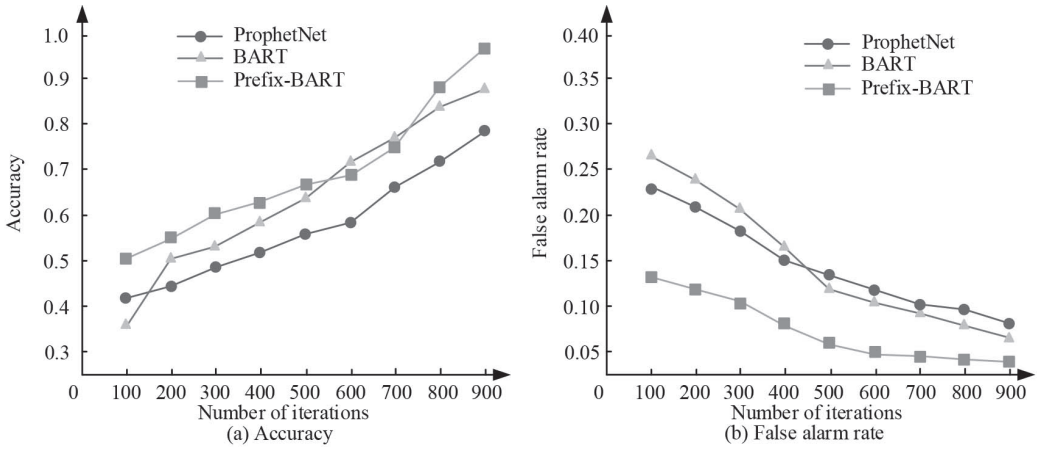

Figure 9 Comparison of accuracy and false alarm rate of various translation models.

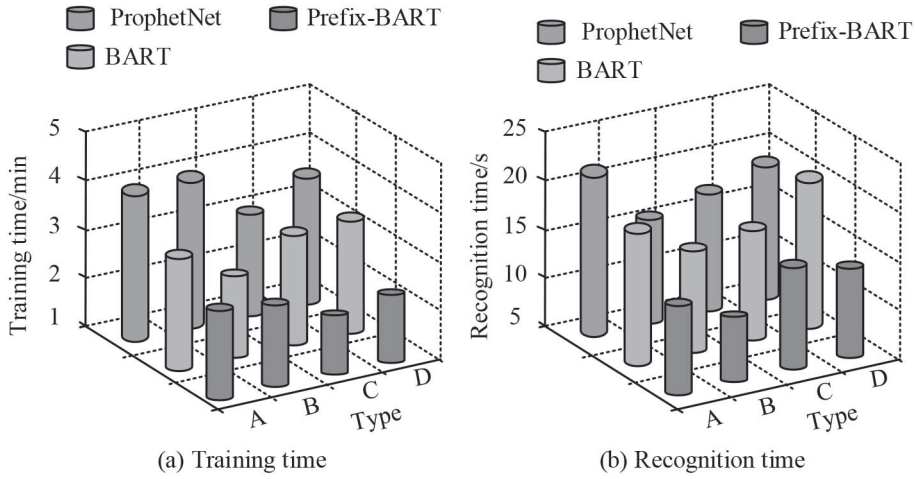

Figure 10 Analysis of the operational efficiency of each model.

Figure 9(a) shows the accuracy changes of three models at different iteration times, and Figure 9(b) shows the changes in false alarm rates (FARs) at different iteration times. According to Figure 9(a), as the number of iterations increased, the accuracy of all three models increased. Among them, the accuracy of Prefix-BART increased the fastest, approaching 1.0, and reached its highest value after 900 iterations. ProphetNet also had a high accuracy of about 0.85, but its relative growth rate was relatively slow. Although the accuracy of BART gradually increased to 0.75 after 900 iterations, it still did not reach the level of Prefix-BART and ProphetNet. In Figure 9(b), with the increase of iteration times, the FARs of all three models showed a decreasing trend. The FAR of Prefix-BART showed the most significant decrease, approaching 0.05 after 900 iterations, demonstrating extremely high accuracy. The FARs of ProphetNet and BART also gradually decreased with the increase of iteration times, but even after 900 iterations, the FAR of BART was still higher than that of Priorit-BART, about 0.09, while ProphetNet's FAR was slightly lower, close to 0.1. The experiment findings indicated that the proposed Prefix-BART model had excellent performance. Figure 10 shows the results for the efficiency analysis of each model.

Figures 10(a) and 10(b) show the training time and translation time of three different types of models. In Figure 10(a), the training time of ProphetNet was relatively long, especially in types A and B, which was about 3-4 minutes, reflecting its structural complexity and computational burden. The Prefix-BART exhibited high training efficiency, typically lasting around 2 minutes, demonstrating a relatively optimized training process. The training time of BART was also relatively short, close to the Prefix-BART, maintained within 2 minutes, demonstrating high training efficiency. From Figure 10(b), ProphetNet had a significant time increase in identifying complex attack types, especially close to 25 seconds under type D, suggesting its heavy computational burden. The Prefix-BART and BART models exhibited relatively stable recognition times, with Prefix-BART around 15 seconds and BART around 12 seconds, indicating that these two models have high computational efficiency and strong adaptability and stability when dealing with different types of attacks. The experiment outcomes showed that the proposed translation model had excellent performance. The comprehensive performance of each model was analyzed, and the findings are presented in Table 2.

As shown in Table 2, BART had the shortest translation time, only 12 seconds, demonstrating a fast response speed. The Prefix-BART followed closely behind, with a translation time of 16 seconds, while ProphetNet had the longest translation time of 23 seconds, possibly due to its complex

Table 2 Comprehensive performance analysis of the model.

Model	Translation time(s)	Accuracy (%)	Precision (%)
ProphetNet	23	89.5	85.7
Prefix-BART	16	91.2	87.5
BART	12	88.3	83.9
Model	Recall (%)	F1-score	AUC
ProphetNet	91.2	0.88	0.92
Prefix-BART	92.1	0.89	0.94
BART	90.5	0.86	0.90

model structure. In terms of accuracy, Prefix-BART led with 91.2%, slightly higher than ProphetNet's 89.5% and BART's 88.3%. In terms of precision, Prefix-BART also performed the best, reaching 87.5%, higher than ProphetNet's 85.7% and BART's 83.9%. In terms of recall rate, the Prefix-BART model was also superior to the other two models, reaching 92.1%, indicating that it can better identify positive cases. In terms of F1-score and AUC indicators, the Prefix-BART still dominated, with F1-score of 0.89 and AUC of 0.94, higher than ProphetNet's 0.88 and 0.92 and BART's 0.86 and 0.90, respectively. The experiment findings denoted that the Prefix-BART performed well in accuracy, precision, recall, F1-score, and AUC.

5. CONCLUSION

In response to the dual challenges of insufficient coverage of professional corpora and inefficient processing of polysemous words in college English translation, a high-performance translation framework was constructed through dynamic acquisition of multi-source corpora and data-fusion technology. Using a crawler system to capture bilingual language materials in new fields, high-quality training data was generated through sentence alignment and noise filtering. The experiment findings indicated that combining the BERT-WordPiece model to achieve joint optimization of part of speech and syntactic features improved the accuracy of sequence annotation tasks to 0.92. In the translation model, Prefix-BART integrated domain knowledge through prefix optimization, reducing FAR by 44% compared to BART, achieving an accuracy rate of 91.2%, and improving processing efficiency by 33%. BART had the shortest translation time of only 12 seconds, demonstrating a fast response speed. The Prefix-BART followed closely behind, with a translation time of 16 seconds, while ProphetNet had the longest translation time, reaching 23 seconds, possibly due to its complex model structure. In terms of accuracy, Prefix-BART led with 91.2%, slightly higher than ProphetNet's 89.5% and BART's 88.3%. In terms of precision, Prefix-BART also performed the best, reaching 87.5%, higher than ProphetNet's 85.7% and BART's 83.9%. In terms of recall rate, the Prefix-BART model was also superior to the other two models, reaching 92.1%, indicating that it can better identify positive cases. The research outcomes indicated that the proposed model had excellent translation performance. However, there are still limitations to the research, as the experimental data comprised TOEFL exam texts and the generalization ability to unstructured professional literature needs to be validated. The real-time response performance of Prefix-BART may decrease when processing text that is excessively long. Future work could introduce reinforcement learning to optimize the corpus retrieval path, explore lightweight model deployment solutions, and expand to interdisciplinary translation scenarios for small languages.

FUNDING

This work was supported by the 2025 Guangxi Higher Education Undergraduate Teaching Reform Project "AI-Empowered Construction and Practice of a Four-Dimensional Evaluation System for College English Courses in Guangxi" (No.:2025JGA215); and the 2025 Guangxi Philosophy and Social Sciences Research Project "Digital Translation and International Brand Communication: A Collaborative Mechanism for the Pinglu Canal's Intangible Cultural Heritage" (No.: 25WYF460).

HUMAN ETHICS AND CONSENT TO PARTICIPATE

Not applicable.

CLINICAL TRIAL

Not applicable.

REFERENCES

1. Houssein E H, Hammad A, Ali A A. Human emotion recognition from EEG-based brain–computer interface using machine learning: A comprehensive review. Neural Computing and Applications, 2022, 34(15): 12527–12557.
2. Chakrawarti R K, Bansal J, Bansal P. Machine translation model for effective translation of Hindi poetries into English. Journal of Experimental & Theoretical Artificial Intelligence, 2022, 34(1): 95–109.
3. Khatri M R. Integration of natural language processing, self-service platforms, predictive maintenance, and prescriptive analytics for cost reduction, personalization, and real-time insights customer service and operational efficiency. International Journal of Information and Cybersecurity, 2023, 7(9): 1–30.

4. Xiang Y, Chen Y, Ye F H. Enhancing computer-aided translation system with BiLSTM and convolutional neural network using a knowledge graph approach. Journal of Supercomputing, 2024, 80(5): 5847–5869.
5. Liu X, Chen J, Zhang Q T. Exploration of low-resource language-oriented machine translation system of genetic algorithm-optimized hyper-task network under cloud platform technology. Journal of Supercomputing, 2024, 80(3): 3310–3333.
6. Xiumin T, Yifu S, Zheng Y. Exploration of computer-assisted translation technology in translating technical terms in traditional Chinese medicine under the perspective of AI vision. Journal of Artificial Intelligence Practice, 2023, 6(7): 37–42.
7. Lee S M. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. Computer Assisted Language Learning, 2023, 36(1–2): 103–125.
8. Kim S, Baek J, Park J, Ha J, Jung E, Lee H, Kim T. InstaFormer: Multi-Domain Instance-Aware Image-to-Image Translation with Transformer. International Journal of Computer Vision, 2024, 132(4): 1167–1171.
9. Guo Y, Liu T, Zhang A W W. End-to-end translation of human neural activity to speech with a dual-dual generative adversarial network. Knowledge-Based Systems, 2023, 277(10): 1–11.
10. Mao Q, Tseng H Y, Lee H Y, Huang J B, Ma S, Yang M H. Continuous and Diverse Image-to-Image Translation via Signed Attribute Vectors. International Journal of Computer Vision, 2022, 130(2): 517–549.
11. Huang P, Zhao J, Sun S, Lin Y. Knowledge enhanced zero-resource machine translation using image-pivoting. Applied Intelligence, 2022, 53(7): 7484–7496.
12. Sitender, Bawa S, Kumar M, Sangeeta S. A comprehensive survey on machine translation for English, Hindi and Sanskrit languages. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(4): 3441–3474.
13. You H. Elevating English translation strategies of children's picture books through deep learning and artificial intelligence. Engineering Intelligent Systems, 33(1): 69–75, 2025.
14. Chen Y. Analyzing the design of intelligent English translation and teaching model in colleges using data mining. Soft Computing, 2023, 27(19): 14497–14513.
15. Uc-Cetina V, Navarro-Guerrero N, Martin-Gonzalez A, Weber C, Wermter S. Survey on reinforcement learning for language processing. Artificial Intelligence Review, 2023, 56(2): 1543–1575.
16. Alzubaidi M A, Otoom M, Abu Rwaq A M. A novel assistive glove to convert Arabic sign language into speech. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(2): 1–16.
17. Wang N. Personalized learning of college English using knowledge graphs combined with user portraits. Engineering Intelligent Systems, 2025, 33(3): 339–343.
18. Zhong Y, Yue X. On the correction of errors in English grammar by deep learning. Journal of Intelligent Systems, 2022, 31(1): 260–270.
19. Zhao X, Jiang Y. Synchronously improving multi-user English translation ability by using AI. International Journal of Artificial Intelligence Tools, 2022, 31(4): 2240–2247.
20. Simon K, Vicent M, Addah K, Bamutura D, Atwiine B, Nanjebe D, Mukama A O. Comparison of deep learning techniques in detection of sickle cell disease. AIA, 2023, 1(4): 252–259.