Article

Named Entity Recognition for Chinese Texts on Marine Coral Reef Ecosystems Based on the BERT-BiGRU-Att-CRF Model

Danfeng Zhao [1,*] **, Xiaolian Chen** [1] **and Yan Chen** [2]

1   College of Information Technology, Shanghai Ocean University, Shanghai 201306, China;
    cxl19961126@outlook.com
2   College of Information Technology, Shanghai Jian Qiao University, Shanghai 201306, China;
    cy13145643427@163.com
*   Correspondence: dfzhao@shou.edu.cn

Abstract: In addressing the challenges of non-standardization and limited annotation resources in Chinese marine domain texts, particularly with complex entities like long and nested entities in coral reef ecosystem-related texts, existing Named Entity Recognition (NER) methods often fail to capture deep semantic features, leading to inefficiencies and inaccuracies. This study introduces a deep learning model that integrates Bidirectional Encoder Representations from Transformers (BERT), Bidirectional Gated Recurrent Units (BiGRU), and Conditional Random Fields (CRF), enhanced by an attention mechanism, to improve the recognition of complex entity structures. The model utilizes BERT to capture context-relevant character vectors, employs BiGRU to extract global semantic features, incorporates an attention mechanism to focus on key information, and uses CRF to produce optimized label sequences. We constructed a specialized coral reef ecosystem corpus to evaluate the model's performance through a series of experiments. The results demonstrated that our model achieved an F1 score of 86.54%, significantly outperforming existing methods. The contributions of this research are threefold: (1) We designed an efficient named entity recognition framework for marine domain texts, improving the recognition of long and nested entities. (2) By introducing the attention mechanism, we enhanced the model's ability to recognize complex entity structures in coral reef ecosystem texts. (3) This work offers new tools and perspectives for marine domain knowledge graph construction and study, laying a foundation for future research. These advancements propel the development of marine domain text analysis technology and provide valuable references for related research fields.

Keywords: coral reef ecosystem; named entity recognition; BERT; BiGRU; attention mechanism

1. Introduction

Coral reef ecosystems are vital pillars of marine biodiversity and ecological balance, possessing immense ecological, economic, and social value. These ecosystems provide abundant marine resources and habitats, playing a crucial role in fisheries, tourism, and coastal protection. However, human activities, such as global climate change, marine pollution, and overfishing, are severely threatening the survival of coral reefs [1,2]. Therefore, in-depth research and the protection of coral reef ecosystems have become key tasks in the field of marine science. In the realm of Natural Language Processing (NLP), Named Entity Recognition (NER) is a fundamental and crucial task. The goal of NER is to identify specific named entities in a text, such as names of people, places, organizations, dates, and quantities. In the marine domain, NER tasks are particularly challenging due to the lack of explicit word boundaries in Chinese texts, difficulties in tokenization, and the symbolic characteristics of Chinese characters [3], which add semantic weight. Although translating Chinese text into English can facilitate analysis, this approach may introduce errors and lose context-specific meanings [4], especially for domain-specific terms. Therefore, processing Chinese text directly ensures more precise and contextually accurate NER results.

In the field of Natural Language Processing (NLP), Named Entity Recognition (NER) technology is an important tool for extracting key entity information from large volumes of text data. The evolution of NER technology has undergone multiple stages, from rule-based matching to deep learning. Initially, researchers relied on manually crafted rules for entity recognition, which proved inadequate when dealing with complex entity structures and diversity. In contrast, methods based on statistical models such as Hidden Markov Models (HMM) [5], Conditional Random Fields (CRF) [6], and Support Vector Machines (SVM) [7], despite relying on manual feature extraction and classifier training, have achieved some success in entity recognition tasks. However, these methods still face issues with inefficient feature extraction and over-reliance on domain expert knowledge. In the marine domain, Cao et al. [8] proposed a method combining user dictionaries with CRF, enhancing recognition rates by deeply analyzing the features of marine-specific entities. Although this method improved recognition to some extent, it still faced issues related to the completeness of user dictionaries and limitations in feature selection.

In the field of marine science, named entity recognition (NER) techniques rarely rely solely on statistical models, as their accuracy in entity extraction is typically low. With the advancement of deep learning technologies, an increasing number of researchers are integrating statistical models with deep learning methods. This approach not only enhances accuracy but also reduces dependence on manually defined features and external resources. In recent years, NER methods based on deep learning have made significant progress. Particularly, models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) [9] have successfully addressed the issues of gradient vanishing and explosion in long sequence training, becoming the main models for this task. In 2016, Lample et al. [10] introduced the BiLSTM-CRF model, pioneering the integration of statistical models with deep learning techniques for named entity recognition (NER). This approach, in comparison to using a standalone Conditional Random Field (CRF), incorporates contextual meanings related to individual words from both directions, thereby enhancing annotation accuracy. To improve the recognition of alien marine species entities, He et al. [11] proposed the CNN-BiGRU-CRF model. This model incorporates an attention mechanism that mimics the human tendency to focus differently on various aspects. By introducing the attention mechanism, the model selectively emphasizes critical information while de-emphasizing other concurrent inputs. In text processing, this translates to assigning higher weights to key texts and reducing the weights of less relevant texts. The model initially uses BiGRU to learn and represent the contextual information of alien species at the text level, and then employs the attention mechanism to capture the salient semantic features of marine species entities. This approach mitigates the issue of long-distance dependencies in textual data, thereby enhancing the accuracy of marine species entity recognition and providing a reference for the identification of marine organisms and concepts in the future. Recognizing the dependency of the attention mechanism on external factors, He et al. [12] further incorporated the multi-head self-attention mechanism into the recognition of entities in Chinese marine text data. They used knowledge graph embedding vectors and BiLSTM output vectors jointly as input vectors for the self-attention mechanism. This approach considers both the internal correlation of features and long-sequence dependencies, thereby improving the accuracy of alien marine species entity recognition and enhancing the overall entity recognition capability of the corpus. The successful application of these techniques to marine data underscores the promising future of multi-head attention mechanisms in the marine domain. Ma et al. [13] introduced a model that integrates attention mechanisms, an expanded convolutional neural network (IDCNN), and CRF, aimed at enhancing the ability to recognize marine natural product (MNP) entities from unstructured texts. In the biomedical field, a study by Perera et al. [14] demonstrated the application of NER and relationship detection in biomedical information extraction. They adopted a comprehensive approach that effectively extracted entity and relationship information from biomedical literature, indicating that deep learning models have broad applicability in handling specialized domain texts. Moreover, pre-trained language models such as BERT, GPT, and RoBERTa [15–18] have

significantly improved the capture of contextual information and semantic representation through large-scale text data pre-training, further enhancing NER performance.

Despite these advances, applying deep learning models to named entity recognition in the Chinese marine domain still poses numerous challenges. The uniqueness of Chinese text, the complexity of marine-specific terminologies, and the abstract nature of relationships between entities make traditional character-based models inadequate. Moreover, most existing pre-trained models are trained on general corpora and are not specifically optimized for the marine domain, resulting in less than satisfactory performance when dealing with specialized terminology. Therefore, the task of entity recognition in the marine domain requires not only identifying the entities themselves but also understanding the complex relationships between them, which demands higher comprehension capabilities from the models. In light of these challenges, this study introduces a novel model architecture designed to enhance the performance of Chinese marine named entity recognition by integrating deep learning models with greater domain adaptability, Chinese linguistic features, and marine-specific knowledge. Specifically addressing the recognition of long and nested entities in coral reef ecosystem texts, our model architecture is capable of effectively identifying these complex entities while adapting to the characteristics of Chinese text and the complexity of marine science. Experimental validation on publicly available marine science coral reef ecosystem text datasets demonstrates the effectiveness and superiority of our proposed model in the task of Chinese marine named entity recognition.

This research contributes not only a new methodological framework to address the challenges in Chinese marine named entity recognition but also provides technical support for the digital transformation and knowledge mining in marine science. These efforts are expected to further advance marine science development, promote the sustainable use of marine resources, and provide robust information support for marine ecological protection and management.

The rest of this article is organized as follows. Section 2 introduces the BERT-BiGRU-Att-CRF model proposed in this paper. Section 3 describes the experimental datasets, network model evaluation metrics, experimental results, and analysis. Section 4 presents the Discussion. Finally, the conclusion is presented in Section 5.

2. A BERT-BiGRU-Att-CRF Chinese Coral Reef Ecosystem Named Entity Recognition Model Combined with Attention Mechanism

2.1. Overall Framework of Model

The BERT-BiGRU-Att-CRF Chinese coral reef ecosystem named entity recognition model, incorporating attention mechanisms, comprises four main components: an embedding layer, a text encoding layer, an attention layer, and a decoding output layer. Initially, we utilize the BERT pre-trained model to obtain rich character-level contextual embedding vectors. BERT is capable of dynamically capturing the semantics of words in context, which is crucial for understanding complex terms within marine ecosystems. Subsequently, we introduce Bidirectional Gated Recurrent Units (BiGRUs) to further extract global features of the text. Through forward and backward recurrent units, the BiGRU effectively captures contextual information, enhancing the model's understanding of semantic shifts in different contexts. To improve the model's focus on critical information, we integrate an attention mechanism that adaptively allocates different attention weights. This enhances the model's ability to extract key semantic features, especially when dealing with structurally complex and interrelated entities. Finally, we employ Conditional Random Fields (CRFs) as the decoding layer to achieve globally optimal label sequences in the sequence labeling task. By considering constraints between adjacent labels and assigning appropriate transition probabilities for each label, CRF ensures the global optimality of the final label sequence. The model architecture is illustrated in Figure 1.
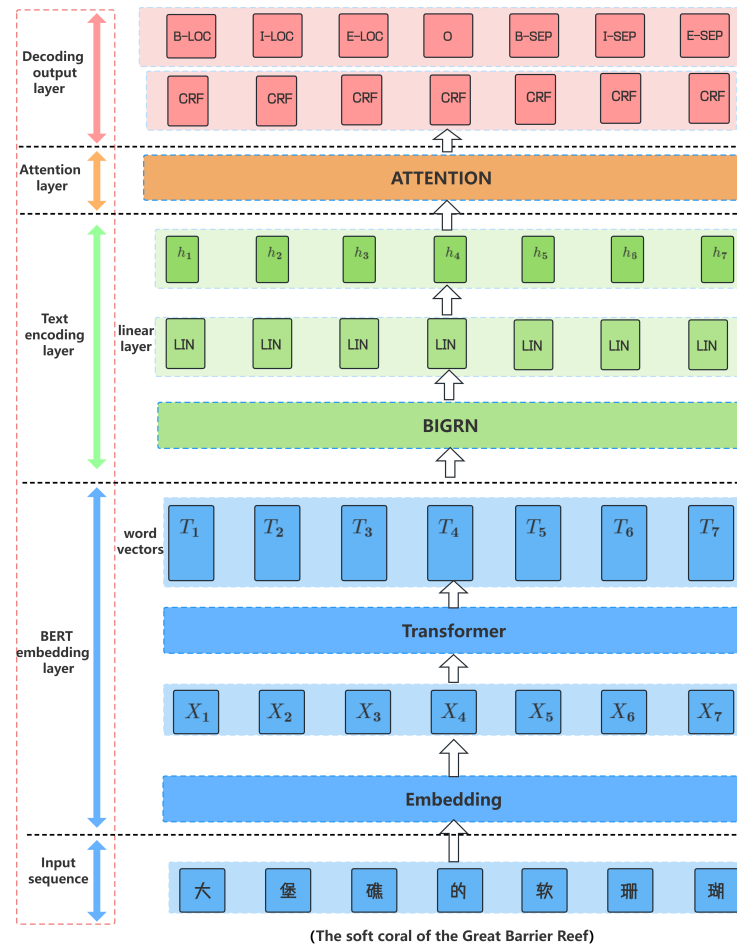
Figure 1. BERT-BiGRU-Att-CRF Chinese Coral Reef Ecosystem Named Entity Recognition Model architecture.

2.2. Embedding Layer: BERT Model

The BERT pre-trained model is a transfer learning method that involves training the model using a large amount of text data unrelated to the target task. Based on this pre-training, the model parameters are fine-tuned using a training set specific to the target domain, thereby accomplishing particular natural language processing tasks. In this paper, we use the ChineseBERT model, which is pre-trained on the Chinese Wikipedia corpus, and fine-tune the model parameters using general corpora from PKU and a self-constructed text dataset of the coral reef ecosystem to optimize its performance in the specific task. By leveraging bidirectional context information in pre-training, BERT better captures semantics and context, effectively addressing polysemy. BERT's use of the Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks during pre-training not only enhances the quality of word vectors but also strengthens the model's contextual understanding [19]. These innovations have led BERT to excel in various natural language processing tasks, underscoring its significance in the field.

BERT's input design includes unique innovations, incorporating word vectors, segment vectors, and position vectors to capture comprehensive word information, as illustrated in Figure 2. In the task of named entity recognition for marine coral reef ecosystems, special tokens such as [CLS] and [SEP] provide clear sentence boundaries and entity separation cues. For instance, the [CLS] token aids in recognizing the beginning of a sentence, while [SEP] is used to differentiate between entities or sentences. These elements work together to enhance the model's understanding of text structure, which is crucial for recognizing proper nouns.

Figure 2. Example diagram of BERT input.

At its core, BERT is built on multiple layers of bidirectional Transformer encoders, which, through self-attention mechanisms and feed-forward neural networks, capture input data features. The introduction of self-attention allows the model to focus on words at different positions within the input sequence, effectively capturing contextual information. This is vital for accurately identifying specialized terms in coral reef ecosystems. By processing input vectors, BERT generates rich semantic representations for texts in the marine ecology domain, laying the groundwork for precise terminology recognition. The BERT network structure is depicted in Figure 3. During training, BERT uses two unsupervised tasks, MLM and NSP [20], to enhance its understanding of language models. MLM enables BERT to learn and predict masked words, deepening its contextual understanding, particularly crucial in marine ecological texts rich in terminology. NSP, on the other hand, trains the model to recognize logical relationships between sentences. The combination of these training tasks provides BERT with a profound understanding of language, making it particularly suitable for complex tasks like marine ecological named entity recognition.

Figure 3. BERT network model.

The flexibility and exceptional contextual understanding capabilities of the BERT model make it an ideal choice for specialized domain tasks, such as named entity recognition in marine coral reef ecosystems.

2.3. Encoder Layer: BiGRU Layer

The Gated Recurrent Unit (GRU) is an improved recurrent neural network (RNN) unit structure [21], demonstrating higher efficiency and better performance than traditional RNNs and LSTMs when dealing with long-term dependencies [22,23]. By introducing

two gating mechanisms—the reset gate and the update gate—GRU effectively addresses the vanishing gradient problem and optimizes the control of information flow. The reset gate determines how much of the past information to retain at the current time step, while the update gate decides how the new information is integrated into the current state. This structural design not only enhances the model's understanding of sequential data but also boosts its generalization capability in small-sample learning by reducing the number of parameters. These attributes make GRU particularly valuable in scenarios with limited resources or where rapid training is required. In practical applications, whether in natural language processing or time series prediction [24,25], GRU has demonstrated its robust modeling capabilities. The internal structure of GRU is illustrated in Figure 4.
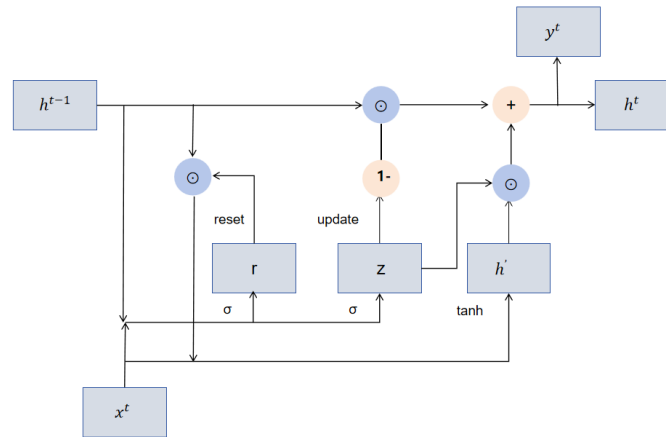
Figure 4. Internal structure of GRU.

The GRU combines the current node input x^t and the state h^{t-1} transmitted from the previous node to produce the current node's output y^t and the hidden state h^t passed to the next node. The internal parameter transmission and update formulas in the network are shown in Equations (1)–(4).

$$r^t = \sigma\left(w^{rx}x^t + w^{rh}h^{t-1} + b^r\right) \tag{1}$$

$$z^t = \sigma\left(w^{zx}x^t + w^{zh}h^{t-1} + b^z\right) \tag{2}$$

$$h' = \tanh\left(w^{xh}x^t + r^t \odot w^{hh}h^{t-1}\right) \tag{3}$$

$$h^t = \left(1 - z^t\right) \odot h^{t-1} + h' \odot z^t \tag{4}$$

In these formulas, σ denotes the *sigmoid* function, which serves as a gate control signal to keep values within the [0,1] range. The closer the gate signal is to 1, the more data are remembered. Conversely, more are forgotten. r is the reset gate control, and z is the update gate control. h' indicates the candidate's hidden state. w^{rx}, w^{rh}, and others are weight matrices, while b^r, b^z, etc., are bias terms. $\odot$ represents the Hadamard product, meaning the element-wise multiplication of matrices. The reset gate computes the data after reset $r^t \odot h^{t-1}$, which is then concatenated with x^t and passed through a *tanh* activation function to keep the output value within $[-1,1]$, resulting in the hidden state h'. Simultaneously, the update gate performs forgetting and selective memory functions, where $\left(1 - z^t\right) \odot h^{t-1}$ selectively forgets the state of previous nodes, and $h' \odot z^t$ selectively remembers the hidden state. In named entity recognition tasks, contextual information is paramount for predicting entity labels. To acquire contextual information from both the past and future within the text, bidirectional gated recurrent unit networks (BiGRUs) [26] are commonly utilized. The hidden state h' of the GRU can only access information from past contexts, not future ones.

In processing text data for Chinese marine coral reef ecosystem studies, named entity recognition plays a vital role. It aims to identify crucial ecological entities within complex natural language texts, like species names, places, ecological phenomena, etc. Given the

unique features of Chinese texts, such as flexible word order and the absence of clear word boundary delimiters, this task poses significant challenges. Traditional unidirectional GRUs might not adequately capture the contextual information around an entity, particularly subsequent information, potentially impacting the precision and completeness of entity recognition.

In the text data processing of Chinese marine coral reef ecosystem studies, Named Entity Recognition (NER) is a key task aimed at identifying critical ecological entities from complex natural language texts, such as species names, locations, ecological phenomena, etc. Due to the characteristics of Chinese text, such as flexible word order and the lack of clear word boundary markers, this task is particularly challenging. Traditional unidirectional GRUs may not adequately capture the contextual information around entities, especially posterior information, which could affect the accuracy and completeness of entity recognition.

By introducing the Bidirectional GRU (BiGRU) network, this study has optimized the Named Entity Recognition task for Chinese marine coral reef ecosystem texts. The BiGRU integrates forward and backward GRU units to capture contextual information from both before and after the entity at each time step. This structure is particularly effective in Chinese text processing, as it allows the model to consider both the modificatory information before the entity and the limiting information after the entity, thereby achieving a more comprehensive understanding of the entity.

Taking the sentence "In the coral reef ecosystem of the South Sea, the blue starfish exhibits extremely high adaptability" as an example, the BiGRU can effectively utilize the information from both "In the coral reef ecosystem of the South Sea" and "exhibits extremely high adaptability" to enhance the recognition of the entity "blue starfish". In the Chinese context, this bidirectional analysis is especially important because the attributes and descriptions of an entity are often distributed on both sides of the entity.

Therefore, by adopting BiGRUs, this study has not only improved the accuracy of entity recognition in Chinese marine coral reef ecosystem texts but also enhanced the model's versatility and flexibility in processing Chinese natural language. This demonstrates that BiGRUs have significant application value and potential in the task of Named Entity Recognition in the Chinese marine ecological field.

The structure of the BiGRU is illustrated in Figure 5. The BiGRU (Bidirectional Gated Recurrent Unit) leverages both forward and backward GRUs for the extraction of contextual information features and weights; sums the outputs; and then passes them through a linear layer to map a d-dimensional vector to an m-dimensional vector, resulting in the final output label vector list of the BiGRU network, $H = \{h_1, h_2, \ldots, h_n\}$, where $H \in \mathbb{R}^{n \times m}$, with n being the length of the text sequence and m being the number of entity type labels.
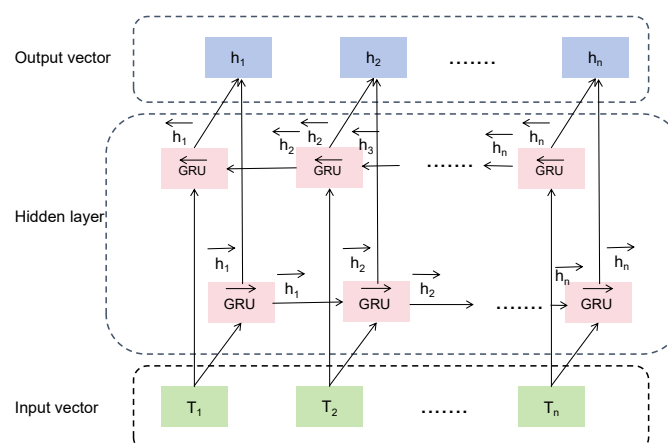
Figure 5. BiGRU structure.

2.4. Attention Layer

When dealing with Chinese text data in marine coral reef ecosystem studies, researchers often face the challenge of long sentences that contain vast amounts of information. Although the Bidirectional GRU (Bi-GRU) network can comprehensively consider the contextual information of the text, it has limitations in emphasizing specific information that is directly relevant to the target task. Therefore, to enhance the model's focus on parts of the context relevant to the current task and thereby extract more effective semantic features, this study introduces an attention layer on top of the Bi-GRU network. Specifically, the attention mechanism added to the Bi-GRU network can assign different weights to various pieces of contextual information. For information closely related to Named Entity Recognition in marine coral reef ecosystems, the model will allocate more attention, while for less relevant information, it will assign less attention. This mechanism enables the model to process information related to specific tasks more effectively, thus achieving higher performance in recognizing named entities within marine ecosystems.

By combining Bi-GRUs with the attention mechanism, the model can not only capture long-distance dependencies but also emphasize the importance of relevant features by assigning them higher weights [27]. Such an approach helps the model understand and process named entities in marine coral reef ecosystem texts more accurately and efficiently, especially in handling complex Chinese texts, significantly enhancing recognition accuracy and efficiency.

The computation steps of the attention mechanism are as follows: define $\{x_1, x_2, \ldots, x_M\}$ as the sequence of joint feature vectors input into the BiGRN network and $S = \{s_1, s_2, \ldots, s_w\}$ as the sequence of joint vectors output by the BiGRN neural network, where α_{mw} represents the normalized weights, and β_{mj} represents the attention contribution matrix, i.e., the weights assigned to the feature vectors by the attention mechanism. The specific computation formulas are shown in Equations (5) and (6):

$$\alpha_{mw} = \frac{\exp \beta_{mw}}{\sum_{k=1}^{M} \exp \beta_{mk}} \tag{5}$$

$$\beta_{mj} = c_\alpha^M \tan(w_\alpha v_{m-1} + u_\alpha s_w) \tag{6}$$

where: c, w, u are the weight matrices; v_{m-1} represents the state of the attention mechanism at the previous moment. The final output state of the attention mechanism is as shown in Equation (7):

$$V_w = \sum_{w=1}^{M} \alpha_{mw} S_w \tag{7}$$

2.5. Decoding CRF Layer

The Conditional Random Field (CRF) is a model widely applied to sequence labeling tasks [28]. In the field of natural language processing, particularly in Named Entity Recognition (NER) tasks, the introduction of a CRF layer can significantly enhance the model's performance. This improvement is attributed to CRF's ability to incorporate additional constraints in sequence prediction, aiding the model in more accurately capturing the dependencies among labels. By incorporating a CRF layer on top of the BiGRU–Attention model, a robust framework for Named Entity Recognition can be established [29]. While the BiGRU–Attention model effectively addresses the problem of long-distance dependencies by focusing on important information through the attention mechanism, it does not inherently enforce constraints on the rationality among labels. At this juncture, the role of the CRF layer becomes particularly critical: it not only considers the best prediction for individual labels but also optimizes the joint probability of the entire label sequence, ensuring that the final predicted sequence is globally optimal.

For example, in the recognition of named entities within the Chinese marine coral reef ecosystem, the CRF can utilize constraints to ensure the correct identification of entity

boundaries. It can force the model to follow specific annotation rules, such as a "B-" tag for the beginning of an entity not being directly followed by an "I-" tag for the inside of an unrelated entity. In this manner, the CRF layer helps the model correctly predict the relationships among adjacent tags, avoiding label sequences that do not conform to actual language patterns.

Positioning the CRF layer as the output layer of the BiGRU–Attention model [30] implies that in determining the label for each word, the model considers not only the features and contextual information of the word itself but also the constraints of its adjacent labels. This design significantly improves the accuracy and robustness of the model in complex text data, especially in the task of Named Entity Recognition within Chinese marine coral reef ecosystem texts. The core function of CRF is to model the dependencies among labels through a transition score matrix, thereby outputting a globally optimal and reasonable label sequence. Receives an observation sequence $\{X_1, X_2, \ldots, X_n\}$, and outputs a state sequence $\{Y_1, Y_2, \ldots, Y_n\}$ through probability calculation. The calculation method involves computing the sentence label score based on the state score (emission score) outputted by the BiGRU–Attention model and the transition score. The calculation Formula (8) is as follows:

$$\text{score}(x, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n} A_{y_i, y_{i+1}} \tag{8}$$

P_{i,y_i} represents the score of predicting the ith character as the y_ith label; $A_{y_i, y_{i+1}}$ represents the score of transitioning from the y_ith label to the y_{i+1}th label.

3. Experiment and Results Analysis

3.1. Data Processing and Annotation

3.1.1. Data Preprocessing

To address the data scarcity issue in coral reef ecosystem research, this study implemented a series of steps to construct a high-quality coral reef ecosystem corpus, as illustrated in Figure 6. This process includes data collection, cleaning, annotation, and data quality control, aimed at ensuring the accuracy and usability of the corpus.
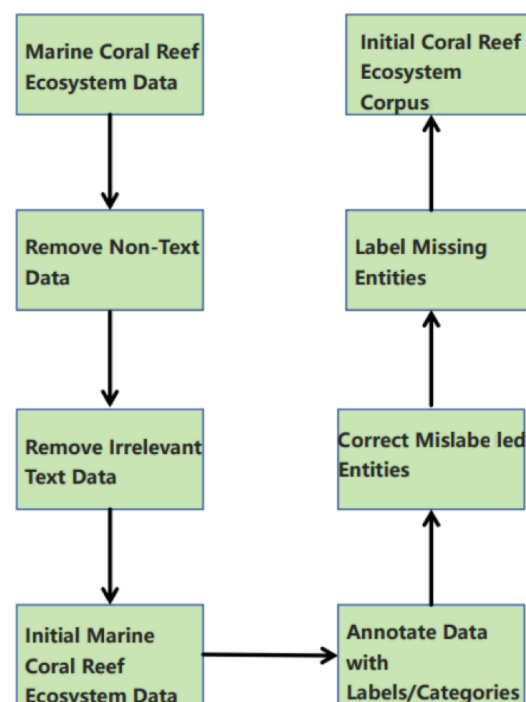

Figure 6. Flowchart for constructing a high-quality marine coral reef ecosystem data corpus.

Initially, our research focused on the Coral Reef System Status Bulletins published between 1999 and 2022, which formed our preliminary corpus. This time period was selected because it encompasses significant environmental change events, which is crucial for understanding the evolution of coral reef ecosystems. These bulletins serve as vital resources for comprehending the changes in coral reef ecosystem statuses. Next, to enhance data quality, we used regular expression techniques to precisely remove spaces, emoticons, and other irrelevant content from the text, ensuring data cleanliness and professionalism. Additionally, non-textual content that was not beneficial to the research was excluded to ensure the data's relevance and specificity. At this stage, we obtained approximately 180,000 characters of purified text data, laying the foundation for subsequent annotation work. Following this, based on extensive literature research and consultations with domain experts, we developed a specialized labeling system and meticulously annotated the data to ensure the professionalism and accuracy of the annotations. The main categories of annotation included coral species, biological species, geographical locations, environmental factors, etc., aiming to comprehensively capture the key information of the coral reef ecosystem. To enhance the accuracy and consistency of the annotations, we employed a statistical strategy to correct mislabeled entities and used text matching techniques to supplement missed entities, ensuring all relevant entities were accurately identified. Finally, we validated the accuracy and consistency of the annotation results through random sampling and comparison with the original text, ensuring high data quality standards. This step is crucial in assuring the quality of the corpus, directly impacting the effectiveness of model training.

Through the efforts described above, we successfully created a valuable resource for coral reef ecosystem research. To foster research collaboration and knowledge sharing, we meticulously documented the annotation rules and methods and plan to share the corpus and its annotated data. This will significantly advance coral reef ecosystem research.

3.1.2. Data Annotation Guidelines

The sequence labeling method plays a pivotal role in entity annotation, among which BIO and BIOES [31] are commonly used approaches. The BIO method, which is simple and intuitive, uses the B (beginning of entity) and I (inside of entity) labels to indicate the entity's location but does not precisely identify the entity's end. In contrast, the BIOES method, by adding E (end of entity) and S (single-character entity) labels, offers more detailed boundary information of the entity, particularly suitable for the discontinuous entities found in marine coral reef ecosystems, such as single-character coral species or geographical locations. Consequently, this study adopted the BIOES method, as demonstrated in Table 1, aiming for more precise entity annotation in coral reef ecosystem research.

Table 1. Label type definition.

Tag	Full Name	Description
B	Begin	Position at the beginning of an entity
I	Inside	Position inside an entity
E	End	Position at the end of an entity
S	Single	Single-character entity
O	Other	Any part that is not an entity (including punctuation, etc.)

In this study, we employed the BIOES five-tag sequence labeling approach, using the English abbreviations SPE (coral species), BSP (biological species), LOC (geographical location), and ENV (environmental factors) to represent different entity categories. Figure 7 shows the annotation format of the coral reef ecosystem named entity recognition experiment dataset (part). This method aims to refine entity recognition, especially in identifying discontinuous entities within the marine coral reef ecosystem. The study annotated a total of 4865 entities, as shown in Table 2, including 1946 coral species entities, 1459 biological

species entities, 973 geographical location entities, and 487 environmental factor entities. These data reflect the complexity of the coral reef ecosystem, particularly the diversity of coral species. The larger number of coral species entities could be attributed to the richness and complexity of the marine coral reef ecosystem. This finding underscores the necessity of accurate identification and classification of coral species, which is crucial not only for understanding the biodiversity of coral reef ecosystems but also for devising effective conservation strategies. Conversely, the smaller number of geographical location and environmental factor entities might result from the research focus, data collection methods, and the challenges encountered during the annotation process.

Figure 7. Annotation format of coral reef ecosystem Named Entity Recognition experimental dataset (part).

Table 2. Statistics of the number of annotated entities in marine coral reef ecosystems.

Entity Category	Number of Annotated Entities
Coral species	1946
Biological species	1459
Geographic locations	973
Environmental factors	487

Future research should adopt more diverse data collection and analysis methods to ensure the comprehensive annotation and analysis of key entities like geographical locations and environmental factors, further facilitating in-depth studies of coral reef ecosystems.

3.2. Network Model Evaluation Metrics

When evaluating the performance of Named Entity Recognition models for marine coral reef ecosystems, this study utilizes three core metrics: Precision (P), Recall (R), and the F1 Score. These metrics collectively form a comprehensive evaluation system to quantify the model's performance. Precision (P) reflects the proportion of named entities correctly identified by the model out of all the entities it identified. Recall (R) measures the proportion

of named entities correctly identified by the model out of all actually named entities. The F1 Score is the harmonic mean of Precision and Recall, providing a balanced reflection of the model's accuracy and robustness. The formulas for these three metrics are shown in Equations (9)–(11):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{10}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{11}$$

Here, "TP" (True Positive) refers to the number of samples correctly predicted as positive, "FP" (False Positive) refers to the number of samples incorrectly predicted as positive, and "FN" (False Negative) refers to the number of samples incorrectly predicted as negative.

Through the comprehensive application of these three metrics, the performance of the Named Entity Recognition model for marine coral reef ecosystems can be thoroughly evaluated, ensuring both comprehensive and precise model assessment.

### 3.3. Experimental Results and Analysis

This study introduces a deep learning model that integrates BERT, BiGRU, and CRF, enhanced by an attention mechanism, to improve the recognition of complex entity structures. The model utilizes BERT to capture context-relevant character vectors, employs BiGRU to extract global semantic features, incorporates an attention mechanism to focus on key information, and uses CRF to produce optimized label sequences. We constructed a specialized coral reef ecosystem corpus to evaluate the model's performance through a series of experiments. To validate the effectiveness of the proposed BERT+BiGRU+Att+CRF model, two sets of comparative experiments were conducted, as detailed in Tables 3 and 4. The first set included the proposed BERT+BiGRU+Att+CRF model, along with BERT+BiLSTM+Att+CRF, BERT+BiGRU+CRF, BERT+BiLSTM+CRF, BiGRU+CRF, BiLSTM+CRF and CRF models for comparison. The results, as seen in Table 3 and Figure 8, compare the impact of different algorithmic models on named entity recognition. It was observed that BiGRU, in comparison to BiLSTM, shows slightly superior performance without the integration of other technologies (Model 2 vs. Model 3). Introducing the BERT pre-trained model (Models 4 and 5 vs. Models 2 and 3) significantly enhances all performance metrics, highlighting BERT's effectiveness in improving entity recognition. Further, the incorporation of the attention mechanism (Attention) in Models 6 and 7, as compared to Models 4 and 5, markedly improves performance, with Model 7 demonstrating enhancements in all metrics. This analysis suggests that BiGRU, relative to BiLSTM, may provide a more favorable performance–time cost ratio, especially when combined with the BERT pre-trained model and attention mechanism. The BERT pre-trained model, by leveraging semantic information across different contexts to generate dynamic word embeddings, effectively addresses polysemy. The attention mechanism further boosts the model's ability to distill semantic features, thereby enhancing entity recognition performance. Notably, Model 7 (BERT-BiGRU+Att+CRF) not only delivers robust performance but also reduces training time by employing BiGRU instead of BiLSTM, offering savings in time costs.

Table 3. Comparison of overall model recognition performance (%).

Index	Model	Precision (P)	Recall (R)	F1 Score
1	CRF	75.23	73.65	74.44
2	BiLSTM+CRF	81.03	82.05	81.62
3	BiGRU+CRF	81.56	82.33	81.94
4	BERT+BiLSTM+CRF	83.95	83.81	83.88
5	BERT+BiGRU+CRF	84.14	84.09	84.11
6	BERT-BiLSTM-Att-CRF	85.96	85.68	86.06
7	BERT-BiGRU-Att-CRF	86.25	86.77	86.54

Table 4. Comparison of overall recognition performance with mainstream models (%).

Index	Model	Precision (P)	Recall (R)	F1 Score
1	IDCNN+CRF	81.14	79.90	80.57
2	BERT+IDCNN+CRF	85.29	85.17	85.22
3	BERT-BiGRU-Att-CRF	86.25	86.77	86.54
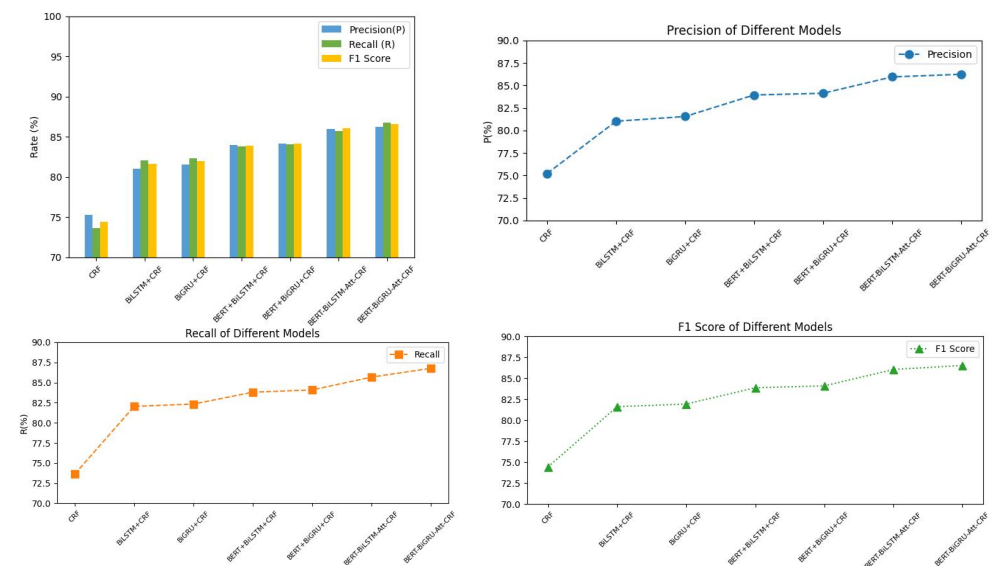
Figure 8. Comparison of overall model recognition performance (%).

The second set of comparative experiments involved the proposed model and the current mainstream named entity recognition models, namely, the IDCNN (Iterative Dilated Convolution)+CRF model and the BERT+IDCNN+CRF model. The overall recognition performance of each model in the second set of experiments is shown in Table 4 and Figure 9. These results indicate that the performance of the IDCNN+CRF combination, when augmented with the BERT pre-trained model, significantly surpasses that of the standalone IDCNN+CRF model. However, the proposed BERT+BiGRU+Att+CRF model further improves across all three metrics, demonstrating the highest Precision, Recall, and F1 Scores. This indicates that the model proposed in this paper, BERT+BiGRU+Att+CRF, outperforms mainstream named entity recognition models in terms of entity recognition effectiveness.

Analyzing the results of these two sets of experiments, it is evident that the BERT+BiGRU+Att+CRF model is effective in the task of named entity recognition within marine coral reef ecosystem contexts, especially when compared against the mainstream IDCNN+CRF model. This superiority in recognition accuracy can be attributed to the richer context word vectors provided by the BERT model, the more effective sequence processing capability of the BiGRU units, and the attention mechanism's capacity to weigh important information. Additionally, the CRF layer ensures that the predicted label sequence is globally optimal, enhancing the accuracy of boundary recognition. The synergistic effect among

these four components—BERT, BiGRU, the attention mechanism, and CRF—enables the model to more precisely identify and classify named entities in texts.
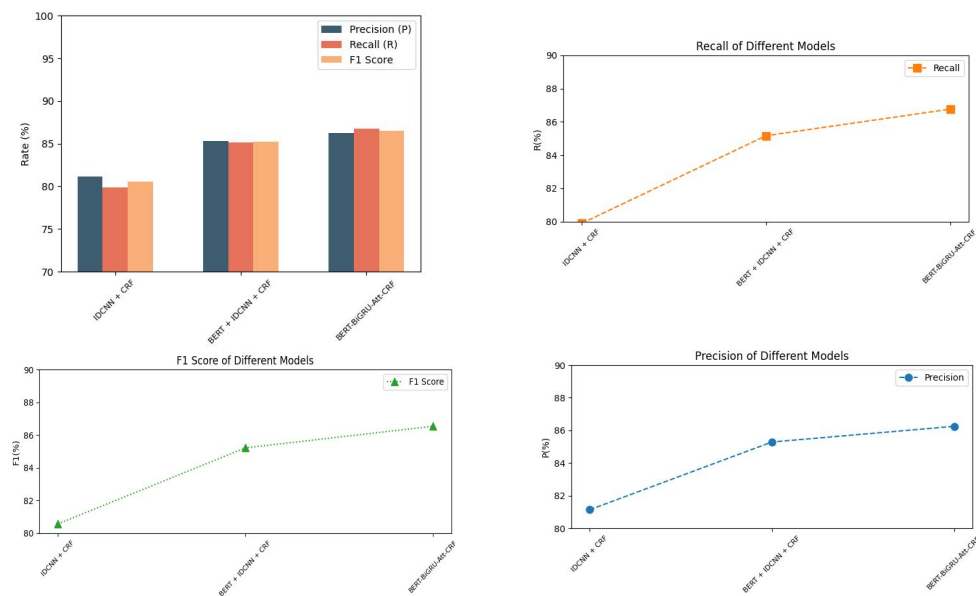
Figure 9. Comparison of overall recognition performance with mainstream models (%).

4. Discussion

4.1. Model Performance and Structural Advantages

The deep learning model proposed in this study, which integrates BERT, BiGRU, attention mechanism, and CRF, performs excellently in the task of named entity recognition in Chinese marine coral reef ecosystem texts, achieving an F1 Score of 86.54%, significantly surpassing existing methods. The model's outstanding performance is primarily attributed to its unique structural design. Firstly, by utilizing the BERT model, our model can capture rich, context-relevant word vectors, providing a solid foundation for understanding the meaning of the text [32]. Then, the use of BiGRU further enhances the model's ability to process sequence dependencies, enabling the model to effectively capture the global semantics at the sentence level [33]. Moreover, the incorporation of the attention mechanism allows the model to focus on key information related to coral reefs, significantly improving the accuracy of entity recognition [34]. Finally, the application of the CRF layer optimizes the output of the label sequence, further improving the model's performance by considering the transition probabilities between labels. This structural innovation is specially optimized for the recognition of long entities and nested entities in marine coral reef ecosystem texts. Ref. [35] significantly enhances the ability to recognize these complex entity structures, addressing the challenges faced by existing named entity recognition methods when dealing with complex marine texts.

These findings directly respond to the research questions by demonstrating that our model can effectively handle the unique characteristics of marine domain texts, particularly the complexity and variability of entity structures. The model's superior performance highlights its potential as a powerful tool for NER in specialized fields.

4.2. Limitations and Future Work

Despite the significant achievements of this study, there are still some limitations. Firstly, the model's performance is largely dependent on the pre-trained BERT model, which may limit its generalization ability on some specific marine texts [36]. Secondly, the current training dataset is relatively limited in size, which may affect the model's robustness [37]. Future work can focus on exploring more advanced pre-trained models to further improve the model's performance and generalization ability. Additionally, expanding the training dataset, especially by increasing the samples of different types of

marine texts, is an important direction for enhancing the model's robustness [38,39]. These limitations suggest that while our model performs well within the scope of the current study, broader applications across diverse marine texts require further enhancement and validation. Addressing these limitations will help in refining the model and expanding its applicability in the marine domain.

4.3. Practical Application Prospects

The achievements of this study provide new tools and perspectives for the construction and research of knowledge graphs in the marine domain, offering technical support for digital transformation and knowledge mining in the field of marine science. These efforts are expected to further promote the development of marine science, facilitate the sustainable utilization of marine resources, and provide strong information support for marine ecological protection and management [40]. This study contributes significantly to the broader field of marine coral reef ecosystem research by providing a robust methodological framework for NER, which is essential for data-driven ecological studies and management. By improving the accuracy and efficiency of extracting key information from textual data, our research supports better decision-making and strategic planning in marine conservation and resource management.

5. Conclusions

This study addresses the challenges of non-standardization and limited annotation resources in Chinese marine domain texts, particularly with complex entities like long and nested entities in coral reef ecosystem-related texts. Traditional NER methods often fail to capture deep semantic features, leading to inefficiencies and inaccuracies. To overcome these challenges, we present a Named Entity Recognition (NER) model, termed the BERT+BiGRU+Att+CRF model, specifically designed for marine coral reef ecosystems. The BERT+BiGRU+Att+CRF model leverages BERT to acquire rich contextual information and utilizes BiGRUs to effectively process sequence dependencies, thereby capturing sentence-level global semantics more accurately. The introduction of an attention mechanism enhances the model's focus on key features, which is particularly important for extracting complex semantic features related to coral reefs. In the final stage, the CRF layer outputs the optimal sequence of entity labels by considering the transition probabilities between tags. Through a series of experiments on a specialized coral reef ecosystem corpus, the BERT+BiGRU+Att+CRF model achieved an F1 Score of 86.54%, significantly outperforming other mainstream models. This demonstrates the model's superiority in the task of entity recognition within marine coral reef ecosystems. Our research contributes to the field in three key ways: (1) Efficient NER framework: We designed an efficient NER framework for marine domain texts, which significantly improves the recognition of long and nested entities. (2) Enhanced model with attention mechanism: By integrating the attention mechanism, we enhanced the model's ability to recognize complex entity structures, addressing the limitations of existing methods in capturing deep semantic features. (3) Foundation for marine knowledge graphs: This work provides new tools and perspectives for the construction of knowledge graphs in the marine domain, laying a solid foundation for future research and advancing the development of marine domain text analysis technology. In future work, we aim to explore more advanced pre-trained models to further enhance the performance of the BERT+BiGRU+Att+CRF model and reduce training time. We also plan to expand the training dataset to boost the model's robustness and generalization ability. Additionally, we will consider applying the model to a broader range of NER tasks to verify its applicability across different domains, thereby contributing valuable references for related research fields. To advance textual analysis in both English and Chinese, it is essential to develop language-specific tools that consider the unique linguistic features of each language. Future research should focus on creating more sophisticated algorithms and models that can accurately process and analyze texts in different languages.

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

References

1. Hughes, T.P.; Barnes, M.L.; Bellwood, D.R.; Cinner, J.E.; Cumming, G.S.; Jackson, J.B.; Kleypas, J.; Van De Leemput, I.A.; Lough, J.M.; Morrison, T.H.; et al. Coral reefs in the Anthropocene. *Nature* **2017**, *546*, 82–90. [CrossRef] [PubMed]
2. Zhao, D.; Lou, Y.; Song, W.; Huang, D.; Wang, X. Stability analysis of reef fish communities based on symbiotic graph model. *Aquac. Fish.* **2023**, *in press*. [CrossRef]
3. Liu, P.; Guo, Y.; Wang, F.; Li, G. Chinese named entity recognition: The state of the art. *Neurocomputing* **2022**, *473*, 37–53. [CrossRef]
4. Liu, C.; Zhang, W.; Zhao, Y.; Luu, A.T.; Bing, L. Is translation all you need? A study on solving multilingual tasks with large language models. *arXiv* **2024**, arXiv:2403.10258.
5. Morwal, S.; Jahan, N.; Chopra, D. Named entity recognition using hidden Markov model (HMM). *Int. J. Nat. Lang. Comput.* **2012**, *1*, 15–23. [CrossRef]
6. Song, S.; Zhang, N.; Huang, H. Named entity recognition based on conditional random fields. *Clust. Comput.* **2019**, *22*, 5195–5206. [CrossRef]
7. Ekbal, A.; Bandyopadhyay, S. Named entity recognition using support vector machine: A language independent approach. *Int. J. Electr. Comput. Eng.* **2010**, *4*, 589–604.
8. Cao, X.; Yang, Y. Research on Chinese Named Entity Recognition in the Marine Field. In Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 21–23 December 2018; pp. 1–7.
9. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
10. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. *arXiv* **2016**, arXiv:1603.01360.
11. He, L.; Zhang, Y.; Ba, H. Named entity recognition of exotic marine organisms based on attention mechanism and deep learning network. *J. Dalian Ocean. Univ.* **2021**, *36*, 503–509.
12. He, S.; Sun, D.; Wang, Z. Named entity recognition for Chinese marine text with knowledge-based self-attention. In *Multimedia Tools and Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–15.
13. Ma, X.; Yu, R.; Gao, C.; Wei, Z.; Xia, Y.; Wang, X.; Liu, H. Research on named entity recognition method of marine natural products based on attention mechanism. *Front. Chem.* **2023**, *11*, 958002. [CrossRef] [PubMed]
14. Perera, N.; Dehmer, M.; Emmert-Streib, F. Named entity recognition and relation detection for biomedical information extraction. *Front. Cell Dev. Biol.* **2020**, *8*, 673. [CrossRef] [PubMed]
15. Tenney, I.; Das, D.; Pavlick, E. BERT rediscovers the classical NLP pipeline. *arXiv* **2019**, arXiv:1905.05950.
16. Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; Tang, J. GPT understands, too. *AI Open* **2023**, *in press*. [CrossRef]
17. Wu, Y.; Huang, J.; Xu, C.; Zheng, H.; Zhang, L.; Wan, J. Research on named entity recognition of electronic medical records based on roberta and radical-level feature. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1–10.
18. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; NIPS Foundation: Long Beach, CA, USA, 2017; Volume 30.

21. Dey, R.; Salem, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 6–9 August 2017; pp. 1597–1600.

22. Shewalkar, A.; Nyavanandi, D.; Ludwig, S.A. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *J. Artif. Intell. Soft Comput. Res.* **2019**, *9*, 235–245. [CrossRef]

23. Yin, W.; Kann, K.; Yu, M.; Schütze, H. Comparative study of CNN and RNN for natural language processing. *arXiv* **2017**, arXiv:1702.01923.

24. Zulqarnain, M.; Ghazali, R.; Ghouse, M.G.; Mushtaq, M.F. Efficient processing of GRU based on word embedding for text classification. *Int. J. Inform. Vis.* **2019**, *3*, 377–383. [CrossRef]

25. Zhang, X.; Shen, F.; Zhao, J.; Yang, G. Time series forecasting using GRU neural network with multi-lag after decomposition. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, 14–18 November 2017, Proceedings, Part V 24*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 523–532.

26. She, D.; Jia, M. A BiGRU method for remaining useful life prediction of machinery. *Measurement* **2021**, *167*, 108277. [CrossRef]

27. Zhang, Y.; Chen, Y.; Yu, S.; Gu, X.; Song, M.; Peng, Y.; Chen, J.; Liu, Q. Bi-GRU relation extraction model based on keywords attention. *Data Intell.* **2022**, *4*, 552–572. [CrossRef]

28. Souza, F.; Nogueira, R.; Lotufo, R. Portuguese named entity recognition using BERT-CRF. *arXiv* **2019**, arXiv:1909.10649.

29. Liu, W.; Hu, Z.; Zhang, J.; Liu, X.; Lin, F. Optimized Named Entity Recognition of Electric Power Field Based on Word-Struct BiGRU. In Proceedings of the 2021 IEEE Sustainable Power and Energy Conference (iSPEC), Nanjing, China, 23–25 December 2021; pp. 3696–3701.

30. Cai, G.; Su, X.; Wu, T. Causality Extraction of Fused Character Features with BiGRU-Attention-CRF. *Int. Core J. Eng.* **2023**, *9*, 47–59. [CrossRef]

31. Ke, J.; Wang, W.; Chen, X.; Gou, J.; Gao, Y.; Jin, S. Medical entity recognition and knowledge map relationship analysis of Chinese EMRs based on improved BiLSTM-CRF. *Comput. Electr. Eng.* **2023**, *108*, 108709. [CrossRef]

32. Jia, C.; Shi, Y.; Yang, Q.; Zhang, Y. Entity enhanced BERT pre-training for Chinese NER. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6384–6396.

33. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.

34. Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; Li, H. Modeling coverage for neural machine translation. *arXiv* **2016**, arXiv:1601.04811.

35. Lafferty, J.; McCallum, A.; Pereira, F. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*; Icml: Williamstown, MA, USA, 2001; Volume 1, p. 3.

36. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, 18–20 October 2019, Proceedings 18*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 194–206.

37. Ratner, A.J.; De Sa, C.M.; Wu, S.; Selsam, D.; Ré, C. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.

38. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* **2017**, arXiv:1706.05098.

39. Zhao, D.; Yang, X.; Song, W.; Zhang, W.; Huang, D. Visibility graph analysis of the sea surface temperature irreversibility during El Ni no events. *Nonlinear Dyn.* **2023**, *111*, 17393–17409. [CrossRef]

40. Hedley, J.D.; Roelfsema, C.M.; Chollett, I.; Harborne, A.R.; Heron, S.F.; Weeks, S.; Skirving, W.J.; Strong, A.E.; Eakin, C.M.; Christensen, T.R.; et al. Remote sensing of coral reefs for monitoring and management: A review. *Remote Sens.* **2016**, *8*, 118. [CrossRef]