

Article

Domain-Specific Named Entity Recognition from Chinese Building Fire Protection Design Codes for Automated Compliance Checking

Lu Gao ^{1,*} , Dejun Qiao ¹, Hong Zhang ² and Liang Zhao ^{3,*}¹ School of Software, Handan University, Handan 056005, China² School of Architecture, Southeast University, Nanjing 210096, China³ School of Energy and Environment Engineering, Hebei University of Engineering, Handan 056038, China

* Correspondence: hdxgao@hdc.edu.cn (L.G.); zhaoliang@hebeu.edu.cn (L.Z.)

Abstract

Automated compliance checking for building fire protection design requires accurate extraction of domain-specific entities from technical code provisions. However, building fire protection codes contain highly specialized terminology, hierarchical clause structures, and complex regulatory semantics, which make manual information extraction time-consuming and error-prone. This study develops a domain-specific named entity recognition approach for Chinese building fire protection design codes. An expert-annotated dataset was constructed from five representative codes, containing 2748 annotated provisions and 13,877 entity mentions across nine entity categories. A RoBERTa-BiLSTM-CRF model was then developed to capture contextual semantic representations, bidirectional sequence dependencies, and label-transition constraints. Results suggest that the proposed model achieves 88.34%, 88.31%, and 88.32% for precision, recall, and F1-score, respectively. The extracted entities can be organized into structured and traceable records, providing a foundation for downstream building fire protection knowledge management and automated compliance checking.

Keywords: building fire protection codes; named entity recognition; pretrained language model; automated compliance checking; information extraction

1. Introduction

With accelerating automation across the construction industry, building fire protection design codes (BFPDCs) play a crucial role in modern construction projects [1]. As a fundamental standard, BFPDCs are highly specialized and essential to the design, construction, and acceptance processes of buildings to ensure fire safety [2,3]. Existing BFPDCs are generally presented as long texts, and manually analyzing and extracting key information from the codes is time-consuming and error-prone [4]. Therefore, how to automatically and accurately extract key information from BFPDCs to improve the efficiency of the building design process has become a crucial research question [5].

Named entity recognition (NER) is a foundational task in natural language processing (NLP), which extracts targeted entities from unstructured text, forming structured knowledge representations and realizing the transformation from “data” to “knowledge” [6,7]. In BFPDCs, named entities mainly include domain-specific entities such as “building type,” “space type,” and “code clauses” [8]. However, the specific terms used in architecture are



Academic Editor: Jaewook Jeong

Received: 9 June 2026

Revised: 5 July 2026

Accepted: 8 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

unique, and traditional NER methods, such as rule-based matching or feature engineering-based machine learning methods, tend to struggle with field-specific texts [9]. The main challenges are as follows: (1) BFPDCs include numerous domain-specific terms and complex contextual logic, and (2) the formats and structures of code texts are diverse, making it difficult for traditional methods to generalize [10,11].

In recent years, automated compliance checking has increasingly been integrated with digital construction technologies such as Building Information Modeling (BIM), Artificial Intelligence (AI), and Digital Twin (DT). In this context, regulatory provisions need to be transformed into structured and machine-interpretable knowledge so that they can be linked with building model information and used for downstream decision support. For example, Szymczak-Graczyk et al. developed a DT-based model for predicting the hygrothermal performance of building materials from moisture permeability tests, demonstrating the value of integrating experimental data and computational models for intelligent decision-making in building engineering [12]. Therefore, domain-specific NER can be regarded as a foundational step for transforming unstructured fire protection regulations into structured knowledge units for BIM-based compliance checking and AI-supported compliance checking.

Considerable progress has been made in NLP techniques based on pretrained language models [13,14]. Pretrained models gather detailed information related to language and meaning by learning from large sets of general text, which can be adjusted for specific tasks to enhance the results [15]. The pretrained model-based method provides substantial benefits compared with the conventional methods. These advantages include its capacity to autonomously acquire context-relevant semantic representations without depending on hand-crafted features and adaptability to domain-specific tasks via fine-tuning, which enhances performance on specialized corpora [16]. However, bidirectional Transformer encoders (BERTs), when employed as feature encoders, require integration with task-relevant structured models to achieve better sequence labeling performance [17,18]. Deep learning-based pretrained models and traditional sequence labeling methods have been widely used for NER tasks [19,20]. However, the combination of these methods to construct an efficient model to resolve the entity extraction problem in BFPDCs has not been fully investigated.

To address this challenge, this paper proposes a pre-trained language model-based entity extraction approach for BFPDCs, aiming to improve recognition accuracy and domain transferability. The major contributions are summarized below:

- (1) A domain-specific annotated dataset was constructed from five Chinese building fire protection design codes, covering 2748 provisions and 13,877 entity mentions across nine entity categories.
- (2) A pretrained model-based entity extraction method is developed for fine-grained entity recognition in highly structured and terminology-intensive code texts. The model combines contextual semantic representation, bidirectional sequence modeling, and label-transition constraints to improve entity boundary recognition and category prediction.
- (3) Experiments show that the proposed method performs well, and a comparative analysis with traditional methods further supports its effectiveness.

This paper is organized as follows: Section 2 surveys related work. Section 3 details our proposed methodology and dataset. Section 4 reports experimental findings and analysis. Section 5 compares our approach with representative deep learning baselines, and Section 6 concludes the paper with a summary of findings and future directions.

2. Literature Review

2.1. Building Fire Protection Design Codes

The BFPDCs are a crucial regulatory and technical document in the building construction industry. Its main purpose is to ensure that the constructed building provides adequate safety in case of fire. As a core component of building design, BFPDCs provide detailed specific requirements for fire protection facilities, equipment, fire protection measures, and fire protection design [21]. With the continuous development and automation of the construction industry, these codes have become more detailed and complex in practice [22,23].

The specialization and complexity of BFPDCs bring challenges to the identification of named entities [24]. First, the code language contains an abundance of specialized terms and industry definitions, such as firewall, fire water source, and fire compartment, which have strict definitions and specific usage scenarios and are uncommon in ordinary contexts. Second, the standards and requirements involved in code texts are often complex in terms of technical content as well as multilevel code structure and interrelated provisions [25]. Manual methods are often inefficient and error-prone when attempting to understand and extract these complex rules and standards. Therefore, accurately recognizing and understanding these criteria and requirements automatically has become a central challenge in NER.

Therefore, NER technology has remarkable application value in building fire design specifications, which can automatically identify key information from complex text, considerably lessen the need for manual processing, and enhance the efficiency of information processing [26]. As NLP technologies advance, NER has become an effective tool for the text analysis of BFPDCs, enabling efficient and accurate information extraction and promoting the intelligent process of the building design industry [27,28].

2.2. Named Entity Recognition

Rule-based NER methods rely on manually formulated rules and patterns to recognize named entities [29]. They can achieve high accuracy in specific domains when the domain knowledge is more explicit [30]. However, the design and maintenance of rule-based NER methods involve a large number of domain experts, which increases their development cost, and the rules need to be updated and adjusted accordingly with the change in domain knowledge [31]. Rules are generally designed for specific types of entities, which weaken their ability to recognize novel or unseen entities [32]. Consequently, rule-based approaches perform poorly for large amounts of uncertain and changing natural language text.

Unlike rule-based NER methods, statistical models can automatically learn patterns and laws of entity recognition from data by utilizing large-scale labeled datasets and machine learning algorithms. Common statistical models, such as Hidden Markov Model (HMM) [33], CRF [34], and Maximum Entropy Markov Model (MEMM) [35], have been widely used in NER tasks. Despite being simple and effective, HMM results in limited performance when dealing with complex contexts owing to its difficulty in capturing long-range dependencies in sequences [36]. Compared with HMM, CRF can estimate globally normalized conditional probabilities and model label transition dependencies more effectively, thus performing better in sequence labeling tasks [37]. The advantage of CRF lies in its joint decoding over the entire label sequence, which enforces global sequence-level consistency and reduces invalid label combinations, making it suitable for tasks that require structured prediction with global constraints [38].

Deep learning techniques eliminate the need for laborious feature engineering compared with the machine learning-based techniques [39]. Multilayer neural networks can

abstract data, enabling deep learning models to automatically identify hidden properties. NER tasks have advanced significantly, particularly with the use of convolutional neural networks and LSTM networks [40,41]. Alsuwat et al. [42] proposed a model for acquiring temporal features based on BiLSTM networks, which models contextual dependencies in both directions. Zhang et al. [43] introduced the BiLSTM-CRF model, which fully leverages the benefits of BiLSTM in modeling sequential information and optimizes the label sequence using the CRF layer to further enhance NER performance. Nevertheless, deep learning techniques based on BiLSTM and CRF still find it challenging to acquire large-scale annotated corpora for model training and consume significant computational resources [44,45].

2.3. Pre-Trained Language Models

Pretrained BERT models have made substantial advances in NLP and demonstrated exceptional performance in a variety of NLP applications [46,47]. BERT has given an outstanding performance in information extraction [48], text categorization [49], and automated question answering owing to its robust context modeling skills in sequence annotation [50]. Pretraining on a large-scale corpus enables BERT to capture richer semantic information, providing strong support for domain-specific NLP tasks, particularly in the research areas of domain knowledge understanding and data enhancement [51,52].

BERT utilizes the bidirectional attention mechanism to capture contextual relationships, but independent token-level predictions may lead to inconsistent boundaries for multi-word entities [53]. This issue can be alleviated by introducing BiLSTM to further model sequential dependencies and enhance entity boundary recognition [54]. Moreover, adding a CRF layer to BERT can ensure the completeness of entity boundaries by modeling dependencies between adjacent labels [55].

Even though BERT is trained on a wide range of text and it excels at understanding language, it does not always effectively handle specialized fields, such as architectural engineering and fire protection design, as it misses some specific terms and grammar, leading to misunderstandings [3]. RoBERTa, which is more fully structurally optimized, is chosen as the base model to enhance the adaptability of the pretrained model in prescriptive texts, and domain-specific fine-tuning strategies are introduced [56,57]. The performance of the model in legal, biomedical, and engineering named entity recognition tasks has been shown to be remarkably improved by fine-tuning domain-specific corpora.

In summary, by introducing modules such as BiLSTM and CRF, these difficulties of the RoBERTa model in NER can be effectively solved, and the accuracy and robustness of the model can be enhanced [58]. Specifically, BiLSTM helps capture long-distance dependencies in text, while CRF can model the dependencies between labels and optimize the boundaries of entities and the order of labels, making RoBERTa more accurate and effective in NER tasks [59,60]. Therefore, the proposed model, which is based on RoBERTa, BiLSTM, and CRF, uses RoBERTa to understand context, combines with BiLSTM to recognize long-distance relationships, and uses the CRF layer to understand how labels relate to each other, effectively improving the accuracy of NER in BFPDCs.

3. Methodology

Figure 1 presents the overall framework for NER in BFPDCs. The overall workflow is organized into three steps: data collection and preprocessing, entity classification and labeling, and the entity extraction model. First, data collection and preprocessing are fundamental to the entire entity extraction process. A large amount of text data was collected from BFPDCs, and the data underwent word segmentation and other processing. The complex building fire protection design knowledge was structured into an easily manageable

JSON format. Second, entity classification and labeling involve classifying different types of entities and providing detailed BIO annotations to the text data, ensuring that the data provides standardized input for subsequent model training. Finally, in the entity recognition and extraction stage, this paper establishes a RoBERTa-BiLSTM-CRF model, which achieves accurate entity extraction from BFPDCs by jointly leveraging the encoding and decoding layers.

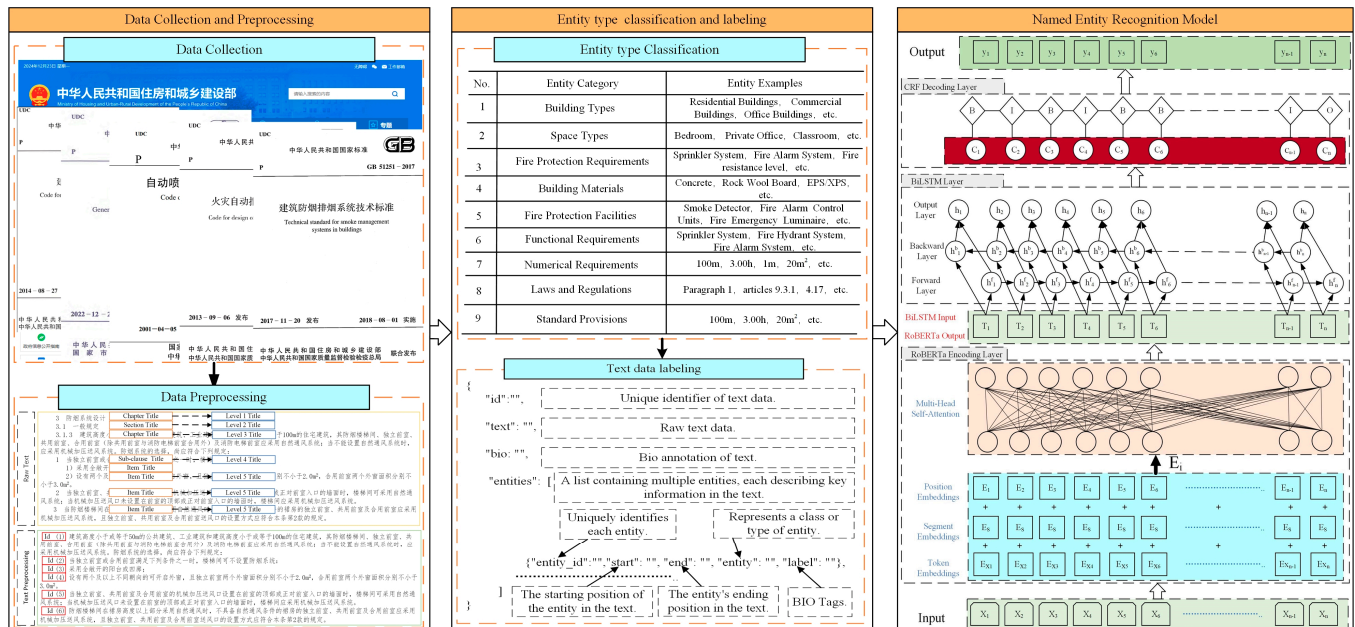


Figure 1. Pretrained model-based NER framework for BFPDCs.

3.1. Data Collection and Preprocessing

3.1.1. Data Collection

In NER, high-quality data collection is fundamental to building effective models. This is especially true for entity extraction from codes related to building fire protection based on pretrained models, where the quality of labeled data directly determines model performance. Herein, we target entity extraction on BFPDCs. Therefore, data collection should comply with the relevant requirements of the building fire protection design. Table 1 presents the current major BFPDCs considered.

The data in this study originate from the current effective codes for building fire protection design and representative standards in various specialized fields in the building design process. The reasons behind selecting them are as follows: (1) These codes are the fundamental documents for fire protection design in China's building design industry. These codes, widely used in operations for designing fire protection in buildings, are authoritative and legally binding. Therefore, the information contained in these codes is quite useful for identifying entities, (2) These codes cover several important aspects of building fire protection design, such as alarm systems, fire extinguishing systems, smoke control, and exhaust. The model can accurately extract different objects using these documents as a source of information, ensuring that all key parts of building fire protection designs are included, and (3) a wide range of technical provisions, standard requirements, system components, and equipment types are included in BFPDCs. By extracting entities in these standard texts, rich training samples can be provided for the model, thereby effectively improving its ability to recognize terms in the field of building fire protection.

Table 1. Primary data sources for BFPDCs.

No.	Code for Fire Protection	Description
1	Code for fire protection design of buildings (GB50016-2014) [61]	This is a basic and comprehensive technical regulation that primarily regulates the fire protection design of buildings. It is applicable to newly built, expanded, or renovated industrial and civil buildings, and it aims to ensure that the buildings can protect the safety of life and property to the greatest extent in case of fire.
2	General code for fire protection of buildings and constructions (GB 55037-2022) [62]	This specification is a mandatory engineering construction specification that covers all aspects of building fire protection and requires the strict implementation of all fire protection design provisions in engineering construction. This is a comprehensive requirement for building fire protection technology to ensure that building design meets safety specifications.
3	Code for design of sprinkler systems (GB50084-2017) [63]	This specification specifies in detail the design requirements and technical standards for automatic sprinkler fire extinguishing systems, with the aim of ensuring that the system can effectively extinguish fires when they occur.
4	Code for design of automatic fire alarm system (GB50116-2013) [64]	This specification aims to ensure that when a fire catches, the source of the fire can be discovered and the alarm can be promptly sounded through the alarm system so that effective measures can be taken.
5	Technical standard for smoke management systems in buildings (GB 51251-2017) [65]	This standard provides technical requirements for the design, construction, and acceptance of smoke prevention and exhaust facilities in buildings, ensuring that when it catches fire, the smoke in the building can be promptly cleared and smooth evacuation of personnel and firefighters can be ensured.

3.1.2. Data Preprocessing

Note that Table 1 only lists authoritative guidance documents from BFPDCs, which are highly representative when constructing the entity recognition dataset. As the design code texts are generally presented in a semi-structured format, appropriate preprocessing is required to improve the accuracy of the subsequent NER task. Our preprocessing procedure builds on previously published methods.

The structure level of the BFPDCs shown in Figure 2 generally includes chapters, sections, clauses, subclauses, and items. Each level represents a different degree of abstraction, ranging from conceptual provisions to specific operational requirements. Among them, the chapter is the highest level in a building design code and generally represents the main part of the code. Sections further refine the content of a chapter, breaking down the broad areas into multiple more specific parts, each of which generally involves a category of related design requirements or standards. Clauses are the core units of content in a specific section, each defining a specific design requirement or technical standard. Sub-clauses are usually further refinements of clauses. The item is the smallest unit in the code and is used to further refine the specific operational requirements in the clause or subclause or to specify specific technical parameters and design requirements. Therefore, in this study, the text data such as clauses, subclauses, and items are regarded as independent data units, and each unit is assigned a unique identifier for labeling.

3.2. Entity Type Classification and Labeling

3.2.1. Entity Type Classification

The entities in BFPDCs are classified based on pertinent provisions in China and the practical requirements of building design and fire engineering. In accordance with the many phases of building design, functional specifications, technological criteria, and

regulatory standards, fire protection measures are effectively planned and administered throughout all stages. Based on the semantic descriptions and document architecture of diverse information within the code, the entity classifications are reported in Table 2.

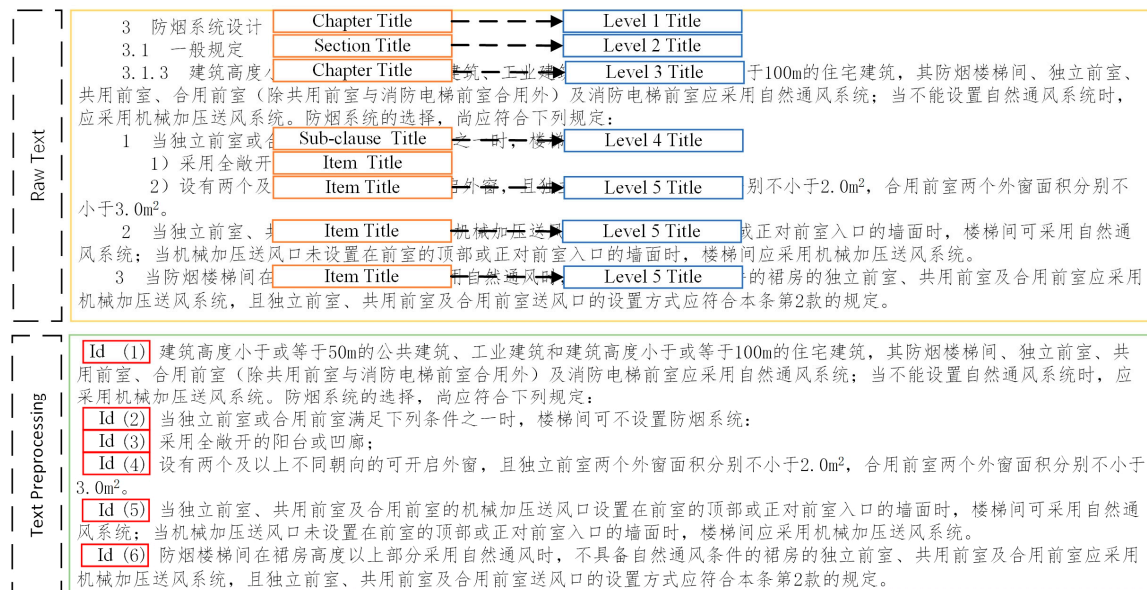


Figure 2. Text data structure.

Table 2. Entities classification of BFPDCs.

No.	Entities	Label	Description
1	Building Types	BT	This is a method of classifying buildings as a whole according to standards such as building function, structural form, risk level, or design purpose.
2	Space Types	ST	Specific functional usages divide subareas within a building.
3	Fire Protection Requirements	FPR	Fire Protection Requirements refer to regulatory requirements that specify fire-safety protection measures, configuration conditions, or protective provisions that should be provided or satisfied in building fire protection design.
4	Building Materials	BM	Building materials refer to all types of materials used for constructing and decorating buildings. According to fire safety requirements, building materials are primarily categorized and regulated based on their combustion and fire resistance properties.
5	Fire Protection Facilities	FPF	Fire protection facilities are systematic equipment used in building design to prevent, control, and extinguish fire.
6	Functional Requirements	FR	Functional Requirements refer to the operational, control, or performance functions that a specific fire protection facility, system, component, or space should possess to support fire prevention, evacuation, smoke control, or firefighting.
7	Numerical Requirements	NR	Numerical requirements are the standardized regulations for the spatial dimensions of various key design elements in building design to ensure that the requirements of the building in terms of functionality, safety, and operability are met.
8	Laws and Regulations	LR	The design for building fire protection needs to be carried out in accordance with national and local laws and regulations to ensure that the design plan meets statutory standards and safety requirements.
9	Standard Provisions	SC	During the design process for building fire protection, standard provisions refer to other clauses in this code or other pertinent codes.

3.2.2. Text Data Labeling

After completing the preprocessing of BFPDCs, the hierarchical structure of the codes (chapter-section-clause/item) is converted into model-ready samples. The specific approach is as follows: First, the original number of each provision (e.g., chapter/section/clause number) is retained, and then the provision is divided into sentence-level/item-level segments according to the punctuation and item separators commonly used in Chinese technical standards (e.g., “,”, “.”, “:”, and numbered items). To avoid information leakage, all segments from the same provision identifier are forcibly assigned to the same data subset during data partitioning. Each segment adopts the BIO annotation scheme at the entity span level, where B is assigned to the first token of the entity span, I to subsequent tokens within the same entity, and O to non-entity tokens. This design preserves the traceability of the extraction results back to the original provisions, providing a basis for evidence location in downstream compliance review.

To improve annotation consistency, this paper compiles annotation guidelines for fire protection codes, clarifying rules for entity boundary determination, handling of nested/overlapping cases, and normalization rules. For ambiguous samples, domain experts and annotators iteratively discuss and update the guidelines accordingly, and a consistency review is performed on all data. To further ensure annotation reliability, a subset of the corpus was independently checked by two annotators according to the predefined annotation guidelines. Cohen’s Kappa was calculated at the BIO-label level, achieving a value of 0.86, which indicates good inter-annotator agreement. Inconsistent labels were mainly related to semantically similar categories, such as FPR, FR, and BM, and these cases were further reviewed and finalized through discussion with domain experts. The structure of the annotated data is shown in Figure 3.

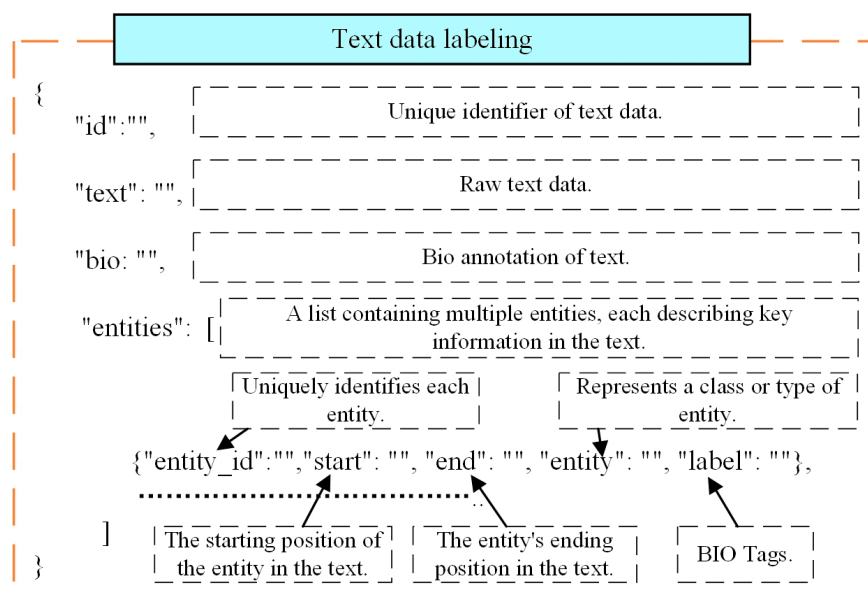


Figure 3. Text data labeling.

3.3. Named Entity Recognition Model

Herein, a multilayer neural network structure integrating RoBERTa, BiLSTM, and CRF is constructed to cope with the complex structure and terminology-intensive NER task in the CFPDB texts. The architecture is illustrated in Figure 4. In this model, RoBERTa serves as the encoder to construct a deep language representation based on a large-scale pretrained corpus, which effectively captures fine-grained semantic features in the context in which the entity is embedded. The BiLSTM module is used to model the forward and

backward bidirectional dependencies in sequences, which enhances the extraction ability of long distance-dependent entities. The CRF decoding layer introduces global constraints between labels in the output phase, which effectively avoids the generation of illegal label sequences. Upon combining these three modules, the model can better capture the semantic structure of complex regulatory texts and has greatly enhanced consistency and accuracy in predicting entity labels. The structure performs well in the typical professional language and highly structural context of building codes, which lays a technical foundation for the automatic extraction and management of fire protection knowledge.

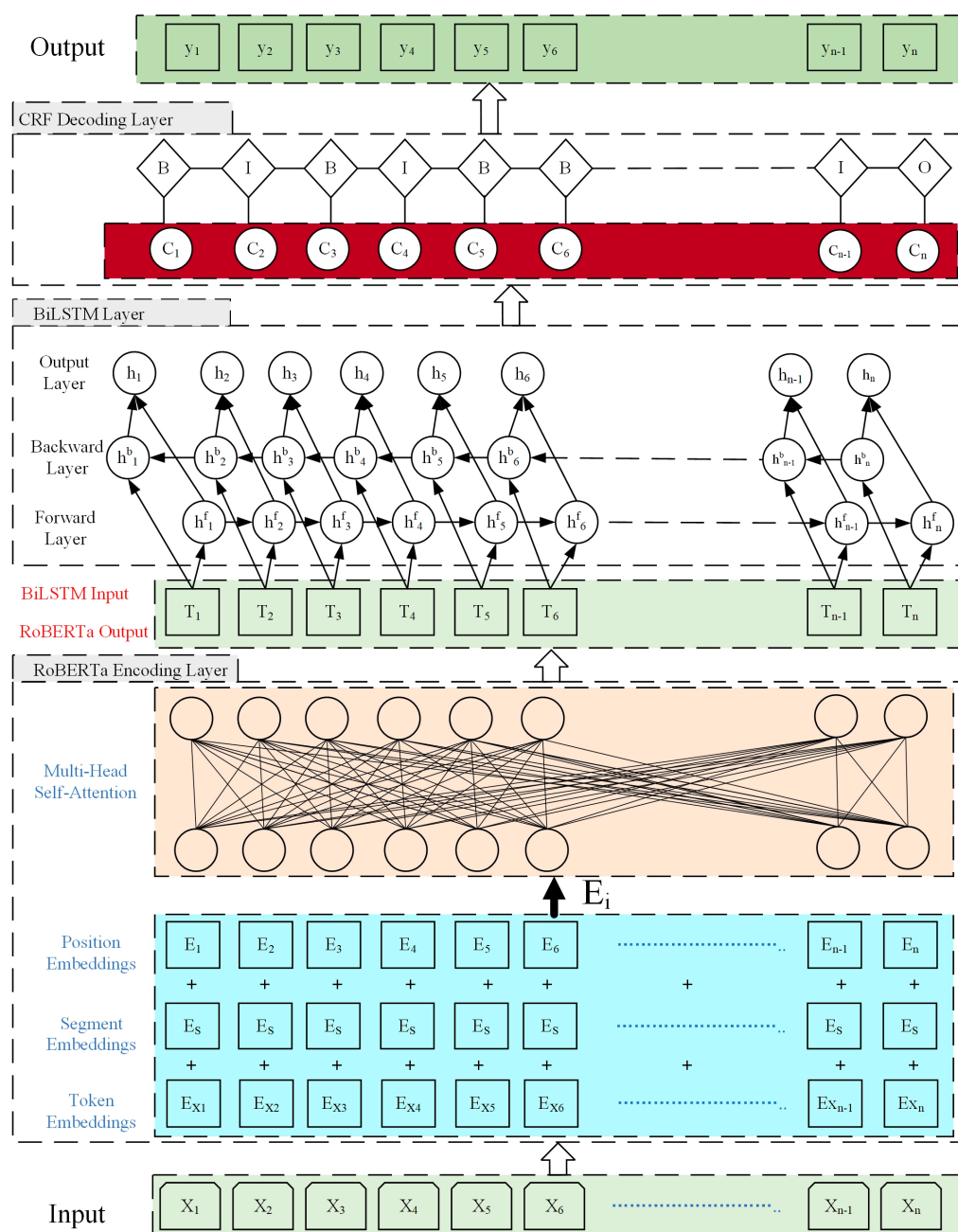


Figure 4. Framework of the RoBERTa-BiLSTM-CRF entity extraction model.

3.3.1. RoBERTa Encoding Layer

Herein, RoBERTa is employed as the text encoder for entity extraction on the BFPDCs, which is primarily responsible for extracting the contextual semantic representations at the lexical and sentence levels. Figure 4 illustrates the core architecture of the model. RoBERTa

is developed based on the Transformer architecture and comprehensively models the contextual dependencies via the multi-head attention mechanism to generate deeply semantic feature representations that effectively support subsequent entity recognition.

Given a sequence $X = \{x_1, x_2, \dots, x_n\}$, each token is represented by a fused embedding formed from the token, position, and segment embeddings, as expressed in Equation (1).

$$E_i = \text{TokenEmb}(x_i) + \text{PositionEmb}(i) + \text{SegmentEmb}(x_i) \quad (1)$$

where $\text{TokenEmb}(x_i)$ is the token embedding, $\text{PositionEmb}(i)$ encodes the positional information of token at position i through learned positional embeddings, and $\text{SegmentEmb}(x_i)$ encodes segment information, which is a constant zero vector in RoBERTa. Let $E = \{E_1, E_2, \dots, E_n\}$ denote the embedding matrix of the whole sequence.

The Transformer architecture in RoBERTa comprises multiple layers, each containing a multi-head self-attention mechanism and a feed-forward neural network. For the j -th attention head, the query, key, and values are obtained by linear projections:

$$Q^{(j)} = EW_Q^{(j)}, K^{(j)} = EW_K^{(j)}, V^{(j)} = EW_V^{(j)} \quad (2)$$

where $W_Q^{(j)}, W_K^{(j)}, W_V^{(j)}$ denote projection matrices and d_k is the key/query dimension per head. The scaled dot-product self-attention output of head j is computed as Equation (3).

$$\text{head}_j = \text{softmax} \left(\frac{Q^{(j)} (K^{(j)})^T}{\sqrt{d_k}} \right) V^{(j)} \quad (3)$$

RoBERTa performs attention in parallel with h heads and concatenates their outputs to obtain the result:

$$M = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) W_o \quad (4)$$

where h denotes the number of attention heads, head_j denotes the j -th self-attention head, and W_o represents the output projection matrix.

After the L -layer Transformer, the final output sequence representation is expressed as Equation (5):

$$T = \{t_1, t_2, \dots, t_n\} = \text{RoBERTa}(X) \quad (5)$$

where T represents the contextualized representations, which serve as input to the subsequent BiLSTM layer.

3.3.2. BiLSTM Layer

The BiLSTM layer is used to encode bidirectional context for the sequence, enabling the model to leverage both preceding and succeeding contextual dependencies. It takes the RoBERTa output $T = \{t_1, t_2, \dots, t_n\}$ as input. The forward LSTM reads the sequence from left to right and updates its hidden state as

$$\vec{h}_i = \text{LSTM}_f \left(t_i, \vec{h}_{i-1} \right) \quad (6)$$

where $\vec{h}_i$ denotes the forward hidden state at time step i , $\vec{h}_{i-1}$ is the previous state, t_i is the current input, and $\text{LSTM}_f(\cdot)$ denotes the forward LSTM cell function with standard gating mechanisms.

Similarly, the backward LSTM processes the sequence in reverse order, computing $\overleftarrow{h}_i$ from the subsequent state $\overleftarrow{h}_{i+1}$ and the current input t_i , as shown in Equation (7):

$$\overleftarrow{h}_i = \text{LSTM}_b(t_i, \overleftarrow{h}_{i+1}) \quad (7)$$

where $\overleftarrow{h}_i$ is the backward hidden state and $\text{LSTM}_b(\cdot)$ follows the same gating mechanism as the forward LSTM.

At each position, the forward and backward hidden states are concatenated to form the bidirectional representation, as expressed in Equation (8):

$$h_i^{\text{BiLSTM}} = \begin{bmatrix} \overrightarrow{h}_i; \overleftarrow{h}_i \end{bmatrix} \quad (8)$$

where $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$ are the representations obtained from the two directional LSTM passes, respectively. $[\cdot; \cdot]$ denotes vector concatenation.

The resulting sequence $H^{\text{BiLSTM}} = \{h_1^{\text{BiLSTM}}, h_2^{\text{BiLSTM}}, \dots, h_n^{\text{BiLSTM}}\}$ provides enriched bidirectional features and is fed into the subsequent CRF layer for structured label decoding.

3.3.3. CRF Decoding Layer

The CRF layer is adopted to model dependencies among adjacent labels and to generate a globally optimal tag sequence. Unlike token-wise classification, linear-chain CRF jointly decodes the entire sequence by considering both emission scores from the BiLSTM and transition scores between labels, which enforces valid label transitions and improves sequence consistency.

Given the BiLSTM outputs H^{BiLSTM} , a linear layer produces the emission score vector at each position i :

$$p_i = Wh_i^{\text{BiLSTM}} + b, \quad p_i \in \mathbb{R}^K \quad (9)$$

where K is the number of labels, W and b are trainable model parameters, and $p_{i,k}$ denotes the emission score of assigning label k to position i .

Let $A \in \mathbb{R}^{K \times K}$ be the trainable transition matrix, where $A_{a,b}$ represents the transition score from label a to b . For a label sequence $y = \{y_1, y_2, \dots, y_n\}$ given input X , the CRF defines a sequence score as

$$s(X, y) = \sum_{i=1}^n (A_{y_{i-1}, y_i} + p_{i, y_i}) \quad (10)$$

where y_0 denotes a special start label. The conditional probability of y given X is

$$P(y|X) = \frac{\exp(s(X, y))}{\sum_{y'} \exp(s(X, y'))} \quad (11)$$

where the denominator sums over all possible label sequences, which can be efficiently computed using the forward algorithm.

During training, the objective is to maximize the log-likelihood of the correct label sequence. During inference, the optimal label sequence y^* is obtained by finding the sequence that maximizes the conditional probability, as shown in Equation (12):

$$y^* = \underset{y}{\operatorname{argmax}} P(y, X) = \underset{y}{\operatorname{argmax}} s(X, y) \quad (12)$$

where y^* represents the predicted label sequence. The decoding process is efficiently performed using the Viterbi algorithm, which finds the highest-scoring path through dynamic programming.

4. Experiment

4.1. Dataset Construction

This study aims to construct an entity extraction dataset from BFPDCs. The chosen BFPDCs are presented in Table 1. The collected code texts were first preprocessed, followed by entity classification and manual annotation. In total, 2748 annotated samples were obtained, with sample lengths ranging from 3 to 233 characters (Table 3). In the experimental evaluation, a ratio of 8:1:1 was used to divide the training set, validation set, and test set.

Table 3. Statistics of the CFPDB dataset.

No.	Code for Fire Protection	Number of Provisions	Length	
			Min	Max
1	Code for fire protection design of buildings (GB 50016-2014) [61]	883	7	223
2	General code for fire protection of buildings and constructions (GB 55037-2022) [62]	727	3	184
3	Code for design of sprinkler systems (GB 50084-2017) [63]	254	7	174
4	Code for design of automatic fire alarm system (GB 50116-2013) [64]	439	5	229
5	Technical standard for smoke management systems in buildings (GB 51251-2017) [65]	445	5	229

The reasons for not removing regulatory function words and normative expressions while developing the building fire protection design code dataset are primarily based on the following points: (1) The text of BFPDCs has a strong structure and regularity, and regulatory function words and normative expressions play an important grammatical function in it [66]; (2) as a technical document, the language of BFPDCs is rigorous and standardized, and the regulatory function words and normative expressions are often used to connect the technical clauses, definitions, and conventions, which has an indispensable grammatical role [67]; and (3) models such as RoBERTa have strong contextual understanding and can effectively recognize entities without removing regulatory function words and normative expressions. Therefore, retaining the regulatory function words and normative expressions may not have a considerable negative impact on the prediction accuracy. The entity statistics are presented in Table 4, which includes 13,877 labeled entities distributed across nine categories.

4.2. Evaluation Metrics

This study conducts evaluation at the entity level and adopts three standard metrics commonly used for named entity recognition: Precision, Recall, and F1-score. Following Equations (13) and (14), Precision reflects the correctness of the model's predicted entity mentions, while Recall quantifies how many gold-standard entity mentions are successfully recovered.

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

where TP , FP , and FN denote True Positives, False Positives, and False Negatives, respectively.

The F1 score, computed by Equation (15), aggregates Precision and Recall via the harmonic mean to summarize their trade-off, as calculated in Equation (15):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (15)$$

F1 lies in $[0, 1]$ and increases when both precision and recall are high, offering a comprehensive performance metric. Unless otherwise specified, all results reported in this paper are micro-averaged scores computed over all entity instances.

Table 4. Entity category statistics.

No.	Code for Fire Protection	Number of Entities									
		BT	ST	FPR	BM	FPF	FR	NR	LR	SC	Total
1	Code for fire protection design of buildings	712	818	686	132	1298	584	1030	39	139	5492
2	General code for fire protection of buildings and constructions	531	1051	372	158	617	297	510	0	36	3572
3	Code for design of sprinkler systems	34	111	30	36	552	249	170	4	83	1269
4	Code for design of automatic fire alarm system	19	328	112	34	832	176	217	5	40	1763
5	Technical standard for smoke management systems in buildings	20	304	299	57	740	130	239	5	41	1765
	Total	1316	2612	1499	417	4039	1436	2166	53	339	13,877

4.3. Experimental Results

In this experiment, key hyperparameters considered during model training include max_len, learning rate, batch size, and epoch. Furthermore, the partitioning of the training, validation, and test sets significantly impacts model performance. Combining the statistical results in Table 3 with the characteristics of text structure, the max_len is set to 240. To evaluate the effects of different parameters on model performance, a univariate control strategy is adopted, where each hyperparameter is varied while keeping the others fixed, and all configurations are evaluated on the same fixed hold-out split. To guarantee that the model has generalization capacity on unseen samples, the dataset is divided into an 8:1:1 ratio of training, validation, and test sets.

Table 5 summarizes the average performance of the RoBERTa-BiLSTM-CRF model in recognizing named entities in BFPDCs under different hyperparameter settings. Results suggest that the parameters of the learning rate, batch size, and epoch notably affect the performance of the model. In particular, the model demonstrated the best performance on average when the learning rate was set to 2×10^{-5} , the batch size to 8, and the epoch to 90. These results provide empirical support for model parameter optimization and provide a reference basis for subsequent model training.

Table 5. Performance under different parameters.

NO.	Parameter	Value	Evaluation Metrics		
			Precision	Recall	F1
1	Learning Rate	1×10^{-5}	0.8757	0.8768	0.8763
2		2×10^{-5}	0.8834	0.8831	0.8832
3		3×10^{-5}	0.8742	0.8789	0.8765
4		4×10^{-5}	0.8724	0.8789	0.8756
5		5×10^{-5}	0.876	0.8818	0.8789

Table 5. Cont.

NO.	Parameter	Value	Evaluation Metrics		
			Precision	Recall	F1
6	Batch Size	8	0.8834	0.8831	0.8832
7		16	0.8653	0.8718	0.8686
8		24	0.8476	0.8606	0.8542
9		32	0.8699	0.8764	0.8731
10		40	0.8607	0.8714	0.866
11		48	0.8515	0.8664	0.8589
12		56	0.8425	0.8615	0.8518
13		64	0.8335	0.8566	0.8449
14	Epoch	10	0.8247	0.8419	0.8332
15		20	0.8527	0.8577	0.8552
16		30	0.8634	0.871	0.8672
17		40	0.8573	0.8747	0.8659
18		50	0.8695	0.8764	0.873
19		60	0.8695	0.8764	0.873
20		70	0.8714	0.8768	0.8741
21		80	0.8714	0.8768	0.8741
22		90	0.8834	0.8831	0.8832
23		100	0.8775	0.8826	0.8801

5. Discussion

5.1. Model Performance Analysis

In the NER task, hyperparameter settings significantly impact the model performance. The learning rate, batch size, and epoch are selected for sensitivity analysis to comprehensively evaluate the model's robustness and generalizability. This is to explore the fluctuation of model performance under different configurations to provide theoretical support and practical reference for subsequent model optimization.

Learning rate significantly affects how quickly and smoothly the model converges. We therefore tested five values (1×10^{-5} to 5×10^{-5}) while keeping all other hyperparameters fixed (Figure 5). The learning rate of 2×10^{-5} achieved the best performance, with the highest F1 score of 0.8832, and exhibited stable convergence. A larger learning rate may lead to unstable optimization, whereas a smaller learning rate typically results in slower convergence and may require more training steps. Therefore, we use 2×10^{-5} as the learning rate in subsequent experiments.

Figure 6a indicates that adopting batch size 8 achieves a higher F1 score, reaching 0.8832. When the batch size increases to 64, the F1 score drops considerably. This may be because larger batch sizes reduce the frequency and stochasticity of gradient updates, which can weaken the model's adaptability to fine-grained entity boundaries in a relatively small domain-specific corpus. As illustrated in Figure 6b, the F1 improves with more training epochs; however, when epochs exceed 90, the performance of a few batch-size configurations starts to decrease, suggesting the onset of overfitting. This indicates that excessive training may cause the model to further fit domain-specific expressions and annotation patterns in the training data while reducing its generalization ability on unseen samples. Considering both performance and training efficiency, we adopt 90 epochs in this study.

In the NER task, the partitioning strategy of the dataset has a considerable impact on the training effect. This study adopts three ratio settings to evaluate the impact of different portioning schemes, and Figure 7 summarizes the comparative results. The results show that adopting the 8:1:1 split technique enables the model to achieve the highest F1 score. This ratio provides sufficient training data while reserving appropriate subsets for valida-

tion and testing. Both alternative partitioning schemes (7:1.5:1.5 and 6:2:2) result in lower performance than 8:1:1, possibly because they allocate fewer samples for training. Overall, the 8:1:1 data partitioning method performs well for this task and is recommended for subsequent experiments.

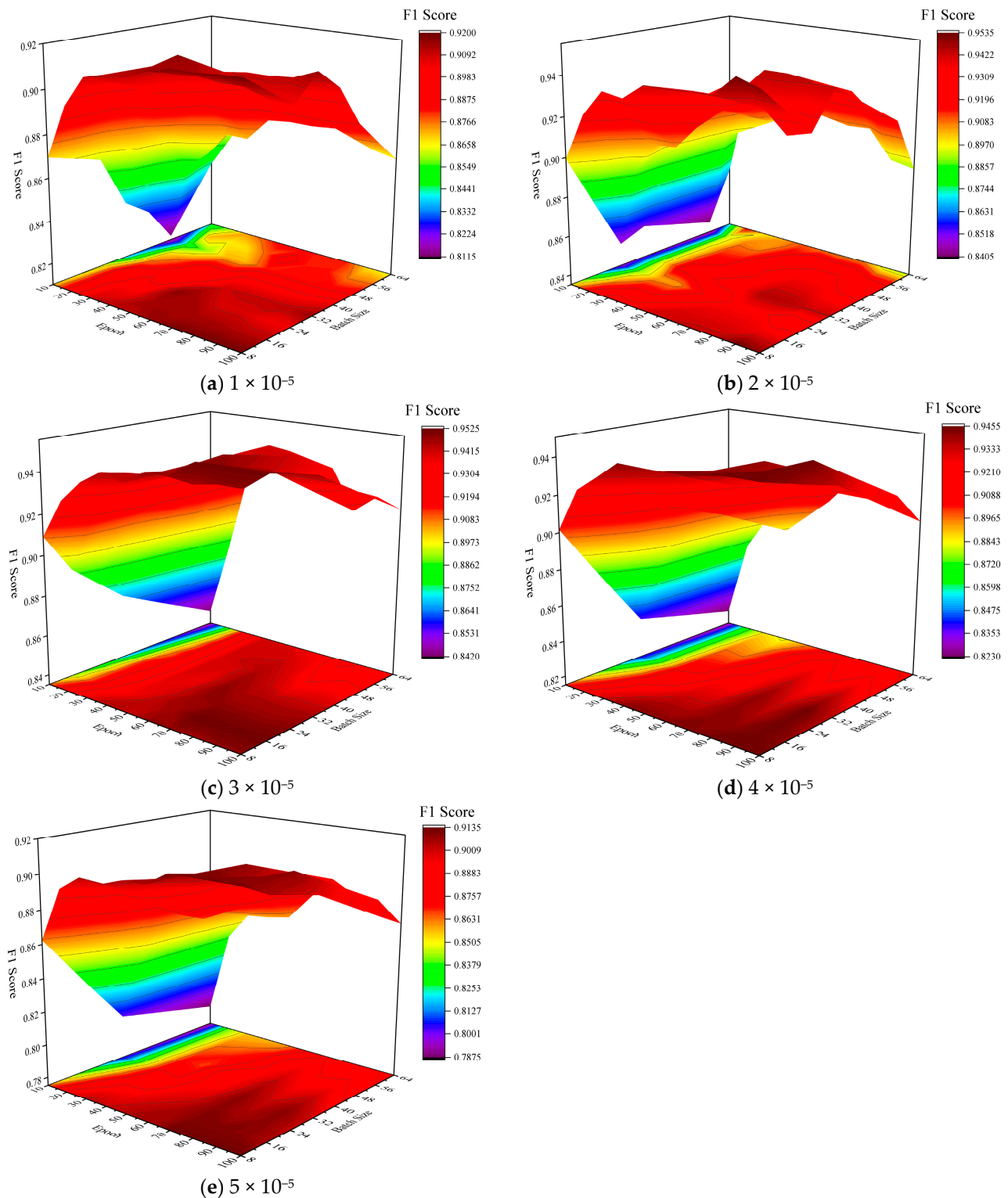


Figure 5. Impact of learning rate on the model.

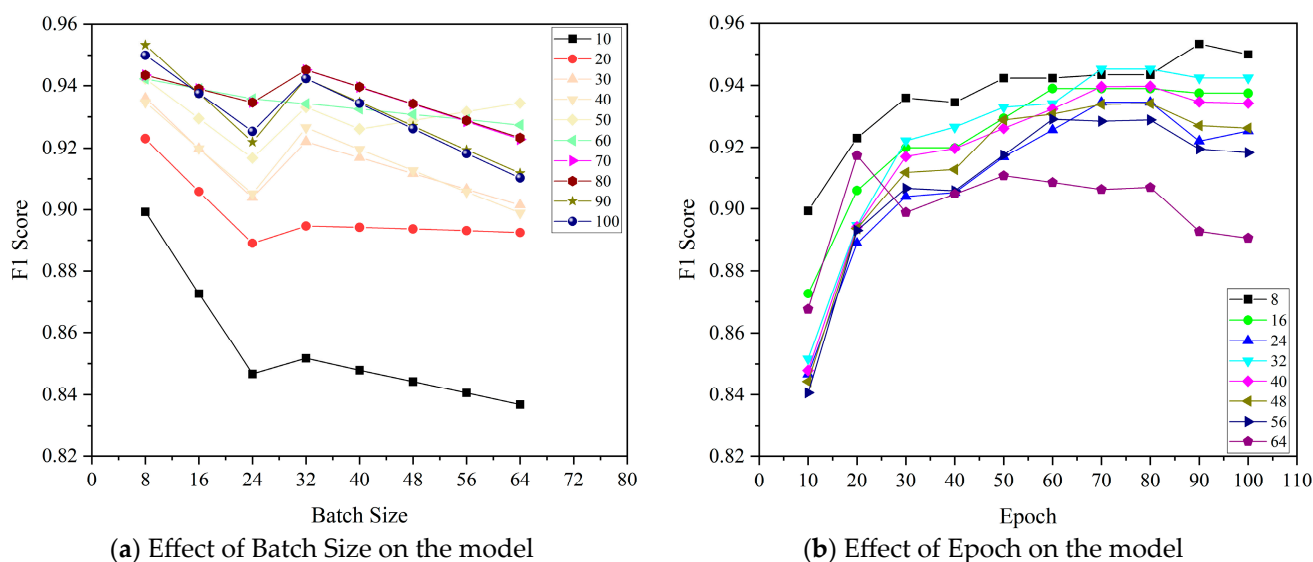


Figure 6. Effect of Batch Size and Epoch on the model.

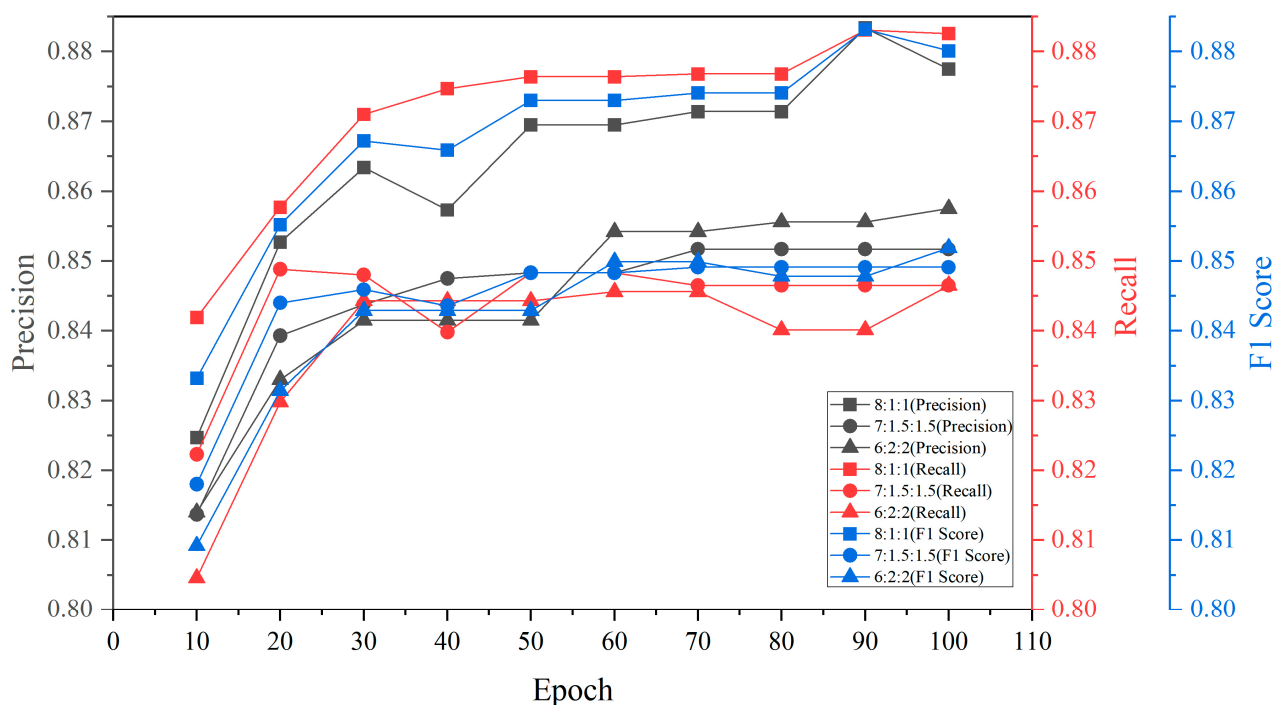


Figure 7. Impact of dataset partitioning on modeling.

5.2. Ablation Experiment and Model Effectiveness Analysis

5.2.1. Comparison of the Adaptability of Various BERT Models

This study compares six mainstream models, including the standard BERT, RoBERTa, MacBERT, ChineseBERT, mBERT, and TinyBERT, to assess the efficacy of BERT and its variants in the NER task of BFPDCs. Figure 8a depicts the experimental findings for each model. The results indicate that RoBERTa delivers the best performance, achieving an F1 score of 0.8832, which is higher than that of the standard BERT baseline. This result suggests that RoBERTa's pretraining strategy and representations may be more effective for this task. Furthermore, MacBERT performs better (F1 score = 0.8736), ranking second only to RoBERTa. In comparison, ChineseBERT attains an F1 score of 0.8394, indicating that its performance is relatively lower on this specialized dataset. Although the lightweight

TinyBERT model yields a lower F1 score (0.8501), its smaller model size and faster inference potential make it an attractive option for deployment in resource-constrained scenarios.

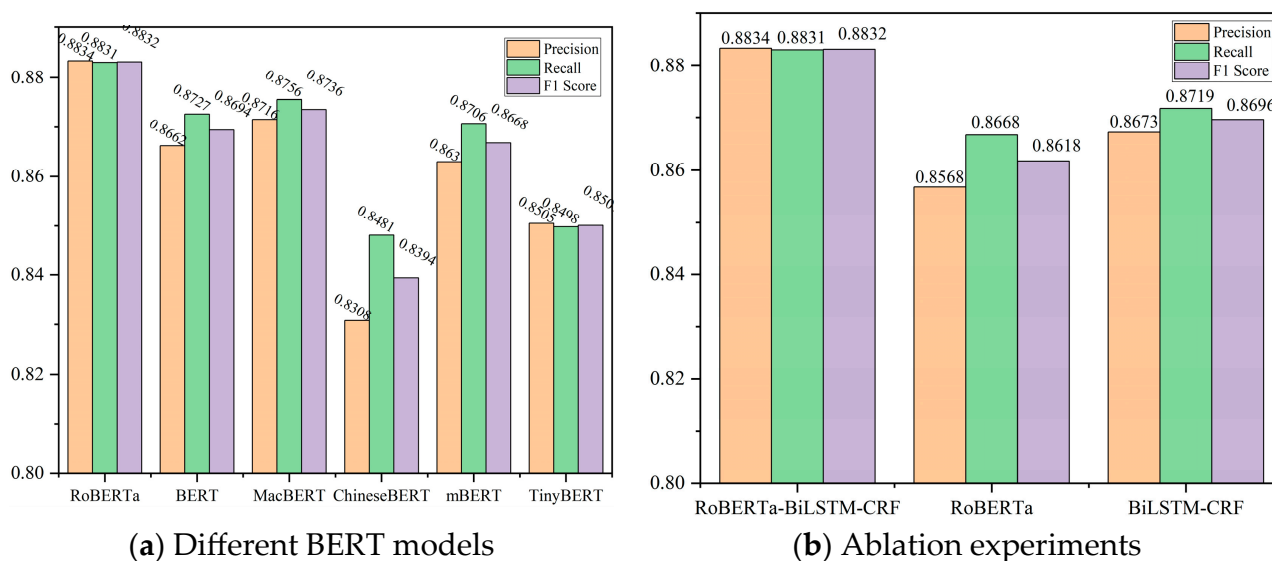


Figure 8. Effectiveness analysis of RoBERTa.

To further understand how RoBERTa interacts with other modeling components, this study proposes experiments that eliminate specific structures and examine how well the combination of RoBERTa-BiLSTM-CRF, RoBERTa, and BiLSTM-CRF performs. As presented in Figure 8b, the RoBERTa-BiLSTM-CRF arrangement is the most effective at recognizing entity boundaries and interpreting meanings.

The F1 score for the structure with only RoBERTa is 0.8618, indicating proficient contextual semantic modeling capabilities; nevertheless, it lacks explicit modeling of label dependencies, resulting in a relatively ambiguous recognition entity boundary. Similarly, while BiLSTM-CRF achieves a recall of 0.8719, the absence of pretraining semantic support may result in a precision of 0.8568, yielding an F1 score of 0.8696, thereby highlighting the important role of pretrained models in enhancing entity extraction performance.

In summary, the synergistic combination of RoBERTa and BiLSTM-CRF remarkably improves the ability to model long-distance dependencies and professional terminology structures in BFPDCs texts.

5.2.2. Analysis of the Role of BiLSTM with Parameter Tuning

BiLSTM is introduced as a downstream module to further enhance the sequence modeling capability of the proposed model. Built on top of contextual embeddings, BiLSTM can better capture word order and bidirectional dependency relationships, which is suitable for sequence labeling in building fire protection codes with complex structures. To validate the performance of BiLSTM hyperparameters, this study designs parameter ablation experiments, including num_layers, hidden_size, and dropout, and Table 6 reports the results.

Experiments show that in a single-layer BiLSTM, when the hidden_size is 256 and the dropout is 0.3, the model achieves superior performance, suggesting that an appropriate model capacity and regularization are helpful in this setting. The two-layer BiLSTM structure achieves the highest F1 score in this group of experiments with a dropout of 0.3 and a hidden_size of 256, indicating that a deeper BiLSTM can be beneficial, but its performance depends on suitable regularization and model size. When the number of layers increases but the dropout is too small or the hidden size is too large, the model exhibits

performance fluctuations in some experimental groups, reflecting the sensitivity of the BiLSTM structure to hyperparameters in complex tasks.

Table 6. Model performance under different BiLSTM parameter configurations.

Num_Layers	Dropout	Hidden_Size	Precision	Recall	F1 Score
1	0.2	64	0.8702	0.8768	0.8735
1	0.2	128	0.8705	0.8731	0.8718
1	0.2	256	0.8696	0.8685	0.8689
1	0.3	64	0.8775	0.8672	0.8723
1	0.3	128	0.8685	0.8763	0.8724
1	0.3	256	0.8752	0.8785	0.8768
2	0.2	64	0.8708	0.8776	0.8742
2	0.2	128	0.8734	0.8731	0.8733
2	0.2	256	0.8668	0.8772	0.8720
2	0.3	64	0.8630	0.8731	0.8680
2	0.3	128	0.8811	0.8822	0.8816
2	0.3	256	0.8834	0.8831	0.8832

Figure 9a shows the model performance on the NER task under different BiLSTM parameter settings. The results indicate that the model performance generally improves as num_layers and hidden_size increase within the tested range. When dropout is applied, the performance is more stable across settings. The best performance is achieved with num_layers = 2, dropout = 0.3, and hidden_size = 256, reaching an F1 score of 0.8832.

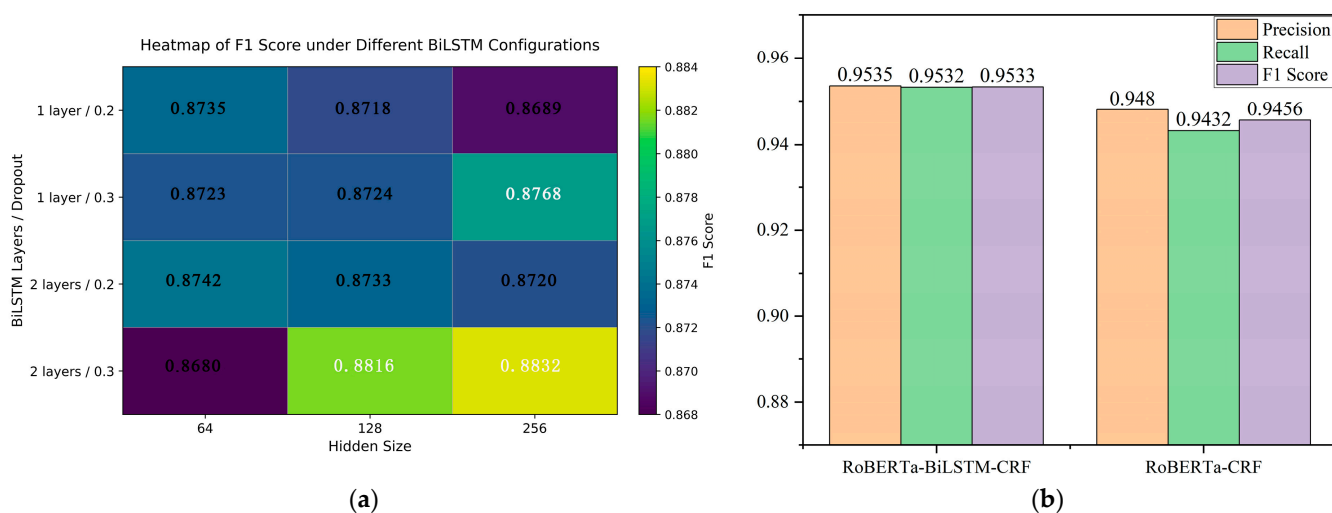


Figure 9. BiLSTM effectiveness analysis. (a) F1 score heatmap for different combinations of hidden size, dropout, and number of layers; (b) BiLSTM ablation experiments.

Figure 9b further analyzes the contribution of the BiLSTM component via structural ablation. Compared with the baseline model (RoBERTa-CRF), which only includes RoBERTa and CRF, the RoBERTa-BiLSTM-CRF model performs better in Precision (0.8834), Recall (0.8831), and F1 score (0.8832). Although the overall performance of the RoBERTa-CRF model remains strong, it lacks additional order-sensitive sequential modeling in this setting. In particular, when complex contextual linkages exist between entities, their recall is lower, at 0.8738.

Comprehensive results indicate that introducing the BiLSTM layer enhances context modeling capability, particularly for texts containing many professional terms and long sentence dependencies related to BFPDCs. Its dual-channel modeling mechanism

of forward and backward semantics can make entity predictions both more accurate and more stable.

5.2.3. Comparison with Traditional NER Baselines

To further evaluate the effectiveness of the proposed model, traditional NER baselines were introduced for comparison, including a rule-based method, HMM, and independent CRF. Unlike pretrained language models such as BERT and RoBERTa, HMM and CRF are traditional statistical sequence labeling models trained directly on the annotated corpus. The rule-based method was constructed using domain dictionaries, trigger expressions, and regular expressions. Specifically, domain dictionaries were developed for common building types, space types, fire protection facilities, building materials, and regulatory expressions, while regular expressions were used to identify numerical requirements and clause references. All baseline models used the same dataset split and evaluation metrics as the proposed model.

As shown in Table 7, the proposed RoBERTa-BiLSTM-CRF model achieves the best overall performance. The rule-based method achieves relatively high recall because it can identify entities with stable terminology and explicit patterns, but its precision is limited when entity boundaries are long or context-dependent. The HMM baseline obtains comparable but slightly lower overall performance, indicating that character-level emission and transition probabilities are insufficient for modeling complex regulatory semantics. Independent CRF performs better than rule-based and HMM baselines by using handcrafted contextual features and structured sequence decoding, but it still lags behind the proposed model. In contrast, RoBERTa-BiLSTM-CRF benefits from pretrained contextual representations, bidirectional sequence modeling, and CRF-based label transition constraints, enabling more accurate recognition of domain-specific entities in building fire protection design codes.

Table 7. Comparison with traditional NER baselines.

Model	Precision	Recall	F1 Score
Rule-based method	0.4637	0.5487	0.5026
HMM	0.5071	0.4865	0.4965
Independent CRF	0.7608	0.7150	0.7372
RoBERTa-BiLSTM-CRF	0.8834	0.8831	0.8832

5.3. Recognition Challenges for Fine-Grained Entities

Although the model demonstrates high performance in the overall NER task, which indicates that it has excellent generalization ability in dealing with most of the entity categories, it still faces certain challenges dealing with fine-grained entity categories, particularly in the context of fuzzy semantic boundaries and functional overlap where the extraction accuracy is degraded.

Table 8 indicates that for entity types “FPR” and “FR,” the F1 scores are 0.8037 and 0.8106, respectively, which are much lower than those for clearer categories such as “SC” (F1 = 0.9811) or “NR” (F1 = 0.9554). This result suggests that when dealing with entity categories that have unclear definitions, the proposed model struggles to understand meaning and pick up important context.

According to the category-level results in Table 8, BM, FPR, and FR are the most difficult entity categories to recognize, with F1 scores of 0.8025, 0.8037, and 0.8106, respectively. These categories are more challenging because their semantic boundaries are relatively ambiguous: BM entities are often combined with fire-resistance properties and numerical constraints, while FPR and FR frequently share similar normative expressions such as

“shall be provided” and “shall meet”. In addition, long clauses containing multiple parallel spaces, facilities, or requirements may lead to omission or boundary errors. These findings suggest that future work should further refine annotation rules for overlapping categories and incorporate domain knowledge to improve fine-grained entity recognition.

Table 8. Model performance of different entity types.

Entity Types	Precision	Recall	F1 Score
NR	0.9479	0.9630	0.9554
BT	0.8943	0.9227	0.9083
FPF	0.8865	0.8804	0.8834
ST	0.8623	0.8843	0.8731
BM	0.8025	0.8025	0.8025
FPR	0.8208	0.7873	0.8037
SC	0.9750	0.9873	0.9811
FR	0.8024	0.8189	0.8106
LR	0.9231	0.9231	0.9231

Table 9 presents representative examples of entity recognition challenges in building fire protection design codes. These examples show that the proposed model performs well for entities with clear boundaries and stable terminology, but recognition becomes more difficult when multiple technical meanings are combined in one clause or when different entity categories share similar regulatory expressions.

Overall, the examples indicate that most recognition challenges are related to entity boundary ambiguity, category overlap, and long clauses containing multiple parallel entities. These cases further support the need to refine annotation guidelines and incorporate domain knowledge and boundary-aware constraints in future work.

Table 10 presents the performance of the model under different CFPDB documents. The model exhibits relatively consistent performance in the first three codes, with F1 scores higher than 0.9255. However, the performance of the model drops considerably in the code for design of automatic fire alarm system and the technical standard for smoke management systems in buildings. This performance fluctuation may be related to the differences in text style, term density, and sentence complexity, suggesting that cross-code generalization remains challenging under the current setting.

Although the last two codes show lower F1 scores, the overall micro-averaged performance remains high because entity distributions differ across codes.

Table 11 further explains the cross-code performance differences using derived statistical indicators. The automatic fire alarm system code and the smoke management system standard have high FPF proportions, reaching 47.2% and 40.3%, respectively. Their dense facility-oriented terminology, control relationships, and functional descriptions make entity boundary detection and category classification more difficult, which may partly explain their lower F1 scores.

Table 9. Representative examples of entity recognition challenges.

Example Clause Fragment	Gold-Standard Entity	Error Type	Possible Reason
高层建筑内的中庭回廊应设置自动喷水灭火系统和火灾自动报警系统;	高层建筑/BT; 中庭回廊/ST; 自动喷水灭火系统/FPF; 火灾自动报警系统/FPF	Correct recognition	The entities are common fire protection terms with clear boundaries, so the model can recognize them correctly.

Table 9. Cont.

Example Clause Fragment	Gold-Standard Entity	Error Type	Possible Reason
二级耐火等级多层厂房和多层仓库内采用预应力钢筋混凝土的楼板, 其耐火极限不应低于0.75 h.	二级/NR; 耐火等级/FPR; 多层厂房/BT; 多层仓库/BT; 预应力钢筋混凝土的楼板/BM; 耐火极限/FPR; 0.75 h/NR	Boundary error	Material entities, fire-resistance requirements, and numerical constraints appear in the same clause, which may lead to boundary confusion.
湿式系统, 干式系统应在开放一只洒水喷头后自动启动, 预作用系统, 雨淋系统和水幕系统应根据其类型由火灾探测器, 闭式洒水喷头作为探测元件, 报警后自动启动;	湿式系统/FPF; 干式系统/FPF; 洒水喷头/FPF; 自动启动/FR; 预作用系统/FPF; 雨淋系统/FPF; 水幕系统/FPF; 火灾探测器/FPF; 闭式洒水喷头/FPF; 探测元件/FPF; 自动启动/FR	Category confusion	Functional expressions such as “自动启动” may be confused with general fire protection requirements because they are embedded in normative clauses.
系统应由火灾探测器, 手动火灾报警按钮, 火灾声光警报器, 消防应急广播, 消防专用电话, 消防控制室图形显示装置, 火灾报警控制器, 消防联动控制器等组成.	火灾探测器/FPF; 手动火灾报警按钮/FPF; 火灾声光警报器/FPF; 消防应急广播/FPF; 消防专用电话/FPF; 消防控制室图形显示装置/FPF; 火灾报警控制器/FPF; 消防联动控制器/FPF	Omission error	Multiple FPF entities occur in a long parallel structure, increasing the risk of missing one or more facility entities.

Table 10. Performance of models with different codes.

No.	Code for Fire Protection	Precision	Recall	F1 Score
1	Code for fire protection design of buildings	0.9263	0.9309	0.9286
2	General code for fire protection of buildings and constructions	0.9239	0.9271	0.9255
3	Code for design of sprinkler systems	0.9333	0.9189	0.9261
4	Code for design of automatic fire alarm system	0.7500	0.7540	0.7520
5	Technical standard for smoke management systems in buildings	0.7401	0.7649	0.7523

Table 11. Derived statistical indicators of different fire protection codes.

No.	Code for Fire Protection	Mean Length	Entities per Provision	Entities per 100 Characters	Dominant Entity Type	FPF Proportion
1	Code for fire protection design of buildings	53.62	6.16	11.48	FPF	23.90%
2	General code for fire protection of buildings and constructions	40.64	4.91	12.09	ST	17.30%
3	Code for design of sprinkler systems	49.18	5	10.16	FPF	43.50%
4	Code for design of automatic fire alarm system	46.79	4.02	8.58	FPF	47.20%
5	Technical standard for smoke management systems in buildings	46.53	4.12	8.86	FPF	40.30%

5.4. Comparison with Existing Literature Methods

To further contextualize the performance of the proposed model, several representative NER studies discussed in the related work are summarized in Table 12. These studies cover technical-domain NER, Chinese construction document NER, construction risk text mining, general Chinese NER, and Chinese medical NER. Since the compared studies are

based on different corpora, entity schemas, annotation strategies, and application scenarios, the comparison is intended to provide a performance reference rather than a strictly controlled benchmark comparison.

Table 12. Comparison of named entity recognition tasks in different application fields.

No.	Reference	Method	Application Field/Corpus	Precision (%)	Recall (%)	F1 Score (%)
1	Chen et al. [9]	RoBERTa-Attention-FL	Power dispatch domain texts	92.53	88.12	90.35
2	Zhang et al. [10]	CRF with domain features	Chinese construction documents	—	—	87.9
3	Shishehgharkhaneh et al. [11]	RoBERTa-based Transformer	Construction supply chain risk news	93.41	80.23	85.8
4	Xu et al. [14]	GoalBERT	General Chinese NER/CLUENER2020	80.66	80.31	80.42
5	Chen et al. [16]	BERT-BiLSTM-CRF	Chinese medical NER	77	79	78
6	This study	Proposed domain-specific NER model	Chinese building fire protection design codes	88.34	88.31	88.32

Note: “—” indicates that the corresponding value was not explicitly reported in the accessible source.

As shown in Table 12, the proposed model achieves an F1 score of 88.32% on Chinese building fire protection design codes. This result is lower than the power dispatch NER model reported by Chen et al. [9], which may be attributed to the relatively stable terminology and clearer entity boundaries in power dispatch texts. However, the proposed model slightly outperforms the CRF-based Chinese construction document NER method [10] and achieves better F1 performance than the construction supply chain risk NER model [11], the general Chinese NER model [14], and the Chinese medical NER model [16].

These results indicate that the proposed model achieves competitive performance under a specialized regulatory-text scenario. Compared with general-domain or event-oriented texts, Chinese building fire protection design codes contain dense technical terminology, hierarchical clause structures, numerical requirements, and compliance-oriented semantic constraints. These characteristics make entity boundary recognition and category classification more challenging. Therefore, the proposed RoBERTa-BiLSTM-CRF architecture provides an effective solution for extracting structured entities from fire protection design codes and supports downstream automated compliance checking.

Beyond entity recognition performance, the extracted entities can also provide structured knowledge units for downstream decision support in automated compliance checking. For example, building type, space type, fire protection facility, numerical requirement, and standard provision entities can support clause retrieval, rule formalization, BIM-based checking, and evidence tracing, which is consistent with recent decision-support studies showing that structured engineering knowledge can be transformed into practical engineering decision-making tools [68].

In a practical automated compliance checking workflow, the proposed NER model serves as an upstream information extraction module. The extracted entities can be normalized and mapped to BIM information, where space type entities correspond to BIM Room or Space objects, fire protection facility entities correspond to fire safety components, and numerical requirement entities can be converted into computable rule constraints. Downstream rule-checking modules can then compare BIM-derived information with structured regulatory requirements and generate compliance results with traceable code evidence.

Although this study focuses on Chinese building fire protection design codes, the proposed framework has potential transferability to other building code systems. When applied to international codes such as NFPA, IBC, or Eurocodes, the entity schema, technical terminology, measurement units, regulatory expressions, and clause structures should be adjusted according to the target code system. In addition, annotated samples from the target corpus are required for model fine-tuning, and multilingual applications may further require cross-lingual terminology alignment and semantic adaptation.

6. Conclusions

Herein, a multilayer neural network model integrating RoBERTa, BiLSTM, and CRF was proposed to address the challenges of the NER task in codes for designing fire protection in buildings. The model could effectively model the terminology and complex semantic structures in the specification text and outperformed the traditional methods across all metrics, demonstrating excellent recognition performance.

Specifically, RoBERTa enhances the model's ability to capture terms in the domain of building fire protection through deep contextual semantic learning. BiLSTM captured the forward and backward dependencies of entities in the text and enhanced order-sensitive sequence modeling. The CRF layer further optimized the structural consistency of the label sequences, which can improve the accuracy in recognizing entity boundaries. The model reached 88.34% precision, 88.31% recall, and 88.32% F1 score on the fire specification dataset, demonstrating its effectiveness in complex scenarios. Beyond model performance, the extracted entities can serve as basic knowledge units for transforming unstructured regulatory clauses into structured and traceable code records, thereby supporting downstream applications such as clause retrieval, evidence-based compliance checking, and BIM- or document-based compliance checking.

Although this study achieved encouraging results, several limitations remain. Future work will focus on three aspects: improving the recognition of fine-grained entities with ambiguous boundaries, such as FPR, FR, and BM; enhancing cross-code generalization for specialized documents such as automatic fire alarm and smoke management standards through domain adaptation or incremental fine-tuning; and supporting practical deployment by linking extracted entities with BIM spaces, fire safety components, computable constraints, and traceable regulatory evidence. In addition, error propagation from NER to rule checking, model updating after code revisions, and integration with broader BIM-, AI-, IoT-, and DT-based fire safety management systems should be further investigated. Such integration may support downstream applications such as dynamic fire safety assessment and intelligent evacuation decision-making [69].

Author Contributions: Conceptualization, L.G. and L.Z.; methodology, L.G., H.Z. and L.Z.; software, L.G. and D.Q.; validation, D.Q.; formal analysis, L.G. and L.Z.; data curation, L.G., H.Z. and L.Z.; writing—original draft, L.G. and L.Z.; writing—review and editing, L.G. and H.Z.; visualization, L.G.; funding acquisition, L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Special Project on AI-Empowered Educational Reform in Universities of Hebei Province (Grant No. 2025RGZN065), and Handan City Science and Technology Research and Development Plan Project (Grant No. 1821203003ZC).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Sun, X.; Cai, N.; Zhang, W. Discussing the development of domestic and foreign fire protection technical regulation and fire protection technical standard systems. *J. Saf. Sci. Resil.* **2023**, *4*, 26–29. [\[CrossRef\]](#)
2. Besiroglu, S.; Serteser, N. Investigation of building fire safety measures in building information modelling (BIM) environment. *Int. J. Archit. Comput.* **2026**, *24*, 358–377.
3. Zheng, Z.; Lu, X.Z.; Chen, K.Y.; Zhou, Y.C.; Lin, J.R. Pretrained domain-specific language model for natural language processing tasks in the AEC domain. *Comput. Ind.* **2022**, *142*, 103733. [\[CrossRef\]](#)
4. Zheng, Z.; Zhou, Y.; Lu, X.; Lin, J. Knowledge-informed semantic alignment and rule interpretation for automated compliance checking. *Autom. Constr.* **2022**, *142*, 104524. [\[CrossRef\]](#)
5. Ma, Y.; Wu, H.; Zhao, L.; Wei, W.; Meng, T. Research progress on key technologies of fire design automatic review based on BIM. *J. Civ. Archit. Eng. Inf. Technol.* **2022**, *14*, 131–142. (In Chinese)
6. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Long and Short Papers; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 4171–4186.
7. Kim, J.H.; Park, S.W.; Kim, J.Y.; Park, J.; Jung, S.H.; Sim, C.B. RoBERTa-CoA: RoBERTa-based effective finetuning method using co-attention. *IEEE Access* **2023**, *11*, 120292–120303. [\[CrossRef\]](#)
8. Chen, Y.; Jiang, H.X. Optimizing automated compliance checking with ontology-enhanced natural language processing: Case in the fire safety domain. *J. Environ. Manag.* **2024**, *371*, 123320. [\[CrossRef\]](#)
9. Chen, Y.; Lin, D.; Meng, Q.; Liang, Z.; Tan, Z. Named Entity Identification in the Power Dispatch Domain Based on RoBERTa-Attention-FL Model. *Energies* **2023**, *16*, 4654. [\[CrossRef\]](#)
10. Zhang, Q.; Xue, C.; Su, X.; Zhou, P.; Wang, X.; Zhang, J. Named Entity Recognition for Chinese Construction Documents Based on Conditional Random Field. *Front. Eng. Manag.* **2023**, *10*, 237–249.
11. Shishehgarkhaneh, M.B.; Moehler, R.C.; Fang, Y.; Hijazi, A.A.; Aboutorab, H. Transformer-Based Named Entity Recognition in Construction Supply Chain Risk Management in Australia. *IEEE Access* **2024**, *12*, 41829–41851. [\[CrossRef\]](#)
12. Szymczak-Graczyk, A.; Korentz, J.; Garbowski, T. Digital Twin Model for Predicting Hygrothermal Performance of Building Materials from Moisture Permeability Tests. *Materials* **2025**, *18*, 4360. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Ding, Y.; Ma, J.; Luo, X. Applications of natural language processing in construction. *Autom. Constr.* **2022**, *136*, 104169. [\[CrossRef\]](#)
14. Xu, Y.; Tan, X.; Wang, M.; Zhang, W. GoalBERT: A Lightweight Named-Entity Recognition Model Based on Multiple Fusion. *Appl. Sci.* **2024**, *14*, 11003. [\[CrossRef\]](#)
15. Chandra, C.; Ojima, Y.; Bendarkar, M.V.; Mavris, D.N. Aviation-BERT-NER: Named entity recognition for aviation safety reports. *Aerospace* **2024**, *11*, 890. [\[CrossRef\]](#)
16. Chen, W.; Qiu, P.; Cauteruccio, F. MedNER: A Service-Oriented Framework for Chinese Medical Named-Entity Recognition with Real-World Application. *Big Data Cogn. Comput.* **2024**, *8*, 86. [\[CrossRef\]](#)
17. Agrawal, A.; Tripathi, S.; Vardhan, M.; Sihag, V.; Choudhary, G.; Dragoni, N. BERT-based transfer-learning approach for nested named-entity recognition using joint labeling. *Appl. Sci.* **2022**, *12*, 976. [\[CrossRef\]](#)
18. Zhai, H.M.; Lv, X.J.; Hou, Z.W.; Tong, X.; Bu, F.L. MLNet: A multi-level multimodal named entity recognition architecture. *Front. Neurobotics* **2023**, *17*, 1181143. [\[CrossRef\]](#)
19. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*; Long Papers; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; Volume 1, pp. 1064–1074.
20. Arslan, S. Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text. *Neural Comput. Appl.* **2024**, *36*, 8371–8382. [\[CrossRef\]](#)
21. Zhang, Y.; Bi, X.Q.; Zuo, Y.F.; Li, P.; Yuan, G.M.; Li, X.J. Magnesium silica gel crystallized in the cell lumen of Chinese fir to construct porous structure for filtering toxic fumes. *Eur. J. Wood Wood Prod.* **2023**, *81*, 987–997. [\[CrossRef\]](#)
22. Covi, P.; Tondini, N.; Lamperti Tornaghi, M.; Molina, F.J.; Pegon, P.; Tsionis, G. Seismic experimental analysis of a full-scale steel building with passive fire protections. *Eng. Struct.* **2024**, *300*, 117203. [\[CrossRef\]](#)
23. Lafhaj, Z.; Rebai, S.; Albalkhy, W.; Hamdi, O.; Mossman, A.; Alves Da Costa, A. Complexity in construction projects: A literature review. *Buildings* **2024**, *14*, 680. [\[CrossRef\]](#)
24. Li, S.H.; Wang, J.L.; Xu, Z. Automated compliance checking for BIM models based on Chinese-NLP and knowledge graph: An integrative conceptual framework. *Eng. Constr. Archit. Manag.* **2024**, *32*, 3832–3856. [\[CrossRef\]](#)
25. Chen, W.T.; Bria, T.A. A review of ontology-based safety management in construction. *Sustainability* **2023**, *15*, 413.
26. Wang, X.; El-Gohary, N. Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements. *Autom. Constr.* **2023**, *147*, 104696. [\[CrossRef\]](#)

27. Singh, J.; Anumba, C.J. Building commissioning process and documentation: A literature review and directions for future research. *Int. J. Constr. Manag.* **2023**, *24*, 75–85. [[CrossRef](#)]
28. Kumar, L.A.; Renuka, D.K. *Deep Learning Approach for Natural Language Processing, Speech, and Computer Vision*; Techniques and Use Cases; CRC Press: Boca Raton, FL, USA, 2023.
29. Wang, Y.Z.; Zhang, C.Z. Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *J. Informetr.* **2020**, *14*, 101091. [[CrossRef](#)] [[PubMed](#)]
30. Zhou, J.; Ma, Z. Named entity recognition for construction documents based on fine-tuning of large language models with low-quality datasets. *Autom. Constr.* **2025**, *174*, 106151. [[CrossRef](#)]
31. Wang, Y.K.; Liu, Q.; Dai, T.T.; Lang, J.J.; Lu, L.; Chen, Y.N. Contrastive learning for nested Chinese named entity recognition via template words. *CAAI Trans. Intell. Technol.* **2025**, *10*, 450–459. [[CrossRef](#)]
32. An, J.; Kim, J.; Sunwoo, L.; Baek, H.; Yoo, S.; Lee, S. De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 82. [[PubMed](#)]
33. Tang, S.X.; Liu, H.X.; Almatared, M.; Abudayyeh, O.; Lei, Z.; Fong, A. Towards automated construction quantity take-off: An integrated approach to information extraction from work descriptions. *Buildings* **2022**, *12*, 354. [[CrossRef](#)]
34. Song, Y.; Jia, Y.H.; Zhang, J. Chinese named entity recognition in the furniture domain based on ERNIE and adversarial learning. *Int. J. Web Inf. Syst.* **2024**, *21*, 353–371. [[CrossRef](#)]
35. Abdi, M.T.; Mohd, M. A comparative study of word representation methods with conditional random fields and maximum entropy Markov for BIO-named entity recognition. *Malays. J. Comput. Sci.* **2018**, *31*, 15–30. [[CrossRef](#)]
36. Liu, C.B.; Fan, C.J.; Wang, Z.J.; Sun, Y.Q. An instance transfer-based approach using enhanced recurrent neural network for domain named entity recognition. *IEEE Access* **2020**, *8*, 45263–45270. [[CrossRef](#)]
37. Abu Marjan, M.; Amagasa, T. Domain-adaptive entity recognition: Unveiling the potential of CSER in cybersecurity and beyond. *Int. J. Mach. Learn. Cybern.* **2024**, *16*, 2849–2867. [[CrossRef](#)]
38. Wadawadagi, R.; Tiwari, S.; Pagi, V. Polarity-aware deep attention network for aspect-based sentiment analysis. *Prog. Artif. Intell.* **2025**, *14*, 33–48.
39. Gao, F.F.; Zhang, L.; Wang, W.F.; Zhang, B.; Liu, W.; Zhang, J.Y.; Xie, L. Named entity recognition for equipment fault diagnosis based on RoBERTa-wwm-ext and deep learning integration. *Electronics* **2024**, *13*, 3935. [[CrossRef](#)]
40. Guo, Q.; Guo, Y. Lexicon enhanced Chinese named entity recognition with pointer network. *Neural Comput. Appl.* **2022**, *34*, 14535–14555. [[CrossRef](#)]
41. Li, L.T.; Xu, W.Z.; Yu, H. Character-level neural network model based on Nadam optimization and its application in clinical concept extraction. *Neurocomputing* **2020**, *414*, 182–190. [[CrossRef](#)]
42. Alsuwat, E.; Alsuwat, H. An improved multi-modal framework for fake news detection using NLP and Bi-LSTM. *J. Supercomput.* **2025**, *81*, 177. [[CrossRef](#)]
43. Zhang, N.; Zhou, C.; Wan, F.; Liu, F.; Wang, Y.; Xu, D. Tunnel construction safety domain named entity recognition based on BERT-BiLSTM-CRF. *China Saf. Sci. J.* **2024**, *34*, 56–63. (In Chinese) [[CrossRef](#)]
44. Yuan, G.Q.; Zhao, X.J.; Li, L.; Zhang, S.; Wei, S.M. Few-shot learning sensitive recognition method based on prototypical network. *Mathematics* **2024**, *12*, 2791. [[CrossRef](#)]
45. Xu, K.; Yang, Z.G.; Kang, P.P.; Wang, Q.; Liu, W.Y. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Comput. Biol. Med.* **2019**, *108*, 122–132. [[CrossRef](#)] [[PubMed](#)]
46. Hu, L.M.; Liu, Z.Y.; Zhao, Z.W.; Hou, L.; Nie, L.Q.; Li, J.Z. A survey of knowledge enhanced pre-trained language models. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 1413–1430. [[CrossRef](#)]
47. Tan, K.L.; Lee, C.P.; Lim, K.M. RoBERTa-GRU: A hybrid deep learning model for enhanced sentiment analysis. *Appl. Sci.* **2023**, *13*, 3915. [[CrossRef](#)]
48. Hoang, D.H.V.; Quoc, H.V.; Hung, B.T. ConBGAT: A novel model combining convolutional neural networks, transformer and graph attention network for information extraction from scanned image. *PeerJ Comput. Sci.* **2024**, *10*, e2536. [[CrossRef](#)]
49. Wang, Y.H.; Gong, C.; Ji, X.D.; Yuan, Q. Text classification for evaluating digital technology adoption maturity based on BERT: Evidence from industrial AI in China. *Technol. Forecast. Soc. Chang.* **2025**, *211*, 123903. [[CrossRef](#)]
50. Tian, D.; Li, M.C.; Ren, Q.B.; Zhang, X.J.; Han, S.; Shen, Y. Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining. *Autom. Constr.* **2023**, *145*, 104670. [[CrossRef](#)]
51. Hwang, H.; Jung, Y.; Lee, C.; Go, W. A nested named entity recognition model robust in few-shot learning environments using label description information. *Appl. Sci.* **2025**, *15*, 8255. [[CrossRef](#)]
52. Gardazi, N.M.; Daud, A.; Malik, M.K.; Bukhari, A.; Alsahfi, T.; Alshemaimri, B. BERT applications in natural language processing: A review. *Artif. Intell. Rev.* **2025**, *58*, 166. [[CrossRef](#)]
53. Tang, X.Y.; Huang, Y.; Xia, M.; Long, C.F. A multi-task BERT-BiLSTM-AM-CRF strategy for Chinese named entity recognition. *Neural Process. Lett.* **2023**, *55*, 1209–1229.

54. Gao, K.; Zhou, J.H.; Chi, Y.X.; Wen, Y.M. TourismNER: A tourism named entity recognition method based on entity boundary joint prediction. *Intell. Syst. Appl.* **2025**, *25*, 200475. [CrossRef]
55. Xu, Y.R.; Mao, C.X.; Wang, Z.Y.; Jin, G.N.; Zhong, L.J.; Qian, T. Semantic-enhanced graph neural network for named entity recognition in ancient Chinese books. *Sci. Rep.* **2024**, *14*, 17488. [CrossRef] [PubMed]
56. Fei, Y.F.; Xu, X.L. GFMRC: A machine reading comprehension model for named entity recognition. *Pattern Recognit. Lett.* **2023**, *172*, 97–105. [CrossRef]
57. Zhong, Y.S.; Goodfellow, S.D. Domain-specific language models pre-trained on construction management systems corpora. *Autom. Constr.* **2024**, *160*, 105316. [CrossRef]
58. Dang, X.C.; Wang, L.; Dong, X.H.; Li, F.F.; Deng, H. Improving low-resource Chinese named entity recognition using bidirectional encoder representation from transformers and lexicon adapter. *Appl. Sci.* **2023**, *13*, 10759. [CrossRef]
59. Li, Y.G.; Luo, L.; Zeng, X.R.; Han, Z. Fine-tuned BERT-BiLSTM-CRF approach for named entity recognition in geological disaster texts. *Earth Sci. Inform.* **2025**, *18*, 368. [CrossRef]
60. Qiu, Q.J.; Tian, M.; Huang, Z.; Xie, Z.; Ma, K.; Tao, L.F.; Xu, D.X. Chinese engineering geological named entity recognition by fusing multi-features and data enhancement using deep learning. *Expert Syst. Appl.* **2024**, *238*, 121925. [CrossRef]
61. Ministry of Housing and Urban-Rural Development of the People's Republic of China. On the Release of Building Design Fire Prevention Code and Other 7 Engineering Construction Standards in English Version of the Announcement. 2020. Available online: https://www.mohurd.gov.cn/gongkai/zc/wjk/art/2020/art_17339_246318.html (accessed on 15 January 2025).
62. Ministry of Housing and Urban-Rural Development of the People's Republic of China. On the Release of National Standards General Code for Fire Protection of Buildings and Constructions. 2023. Available online: https://www.mohurd.gov.cn/gongkai/zc/wjk/art/2023/art_17339_770016.html (accessed on 15 January 2025).
63. Ministry of Housing and Urban-Rural Development of the People's Republic of China. Announcement on the Publication of the National Standard Design Code for Design of Sprinkler Systems. 2017. Available online: https://www.mohurd.gov.cn/gongkai/zc/wjk/art/2017/art_17339_233407.html (accessed on 15 January 2025).
64. Ministry of Housing and Urban-Rural Development of the People's Republic of China. Announcement on the Publication of the National Standard Design Code for Design of Automatic Fire Alarm System. 2015. Available online: https://www.mohurd.gov.cn/gongkai/zc/wjk/art/2015/art_17339_224820.html (accessed on 15 January 2025).
65. Ministry of Housing and Urban-Rural Development of the People's Republic of China. Announcement on the Publication of the National Standard of Technical Standard for Building Smoke Prevention and Exhaust System. 2019. Available online: https://www.mohurd.gov.cn/gongkai/zc/wjk/art/2019/art_17339_240707.html (accessed on 15 January 2025).
66. Wu, C.; Xiao, L.; Guo, Y.; Wang, J.; Ren, Z.; Wang, M.; Yang, Z. Natural language processing for smart construction: Current status and future directions. *Autom. Constr.* **2022**, *134*, 104059. [CrossRef]
67. Zhang, J.; El-Gohary, N.M. Extending building information models semiautomatically using semantic natural language processing techniques. *J. Comput. Civ. Eng.* **2016**, *30*, C4016004. [CrossRef]
68. Walczak, Z.; Ksit, B.; Szymczak-Graczyk, A. Decision-support tools for diagnosing and selecting the optimal method of repairing buildings. *Bull. Pol. Acad. Sci. Tech. Sci.* **2024**, *72*, e151382. [CrossRef]
69. Konopski, P.; Bonenberg, W.; Szymczak-Graczyk, A.; Ksit, B.; Pilch, R. Responsive Architecture in Practice: BIM/DT/AI/IoT for Dynamic Fire Evacuation—A Comparative Case Study Analysis. *Sustainability* **2026**, *18*, 5920. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.