electronics

MDPI

Article

Enhanced Semantic BERT for Named Entity Recognition in Education

Ping Huang *, Huijuan Zhu, Ying Wang, Lili Dai and Lei Zheng

College of Computer and Artificial Intelligence, Nanjing University of Science and Technology Zijin College, Nanjing 210023, China; zhuhuijuan@njust.edu.cn (H.Z.); wangying826@njust.edu.cn (Y.W.); dailili992@njust.edu.cn (L.D.); zhenglei@hhu.edu.cn (L.Z.)

* Correspondence: huangping984@njust.edu.cn; Tel.: +86-025-85786129

Abstract

To address the technical challenges in the educational domain named entity recognition (NER), such as ambiguous entity boundaries and difficulties with nested entity identification, this study proposes an enhanced semantic BERT model (ES-BERT). The model innovatively adopts an education domain, vocabulary-assisted semantic enhancement strategy that (1) applies the term frequency–inverse document frequency (TF-IDF) algorithm to weight domain-specific terms, and (2) fuses the weighted lexical information with character-level features, enabling BERT to generate enriched, domain-aware, character–word hybrid representations. A complete bidirectional long short-term memory-conditional random field (BiLSTM-CRF) recognition framework was established, and a novel focal loss-based joint training method was introduced to optimize the process. The experimental design employed a three-phase validation protocol, as follows: (1) In a comparative evaluation using 5-fold cross-validation on our proprietary computer-education dataset, the proposed ES-BERT model yielded a precision of 90.38%, which is higher than that of the baseline models; (2) Ablation studies confirmed the contribution of domain-vocabulary enhancement to performance improvement; (3) Cross-domain experiments on the 2016 knowledge base question answering datasets and resume benchmark datasets demonstrated outstanding precision of 98.41% and 96.75%, respectively, verifying the model's transfer-learning capability. These comprehensive experimental results substantiate that ES-BERT not only effectively resolves domain-specific NER challenges in education but also exhibits remarkable cross-domain adaptability.

Keywords: NER; BERT; semantic enhancement; BiLSTM; CRF; focal loss

Academic Editor: Arkaitz Zubiaga

Received: 27 August 2025
Revised: 3 October 2025
Accepted: 4 October 2025
Published: 7 October 2025

Citation: Huang, P.; Zhu, H.; Wang, Y.; Dai, L.; Zheng, L. Enhanced Semantic BERT for Named Entity Recognition in Education. *Electronics* **2025**, *14*, 3951. https://doi.org/10.3390/electronics14193951

1. Introduction

Named entity recognition (NER) in the educational domain serves as a core natural language processing (NLP) application, providing crucial support for educational informatization by automatically identifying key entities—such as subject concepts, institutions, and teaching resources—in educational texts [1,2]. By extracting structured information from unstructured data, NER supports advanced applications such as relation extraction [3], question answering [4], and knowledge graph construction [5]. Chinese educational NER faces specific challenges due to linguistic characteristics. The inherent ambiguity of Chinese word boundaries necessitates simultaneous attention to word segmentation and entity detection. Furthermore, Chinese named entities often exhibit a coarser granularity than

words, ranging from single characters to multi-character compounds. Consequently, accurate boundary identification and enriched lexical semantics are critical for improving model performance [6,7].

The prevalence of specialized terminology, abbreviations, and compound words in the Chinese educational domain gives rise to significant challenges, including entity boundary ambiguity and nested entity structures. The phrase "University Physics," for example, represents an indivisible course entity yet is susceptible to misinterpretation as a generic descriptive phrase. Furthermore, compound terms such as "Artificial Intelligence Python Practice" exhibit nesting, where the core entity contains subordinate entities denoting a field and a pedagogical method, respectively. The accurate resolution of boundary ambiguity necessitates precise contextual comprehension, whereas the identification of nested entities requires capabilities for multi-granular semantic recognition. BERT provides a transformative solution through its bidirectional Transformer architecture, which acquires dynamic, context-sensitive word representations via deep pre-training, thereby facilitating nuanced boundary detection and hierarchical analysis. The model's multi-layer attention mechanism additionally enables the learning of features at varying granularities, offering inherent support for nested entity recognition. Together, these capabilities establish a consolidated and powerful semantic modeling foundation for addressing ambiguity and structural complexity in specialized domains.

To address the challenges of ambiguous boundaries and nested entity recognition in the educational domain NER, we propose an enhanced semantic BERT model (ES-BERT). By fusing domain-specific lexical knowledge from educational dictionaries with character-level features, ES-BERT enriches semantic representations. Experiments demonstrate that this fusion strategy mitigates semantic gaps and boundary ambiguity in Chinese character-level NER, achieving improvement in F1 score over baseline models. The main contributions of this study are threefold:

- Model Architecture: Building upon BERT, we propose ES-BERT (enhanced semantic BERT), a hybrid-fusion model that combines domain-specific lexical features with character-level representations. This dual-path design synergistically enhances granular semantic encoding, providing a theoretical foundation for educational NER performance improvement.
- Training Objective: To address class imbalance (non-entity token dominance) and boundary detection challenges, we integrate a novel focal loss-based joint training objective. This approach optimizes label constraints and boundary representations, significantly improving accuracy for boundary-sensitive entities.
- Resource Construction: We manually curate a computer education corpus with 1000 expert-annotated research abstracts, processed through tokenization and rigorous annotation protocols. Additionally, we compile a domain-specific dictionary to mitigate the current lack of linguistic resources in this field.

2. Related Work

The rapid advancement of named entity recognition (NER) has led the research community to favor hybrid architectures that synergistically combine neural networks with statistical learning. A prime example is the BERT+BiLSTM+CRF framework, which accomplishes efficient sequence labeling through a multi-stage complementary process: BERT generates contextually rich character-level embeddings; BiLSTM captures long-range bidirectional dependencies to enhance feature coherence; and CRF optimizes the label sequence globally by modeling transition constraints, thereby preventing invalid predictions. By integrating deep representation learning with structured prediction, this framework exhibits robust adaptability and stability across diverse NER domains. Dong et al. [8] proposed

BERT-WWM-EXT, which incorporates Whole Word Masking during pre-training to enhance hierarchical representation learning. When integrated with BiLSTM-CRF, this approach demonstrated superior boundary detection capabilities in judicial texts. Hu et al. [9] proposed a BERT-BiLSTM-CRF-based named entity recognition method for core educational technology courses to enhance learning efficiency, reporting improved performance over conventional CRF and BiLSTM-CRF models. Xie et al. [10] enhanced the BERT-BiLSTM-CRF paradigm with contextual-semantic enhancement layers, setting new benchmarks on the MSRA and People's Daily corpora, particularly for polysemous entities. During the ongoing evolution of the BERT model, a number of efficient variants have emerged—represented by RoBERTa [11] and ALBERT [12]—which have been widely applied in named entity recognition tasks.

With the advancement of research, semantic enhancement techniques have demonstrated effectiveness in named entity recognition (NER). Numerous studies have explored the integration of character-level, word-level, and semantic information to enhance the performance of Chinese NER tasks. Chen et al. [13] aggregated multiple word-level features associated with characters and incorporated them into BERT for text feature enhancement, successfully addressing Chinese nested named entity recognition. Liu et al. [14] proposed LEBERT, a lexically enhanced BERT model for Chinese sequence labeling. The model integrates external lexical knowledge directly into BERT layers through a lexical adapter and performs deep lexical knowledge fusion, achieving state-of-the-art performance in NER and related tasks. Wu et al. [15] developed an entity recognition model that incorporates external dictionaries for feature enhancement and adversarial training. Their approach constructs character–word pairs from dictionaries, employs a dictionary adapter module for feature fusion, and utilizes FGM (Fast Gradient Method) to enhance model robustness. Sheng et al. [16] designed a lexicon-enhanced BiLSTM-CRF model that combines character and word embeddings at the embedding layer to enrich initial text representations. These enhanced embeddings serve as input to a BiLSTM network, with the CRF layer performing final label decoding. Wang et al. [17] introduced an interactive fusion method for integrating character and word information in Chinese NER. Their method employs an interactive graph structure to merge character and lexical features, resulting in more comprehensive feature representations. To more clearly illustrate the technical approaches of existing semantic enhancement techniques, a summary is provided in Table 1.

Table 1. Research methods for semantic enhancement in Chinese NER.

Method Category	Core Methodology	Reported Performance
swM [13]	Integrates character-related lexical information and uses an MLP classifier to predict the start and end characters of spans for boundary detection.	Improved recognition accuracy in experiments on public datasets.
LEBERT-CRF [14]	Integrates external lexical knowledge into BERT layers through a designed lexicon adapter layer for deep fusion.	Achieved SOTA results on multiple Chinese NER benchmarks.
LEBERT-BiLSTM-CRF [15]	Fuses character and word vector features via a lexicon adapter module and uses FGM to enhance model robustness.	Improved model robustness and recognition performance.
BERT+meanword & BERT+softword [16]	Two lexicon-enhanced vector fusion methods: soft word frequency and average word vector.	Increased F1 score on two datasets.
BERI-SoftLexical-GAT-CRF [17]	Employs an interactive graph structure to integrate character and lexical information for more comprehensive feature representation.	F1 score surpassed previous best models on multiple datasets.

Prior research has demonstrated that semantic enhancement techniques can effectively identify entity boundaries and extract rich semantic information from textual data, yielding measurable improvements in NER performance. To address the research gap in the educational domain NER, we propose ES-BERT, a novel architecture that integrates domain-adapted lexical features with character-level embeddings. Specifically, ES-BERT employs an educa-

tional domain lexicon to generate hybrid character–word representations augmented with domain-specific semantics. The proposed framework simultaneously addresses two key challenges in educational NER, as follows: (1) ambiguous entity boundaries and (2) nested entity recognition difficulties, resulting in substantial performance gains.

3. ES-BERT with Enhanced Semantic Representation

The BERT model has demonstrated remarkable advantages in named entity recognition (NER) tasks, with its contextual understanding capabilities and fine-tuning flexibility establishing it as a state-of-the-art approach [18–20]. As a character-level model, BERT processes each Chinese character as an independent token, thus circumventing word segmentation errors. However, this architecture inherently lacks lexical information utilization—a critical component for capturing inter-character relationships and facilitating boundary detection [21].

To address this limitation, we propose ES-BERT (enhanced semantic BERT, Figure 1), which integrates word-level features with character representations to strengthen inter-character associations. This hierarchical semantic enhancement provides two key benefits, as follows: (1) more precise entity boundary identification and (2) richer contextual semantic extraction, which collectively advance the performance of Chinese NER. The construction process of the ES-BERT model is shown in Algorithm 1.

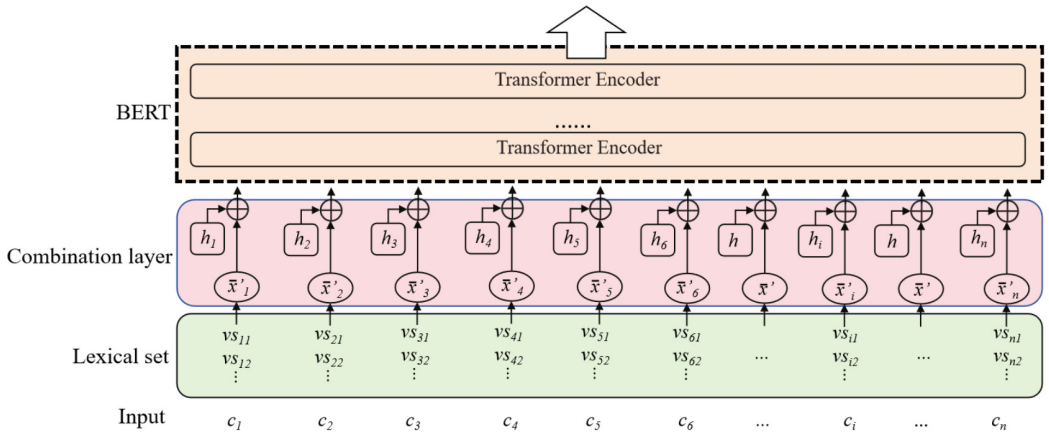

Figure 1. ES-BERT model.

3.1. Domain Lexicon Construction

In the Chinese linguistic context, multi-character lexical units (versus single characters) inherently embody richer semantic priors and provide more definitive boundary indicators for named entity recognition. This phenomenon is particularly pronounced in specialized domains. Domain-specific terminologies exhibit strong predictive power that can effectively guide and refine model predictions.

To optimize lexical semantic representation, our approach incorporates two key components, as follows: (1) core computer science terminologies and (2) domain-specific vocabulary from both tertiary and vocational education contexts. The educational domain lexicon was compiled by selectively extracting relevant terms from Tencent's open-source Chinese word embeddings, focusing particularly on computer science terminology.

Algorithm 1 Enhanced-semantic-BERT Chinese NER algorithm

Input: Chinese sentence $S = \{c_1, c_2, \ldots, c_n\}$, domain lexicon D
Output: Encoded semantic vector sequence $P = [p_1, p_2, \ldots, p_n]$

Phase 1: Character–Lexicon Matching
for $i = 1$ **to** n **do**
  $vs_i \leftarrow \emptyset$ *# Initialize lexicon set for character* c_i
  Generate word combinations centered at c_i in sentence context
  for each word combination w **do**
    if $w \in D$ *# Verify match in domain lexicon* **then**
      $vs_i \leftarrow vs_i \cup \{w\}$
    end if
  end for
end for
$Scs \leftarrow \{(c_1, vs_1), (c_2, vs_2), \ldots, (c_n, vs_n)\}$
Phase 2: Character–Word Vector Fusion
for $i = 1$ **to** n **do**
  $h_i \leftarrow \text{BERT\_Embedding}(c_i)$ *# Character vector: token + segment + position embeddings*
  $x_i \leftarrow \emptyset$ *# Initialize word vector set*
  for each word $w_j \in vs_i$ **do**
    $x_{ij} \leftarrow \text{GetWordEmbedding}(e_j)$ *# Get word vector from domain lexicon*
    $x_i \leftarrow x_i \cup \{x_{ij}\}$
  end for
  TF-IDF Weight Calculation
  for each word vector $x_{ij} \in x_i$ **do**
    $w_{ij} \leftarrow TF(vs_{ij}) \times IDF(vs_{ij})$ *# Calculate TF-IDF weights*
  end for
  Weighted Word Vector Fusion
  $\bar{x}_i \leftarrow \mathbf{0}$ *# Initialize fused word vector*
  for $j = 1$ **to** $|x_i|$ **do**
    $\bar{x}_i \leftarrow \bar{x}_i + w_{ij} \times x_{ij}$
  end for
  Dimension Alignment Transformation
  $\bar{x}'_i \leftarrow W_2 \times \text{ReLU}(W_1 \times \bar{x}_i + b_1) + b_2$
  Final Fusion
  $h'_i \leftarrow h_i + \bar{x}'_i$ *# Obtain character–word fused vector*
end for
Phase 3: BERT Encoding
$H \leftarrow [h'_1, h'_2, \ldots, h'_n]$ *# Construct input sequence*
Self-Attention Mechanism
for each Transformer layer **do**
  $Q \leftarrow \text{Linear}(H) \times W^Q$
  $K \leftarrow \text{Linear}(H) \times W^K$
  $V \leftarrow \text{Linear}(H) \times W^V$
  $\text{Attention} \leftarrow \text{Softmax}(\frac{Q \times K^T}{\sqrt{d_k}}) \times V$
  Multi-Head Attention
  $\text{MultiHead} \leftarrow \text{Concat}(\text{head}_1, \ldots, \text{head}_h) \times W^O$
  where $\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$
  $H \leftarrow \text{LayerNorm}(H + \text{MultiHead})$
  $H \leftarrow \text{LayerNorm}(H + \text{FFN}(H))$ *# Feed-forward network*
end for
$P \leftarrow H$ *# Output encoded semantic vector sequence*
return P

Leveraging this lexicon, we identify character-position-aware lexical candidates within input texts. This design intentionally circumvents error-prone word segmentation, since segmentation precision constitutes a fundamental bottleneck in Chinese NER systems.

Crucially, any segmentation errors would propagate through the recognition cascade, severely compromising entity identification reliability.

To address this fundamental challenge, our framework dynamically constructs domain-specific lexical candidate sets through comprehensive sentence-to-lexicon matching. The core algorithm operates on a character-by-character basis, as follows: For each Chinese character in the input sequence, it evaluates all possible contextual combinations. When a combination matches a domain lexicon entry, the corresponding word is incorporated into the character's lexical candidate set.

Formally, let $S = \{c_1, c_2, \ldots, c_n\}$ denote an n-character Chinese sentence. Our method generates a character–word sequence $S_{cs} = \{(c_1, vs_1), (c_2, vs_2), \ldots, (c_n, vs_n)\}$, where each vs_i represents the potential lexical set containing all domain words that (1) include character c_i, (2) occur in sentence S, and (3) exist in the domain lexicon (see Figure 1). This is implemented using an optimized maximum matching algorithm with $O(n^2)$ time complexity.

3.2. Character–Word Vector Fusion

Upon acquiring the character–word sets (c_i, vs_i), we transform both characters and their associated words into vector representations (h_i, x_i). Specifically, the character vector h_i at position i is derived from BERT's output representation, which aggregates three components, as follows: token embedding, segment embedding, and position embedding. The word vector set $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ contains all pre-trained lexical embeddings corresponding to character c_i, where each x_{ij} is retrieved from our domain-specific lexicon.

To mitigate segmentation errors, we compute TF-IDF weights for candidate words, assigning lower weights to the following: (1) infrequent terms, and (2) out-of-vocabulary words. The weight w_{ij} for the j-th word associated with character c_i is calculated as follows:

$$w_{ij} = TF(vs_{ij}) * IDF(vs_{ij}) \tag{1}$$

where $TF(vs_{ij})$ represents the term frequency (TF) of word vs_{ij} (the j-th word corresponding to character c_i) in the document, $IDF(vs_{ij})$ represents the inverse document frequency (IDF) of word vs_{ij}.

For each character c_i, we compute the weighted sum of all its corresponding word vectors to generate the fused lexical vector $\bar{x}_i$:

$$\bar{x}_i = \sum_{j=1}^{m} w_{ij} x_{ij} \tag{2}$$

To address the dimensionality mismatch between fused lexical vectors and BERT-generated character embeddings, we first apply a nonlinear projection layer for vector alignment prior to secondary fusion. This dimensional transformation ensures feature space compatibility between the two representation types:

$$\bar{x}_i' = W_2(tanh(W_1 \bar{x}_i + b_1)) + b_2 \tag{3}$$

where, $tanh()$ is the activation function, z_i represents the aligned fused lexical vector, $W_1 \in \mathbb{R}^{d_c * d_w}$ denotes the transformation matrix, $W_2 \in \mathbb{R}^{d_c * d_c}$ is the projection matrix, and b_1, b_2 are bias terms, where d_c indicates the character embedding dimension and d_w represents the word vector dimension.

Finally, we perform element-wise summation between the aligned fused lexical vectors and the character vectors from BERT's embedding layer, yielding the combined character–word fusion vectors. These integrated vectors h_i' simultaneously encapsulate both character-level and word-level information, as illustrated in Figure 2.

$$h'_i = h_i + \bar{x}'_i \tag{4}$$

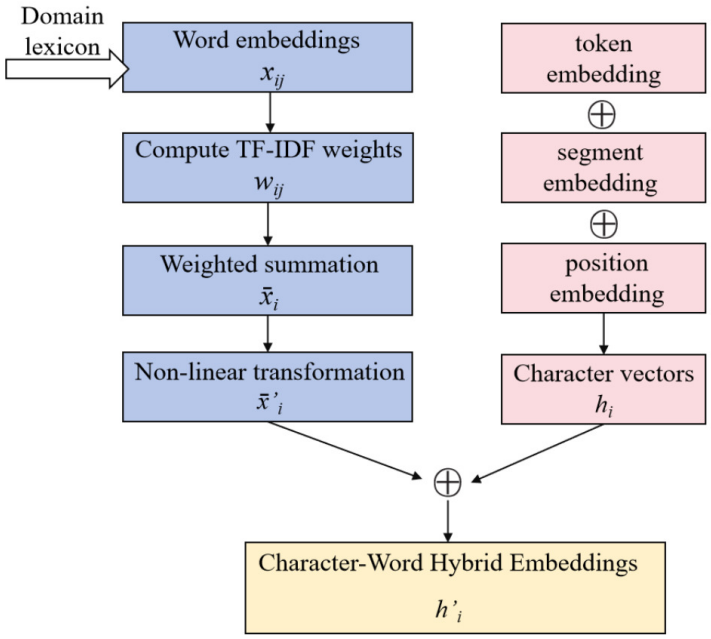

Figure 2. Semantic fusion module.

3.3. BERT-Enhanced Hybrid Representation

The hybrid character–word embeddings $H = [h'_1, h'_2, \ldots, h'_n]$ are subsequently processed through BERT's Transformer encoder stack. The architecture employs 12-head self-attention mechanisms for dynamic modeling of contextual dependencies, enabling each position to attend to and aggregate the most informative features across the sequence. Specifically, this is achieved by computing three vector representations per token position—query (Q), key (K), and value (V), which interact through scaled dot-product attention:

$$\begin{cases} Q = Linear(H)W_Q \\ K = Linear(H)W_K \\ V = Linear(H)W_V \end{cases} \tag{5}$$

Here, $Linear()$ denotes a linear transformation, and W_Q, W_K, and W_V represent trainable weight matrices.

The attention computation is formulated as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Here, $Attention()$ denotes the scaled dot-product attention operation. The $\sqrt{d_k}$ term projects the attention matrix into a standard normal distribution space. $Softmax()$ is the normalized exponential function that ensures the attention weights sum to 1.

The model employs multi-head self-attention through parallel computation across multiple attention heads, followed by concatenation of all head outputs and a final linear projection. This architecture facilitates multi-perspective feature extraction from distinct representation subspaces, enables the acquisition of diverse semantic patterns, and ensures comprehensive contextual understanding.

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{7}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, head_2, \ldots, head_k)W^o \tag{8}$$

Let W_i^Q, W_i^K, and W_i^V denote the query/key/value weight matrices for the *ith* attention head, with W^o being the output projection matrix and $Concat()$ the concatenation operator.

The sequence $P = [p_1, p_2, \ldots, p_n]$, produced by successive Transformer encoder layers, forms contextualized fusion vectors that holistically encode full-text semantics, making them immediately applicable to named entity recognition tasks.

4. ES-BERT Enhanced Named Entity Recognition Model

Our proposed framework for the educational domain NER integrates the enhanced semantic BERT (ES-BERT) model with a BiLSTM-CRF architecture [22–24]. This integration synergistically leverages BERT's deep contextual representations, BiLSTM's capacity for capturing sequential dependencies, and CRF's ability to enforce global label consistency. The model begins with ES-BERT generating semantically enhanced embeddings. These embeddings are then processed by the BiLSTM layer to extract hierarchical contextual features. Finally, the CRF layer decodes the most probabilistically optimal label sequence using the Viterbi algorithm. The entire system is jointly optimized with a focal loss objective to enhance robustness against annotation noise, as illustrated in Figure 3.

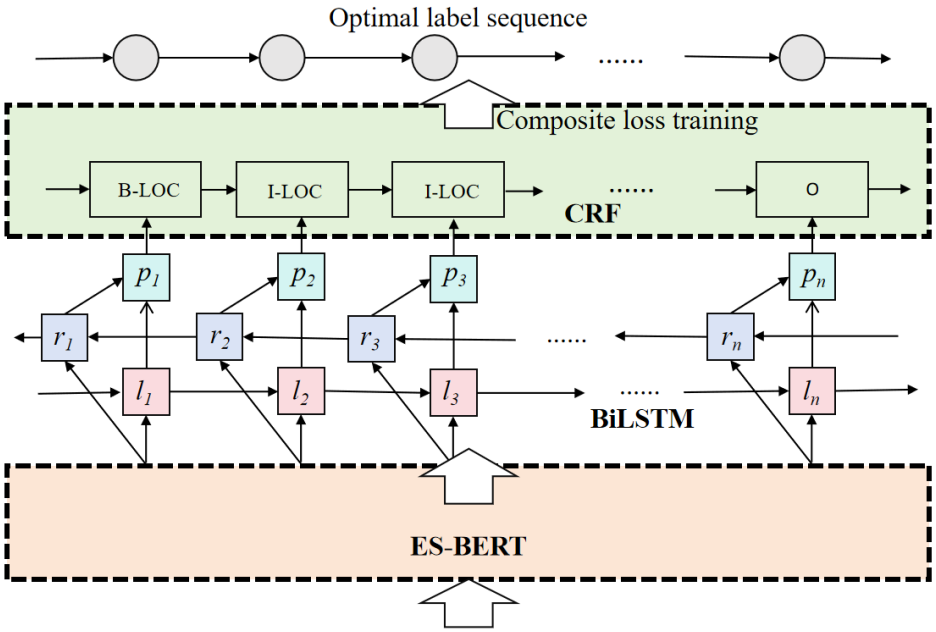

Figure 3. Overall structure of the ES-BERT+BiLSTM+CRF model.

4.1. BiLSTM Layer

Standard LSTM networks process sequences unidirectionally (left-to-right), inherently assigning greater importance to later tokens—a problematic bias for segmentation tasks. Our method constructs bidirectional representations through position-wise concatenation of forward $(\overrightarrow{p_1}, \overrightarrow{p_2}, \ldots, \overrightarrow{p_n})$ and backward $(\overleftarrow{p_1}, \overleftarrow{p_2}, \ldots, \overleftarrow{p_n})$ LSTM outputs, formally defined as $L = (l_1, l_2, \ldots, l_n) \in \mathbb{R}^{n*m}$, where each l_i combines both directional contexts for optimized label prediction.

$$\overrightarrow{l_i} = \overrightarrow{LSTM}(p_i) \tag{9}$$

$$\overleftarrow{l_i} = \overleftarrow{LSTM}(p_i) \tag{10}$$

$$l_i = \overrightarrow{l_i} \oplus \overleftarrow{l_i} \tag{11}$$

Here, $\overrightarrow{LSTM}()$ and $\overleftarrow{LSTM}()$ represent the forward and backward LSTM networks, respectively, capturing historical and future contextual information for the current feature position p_i. The symbol $\oplus$ denotes vector concatenation. In named entity recognition (NER) tasks, the BiLSTM architecture fully utilizes bidirectional sequence context, significantly improving character-level label prediction accuracy.

4.2. CRF Layer

The CRF layer explicitly models both word-level semantic relationships and inter-label transition probabilities. When the BiLSTM layer generates erroneous predictions, the CRF layer corrects these errors by analyzing neighboring label contexts, thereby refining the final output sequence [25]. Formally, given a label sequence $y = (y_1, y_2, \ldots, y_n)$ and a feature vector $L \in \mathbb{R}^n$, the sequence scoring function is defined as follows:

$$score(y, L) = \sum_{i=1}^{n} E_{y_i} * l_i + \sum_{i=2}^{n} T_{y_{i-1}, y_i} \tag{12}$$

The score at each position is determined by (1) the emission matrix E from BiLSTM outputs, which associates each character with corresponding label weights, and (2) the transition matrix T of the CRF that captures label-to-label movement scores, where E_i denotes the weight vector for label i and T_{ij} indicates the transition score from label i to label j.

The Viterbi algorithm's identification of the maximum-probability path is ideally suited for named entity recognition (NER) tasks by globally optimizing sequence likelihood (avoiding local decision pitfalls), enforcing structural constraints through transition probabilities, and capturing long-range dependencies essential for precise entity boundary detection. Formally, given an input sequence L and label space y, the optimal tag sequence $\hat{y}$ is computed as follows:

$$\hat{y} = \underset{y'}{argmax} score(y', L) \tag{13}$$

where y' denotes the complete collection of predicted label sequences.

4.3. Novel Focal Loss-Based Composite Loss

The negative log-likelihood loss (NLL), as the standard CRF loss function, demonstrates strong performance under ideal conditions (clean data, balanced class distribution, and complete annotations).

$$L_{CRF} = -log(\frac{exp(score(y, L))}{\sum_{y' \in y^n} exp(score(y', L))}) \tag{14}$$

However, in named entity recognition (NER) tasks, non-entity ("O") labels account for an overwhelmingly high proportion of tokens, while numerous challenging samples (e.g., boundary-ambiguous or nested entities) coexist. When certain entity categories dominate the dataset with significantly larger sample sizes, the negative log-likelihood (NLL) loss introduces substantial gradient bias. This causes the loss function to be predominantly influenced by high-frequency categories, consequently leading to model predictions that are markedly skewed toward majority classes.

Focal loss, originally designed to handle class imbalance and varying classification difficulty in sequence labeling tasks [26], is enhanced in our framework through Viterbi-decoding integration. The proposed method first computes positional alignment discrepancies between Viterbi-derived paths and ground truth annotations, then generates sequence-aware weighting coefficients β_t. Only samples misclassified under sequential constraints receive elevated weights β_t. Consequently, we reformulate focal loss as follows:

$$L_{V-FL} = -\frac{1}{T}\sum_{t=1}^{T}\beta_t(1-P_t)^{\gamma}log(P_t) \tag{15}$$

$$s.t. \quad \beta_t = \begin{cases} 1+\alpha & if\ \hat{y}_t \neq y_t \\ 1 & otherwise \end{cases}$$

In the formulation, T represents the sequence length, p_t denotes the probability of the true label at position t, and γ is the focusing hyperparameter. To address the class imbalance inherent in this task, where majority-class samples are typically classified with ease, a moderate γ value (e.g., 2.0) is set to suppress gradient contributions from these simple samples. This focuses the optimization on learning discriminative features, thereby enhancing minority-class performance. The Viterbi weighting factor β_t increases when predictions violate sequential constraints, while the hyperparameter $\alpha > 0$ amplifies weights for such misclassified samples. When $\alpha = 0$, the sequence-aware mechanism is inactive; setting $\alpha > 0$ (e.g., 0.75) provides a stronger signal to prioritize correcting errors that disrupt sequential coherence.

During model training, we implement a joint optimization framework that simultaneously preserves the standard CRF negative log-likelihood (L_{CRF}) objective and integrates Viterbi-augmented focal loss (L_{V-FL}) as an adaptive regularization component (Algorithm 2).

$$L = L_{CRF} + \lambda \cdot L_{V-FL} \tag{16}$$

Here, $\lambda \in [0.1, 1.0]$ is a hyperparameter, with the aim of minimizing the loss.

Algorithm 2 Viterbi-enhanced focal loss (V-FL)

Input: Sequence length T;
True label probability P_t at position t;
True label sequence $\mathbf{y}_{\text{true}}$;
Viterbi decoding path $\mathbf{y}_{\text{viterbi}}$.
Output: Total loss $\mathcal{L}_{\text{total}}$

Compute traditional CRF negative log-likelihood loss $\mathcal{L}_{\text{CRF}}$ from Equation (14);
Calculate sequence-aware weighting coefficients:
for $t = 1$ to T **do**
  if $\mathbf{y}_{\text{viterbi}}[t] \neq \mathbf{y}_{\text{true}}[t]$ **then**
    $\beta_t \leftarrow 1 + \alpha$ *# Increase weight for Viterbi prediction errors*
  else
    $\beta_t \leftarrow 1$ *# Base weight for correct predictions*
  end if
end for
Compute Viterbi-based focal loss (V-FL):
$\mathcal{L}_{\text{V-FL}} \leftarrow 0$
for $t = 1$ to T **do**
  $\mathcal{L}_{\text{V-FL}} \leftarrow \mathcal{L}_{\text{V-FL}} - \beta_t \times (1 - P_t)^{\gamma} \times \log(P_t)$
end for
$\mathcal{L}_{\text{V-FL}} \leftarrow \mathcal{L}_{\text{V-FL}} / T$ *# Normalization*
Compute joint optimization total loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CRF}} + \lambda \times \mathcal{L}_{\text{V-FL}}$
return $\mathcal{L}_{\text{total}}$

This combined approach integrates the sequence-level constraints of the CRF loss with the token-level hard example weighting of the V-Focal loss. The CRF loss inherently models label transition patterns to ensure global coherence, whereas the V-Focal loss emphasizes difficult-to-classify tokens. Their joint optimization achieves two objectives, as follows: (1) globally coherent label sequences and (2) enhanced attention to challenging boundary

cases. As a result, the final predictions exhibit both global sequence consistency and high local classification confidence.

5. Experiments

5.1. Experimental Environment

The experimental configuration consisted of an Intel Core i7-8565U processor with 16 GB RAM and 8 GB GPU VRAM, running on Windows 11 with a Python 3.8 runtime environment. The implementation employed PyCharm Professional 2023.2.5 in Anaconda with TensorFlow 2.10.0, and PyTorch 1.13.1, implementing Google's BERT-Base-Chinese pretrained model whose key hyperparameters are detailed in Table 2.

Table 2. Model parameter settings.

Hyperparameters	Values
BERT lr	3×10^{-5}
BiLSTM lr	1×10^{-4}
CRF lr	1×10^{-3}
Batch size	32
Hidden units	768
Encoder layers	12
Attention heads	12
Optimizer	Adam

5.2. Experimental Data

The experimental dataset comprises a custom-built corpus of educational literature in computer-related disciplines (Edu_literature). We collected 1000 pedagogical research papers from CNKI using web crawlers programmed with topic keywords such as "industry-education integration", "OBE concept", and "project-driven approach". The papers included author affiliations, titles, abstracts, and keywords. Entity annotation was performed through a hybrid approach combining machine processing and manual review. The procedure began with an initial annotation using a pre-trained model-based tool, followed by human-assisted verification to identify and resolve ambiguities and errors. A total of 5312 entities across four categories were annotated. The distribution of these categories is shown in Table 3.

Table 3. Categories and quantity of educational literature.

Entity Types	Count
Institution names	1107
Academic majors	575
Course titles	2692
Teaching methods/Pedagogical approaches	938

The dataset employs the standard BIO (begin–inside–outside) annotation scheme, where 'B-X' marks the beginning of entity X, 'I-X' indicates the continuation of entity X, and 'O' denotes non-entity tokens. Annotation examples are shown in Table 4.

Table 4. Annotation method of the BIO entity sequence.

Entity Types	Start	Middle and End	Entity Example
Institution names	B-N	I-N	Peking University, Tsinghua University, Nanjing University, etc.
Academic majors	B-M	I-M	Computer Science and Technology, Software Engineering, Digital Media, etc.
Course titles	B-C	I-C	Principles of operating systems, Data structures, Digital media technology, etc.
Teaching methods/ Pedagogical approaches	B-T	I-T	Project-driven teaching, Integration of industry and education, Flipped classroom, etc.

5.3. Evaluation Metrics

To assess model performance, we employ three standard evaluation metrics—precision, recall, and F1 score—for named entity recognition analysis.

$$Pre = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2 \times Pre \times Recall}{Pre + Recall} \tag{19}$$

where TP denotes the number of correctly identified entities, FP represents the number of non-entities incorrectly identified as entities, and FN indicates the number of entities that failed to be recognized.

5.4. Entity Recognition and Analysis in Education

Owing to the limited size of the Edu_literature dataset, a single train–test split risks unstable evaluation due to partitioning variability. To mitigate this, we employed 5-fold cross-validation for all six comparative models. This method ensures that models are trained and validated on nearly the entire dataset in each cycle, maximizing data utility. Performance metrics were averaged over all folds to yield a more robust estimate of model performance and stability.

The average performance metrics across all folds from the 5-fold cross-validation experiments for the models BiLSTM+CRF, BERT+BiLSTM+CRF [9], ALBERT+BiLSTM+CRF [12], Lattice-LSTM [27], Soft-Lexicon (LSTM) [28], and the model proposed in this study are presented in Table 5.

Table 5. Experimental results on Edu_literature.

Model	Pre	Recall	F1
BiLSTM+CRF	83.16%	85.18%	84.11%
BERT+BiLSTM+CRF	87.11%	86.37%	86.71%
ALBERT+BiLSTM+CRF	87.08%	87.55%	87.30%
Lattice-LSTM	85.47%	86.22%	85.84%
Soft-Lexicon (LSTM)	88.86%	88.04%	88.44%
Our model	90.38%	89.71%	90.04%

As shown in Table 5, the model proposed in this study demonstrates superior performance across all three metrics—precision (90.38%), recall (89.71%), and F1 score (90.04%)—significantly outperforming all other comparative models, which reflects its comprehensive advantage. Compared to the baseline model BiLSTM+CRF (F1 = 84.11%), the proposed model achieves an improvement of nearly 6 percentage points in F1 score, indicating the effectiveness of its architectural enhancements. Among the pre-trained language models, ALBERT+BiLSTM+CRF (F1 = 87.30%) slightly surpasses BERT+BiLSTM+CRF (F1 = 86.71%), suggesting that ALBERT's parameter-sharing mechanism may lead to more stable representation learning. However, both fall short of the proposed model, implying that there remains room for optimization beyond simply employing pre-trained models. Within lexicon-enhanced models, Soft-Lexicon (LSTM) (F1 = 88.44%) performs better than Lattice-LSTM (F1 = 85.84%), validating the effectiveness of soft lexical integration. Nevertheless, the proposed model pushes beyond the performance ceiling of such methods, likely due to more refined semantic enhancement mechanisms—such as weighted integration of domain-specific dictionaries—that enable more accurate boundary detection and type recognition. Overall, the proposed model maintains high precision while balancing recall, with its performance advantage likely stemming from targeted modeling of domain-specific characteristics and deep fusion of character- and word-level information, offering an effective new solution for NER tasks.

To scientifically evaluate whether the performance improvement of our proposed ES-BERT model over the classic BERT+BiLSTM+CRF model is statistically significant, we compared the performance of the two models using a paired t-test based on F1 scores from 5-fold cross-validation. The results are presented in Table 6.

Table 6. *T*-test: paired two-sample for means.

Model	M	SD	Correlation Coefficient	t(4)	*p* (Two-Tailed)
BERT+BiLSTM+CRF	0.867	0.008	0.336	−6.91	0.002
Our model	0.900	0.010			

The paired-sample t-test on the 5-fold cross-validation results (Table 6) revealed that the proposed model (M = 0.900, SD = 0.010) significantly outperformed the BERT+BiLSTM+CRF baseline (M = 0.867, SD = 0.008), t(4) = −6.91, $p = 0.002$. This statistically significant difference indicates that the performance gain afforded by the proposed model is robust and not attributable to random variation. The moderate positive correlation (r = 0.336) between model performances across folds suggests that while both models were influenced by similar data characteristics, the proposed model consistently achieved higher scores. These results provide strong evidence that the proposed model offers a meaningful improvement over the baseline in terms of predictive accuracy.

Based on the comparative experimental results, we further evaluate our model's performance across four entity categories (institutions, majors, courses, and teaching methods/modes) using multiple evaluation metrics, as shown in Figure 4.

Figure 4 demonstrates that our model achieves superior F1 scores for institutions and major entities, while exhibiting relatively lower performance on courses, teaching-method, and teaching-mode categories. In-depth analysis indicates that institutional entities in our self-constructed educational domain dataset exhibit higher frequency and diversity, enabling more comprehensive pattern learning during training. The presence of distinctive lexical markers (e.g., "University"/"College") in institution names further boosts recognition accuracy. The model's high major identification accuracy stems from standardized naming conventions documented in China's Ministry of Education publications, particu-

larly the "Undergraduate Program Catalog" and "Vocational Education Major Catalog", which establish clear naming patterns (e.g., "Computer Science", "Computer Science and Technology"). Course names present greater challenges due to lexical polysemy: terms like "Java Programming", "Web Front-end Development", and "Computer Operating Systems" function ambiguously as both course titles and technical terminology, introducing classification noise that impedes accurate disambiguation. To address these limitations, our future work will focus on the following: (1) dataset expansion and education-domain lexicon enrichment; (2) semantic segmentation precision improvement and named entity boundary detection optimization, ultimately aiming to acquire more precise entity features and boundary information for enhanced recognition performance.

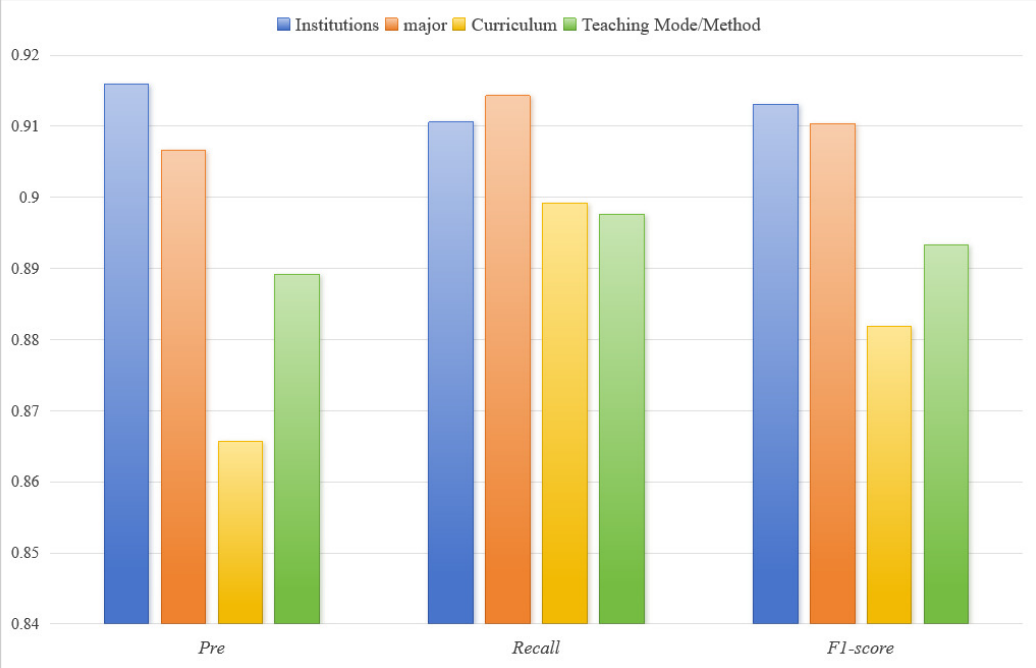

Figure 4. Various entity recognition effects.

5.5. Ablation Study

To evaluate the effectiveness of the ES-BERT model with enhanced semantic representation in named entity recognition, four progressively advanced model variants were designed for comparative analysis, with the objective of assessing the contribution of each module:

- Model 1 (Baseline): Uses exclusively BERT-generated character embeddings as input to the BiLSTM-CRF framework, with negative log-likelihood (NLL) loss as the objective function.
- Model 2: Building upon Model 1, semantic enhancement was incorporated without the use of a domain-specific lexicon. A word segmentation tool was applied to generate tokens, which were then fused with character-level representations to form the input to the model.
- Model 3: Based on Model 2, a domain lexicon was integrated along with the proposed ES-BERT model. Domain-relevant vocabulary was combined with character-level information to construct the input representation.
- Model 4 (Full Model): An enhanced version of Model 3, in which a modified focal loss-based joint loss function was introduced to replace the original loss function.

Figure 5 displays the F1 scores obtained from 20 iterative training runs of each model configuration.

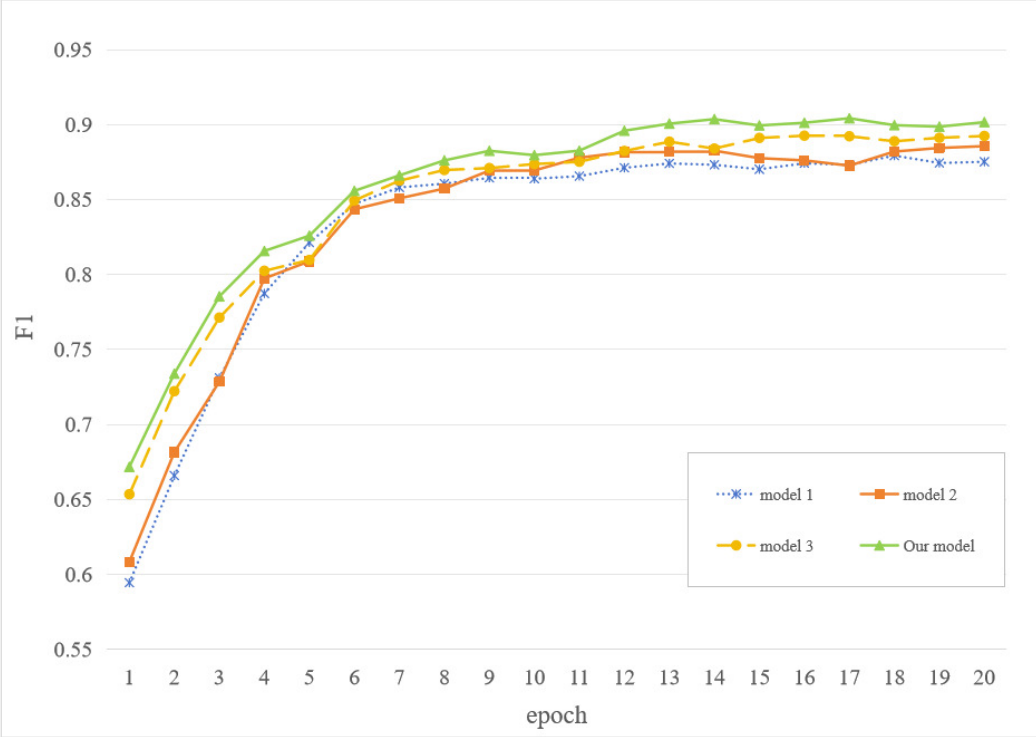

Figure 5. The accuracy of ablation experiments.

Figure 5 demonstrates that Model 2 achieves higher overall F1 scores compared to Model 1 after convergence, confirming the effectiveness of lexical information fusion for recognition enhancement. However, our analysis reveals that word segmentation errors cause error propagation, adversely affecting downstream word recognition and, consequently, limiting further performance improvements.

To address this limitation, Model 3 introduces an innovative domain dictionary matching mechanism. By generating character-level candidate word sets with TF-IDF weighting, it significantly reduces the impact of incorrect segmentation—low-weight erroneous segmentations are effectively filtered out, resulting in F1 score improvement over Model 2.

This study introduces an enhanced model based on Model 3, which obtains the optimal F1 score. The model tackles two principal difficulties in domain-specific NER—vague entity boundaries and specialized terms—through strengthened domain semantic encoding, yielding more informative and precise features that augment boundary recognition. Additionally, a modified focal loss function addresses class imbalance by applying increased weights to persistently misclassified examples under contextual constraints, thus refining the learning focus on challenging instances. The collaborative operation of these modules markedly improves recall, consequently elevating the F1 score.

The ablation study confirms the critical importance of component-level synergistic optimization for performance improvement, providing new methodological insights for multi-feature fusion in NER tasks.

5.6. Model Transferability

Comparative experiments were conducted to validate the cross-domain transferability of the proposed model using the open-source KBQA and Resume datasets. Given the lack of dedicated domain-specific lexicons for these datasets, Tencent's open-source Chinese word vectors served as the common domain dictionary for word segmentation in all Trials.

(1) The knowledge base question answering (KBQA) dataset, sourced from the NLPCC-ICCPOL 2016 shared task, comprises a knowledge base with 24,030 entity-attribute

triplets. We split the corresponding QA pairs into training (14,609 pairs), validation (3945 pairs), and test (3945 pairs) sets.

(2) The Resume dataset, contains 16,565 annotated entities across 8 categories (including educational background, locations, and personal names), providing additional scenarios for evaluating transfer learning capability.

Table 7 presents the model's performance on the specialized KBQA (Knowledge Base Question Answering) task set with a triple-based knowledge base, where our model achieves new state-of-the-art benchmarks of 98.41% precision, 98.35% recall, and 98.36% F1 score. These results conclusively evidence the model's Outstanding capability for accurate recognition of diverse entity categories (including persons, organizations, and temporal expressions) in knowledge-intensive QA scenarios. Further analysis indicates the model's dual advantage: maintaining high precision while achieving improved low-frequency entity recognition rates—a breakthrough stemming from the synergistic integration of our ES-BERT architecture with novel focal loss-based joint training optimization. The experimental outcomes not only confirm the architectural superiority of the model but also present a viable technical solution for entity recognition in knowledge base question answering systems.

Table 7. Experimental results on KBQA.

Model	Pre	Recall	F1
BiLSTM+CRF	94.54%	94.44%	94.49%
BERT+BiLSTM+CRF	96.29%	96.18%	96.24%
ALBERT+BiLSTM+CRF	97.94%	97.90%	97.92%
Lattice-LSTM	95.52%	95.31%	95.41%
Soft-Lexicon (LSTM)	96.90%	97.02%	96.96%
Our model	98.41%	98.35%	98.36%

Experimental results on the Chinese Resume dataset (Table 8) confirm the efficacy of the ES-BERT model. By deeply fusing character-level and domain-specific lexical features, the model effectively captures domain knowledge from resumes and precisely identifies entity boundaries (e.g., person names, education, work experience). It achieves state-of-the-art performance (96.75% precision, 96.31% recall, and 96.53% F1 score), demonstrating the validity of our approach for handling the unique linguistic patterns of Chinese resumes. The architecture's adaptive learning mechanism yields highly competitive boundary detection accuracy.

Table 8. Experimental results on Resume.

Model	Pre	Recall	F1
BiLSTM+CRF	93.10%	93.15%	93.12%
BERT+BiLSTM+CRF	95.06%	95.15%	95.10%
ALBERT+BiLSTM+CRF	95.54%	96.13%	95.83%
Lattice-LSTM	94.04%	94.27%	94.16%
Soft-Lexicon (LSTM)	95.45%	95.58%	95.51%
Our model	96.75%	96.31%	96.53%

Systematic evaluation of the cross-domain transfer experiments demonstrates that when employing the general-purpose Tencent open-source Chinese word vectors as a domain lexicon, the proposed model architecture exhibits strong domain adaptation and

generalization capabilities across two distinct datasets. These results establish a solid foundation for the model's application in cross-domain scenarios.

6. Conclusions

Advances in AI technology are increasing the importance of named entity recognition (NER) for educational applications—such as adaptive learning, intelligent tutoring, and academic research—establishing it as a critical enabling technology for intelligent education systems. A key limitation in educational NER is insufficient lexical semantic awareness, which leads to weak domain-specific representations and imprecise entity boundaries. To address this, we propose ES-BERT, a novel model that introduces enhanced semantic representation. By integrating a domain-specific semantic enhancement mechanism into BERT's architecture, ES-BERT provides an efficient solution for domain-specific NER.

Nevertheless, the study is subject to certain limitations, particularly the restricted scale of the annotated dataset and the limited coverage of the domain lexicon, underscoring the need for more extensive resources. Several architectural directions warrant further investigation in future work, such as extending lexical-semantic enhancement beyond BERT's lower layers to optimize feature integration strategies, developing dynamic weighting mechanisms for lexical representations across different network depths, and systematically examining the effect of focal loss hyperparameters. Advances in these areas are expected to strengthen the model's ability to comprehend complex semantic structures in the educational domain.

Author Contributions: Conceptualization, methodology, software, writing—original draft, writing—review, editing: P.H. and H.Z.; methodology, data curation, writing—original draft, visualization: Y.W. and L.D.; validation, investigation: L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the following projects: the Research Project on Higher Education Teaching Reform in Jiangsu Province (grant number 2025JGYB405), the Jiangsu Province Industry Education Integration First Class Curriculum Project "Virtual Reality Technology and Application" (grant number 03145031), and the School Level Educational Reform Project of Nanjing University of Science and Technology ZiJin College (grant number 20240103005).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author: huangping984@njust.edu.cn.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zheng, S.; Shen, Y. Educational Named Entity Recognition Integrating Word Information and Self-Attention Mechanism. *Softw. Guide* **2024**, *23*, 105–109.
2. Ren, Y.; Su, B.; Yuan, S. Multidimensional Feature Named Entity Recognition Method in Education Domain. *Comput. Eng.* **2024**, *50*, 110–118.
3. Qiao, B.; Zou, Z.; Huang, Y.; Fang, K.; Zhu, X.; Chen, Y. A joint model for entity and relation extraction based on BERT. *Neural Comput. Appl.* **2022**, *34*, 3471–3481. [CrossRef]
4. Han, Q. Construction of Intelligent Question-Answering System to Improve Knowledge Management Ser-vice from the Perspective of Education Informatization. *J. Inf. Knowl. Manag.* **2025**, *24*, 1–22.
5. Li, Y.; Liang, Y.; Yang, R.; Qiu, J.; Zhang, C.; Zhang, X. CourseKG: An Educational Knowledge Graph Based on Course Information for Precision Teaching. *Appl. Sci.* **2024**, *14*, 2710. [CrossRef]
6. Hu, Y.; Chen, Y.; Huang, R.; Qin, Y. CRF-combined boundary assembly method for biomedical named entity recognition. *Appl. Res. Comput.* **2021**, *38*, 2025–2031.

7. Zhang, R.; Dai, L.; Guo, P.; Wang, B. Chinese Nested Named Entity Recognition Algorithm Based on Segmentation Attention and Boundary-aware. *Comput. Sci.* **2023**, *50*, 213–220.
8. Dong, H.; Kong, Y.; Gao, W.; Liu, J. Named entity recognition for public interest litigation based on a deep contextualized pretraining approach. *Sci. Program.* **2022**, *1*, 7682373.1–7682373.14. [CrossRef]
9. Hu, H.; Li, J.; Dong, Z.; Bai, X. Named Entity Recognition Method in Educational Technology Field Based on BERT. *Comput. Technol. Dev.* **2022**, *32*, 164–168.
10. Xie, T.; Yang, J.-A.; Liu, H. Chinese entity recognition based on BERT-BiLSTM-CRF mode. *Comput. Syst. Appl.* **2020**, *29*, 48–55.
11. Lu, Z.; Zhao, W.; Yin, G. Named Entity Recognition for Textual Intelligence Based on RoB-ERTa_BiLSTM_CRF. *J. China Acad. Electron. Inf. Technol.* **2024**, *19*, 442–447.
12. Jin, L.; Zhang, Y.; Yuan, Z.; Gao, S.; Gu, M.; Liu, X. Chinese named entity recognition of transformer bushing faults based on ALBERT-BiLSTM-CRF. *IEEE Trans. Ind. Appl.* **2025**, *61*, 2115–2123. [CrossRef]
13. Chen, S.; Dou, Q.; Tang, H.; Jiang, P. Chinese nested named entity recognition based on vocabulary fusion and span detection. *Appl. Res. Comput.* **2023**, *40*, 2382–2386+2392.
14. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon enhanced chinese sequence labeling using BERT adapter. *Int. Jt. Conf. Nat. Lang. Process. Meet. Assoc. Comput. Linguist.* **2021**, *2021*, 5847–5858.
15. Wu, G.; Fan, C.; Tao, G.; He, Y. Entity recognition of electronic medical records based on LEBERT-BCF. *Comput. Era* **2023**, *2*, 92–97.
16. Sheng, L.; Zhang, Y.; Wu, D. Chinese named entity recognition method based on lexical enhancement. *Mod. Electron. Tech.* **2022**, *45*, 157–162.
17. Wang, Y.; Wang, Z.; Yu, H.; Wang, G.; Lei, D. The interactive fusion of characters and lexical information for Chinese named entity recognition. *Artif. Intell. Rev.* **2024**, *57*, 258.1–258.21. [CrossRef]
18. Zhao, J.; Qian, Y.; Wang, K.; Hou, S.; Chen, J. Survey of chinese named entity recognition research. *Comput. Eng. Appl.* **2024**, *60*, 15–27.
19. Yang, P.; Dong, W. Chinese named entity recognition method based on BERT embedding. *Comput. Eng.* **2020**, *46*, 40–45+52.
20. Huang, S.; Sha, Y.; Li, R. A Chinese named entity recognition method for small-scale dataset based on lexicon and unlabeled data. *Multimed. Tools Appl.* **2023**, *82*, 2185–2206. [CrossRef]
21. Chen, S.; Luo, C.; Ouyang, X.; Li, W. A Semantic-enhanced Chinese Named Entity Recognition Algorithm Based on Dynamic Dictionary Matching. *Radio Eng.* **2021**, *51*, 519–525.
22. Zheng, X.; Li, B.; Feng, Z.; Liu, X. Entity recognition of network sensitive words and variants based on BERT-BiLSTM-CRF. *Comput. Digit. Eng.* **2023**, *51*, 1585–1589.
23. Che, X.; Xu, H.; Pan, M.; Liu, Q.-L. Two-stage learning algorithm for biomedical named entity recognition. *J. Jilin Univ. Eng. Technol. Ed.* **2023**, *8*, 2380–2387.
24. Liu, X.-H.; Xu, R.-Z.; Yang, C.-Y. A Chinese named entity recognition model based on multi-feature fusion embedding. *Comput. Eng. Sci.* **2024**, *46*, 1473–1481.
25. Liao, T.; Gou, Y.; Zhang, S. BERT-BiLSTM-CRF Chinese named entity recognition combined with atten-tion mechanism. *J. Fuyang Norm. Univ. Nat. Sci.* **2021**, *38*, 86–91.
26. Dang, X.; Liu, J.; Dong, X.; Zhu, Z.; Li, F. Named Entity Recognition of Mechanical Equipment Failure for Imbalanced Data. *Comput. Eng.* **2024**, *50*, 104–112.
27. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1554–1564.
28. Ma, R.; Peng, M.; Zhang, Q.; Huang, X. Simplify the usage of lexicon in Chinese NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5951–5960.