applied sciences

MDPI

Article

Named Entity Recognition in Government Audit Texts Based on ChineseBERT and Character-Word Fusion

Baohua Huang *, Yunjie Lin, Si Pang and Long Fu

School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China; 2113391031@st.gxu.edu.cn (Y.L.); 2213393033@st.gxu.edu.cn (S.P.); a1972084551@gmail.com (L.F.)
* Correspondence: bhhuang66@gxu.edu.cn; Tel.: +86-152-9654-4306

Abstract: Named entity recognition of government audit text is a key task of intelligent auditing. Aiming at the problems of scarcity of corpus in the field of governmental auditing, insufficient utilization of traditional character vector word-level information features, and insufficient capturing of auditing entity features, this study builds its own dataset in the field of auditing and proposes the model CW-CBGC for recognizing named entities in governmental auditing text based on ChineseBERT and character-word fusion. First, the ChineseBERT pre-training model is used to extract the character vector that integrates the features of glyph and pinyin, combining with word vectors dynamically constructed by the BERT pre-training model; then, the sequences of character-word fusion vectors are input into the bi-directional gated recurrent neural network (BiGRU) to learn the textual features. Finally, the global optimal sequence label is generated by Conditional Random Field (CRF), and the GHM classification loss function is used in the model training to solve the problem of error evaluation under the conditions of noisy entities and unbalanced number of entities. The F1 value of this study's model on the audit dataset is 97.23%, which is 3.64% higher than the baseline model's F1 value; the F1 value of the model on the public dataset Resume is 96.26%, which is 0.73–2.78% higher than the mainstream model. The experimental results show that the model proposed in this paper can effectively recognize the entities in government audit texts and has certain generalization ability.

Keywords: smart audit; named entity recognition; character-word fusion; GHM loss function; ChineseBERT

Citation: Huang, B.; Lin, Y.; Pang, S.; Fu, L. Named Entity Recognition in Government Audit Texts Based on ChineseBERT and Character-Word Fusion. *Appl. Sci.* **2024**, *14*, 1425. https://doi.org/10.3390/app14041425

Academic Editors: José Ramón Méndez Reboredo and David Ruano-Ordás

Received: 9 January 2024
Revised: 4 February 2024
Accepted: 5 February 2024
Published: 9 February 2024

1. Introduction

With the establishment of China's smart government, the digitization of audit in smart government has witnessed rapid development, leading to a significant increase in the volume of audit data. However, faced with this ever-growing amount of audit information, auditors' efficiency and quality gradually fail to meet the demands of audit tasks [1]. The presence of numerous unstructured data can result in misjudgment and omission by auditors, thereby compromising the reliability and accuracy of audit results. In order to facilitate information comparison for auditors during their tasks, it is crucial to identify named entities from unstructured data. Named Entity Recognition (NER), as a prominent research focus within the natural language processing field, primarily aims at identifying target entities with specific meanings within particular domains from unstructured text [2]. For government audits specifically, NER plays a vital role in identifying relevant entities, such as project names, addresses, amounts involved, timeframes, etc., from audit texts.

Named entity recognition technology originated from the MCU-6 conference in 1995 [3], which first proposed the concept of named entity recognition and made named entity recognition a basic task in natural language processing. After years of development, named entity recognition technology has evolved from the method that initially relies on rules and dictionaries to the method that relies on traditional machine learning models and then to the method that is based on various deep learning technologies. The early rule-and-dictionary-based methods manually construct rule templates that conform to such texts by

combining the language features of the text with relevant dictionaries and finally use manual matching methods to identify entities related to the rule templates. There are disadvantages that require much manpower, poor portability, low reusability, and dependence on rule templates. With the help of traditional machine learning models, researchers regard named entity recognition as a sequence annotation task, and the most common are the hidden Markov model [4], conditional random field model [5,6] and maximum entropy model [7]. Liu et al. [8] improved the traditional hidden Markov model, not only considering the characteristics of the word itself, but also considering the impact of N event states before and after the word, improving the accuracy of model recognition. Although traditional machine learning alleviates the early over-reliance on template rules to some extent, the improvement of model accuracy requires a large amount of feature labeled data to support.

The powerful feature extraction ability of deep neural network does not require feature engineering for text. In recent years, deep learning-based methods have become a popular research direction in named entity recognition tasks [9]. Collobert et al. [10] combined convolutional neural network and conditional random field models to construct a general neural network framework to learn the internal features of a large number of unlabeled data to achieve named entity recognition. The short-term memory effect of traditional convolutional neural network and recurrent neural network is good, but it cannot record the pre- and post-sequence feature information with a long distance, which will lead to gradient attenuation. The works of [11,12] first proposed the model framework combining long short-term memory network, bidirectional long short-term memory network and conditional random field, which alleviates gradient explosion and attenuation by obtaining bidirectional feature information of word vectors. It is applied to named entity recognition with excellent recognition ability and has become the basic framework of most model feature extraction layers and widely used. Lin et al. [13] introduced CNN on the basis of BiLSTM-CRF model framework to capture global context feature information and local key information in parallel and then fused feature information through multi-head attention mechanism to successfully identify relevant entities in the field of subway vehicle equipment, which alleviated the problem of insufficient features in that field. However, the above models use the static word vector generated by the Word2Vec model as input, which cannot represent Chinese polysemous words and affects the recognition effect of the model [14,15].

The pre-training model BERT [16] proposed by Devlin et al. of Google and the subsequent improved pre-training language model can capture character position information and bidirectional feature relationship in the corpus, dynamically generate context-related word vectors, effectively solve the problem of Chinese polysemy, and are widely used in NER tasks. Studies [17,18] have introduced the bidirectional pre-training language model BERT to dynamically obtain word vectors, then capture the context bidirectional feature information and decode and identify related domain entities through BiLSTM-CRF, which solves the problem of low accuracy of biomedical and earthquake emergency text entity recognition. Yang et al. [19] used the BERT pre-training model to obtain word vectors containing position information, extract global and local features with BiLSTM and IDCNN, respectively, and connect them, and identify COVID-19 epidemiological entities through conditional random field decoding, which solves the problem of insufficient local feature extraction. Qian et al. [20] proposed the model framework of MacBERT-BiLSTM-CRF for audit texts and introduced adversarial training, which takes the named entity recognition task as the main task and the Chinese word segmentation task as the auxiliary task and shares the word vector boundary information to help the model identify the entity boundary, which solves the problem of unclear entity boundary recognition in audit texts. In view of the problem that Chinese pre-training models lack two factors of Chinese glyph and pinyin, Sun et al. [21] integrated the visual features of glyph and pronunciation features into the pre-training stage of this model to improve its understanding ability of Chinese grammar and semantics.

Deep learning methods using pre-trained models and incorporating external feature information have continuously achieved better results than traditional models in named entity recognition tasks, gradually becoming a hot topic of current research. Based on external dictionary information and the BERT model, the works of [22,23], respectively, used bidirectional maximum matching strategy and an attention mechanism to integrate external dictionary features into dynamic word vectors, improve the recognition accuracy of the model for rare or unknown entities, and enhance the performance of the model for entity recognition in the agricultural field. Ni et al. [24] created a word-splitting dictionary in the field of automobile production equipment faults, extracted information by using part-of-speech features, and fused the dynamic word vectors generated by BERT to obtain joint word vectors, which solved the problems of internal semantic information loss and single feature extraction of traditional word vectors in named entity recognition in the field of automobile production equipment faults.

Named entity recognition technology has achieved many good results in the fields of medical treatment, agriculture and electric power, helping people to extract key information with specific meaning from massive unstructured texts. Compared with the achievements of named entity recognition technology in other fields, the named entity recognition technology in the field of government audit is not only in its infancy but also faces many problems. In the field of auditing, there are few normative datasets, and traditional models cannot be fully trained, resulting in poor model recognition effect. The audit entity names are long and complex, the character vectors output by the traditional model ignore the glyph and pinyin features of Chinese characters, and the word level information features are not fully utilized, resulting in insufficient capture of audit entity features, leading to low model recognition accuracy.

To solve these problems, a named entity recognition model for government audit text based on ChineseBERT and character-word fusion (CW-CBGC, Character-Word fusion, ChineseBERT, GRU, CRF) is proposed. On the one hand, the audit named entity recognition dataset Audit is constructed based on the professional knowledge of auditors. On the other hand, the ChineseBERT pre-training model is used to extract the character vectors that integrate the glyph and pinyin features of Chinese characters, and they are fused with the character sequence word vectors dynamically constructed by the BERT pre-training model. Finally, the BiGRU-CRF model is used to learn text features based on context information, and the entity loss is reconciled by the Gradient Harmonizing Mechanism (GHM) in the model training stage so as to output the global optimal sequence label with the highest probability.

The rest of this article is organized as follows. Section 2 introduces the CW-CBGC model proposed in this paper. Section 3 describes the experimental datasets, basic experimental settings, experimental results and analysis. Finally, the conclusion is presented in Section 4.

2. Named Entity Recognition Model for Government Audit Text

2.1. Overall Framework of Model

The overall framework of the named entity recognition model CW-CBGC for government audit text is illustrated in Figure 1, comprising three main components: an embedding layer, a feature extraction layer, and a conditional random field layer. This model takes sentences from the audit text as input sequences, with the embedding layer utilizing a fused representation of word and character information. Specifically, the character vector incorporates Chinese pinyin and glyph features through pre-training with ChineseBERT, while the word vector dynamically expresses word representations using BERT. The feature extraction layer uses a bidirectional gated recurrent neural network combined with context information to learn text features; the conditional random field layer considers the dependence and constraints between labels to obtain the global optimal prediction sequence with the maximum probability as the final output of the model.

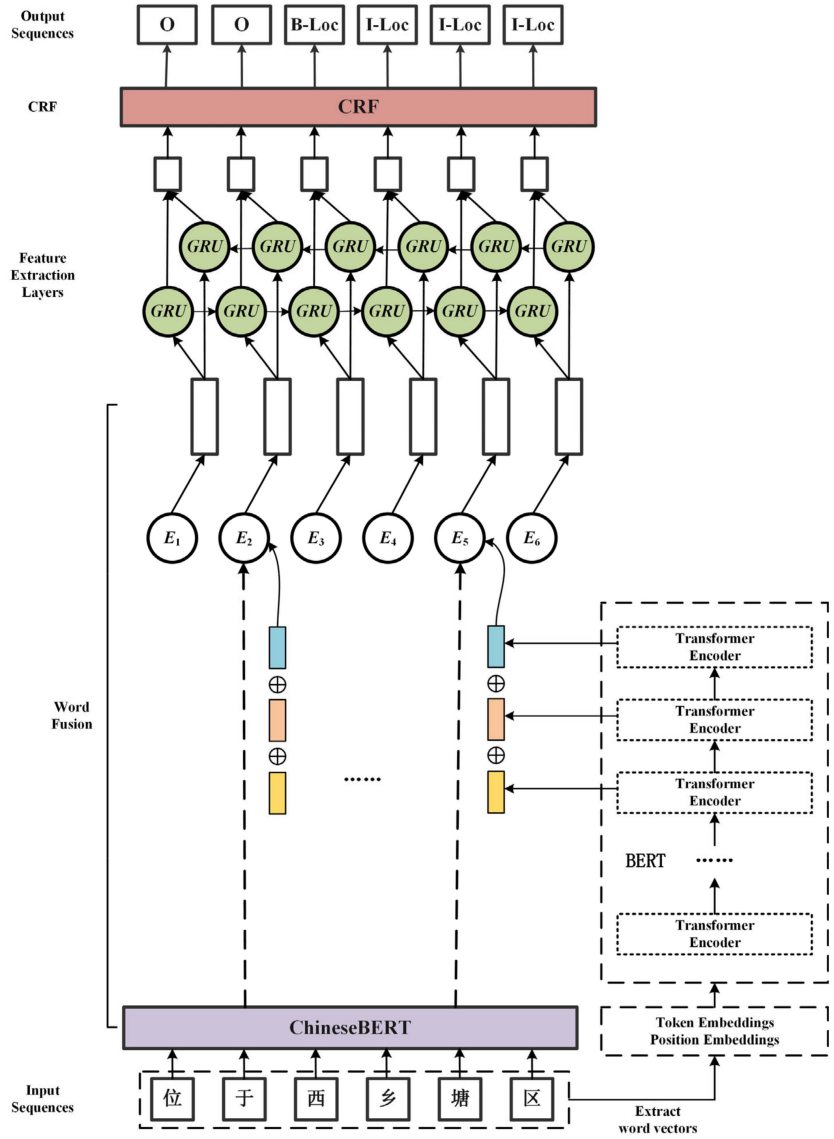

Figure 1. Overall structure of the model. "位于西乡塘区" corresponds to "located in Xixiangtang district".

2.2. Embedded Layer

2.2.1. Character Vector Representation

The traditional character vector representation from the static Word2Vec model to the dynamic BERT model ignores two important features of Chinese characters: glyph and pinyin features. Glyph features can capture the semantic information of Chinese characters, and pinyin features can effectively solve the heterophonic phenomenon of Chinese characters, further enhancing the ability of the model to Chinese natural language processing. Therefore, this study obtains the character vectors that integrate the characteristics of Chinese pinyin and glyph through the ChineseBERT [21] pre-training model. The structure of the ChineseBERT model is shown in Figure 2.

For each Chinese character, the ChineseBERT pre-training model utilizes a fully connected layer to map the character embedding, glyph embedding, and pinyin embedding into a fusion embedding. This fusion embedding is then inputted into a stack of 12 Transformers in the pre-training model. During context feature extraction, the glyph feature captures semantic information of Chinese characters while the pinyin feature addresses the issue of polyphony in single words. By fusing glyph and pinyin features, the charac-

ter vector better represents polysemy and part-of-speech of Chinese characters, thereby enhancing the model's performance in mining audit entity feature information.

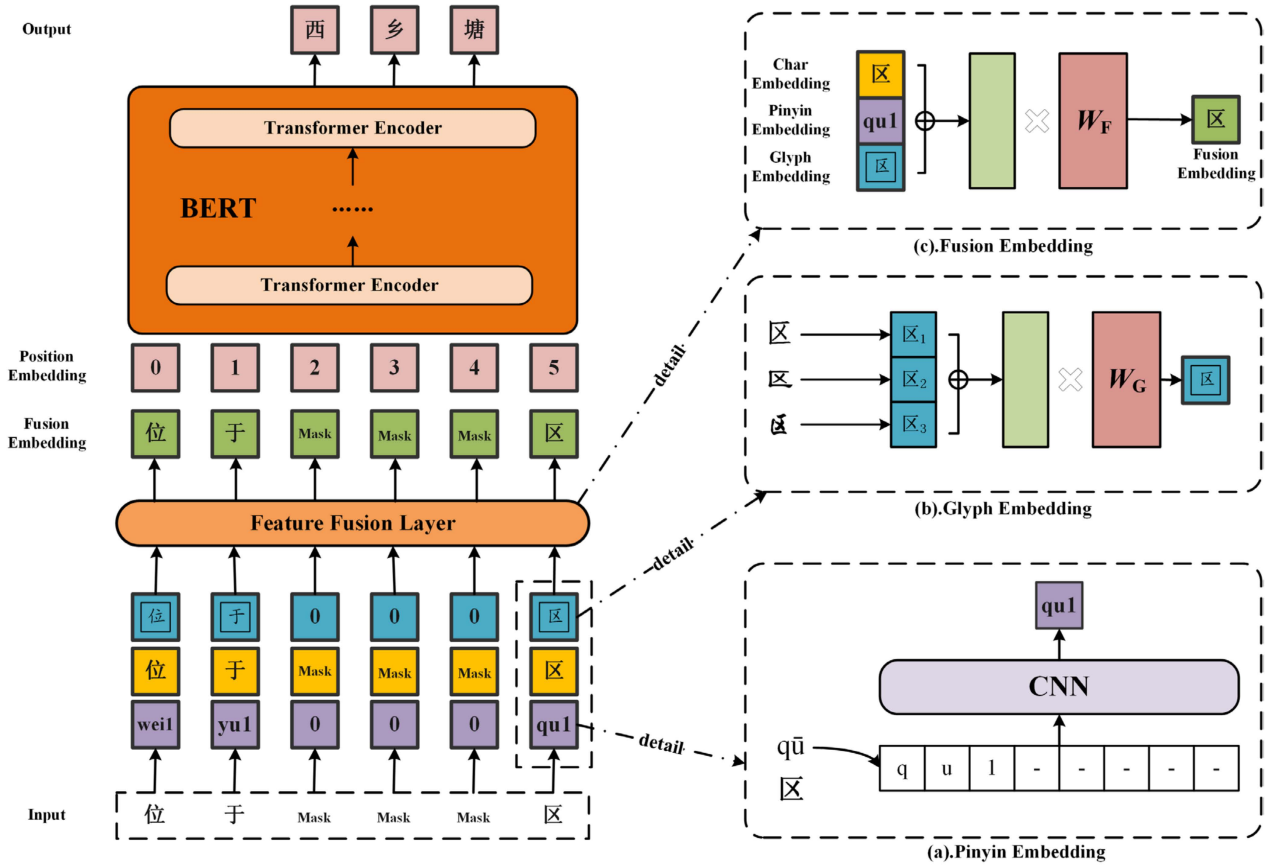

Figure 2. ChineseBERT model structure. "位", "于" and "区" correspond to "Wei", "Yu" and "Qu" respectively. "西乡塘" correspond to "Xixiangtang".

The glyph embedding structure is shown in Figure 2b. For the input Chinese characters, the images of the fonts of FangSong, XingKai and LiShu with the size of 24 × 24 are stitched into a tensor of 24 × 24 × 3, and the glyph embedding is generated through the FC layer after flattening; the pinyin embedding structure is shown in Figure 2a. The pinyin of Chinese characters is converted into a sequence of fixed length eight, in which the four pinyin syllables are represented by the Arabic numerals 1–4 and placed at the end of the sequence. When the actual pinyin sequence is less than eight, the sequence is filled with "-". The sequence is input into the CNN model with a width of 2 and the pinyin embedding is obtained by the maximum pooling method.

For the input audit text sentence, it is defined as the character sequence $S = \{c_1, c_2, \ldots, c_n\}$, where c_i represents the ith Chinese character in the sentence sequence. The ChineseBERT pre-training model is used to obtain the character vector representation of each character in the sentence by integrating Chinese glyph and pinyin features, denoted as $V_C = \{v_1^c, v_2^c, \ldots, v_n^c\}$.

2.2.2. Word Vector Representation

The methods for acquiring word vectors primarily include SoftLexicon dictionary matching, the Word2Vec word vector model, and various other approaches. SoftLexicon constructs word vectors by extensively matching the four sets of BMES words with external dictionaries; however, it lacks generalization and requires specific external dictionaries to be constructed for particular domains. On the other hand, the Word2Vec word vector model can extract comprehensive lexical-level feature information to construct word vec-

tors but fails to address polysemy issues in Chinese. Therefore, this study employs BERT pre-training models to generate dynamic word vectors.

The deep bidirectional pre-training language model BERT [16] is stacked by 12 Transformer networks, and the Encoder network in the Transformer is trained by introducing two tasks, namely random masking (MLM) and next sentence prediction (SEP), in the unsupervised pre-training stage. In the 12-layer network stacked by BERT, the bottom layer captures the surface information features of language; the middle layer captures the syntactic information features; and the high-level network captures the semantic information features and can gradually handle the remote dependency problem. For the audit named entity recognition task, the semantic information features contained in the word vector are crucial to audit entity recognition, while the next sentence prediction task and the surface information features of language have little effect on the entity recognition accuracy. The output of the last three Encoder networks is fused and spliced into a word vector, which carries rich contextual semantic information features, which can further improve the entity recognition accuracy of the model and reduce the impact of redundant information.

Firstly, the audit text sentence is segmented by Jieba word segmentation tool. In order to correspond with the character vector, each word after segmentation is repeated n times, where n is the number of characters contained in Chinese words. The sentence sequence $S = \{ w_1, w_2, \ldots, w_n \}$ of simple word segmentation is input into the BERT pre-training model. In order to fully capture the semantic information features of entities, improve the recognition efficiency of the model and avoid the influence of redundant information on the model, the output of the last three Encoder layers in the high-level network is extracted and fused into the final output word vector representation, namely, $V_w = \{ v_1^w, v_2^w, \ldots, v_n^w \}$. The word vector extraction is shown in Figure 3.

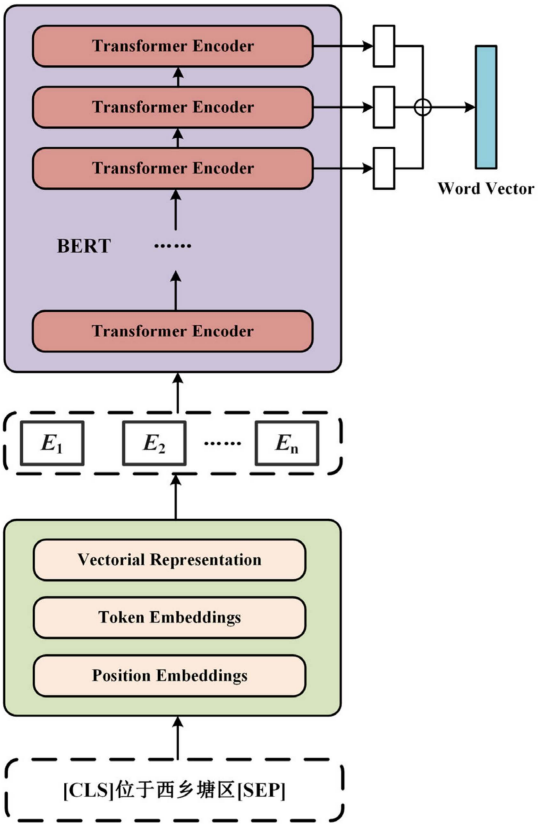

Figure 3. Word vector extraction process. ”位于西乡塘区” corresponds to ”located in Xixiangtang district”.

2.2.3. Character-Word Vector Fusion

After obtaining the character vector representation v_i^c and word vector representation v_i^w, respectively, in order to enrich the lexical feature information of the characters, the character vector representation of each character and the corresponding word vector representation are spliced and fused to form the character-word fusion vector representation, as shown in Equation (1):

$$v_i = concat(v_i^c, v_i^w) \tag{1}$$

The final output sequence of the embedding layer, denoted as $V = \{v_1, v_2, \ldots, v_n\}$, is formed by combining the character-word fusion vector representation corresponding to each character in the input audit text sentence.

2.3. Feature Extraction Layer

There are many addresses, project names and project responsible persons in audit text entities whose lengths exceed 10 characters. Such entity recognition relies on long-distance feature information. The GRU model can better capture the long-distance text feature information of government audit entities, which is helpful to accurately identify audit entities. The GRU [25] model is a special recurrent neural network optimized based on the LSTM [26] model, which has a simple unit structure, less redundant information and high calculation efficiency. The internal structure of the GRU unit is shown in Figure 4.

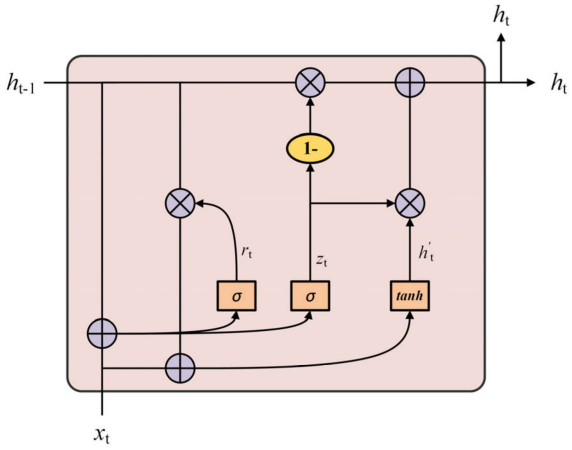

Figure 4. Internal structure of GRU. Where r_t, z_t and h'_t are reset gates, update gates, and candidate hidden states, respectively; and x_t, h_{t-1}, h_t are the current input content, hidden state at the previous moment and current hidden state, respectively. Purple represents mathematical operations between vectors, orange represents functional operations, and yellow represents inverse operations.

The GRU model controls the transmission of information through the gate control unit and the hidden state h_t. The reset gate r_t determines the irrelevant information that needs to be forgotten in the previous hidden state h_{t-1}; the update gate z_t determines the information that needs to be saved into the current hidden state h_t from the previous hidden state h_{t-1} and the candidate hidden state h_t'. The detailed calculation formulas are shown in Equations (2)–(5), where $W^{(r)}$, $U^{(r)}$, $W^{(z)}$, $U^{(z)}$, W and U are the weight matrices of the gate control unit; b, $b^{(r)}$ and $b^{(z)}$ are the bias coefficients; σ is the sigmoid activation function; $\odot$ is the Hadamard product; and x_t is the current input content.

$$r_t = \sigma(W^{(r)} \cdot x_t + U^{(r)} \cdot h_{t-1} + b^{(r)}) \tag{2}$$

$$z_t = \sigma(W^{(z)} \cdot x_t + U^{(z)} \cdot h_{t-1} + b^{(z)}) \tag{3}$$

$$h'_t = tanh(W \cdot x_t + U \cdot (r_t \cdot h_{t-1}) + b) \tag{4}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h'_t \tag{5}$$

The unidirectional GRU model solely considers the transmission of information from forward to back. By incorporating a reverse GRU network to construct a bidirectional recurrent neural network (BiGRU), the current moment's information in both directions can be simultaneously processed, enabling the model to better capture the correlated characteristics between forward and backward in political audit entities. The structure of BiGRU is shown in Figure 5.

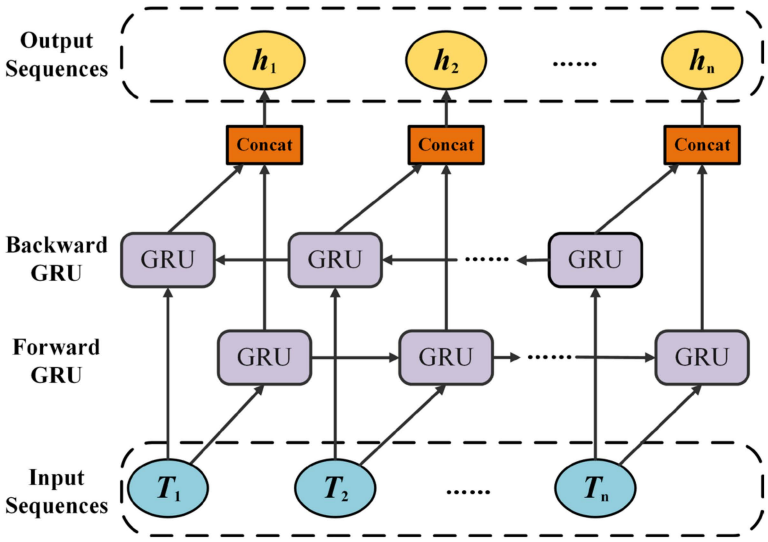

Figure 5. The structure of BiGRU.

The calculation method is shown in Equations (6)–(8), where $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ represent forward and backward sequence feature vectors, respectively.

$$\overrightarrow{h_t} = GRU(x_t; \overrightarrow{h_{t-1}}) \tag{6}$$

$$\overleftarrow{h_t} = GRU(x_t; \overleftarrow{h_{t-1}}) \tag{7}$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{8}$$

2.4. CRF Layer

The output vectors of the feature extraction layer are independent of each other, ignoring the dependency between labels, and the direct linking Softmax layer only outputs the word vectors corresponding to the highest-scoring labels, without considering the constraints. For example, the beginning of a sentence cannot be "I-"—the beginning of a named entity can only be "B-" but not "I-" or "O-", etc., which leads to problems such as annotation bias. CRF is used to learn the dependency relationship between labels and predict and output the reasonable label sequence with the highest global probability.

For CRF for the sequence $X = \{x_1, x_2, \ldots, x_n\}$ output from the feature extraction layer, let its corresponding label sequence be $Y = \{y_1, y_2, \ldots, y_n\}$, and compute the label sequence Y score $S(X,Y)$ by Equation (9):

$$S(X,Y) = \sum_{i=1}^{n} p_{i,y_i} + \sum_{i=0}^{n} A_{y_i,y_{i+1}} \tag{9}$$

where A is the transfer score matrix introduced by CRF; $A_{y_i,y_{i+1}}$ represents the score of label y_i transferred to label y_{i+1}; P is the label score matrix outputted from the feature extraction layer; P_{i,y_i} represents the score that the ith word is the jth label. $S(X,Y)$ is normalized by Softmax function normalized to obtain the probability of the tag sequence Y. Finally, the global prediction sequence score is computed by Viterbi dynamic optimization algorithm and the highest scoring sequence Y^* is used as the final output of the sequence

$X = \{x_1, x_2, \ldots, x_n\}$, which is computed as shown in Equation (10), where $S(X,Y)$ is the score of a path among all paths, Y_X is the set of all possible labeling sequences, and $\widetilde{Y}$ is the real labeling sequence.

$$Y^* = arg_{\widetilde{Y} \in Y_X}(\max S(X, \widetilde{Y})) \tag{10}$$

2.5. Loss Functions

The focus loss function [27] is based on the traditional cross-entropy loss function [28] and introduces the weight parameter α and the modulation factor γ to balance the weights of different categories of entities and difficult-to-classify entities. However, when constructing the audit dataset, there exists a certain level of randomness in the fundamental audit corpus, leading to an imbalanced distribution of entity categories within the dataset. Specifically, certain categories contain a significantly larger number of entities compared to others, the model is insufficient in identifying a few categories of entities. Data annotators lacking knowledge of the auditing domain often misjudge the types of auditing entities, resulting in labeled anomalous and noise entities in the dataset. These difficulties in differentiating entities can adversely affect model evaluation by diverting its attention towards these labeled anomalies and noise instances, leading to convergence issues and potential overfitting. Additionally, achieving optimal results requires simultaneous adjustment of both parameters in the focal loss function. Therefore, this study employs the GHM classification loss function during model training to adjust entity weights and reconcile entity losses, thereby enhancing audit entity recognition accuracy.

The GHM classification loss function characterizes the entity attributes by defining the gradient modulus g_i, which quantifies the difficulty of classifying entity i. The magnitude of g_i indicates the level of complexity in predicting learning outcomes. Equation (11) illustrates the computation, where p represents the model's predicted value for the entity label and p^* denotes the true value of the entity label.

$$g_i = \left| p - p^* \right| = \begin{cases} 1 - p & p^* = 0 \\ p & p^* = 1 \end{cases} \tag{11}$$

Through statistics on the input batch training data, $R(g_i)$ is used to represent the number of entities with similar gradient modulus g_i in the training data, and $l_\varepsilon(g_i)$ is used to represent the gradient interval length of gradient modulus g_i within a certain offset ε. Then, gradient density $GD(g_i)$ can be approximately expressed by Equation (12), which measures the number of entities with gradient modulus g_i within a certain gradient interval.

$$GD(g_i) = \frac{R(g_i)}{l_\varepsilon(g_i)} = \frac{\sum_{k=1}^{N} \delta_\varepsilon(g_k, g_i)}{l_\varepsilon(g_i)} \tag{12}$$

With regards to the focus loss function, a weight equilibrium coefficient β_{conf-i} is introduced for each entity during model training based on the gradient density $GD(g_i)$. This adjustment balances the contribution of each entity to the model gradient update and reconciles entity loss, thereby alleviating any impact from labeled abnormal entities or noise entities on the model gradient. The algorithm is presented in Equation (13), where N represents the number of samples.

$$\beta_{conf-i} = \frac{N}{GD(g_i)} \tag{13}$$

The classification loss function L_{GHM-C} introduces the weight equilibrium coefficient β_{conf-i} to balance the weight of entity categories and reconcile the loss of abnormally labeled entities and noise entities. On the one hand, the gradient density $GD(g_i)$ of the most numerous entity categories is large, and the weight equilibrium coefficient β_{conf-i} is introduced to balance the number of entity categories, adjust the weight of entities, and inhibit the number of entity categories so that the model focuses on the entity categories with

a small number, captures more feature information of the entity categories with a small number, and improves the recognition accuracy of the model. On the other hand, the fitting difficulty of the model to different entities is inconsistent in the training stage, and the error of abnormally labeled entities and noise entities is large and difficult to fit, and the modulus gradient value g_i is large. By introducing the weight equilibrium coefficient β_{conf-i} to multiply the original entity loss by the approximate reciprocal and normalize it, the entity loss is further reconciled, the influence of abnormally labeled entities and noise entities on the model gradient is reduced, the model converges quickly, and the problem of overfitting of the model gradient is alleviated. The calculation method of the classification loss function L_{GHM-C} is shown in Equation (14):

$$L_{GHM-C} = \frac{1}{N} \cdot \sum_{K=1}^{N} \beta_{conf-i} \cdot L_{CE}(p, p^*) = \sum_{K=1}^{N} \frac{L_{CE}(p, p^*)}{GD(g_K)} \quad (14)$$

3. Experiment and Results Analysis

3.1. Experimental Data and Pre-Processing

The self-constructed dataset in the field of audit presented in this paper is derived from the audit text data available on the official website of Xixiangtang People's Government. The initial dataset comprises a substantial amount of unstructured textual information, which has been subjected to cleansing using regular expressions. This process primarily involves eliminating stop words, special symbols, and irrelevant details. Additionally, based on expert recommendations within the audit domain, six distinct audit entities have been identified and annotated sequentially using the BIO annotation method. In this approach, B-X denotes the initiation of entity X, I-X represents its middle or concluding segment, while O signifies other characters unrelated to any specific entity. Consequently, we obtain a standardized dataset specifically tailored for auditing purposes (Audit). Table 1 illustrates the categories associated with these audit entities while Table 2 provides detailed data annotations.

Table 1. Audit domain entity type.

Entity Category	Entity Symbol	Entity	Entity Category
Name	Name	Specific title of audit project	绿色智能制造环保设备生产项目 (Green intelligent manufacturing environmental protection equipment production project)
Address	Loc	Audit project-specific address	西乡塘区秀厢大道36号(Xixiangtang District Xiuxiang Avenue 36)
Willfulness	Prp	Main Project Manager	盛都投资集团有限责任公司(Sheng Du Investment Group limited liability company)
Amount	Fee	Amount to be audited for projects	28.05 万元 (280,500 yuan)
Area	Sco	Audit project area	17.04 hm^2

The self-constructed audit domain dataset is supplemented with the public dataset Resume [14] as experimental data to further evaluate the model's generalization ability across different domains, encompassing eight entity types. In this study, the experimental dataset is partitioned into training set, validation set, and test set in an 8:1:1 ratio. Table 3 presents statistical information regarding the dataset.

Table 2. Sample data labelling.

Text Data	"BIO" Labelling
基(Ji)	O
本(Ben)	O
同(Tong)	O
意(Yi)	O
项(Xiang)	O
目(Mu)	O
位(Wei)	O
于(Yu)	O
西(Xi)	B-Loc
乡(Xiang)	I-Loc
塘(Tang)	I-Loc
区(Qu)	I-Loc

Table 3. Statistical information on DataSets.

Dataset	Category	Training Set	Validation Set	Test Set
Audit	Sentence	5000	625	625
	Entity	6807	753	843
Resume	Sentence	3821	463	477
	Entity	13,438	1497	1630

3.2. Experimental Environment

The experiments presented in this paper were conducted on a computer running Windows 10, equipped with an Intel(R) Xeon(R) Platinum (Intel Corporation, Santa Clara, CA, USA) 8358P @2.60 Ghz CPU, an NVIDIA RTX 3090 GPU (Nvidia Corporation, Santa Clara, CA, USA) with 24 GB video memory, and 64 GB RAM, using CUDA 11.8 and Python 3.8 with PyTorch 2.0.0 framework. During the model training process, we employed the Adam [29] optimizer to adjust parameters while introducing the Dropout [30] mechanism to randomly deactivate neurons for reducing overdependence between them and prevent overfitting; furthermore, a gradient trimming technique was utilized to mitigate the effects of gradient vanishing and explosion. Specific model parameter settings are shown in Table 4.

Table 4. Parameters of the model.

Relevant Parameter	Value
LSTM hidden layer dimension	128
Learning rate	5×10^{-5}
Optimizer	Adam
Batch_size	32
Dropout	0.5
Epochs	50
Clip	5

3.3. Evaluation Indicators

The experiment employs the widely adopted precision, recall, and comprehensive evaluation metric F1 in NER tasks to assess the model's performance. The calculation method is illustrated in Equations (15)–(17):

$$Precision = \frac{Correct\_num}{Predict\_num} \tag{15}$$

$$Recall = \frac{Correct\_num}{Tag\_num} \tag{16}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{17}$$

The variable *Correct_num* represents the count of accurately predicted audit entities, *Predict_num* denotes the count of identified audit entities, and *Tag_num* signifies the total count of labeled audit entities in the dataset.

3.4. Experimental Results and Analysis

3.4.1. Results and Analysis of Experimental Comparison

In order to assess the performance and generalization capability of the CW-CBGC model proposed in this study for identifying audit entities, we compared its accuracy, recall, and F1 value evaluation metrics with those of six other models on both our self-constructed audit dataset (Audit) and a publicly available dataset (Resume). The experimental results are presented in Table 5.

Table 5. Results of comparative experiments.

Experimental Model	Audit Datasets			Resume Datasets		
	Accuracy	Recall	F1	Accuracy	Recall	F1
Word2Vec-BiLSTM-CRF	87.06	94.14	90.46	93.66	93.31	93.48
SoftLexicon	91.25	92.17	91.71	95.30	95.77	95.53
BERT-BiLSTM-CRF	91.78	94.52	93.13	95.75	95.28	95.51
RoBERTa-BiLSTM-CRF	94.69	92.52	93.59	94.96	95.22	95.09
RoBERTa-BiGRU-CRF	95.76	92.79	94.25	95.13	95.25	95.20
RoBERTa-BiGRU-CRF-FL	96.05	94.19	95.11	95.43	95.27	95.35
CW-CBGC	97.38	96.56	97.23	96.35	96.18	96.26

Combined with Table 5, it can be inferred that: (1) the BERT-BiLSTM-CRF model, RoBERTa-BiLSTM-CRF model and this study's model for obtaining dynamic word vectors by the pre-trained model have significant improvement in accuracy and F1 value compared with the BiLSTM-CRF model for obtaining word embeddings using Word2Vec, indicating that the pre-trained model can better capture rich semantic features, effectively solve the polysemous word problem, and enhance the model's ability to recognize entities. (2) The RoBERTa model adopts dynamic masking to learn different semantic information and word-level semantic representations to improve the model entity recognition effect, and the F1 value is slightly improved compared with the BERT model. (3) The Focus Loss Function enhances the model's ability to recognize difficult-to-divide entities through the weighting factor α and the modulation factor γ on the basis of the cross-entropy loss function, and the F1 value of the model improves by 0.86%; however, there are labeled anomalous and noisy entities in the difficult-to-divide entities, and focusing on these types of entities will lead to overfitting of the model, and the GHM loss function balances the weights of the entity categories while reconciling the anomalous or loss of noise entities to prevent model overfitting and further improve the model recognition accuracy of audit entities. (4) The model in this study integrates Chinese glyph pinyin features and word-level information features into the character vector and is superior to the other comparison models in the three evaluation indexes, further proving that the model can effectively identify named entities in the field of government audit, and the model has better recognition performance, with an F1 value of 97.23%. (5) The model proposed in this paper outperforms other comparative models on the publicly available Resume dataset, achieving an impressive F1 score of 96.26%. This result demonstrates the model's remarkable generalization capability across different domains.

3.4.2. Results and Analysis of Experimental Ablation

Considering the advancements made in this study, the ablation experiment validated the efficacy of both the character-word feature fusion strategy and the incorporation of GHM loss function. The ablation experiment was conducted using RoBERTa-BiGRU-CRF

as the baseline model, denoting the introduction of GHM loss function as +GHM and incorporating character-word feature fusion strategy as +CW. The results of the ablation experiment are presented in Table 6.

Table 6. Results of ablation experiments.

Experimental Model	Accuracy	Recall	F1
Baseline models	95.76	92.79	94.25
+CW	97.06	95.20	96.12
+GHM	95.67	95.07	95.36
CW-CBGC	97.38	96.56	97.23

According to Table 6, the incorporation of character-word features fusion strategy in +CW significantly enhances its three major indicators compared to the baseline model. This substantiates that integrating word-level information features effectively improves the accuracy of entity recognition models, further emphasizing their significance for audit text entity recognition tasks. On the other hand, +GHM enhances the loss function compared to the baseline model and exhibits improved recall rate and F1 value. This demonstrates that introducing GHM loss function enhances the model's ability to identify challenging entities, enabling it to recognize more audit entities. However, there is a slight decrease in accuracy primarily influenced by confusion in entity boundary recognition. When both strategies are simultaneously introduced—incorporating character-word feature fusion strategy and GHM classification loss function—this study's model achieves an increase of 1.11% and 1.87%, respectively, in F1 value compared to models with individual enhancements mentioned earlier. These findings further validate that leveraging word-level information features comprehensively improves model accuracy for identifying difficult-to-separate entities when both strategies are employed together. In conclusion, this paper's proposed enhanced strategies have proven effective for improving the performance of our model.

3.4.3. Audit Entity Identification Effectiveness and Analysis

The recognition effect of different models and this model on six types of audit entities is illustrated in Figure 6. Based on the analysis of recognition results, this model demonstrates superior performance in recognizing area and amount entities, achieving an accuracy rate exceeding 95%. These two entity types constitute the largest portion of the dataset and are accompanied by explicit unit prompt words such as "公顷" (hectares) and "万元" (Ten thousand yuan). Consequently, the model effectively captures characteristic information associated with these entities, enabling accurate identification. However, the recognition performance for time and address entities is relatively poor due to limited sample size which hinders comprehensive learning of relevant characteristics, resulting in deviations during recognition. The recognition effectiveness for the remaining two entity types slightly decreases primarily due to factors like lengthy entity names and fuzzy boundary detection; nevertheless, their accuracy remains above 92%. Notably, this model outperforms both comparison models in terms of recognizing most entities, thereby demonstrating its efficacy in identifying various named entities within government audit domains.

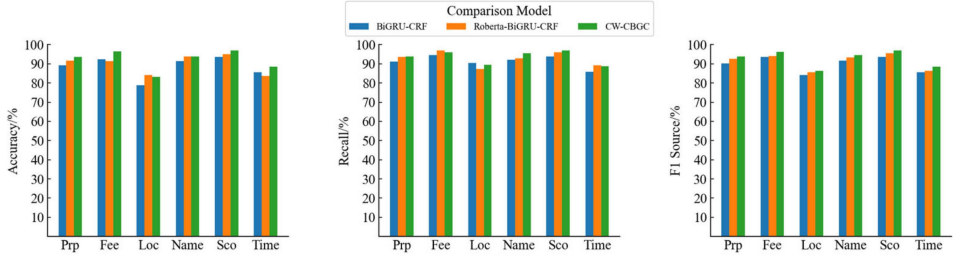

Figure 6. Comparison of entity recognition results of different models.

4. Conclusions

In response to the current scientific and technological landscape and audit requirements, this paper proposes a named entity recognition model for government audit text based on ChineseBERT and character-word fusion. The ChineseBERT pre-training model is utilized to extract character vectors that integrate Chinese glyph and pinyin features, thereby enhancing the grammatical and semantic aspects of the character vectors. These vectors are dynamically fused with word vectors extracted from sentence sequences by the BERT pre-training model. The BiGRU-CRF model is employed to learn text features in conjunction with contextual information, enabling decoding and outputting of globally optimal sequence labels. During the model training stage, entity loss is adjusted using the GHM loss function while simultaneously increasing the weight of minority entities to reconcile labeling errors associated with abnormal or noisy entities. This further enhances the accuracy of identifying audit entities within the model. The F1 values achieved by this model on both our self-built audit dataset (Audit) and a public dataset (Resume) are 97.23% and 96.26%, respectively, surpassing existing mainstream methods in performance evaluation metrics. However, it should be noticed that this model exhibits limited capability in accurately identifying audit entity boundaries as well as recognizing nested named entities effectively. Therefore, future research efforts will focus on developing techniques for precise identification of fuzzy entity boundaries and improving nested named entity recognition.

Author Contributions: Conceptualization, B.H. and Y.L.; methodology, B.H. and Y.L.; software, Y.L.; validation, B.H. and Y.L.; formal analysis, B.H., Y.L., S.P. and L.F.; investigation, S.P. and L.F.; resources, S.P. and L.F.; data curation, Y.L., S.P. and L.F.; writing—original draft preparation, B.H. and Y.L.; writing—review and editing, B.H. and Y.L.; visualization, Y.L.; supervision, Y.L.; project administration, B.H.; funding acquisition, B.H. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China under grant number 61962005.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data underlying this article will be shared upon reasonable request to the corresponding author.

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

References

1. Jiang, N. On State Audit Change and Development in the Age of Artificial Intelligence. *Financ. Account. Mon.* **2022**, *11*, 104–109.
2. Li, D.M.; Luo, S.S.; Zhang, X.P.; Xu, F. A Review of Research on Named Entity Recognition Methods. *J. Front. Comput. Sci. Technol.* **2022**, *16*, 1954–1968.
3. Grishman, R.; Sundheim, B.M. Message Understanding Conference-6: A brief history. In Proceedings of the 16th Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August 1996.
4. Zhang, J.; Shen, D.; Zhou, G.D.; Su, J.; Tan, C.L. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *J. Biomed. Inform.* **2004**, *37*, 411–422. [CrossRef] [PubMed]
5. Lafferty, J.; Mccallum, A.; Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001.
6. Sun, C.J.; Guan, Y.; Wang, X.L.; Lin, L. Rich features based conditional random fields for biological named entities recognition. *Comput. Biol. Med.* **2007**, *37*, 1327–1333. [CrossRef] [PubMed]
7. Chieu, H.L.; Ng, H.T. Named entity recognition with a maximum entropy approach. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, AB, Canada, 31 May–1 June 2003; pp. 160–163.
8. Liu, J. A Chinese Named Entity Recognition Algorithm Based on Improved Hidden Markov Models. *J. Taiyuan Norm. Univ. (Nat. Sci. Ed.)* **2009**, *8*, 80–83+90.
9. Zhang, W.Q. Deep Learning-Based Recognition of Named Entities in Zhuang Language. Master's Thesis, Guangxi Normal University, Guilin, China, 2022.

10. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
11. Hammerton, J. Named entity recognition with long short-term memory. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Edmonton, AB, Canada, 31 May–1 June 2003; pp. 172–175.
12. Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; Dyer, C. Neural architectures for named entity recognition. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, San Diego, CA, USA, 12–17 June 2016; pp. 260–270.
13. Lin, J.T.; Liu, E.D. Research on Named Entity Recognition Method of Metro On-Board Equipment Based on Multiheaded Self-Attention Mechanism and CNN-BiLSTM-CRF. *Comput. Intell. Neurosci.* **2022**, *2022*, 6374988. [CrossRef] [PubMed]
14. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 1554–1564.
15. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*; MIT Press: Lake Tahoe, CA, USA, 2013; pp. 3113–3119.
16. 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, Minneapolis, MN, USA, 2–7 June 2019.
17. Xu, L.; Li, J.H. Biomedical Named Entity Recognition Based on BERT and BiLSTM-CRF. *Comput. Eng. Sci.* **2021**, *43*, 1873–1879.
18. Wang, Z.H.; Huang, M.; Li, C.X.; Feng, J.L.; Liu, S.; Yang, G. Intelligent Recognition of Key Earthquake Emergency Chinese Information Based on the Optimized BERT-BiLSTM-CRF Algorithm. *Appl. Sci.* **2023**, *13*, 3024. [CrossRef]
19. Yang, C.L.; Sheng, L.; Wei, Z.C.; Wang, W. Chinese Named Entity Recognition of Epidemiological Investigation of Information on COVID-19 Based on BERT. *IEEE Access* **2022**, *10*, 104156–104168. [CrossRef]
20. Qian, T.Y.; Chen, Y.F.; Pang, B.W. Audit Text Named Entity Recognition Based on MacBERT and Adversarial Training. *Comput. Sci.* **2023**, *50*, 93–98.
21. Sun, Z.J.; Li, X.Y.; Sun, X.F.; Meng, Y.X.; Ao, X.; He, Q.; Wu, F.; Li, J.W. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 1–6 August 2021; pp. 2065–2075.
22. Li, S.T.; Zhang, M.M.; Liu, B. A Study on Named Entity Recognition in Kiwifruit Cultivation Domain by Incorporating Word Semantic Information. *Trans. Chin. Soc. Agric. Mach.* **2022**, *53*, 323–331.
23. Zhao, P.F.; Zhao, C.J.; Wu, H.R.; Wang, W. BERT-based multi-feature fusion for agricultural named entity recognition. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 112–118.
24. Ni, J.; Wang, Y.J.; Zhao, B. Named Entity Recognition for Automotive Production Equipment Failure Domain by Fusing Header Features and BERT. *J. Chin. Comput. Syst.* **2003**, 1–7. Available online: http://kns.cnki.net/kcms/detail/21.1106.tp.20230413.1826.031.html (accessed on 11 September 2023).
25. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
26. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.M.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
28. Janocha, K.; Czarnecki, W.M. On loss functions for deep neural networks in classification. *arXiv* **2016**, arXiv:1702.05659. [CrossRef]
29. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
30. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.