IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 September 2022, accepted 24 September 2022, date of publication 26 September 2022,
date of current version 6 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210119

APPLIED RESEARCH

Chinese Named Entity Recognition of Epidemiological Investigation of Information on COVID-19 Based on BERT

CHONGLUO YANG, LONG SHENG, ZHONGCHENG WEI, AND WEI WANG
School of Information and Electrical Engineering, Hebei University of Engineering, Handan, Hebei 056038, China
Hebei Key Laboratory of Security Protection Information Sensing and Processing, Hebei University of Engineering, Handan, Hebei 056038, China

Corresponding author: Long Sheng (shenglong@hebeu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62071071, and in part by the Science and Technology Research Project of Higher Education Institutions of Hebei Province under Grant QN2020193 and Grant ZD2020171.

ABSTRACT The named entity recognition based on the epidemiological investigation of information on COVID-19 can help analyze the source and route of transmission of the epidemic to control the spread of the epidemic better. Therefore, this paper proposes a Chinese named entity recognition model BERT-BiLSTM-IDCNN-ELU-CRF (BBIEC) based on the epidemiological investigation of information on COVID-19 of the BERT pre-training model. The model first processes the unlabeled epidemiological investigation of information on COVID-19 into the character-level corpus and annotates it with artificial entities according to the BIOES character-level labeling system and then uses the BERT pre-training model to obtain the word vector with position information; then, through the bidirectional long-short term memory neural network (BiLSTM) and the improved iterated dilated convolutional neural network (IDCNN) extract global context and local features from the generated word vectors and concatenate them serially; output all possible label sequences to the conditional random field (CRF); finally pass the condition random The airport decodes and generates the entity tag sequence. The experimental results show that the model is better than other traditional models in recognizing the entity of the epidemiological investigation of information on COVID-19.

INDEX TERMS Chinese named entity recognition, the epidemiological investigation of information on COVID-19, bidirectional encoder representations from transformer, bidirectional long-short term memory network, iterated dilated convolutional neural network, conditional random field.

I. INTRODUCTION

Since the outbreak in late 2019, COVID-19 has spread rapidly around the world, and it has become a global threat [1]. The epidemiological investigation of information COVID-19 released by the National and Provincial Health Commission contains vital information that plays an essential role in the control of the current outbreak and the prevention of future outbreaks in China, such as the location of the patient route, the time of movement and the means of a vehicle. Therefore, how to quickly and accurately allow computers to find the aforementioned critical information from the epidemiological investigation of information becomes an urgent problem.

The core task of named entity recognition(NER) is to extract entities from natural language text [2].NER is not only a core task for information extraction [3], but also essential for some natural language processing (NLP) tasks [4], such as machine translation [5], text understanding [6] and knowledge graph construction [7].Therefore building a named entity recognition model for the epidemiological investigation of information on COVID-19 can lay the foundation for the following entity relationship extraction.

Most of the named entity recognition in Chinese is based on three entities: person name, place name, and organization name. The instance types in the epidemiological investigation contain the above three entities and more special entities such as patient's transportation and number, and patient's body temperature, which need to be recognized as entities.

To solve the above problems, this paper proposes the BBIEC model by collecting the epidemiological investigation of information on COVID-19 and constructing the dataset with it. It concluded that the BBIEC model could better

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

understand entity boundary information than the traditional named entity model. Compared with a single neural network, the dual neural network of BiLSTM and IDCN can better identify global contextual and local features and improve the metrics of named entity recognition. The BBIEC model proposed in this paper can provide specific ideas, references, and solutions for subsequent research on other COVID-19-related fields.

Therefore, we developed a NER model called BBIEC based on the characteristics of the epidemiological investigation of information on COVID-19. The main contributions of this work are summarized as follows.

(1) Based on the problem that the features of text entities in the epidemiological investigation of information on COVID-19 are different from the general named entity recognition, a BERT pre-training model is used in the epidemiological investigation of information on COVID-19 to dynamically generate the COVID-19 word vectors against COVID-19 words based on the input, which is more suitable for prediction of the COVID-19 corpus.

(2) Based on the problem that the text in the epidemiological investigation of information on COVID-19 is too long to extract features effectively, the text information is obtained from both global and local aspects by BiLSTM and IDCN models, and the recognition effects of both kinds of information are compared in serial mode and parallel mode of the model to select the feature fusion method with better recognition effects.

(3) Based on the problems of high text similarity and neuron necrosis of Relu activation function in the epidemiological investigation of information on COVID-19, we use the elu activation function instead of the ReLu activation function to activate some dormant neurons, which solves the above two problems and further improves the feature extraction ability of the model.

The experimental results on the COVID corpus show that the recall and F1 values of the model surpass most of the models, reaching 0.9561 and 0.9521.

The rest of this paper is structured as follows. We review the development of NER models in Section 2. Then, we develop the BBIEC model in Section 3 and conduct analytical experiments on text classification in Section 4. Finally, concluding remarks are presented in Section 5.

II. RELATED WORKS

There are three main approaches for Chinese named entity recognition: rule-based approach, statistical machine learning-based approach, and deep learning-based approach.

A. RULE-BASED APPROACH

Rule-based methods construct rules or dictionaries by hand, with rule or dictionary and string matching as the primary means. This method requires a high level of personnel to construct rules or dictionaries, which is not only time-consuming and laborious, but also prone to errors due to subjective factors and requires the construction of different rules or dictionaries for different domains, and has poor portability.

B. STATISTICAL MACHINE LEARNING-BASED APPROACH

The two main approaches based on statistical machine learning are the classification model approach and the sequence model approach. For example, Ju *et al.* [8] used Support Vector Machine (SVM) to implement named entity recognition task for biomedical texts; Ekbal *et al.* [9] used the maximum entropy (ME) framework to construct many classifiers based on different representations of a set of features, applied to named entity recognition for medical use in the biomedical field; Niu *et al.* [10] discussed the comparative experiments of HMM and MEM in the same environment, analyzed the characteristics of the two models and their applications in named entity recognition, and pointed out the advantages and disadvantages of the two models; Hu *et al.* [11] investigated CRF-based recognition of Chinese named entities by implementing three main named entities: person, location, and organization recognition, at the word and character levels, respectively. The performance of two-level models is experimentally compared.

C. DEEP LEARNING-BASED APPROACH

In recent years, with the continuous development of hardware devices, deep learning methods with high computing performance requirements have become mainstream. In NER tasks, deep learning-based methods perform better than rule-based or dictionary-based and statistical machine-learning-based methods, which do not require the setting of artificial features, and neural networks can automatically learn features from the dataset, so various neural network-based models are applied to NER tasks. For example, Huang *et al.* [12] used a BiLSTM-CRF model using a bidirectional long- and short-term neural network model combined with CRF layers for the named entity recognition task. Strubell *et al.* [13] proposed a CNN with better large context and structured prediction than traditional CNNs called IDCNN, which is a faster NER for replacing the BiLSTM scheme; Yang *et al.* [14] proposed an improved transformer-BiLSTM-CRF model applied to the text domain of substation knowledge to achieve entity recognition of substation knowledge more effectively; An *et al.* [15]proposed a bidirectional long term memory conditional random domain model (MUSA-biLSTM-CRF) based on multi-headed self-attentiveness. The model is able to capture the weighting relationship between Chinese characters and multi-level semantic feature information more effectively by introducing multi-headed self-attentiveness and incorporating a medical lexicon; Jiang *et al.* [16]proposed a BiLSTM-IDCNN-CRF model based on word embedding, combining the BiLSTM network and IDCNN networks to obtain features with different granularity.

The research focus on named entity recognition is quite different between Chinese and English. Chinese does not have spaces to segment different words, and character and word models have a greater effect on entity recognition in Chinese

contexts. Li *et al.* [17] concluded that word-based models consistently outperformed word-based models by comparing the effects of words and characters on language modeling, machine translation, sentence matching, and text classification models. Liu *et al.* [18] concluded that word vector-based entity recognition is more accurate by comparing the effect of word vector-based and word vector-based entity recognition. However, at the same time, word vectors cannot solve the problem of polysemy. Therefore, Google's Devlin *et al.* [19] proposed a BERT pre-trained language model, which is a deep learning model based on Transformer bidirectional encoder, which pre-trains a bidirectional language model through a large text corpus to capture the bidirectional relationships in utterances and generate contextually relevant word vectors to effectively solve the above problem. The model combined with BERT has achieved excellent results in named entity recognition in different domains. For example, Li *et al.* [20] proposed an identification method combining the Bert model with BiLstm-CRF, which was applied to a literature study on thyroid secretion summarization to obtain a high identification rate; Tang *et al.* [21] proposed a BERT-LCRF model for named entity recognition using a pre-trained language model BERT for feature extraction of clock domain text, and then a linear chain conditional random field (Linear-CRF) method for the NER task; Gao *et al.* [22] added an attention mechanism to the BERT-BiLSTM-CRF model to increase the local extraction capability of the model; Gan *et al.* [23] used a Chinese named entity recognition method based on the BERT-Transformer-BiLSTM-CRF model to address the large number of pronouns and polysemous words; Li *et al.* [24] proposed a BERT-IDCNN-CRF based model for the problem of too many BERT training parameters and too long training time, which is faster and more responsive; Wu *et al.* [25] proposed an improved NER model that uses BERT as a pre-training layer and a BiLSTM network as a coding layer to improve the performance of NER in railroad construction using the feature extraction capability of CNNs; Chang *et al.* [26] proposed a Bert-based named entity recognition method and built a BERT-BiLSTM-IDCNN-CRF model. The trained word vectors were then fed into a BiLSTM and an IDCNN for feature extraction.

At present, there are still some problems in the field of epidemiological investigation of information on COVID-19: the lines of epidemiological investigation of information on COVID-19 published by national and local health committees are not the same; the flow transfer information needs to redefine new entities; there is no mature and large amount of annotated data for epidemiological investigation of information on COVID-19 for the time being, and the manual annotation cost is high; the named entity models in other fields are difficult to be perfectly applied to the field of epidemiological investigation of information on COVID-19. To this end, this paper proposes the BBIEC model by embedding the improved IDCNN into the BiLSTM-CRF network and coupling it with the BERT pre-training model.

III. PROPOSED METHOD

This paper mentions that the core of Chinese named entity recognition based on BERT epidemiological investigation of information on COVID-19 proposed is the construction of the BBIEC neural network model, whose overall structure is shown in Figure 1. Moreover, its function can be divided into three layers: the BERT pre-training layer, BiLSTM-IDCNN-ELU neural network layer, and the CRF inference layer.

To represent our model more clearly and intuitively, the detailed procedures of our proposed BBIEC model are given in visualized algorithm format, as shown in Table 1.

A. BERT PRE-TRAINING LAYER

1) CLASSIFICATION LEVEL OF THE CORPUS

Before inputting the original corpus into the BERT pre-training model, we need to select the division level of the corpus, which is divided into two levels: word level and character level.

For example, "G96次 高 铁" means a high-speed train with the number G96, which contains two entities in six characters, "G96次" is a NUM entity, and "高铁" is a VEH entity, according to the word level, it can only mark the entity type, but cannot determine the boundary of recognition. In contrast, according to the character level, it can not only mark the entity type but also add the position mark for the character; for example, the entity mark of "G96次" is "G, B-NUM," "9, I-NUM", "6, I-NUM", "次, E-NUM", i.e., the entity marker mentioned in Section 1.1 BIOES tagging method, the inclusion of location tagging can distinguish the boundaries of entities more effectively; meanwhile the effect of entity recognition based on character vectors and word vectors concludes that the effect of entity recognition based on character vectors is more accurate [18]. Therefore, the character-level corpus division is chosen in this paper.

Let the sequence of the input model be $X = \{X_1, X_2, \ldots, X_n\}$, where X_i is the corpus processed into a single Chinese character.

2) BERT EMBEDDING LAYER

In the COVID-19 corpus, the same character in different positions of the sequence can represent different meanings, that is, the problem of character polysemy. For example, "拨打120急救热线", which means calling 120 emergency hotline, "患者被120送至医院治疗", which means the patient was taken to hospital by 120 ambulances for treatment, where "120" is in different positions, the meanings are entirely different, the former represents the contact information of the emergency. The former represents the contact information of the emergency organization, and the latter represents the emergency organization itself.

To better distinguish the entities of this type of data, this paper uses the idea of location coding to add different codes for the same characters in different locations to solve the problem of character polysemy. Under the above idea, the BERT embedding layer is chosen to add position

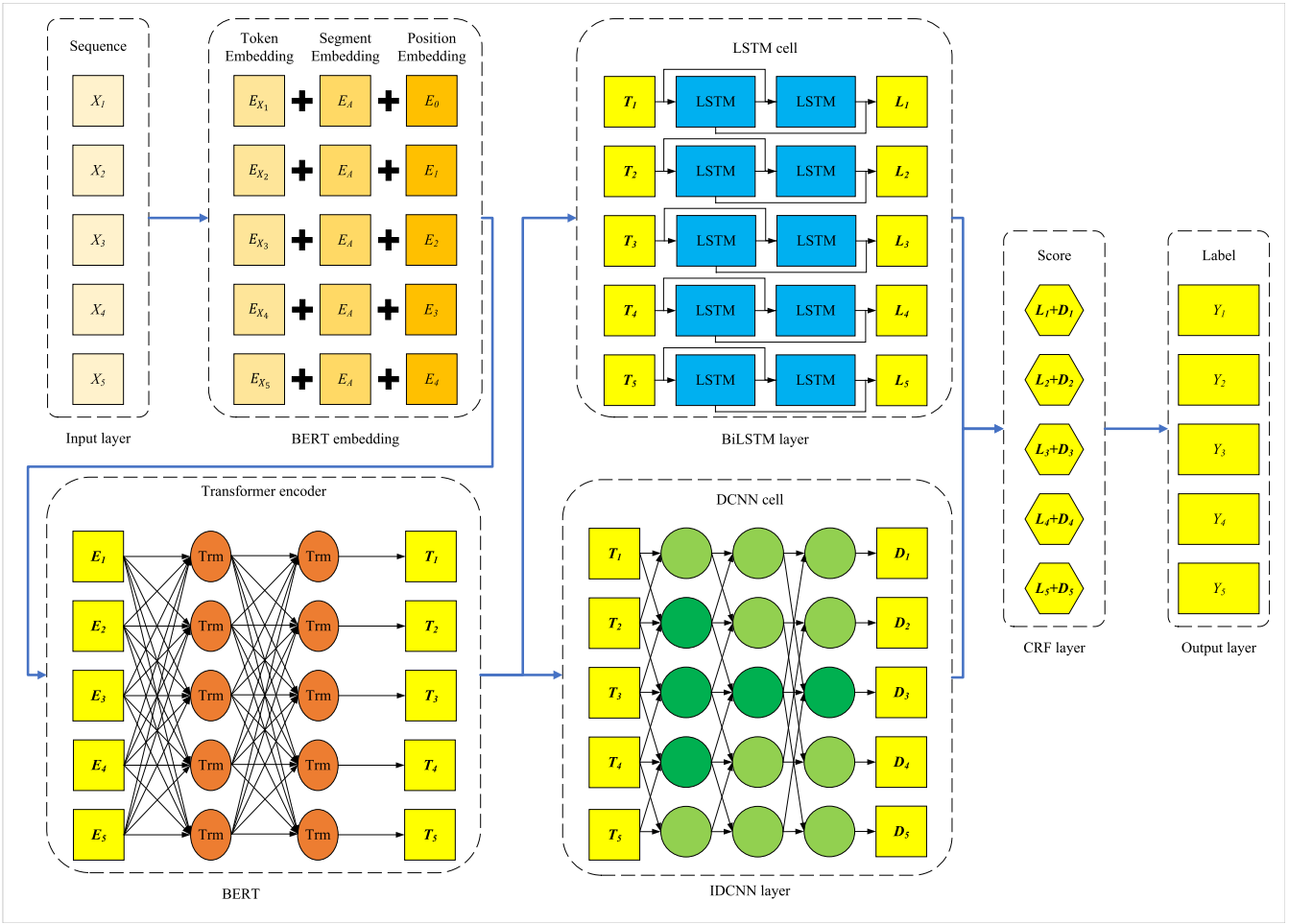


FIGURE 1. BBIEC model structure.

information to the characters, and its structure consists of Token Embeddings, Segment Embeddings, and Position Embeddings, respectively. Among them, the Position Embeddings layer uses the sine and cosine functions to encode the position information of characters into a feature matrix. The same character can express different semantics under the feature matrix of different positions, solving the character polysemy problem.

The character-level corpus is input into the BERT embedding layer, and the sequences are processed and summed by the three embedding layers, and then are converted into embedding vectors $\boldsymbol{E} = \{\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_n\}$ and output to the Trm layer of BERT, which represents a Transformer [27] encoder part. the BERT embedding layer is shown in Figure 2.

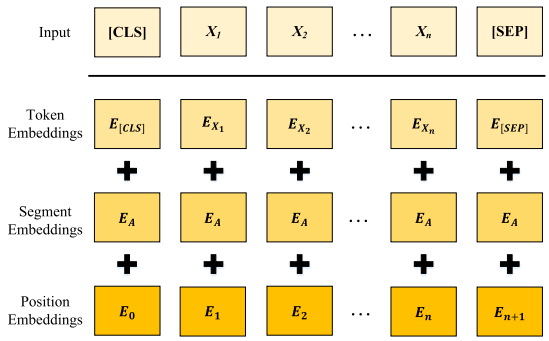


FIGURE 2. BERT-embedding.

3) BERT OVERALL STRUCTURE

The BERT used in this paper consists of 12 Trm layers. The input of the BERT is the embedding vector $\boldsymbol{E} = \{\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_n\}$ output from the embedding layer. The output is $\boldsymbol{T} = \{\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_n\}$. each of the Trm layers consists of a multi-headed self-attentive mechanism layer, a residual connection and normalization layer, and a fully connected feedforward neural network layer, respectively. After 12 Trm layers are encoded, they are normalized by the Softmax layer to the vector $\boldsymbol{T}$ output. The structure of BERT is shown in Figure 3.

B. BiLSTM-IDCNN-ELU NEURAL NETWORK LAYERS

1) BiLSTM NEURAL NETWORK LAYERS

A recurrent Neural Network [28] can take the output of the previous time slice as the input of the next time slice so that

TABLE 1. Algorithm description.

Algorithm:BBIEC model. The Chinese named entity recognition of epidemiological investigation of information on COVID-19 on feature fusion. **Input:**$X = \{X_1, X_2, ..., X_n\}$, a set of epidemiological investigation of information on COVID-19. **Output:**$Y = \{Y_1, Y_2, ..., Y_n\}$, a label set of epidemiological investigation of information on COVID-19.
Methods:(The specific processing steps of the algorithm are introduced as follows) **Begin** 1.$E = BERT\ embedding(X)$./*Transformation of characters into character vectors through three embedding layers in BERT embedding.*/ 2.$T = BERT(E)$./*Pre-train the model with the Trm layer in BERT.*/ 3.$L = BiLSTM(T)$./*Deploy BiLSTM neural network for extracting global feature information.*/ 4.$D = IDCNN(T)$./*Deploy three-layer IDCN neural network to extract global information, and use the elu activation function to increase the feature extraction capability of the model.*/ 5.$Y = CRF(L + D)$./*The features extracted by the two neural networks are fused, and then the CRF is used to verify, classify and output entity labels.*/ **End**

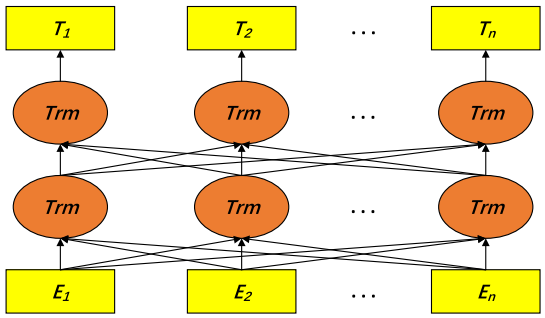


FIGURE 3. BERT structure.

it can handle temporal sequences well. The epidemiological investigation information of the COVID-19 corpus is a temporal sequence. However, its corpus length is long, and some sentences are more than 200 characters long, which is more likely to generate a long-distance dependency problem using RNN. This will lead to the model unable to learn global contextual features effectively, resulting in poor entity recognition.

Long Short-Term Memory Neural Networks [29] is a special kind of RNN. Its neurons have three parts: input gate, forgetting gate, and output gate. The input gate controls which information is input, the forgetting gate controls which information is forgotten in the neuron, and the output gate controls which information is output. The specific formulas of the three gating units are shown in Eqs.

$$\text{input}_i = \sigma(\boldsymbol{W}_{\text{input}} \cdot [\boldsymbol{T}_i, \boldsymbol{L}_{i-1}, \boldsymbol{C}_{i-1}] + \boldsymbol{b}_{\text{input}}). \quad (1)$$

$$\text{forget}_i = \sigma(\boldsymbol{W}_{\text{forget}} \cdot [\boldsymbol{T}_i, \boldsymbol{L}_{i-1}, \boldsymbol{C}_{i-1}] + \boldsymbol{b}_{\text{forget}}). \quad (2)$$

$$\text{output}_i = \sigma(\boldsymbol{W}_{\text{output}} \cdot [\boldsymbol{T}_i, \boldsymbol{L}_{i-1}, \boldsymbol{C}_{i-1}] + \boldsymbol{b}_{\text{output}}). \quad (3)$$

where: input_i,forget_i,$\text{output}_i$are the states of the input gate, forget gate, and output gate, respectively; σ is the Sigmoid activation function; $\boldsymbol{W}$ is the weight matrix of the three gating units; $\boldsymbol{b}$ is the bias term; $\boldsymbol{T}$ is the input vector at the time i; $\boldsymbol{T}_{i-1}$ is the hidden layer state of the LSTM unit at the previous moment; $\boldsymbol{C}_{i-1}$ is the memory information in the LSTM unit at the previous moment. The memory information is updated as shown in Equation (4).

$$\boldsymbol{L}_{i-1} = \text{output}_i \cdot \tanh(\boldsymbol{C}_i). \quad (4)$$

LSTM neurons pass memory information and hidden layer state information between them. Its control of feature circulation and loss through the gate mentioned above effectively solves the long-distance dependency problem, so LSTM is chosen as the base neural network in this paper.

However, the one-way LSTM network can only learn the historical features and not the future features, so this paper adopts the BiLSTM network, which is a bidirectional LSTM, to avoid the problem of losing the historical features due to the long sentences by splicing the features in both directions, to better learn the contextual features and solve the long-distance dependency problem.

The BiLSTM neural network layer accepts the output of the BERT pre-training layer, and the output of the BiLSTM neural network is obtained as $\boldsymbol{L} = \{\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_n\}$ and input to the next neural network. The structure diagram of BiLSTM is shown in Figure 4.

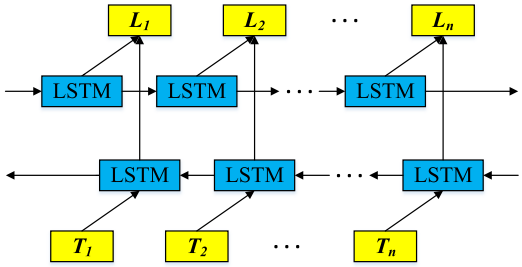


FIGURE 4. BiLSTM structure.

2) IDCNN-ELU NEURAL NETWORK LAYERS

In the epidemiological investigation of information on COVID-19 corpus with long sentence length, although the BiSLTM model can extract the global contextual features well, it may ignore the essential local features in the sentence. Convolutional Neural Networks can extract local features through convolutional operations.

Dilated Convolutional Neural Networks (DCNN) [30] is a particular type of CNN with a convolution kernel that adds a dilation distance d, which increases the perceptual field and

can learn local features better. The internal implementation of the convolution operation is shown in Equation (5).

$$\widetilde{\boldsymbol{D}}_{i-1} = \sum_{t=-l}^{l} \boldsymbol{K}_t \cdot \boldsymbol{T}_{i+t\cdot d} + \boldsymbol{b}. \tag{5}$$

where: $\boldsymbol{K}$ is the weight matrix of the convolution kernel, l is the window length of the convolution kernel, $\boldsymbol{T}$ is the vector of BERT inputs, and b is the bias term.the structure of DCNN is shown in Figure 5.

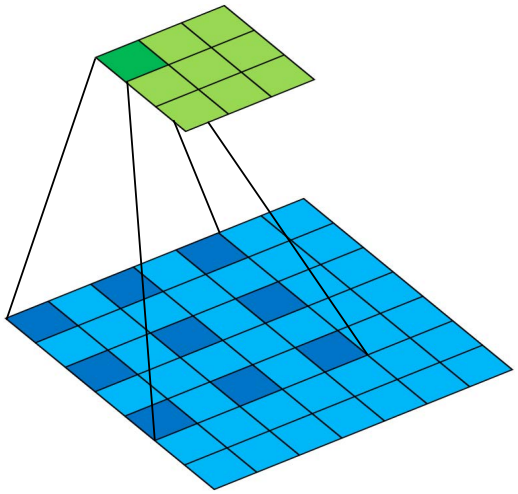


FIGURE 5. DCNN structure.

Iterated Dilated Convolutional Neural Networks (IDCNN) are composed of multiple layers of DCNNs with different dilation widths, followed by the calculation of the feature vector of the current dilation convolution using the previous layer of dilation convolution. The calculation is shown in Equation (6).

$$\widetilde{\boldsymbol{D}}_i = \sigma(\boldsymbol{H}_{i-1} \cdot \widetilde{\boldsymbol{D}}_{i-1}). \tag{6}$$

where: elu (Exponential Linear Unit) activation function, $\boldsymbol{H}_i$ is the layer i expanded convolutional neural network, and $\widetilde{\boldsymbol{D}}_i$ is the feature vector learned by the layer i convolutional network.

The elu activation function [31] is an improved activation function for the negative part of relu [32]. as shown in Equation (7).

$$\text{elu}(x) = \begin{cases} x, & x > 0 \\ a(e^x - 1), & other\ values \end{cases} \tag{7}$$

Elu activation function uses an exponential calculation-like output for $x < 0$, which solves the problem that some neurons in the relu activation function cannot be activated. The images of the relu activation function and the elu activation function are shown in Figure 6.

IDCNN further strengthens the acquisition of local features relative to DCNN. IDCNN model addresses the problem that BiLSTM would lack local features, and the essential features in the corpus are extracted by convolution operation

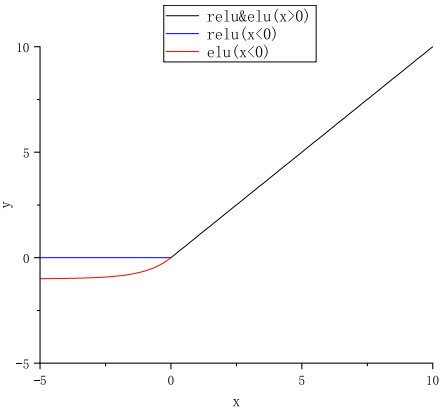


FIGURE 6. Relu and elu.

to ensure the integrity of local features, So this paper joins BILSTM model and IDCNN model in parallel. The IDCNN neural network layer also accepts the output of the BERT pre-training layer, and the output of the IDCNN neural network is obtained as $D = \{D_1, D_2, \ldots, D_n\}$ and input to the next neural network. The structure diagram of IDCNN is shown in Figure 7.

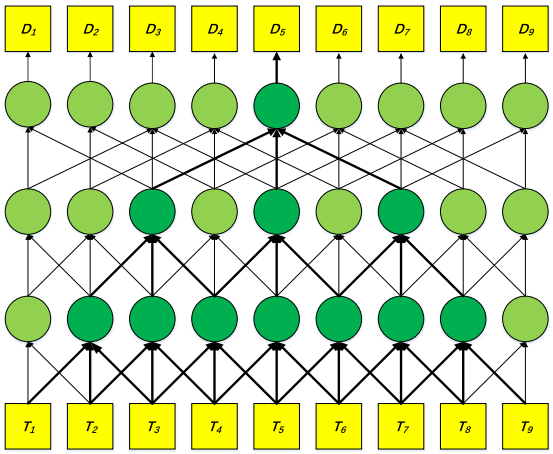


FIGURE 7. IDCNN structure.

C. CRF LAYERS

The BERT-BiLSTM-IDCNN model has been trained to output specific scores for each label, and the largest of these scores is selected as the output label. However, the final score may not be precisely correct. There may be label location errors, e.g., ''Handan, Hebei Province,'' as a location entity. The ''river'' and ''city'' ask for the beginning and the end of the location, respectively, and the rest of the characters are in the middle of the location. If a different wrong location ends at the beginning and middle of the location, the predicted label does not conform to the BIOES annotation system.

CRF can help to modify the above error by introducing constraints directly on the labels. Given that the sequence of the input model is $X = \{X_1, X_2, \ldots, X_n\}$, its corresponding

predicted sequence is $Y = \{Y_1, Y_2, \ldots, Y_n\}$. The score of the label sequence is shown in Equation (8).

$$\text{score}(X, Y) = \sum_{i=-l}^{n} \widetilde{W}_{Y_i, Y_{i+1}} + P_{i+1, Y_{i+1}}. \tag{8}$$

where:$\widetilde{W}$ is the transfer score matrix, $\widetilde{W}_{Y_i, Y_{i+1}}$ is the transfer score of Y_i transferred to Y_{i+1}; P is the score matrix of the upper output, $P_{i+1, Y_{i+1}}$ is the score of the label Y_{i+1} corresponding to the $i + 1$th word of the output sequence. The probability of tag sequence Y generation is shown in Equation (9).

$$P(X|Y) = \frac{e^{\text{score}(X,Y)}}{\sum_{\tilde{Y} \in Y_X} e^{\text{score}(X,\tilde{Y})}}. \tag{9}$$

where: $\tilde{Y}$ denotes all possible labeled sequences. Finally, in the decoding stage, the optimal path is solved using the Vibit algorithm. The Vibit algorithm is calculated as shown in Equation (10).

$$Y^* = \text{argmax}(ln(P(Y|X))). \tag{10}$$

After the above process, CRF can effectively check the labels of columns and improve recognition accuracy. Therefore, CRF is used as an inference layer to avoid label position errors. As a result, the sequence X of the input model is predicted by the BBIEC model to obtain the sequence Y^*.

IV. EXPERIMENT DESIGN

A. DATASET

1) DATASET LABELING

In this paper, the dataset was extracted from the epidemiological investigation of information on COVID-19 published on the structured data of the trajectories of patients diagnosed with the epidemiological investigation of information on COVID-19 [33] released by the national and local health care commissions, major news portals, and Beijing Advanced Innovation Center for Big Data and Brain Computing, of which about 200,000 words of raw data were taken and manually annotated.

The design principle of the entities in the named entity recognition task is to be able to represent the key information in the original text effectively. The epidemiological investigation of information on COVID-19 of named entity identification task is to control the spread of the epidemic, and The spread of the epidemic is mainly caused by direct or indirect human-to-human contact [34], and to control the spread of the epidemic, we need to know the movement trajectory of patients, so the entity design of this paper is centered on the COVID-19 patients, and because the original trajectory text format of the literature [33] is roughly ''someone did something at a certain time and place'', also contains some modifying ingredients(body temperature, transportation, etc.), so based on the above analysis, we abstractly designed the definitions of nine entities with COVID-19 patients as the core entity, including ''Patient name (PER),'' abbreviated as ''Patient''; ''Location of patient's route of residence (LOC),'' abbreviated as ''Location''; ''Organization (ORG''; ''Vehicle used by the patient (VEH),'' abbreviated as ''Vehicle''; ''Telephone number (TEL)''; ''Patient temperature (TEMP)''; ''Number of vehicles (NUM) ''; ''Time (TIME)'' and ''Date (DATE).'' A closed-source named entity recognition dataset for the epidemiological investigation of information on COVID-19 is constructed based on the above nine entities [35]. The specific label meanings of the dataset and the examples are shown in Table 2.

TABLE 2. Label meaning and examples.

Label	Example	Example explanation 87
PER	"李某"	Someone with the last name Lee
LOC	"河北省邯郸市"	Handan City, Hebei Province
ORG	"邯郸市卫健委"	Handan Health Commission
VEH	"私家车"	Private Car
TEL	"137****9845"	Telephone number in China
TEMP	"36.5°C"	36.5 degrees Celsius
NUM	"冀D12345"	Handan City motor vehicle license plate number
TIME	"8时23分"	8:23 am
DATE	"7月24日"	July 24th

2) DATA PRE-PROCESSING

For data that were found manually added from the national and local health care commissions and major news portals, they were changed to a similar expression to that of literature [33]; counting the lengths of all the original data in which 82% of the sentences are in the interval [150,250], and removing those with lengths less than 50 and greater than 500; using the format of the classic Chinese named entity recognition dataset: People's Daily dataset [36] as the construction standard, the original text of the new crown is processed into the form of ''character LPH (label placeholder)'', which is called a single column of unlabeled entity data, where LPH is the label placeholder, and there is a space between the character and There is a space sign between the character and the LPH. After slicing the original data into the above format, the entity information present in the text is found according to the design rules declared in IV.A.1, and the LPH is replaced with the entity corresponding to the character. The pre-processing of the dataset is completed.

3) DATASET COMPOSITION

The original data of the epidemiological investigation of information on COVID-19 was collated and filtered to obtain 1026 data by the preprocessing operation above. The total was divided into a training set, a validation set and a test set according to the ratio of 8:1:1 characters. As shown in Figure 8, it represents the number and percentage of each entity in the dataset.

TABLE 3. Example of a sentence label.

Characters									
1	月	2	日	，	自	驾	车	从	邯
郸	到	邢	台	。					
Labels									
B-DATE	I-DATE	I-DATE	E-DATE	O	O	B-VEH	E-VEH	O	B-LOC
E-LOC	O	B-LOC	E-LOC	O					

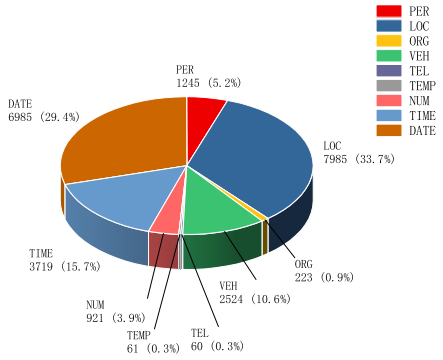


FIGURE 8. Dataset composition.

B. LABELING RULES

In the named entity recognition task, there are two main methods for character-level entity labeling: the BIO tagging method and the BIOES tagging method, respectively [37]. In this paper, we use the BIOES tagging method, which differs from the BIO tagging method in that there are only three tag types, and its method is to introduce an end tag for each entity with a character length greater than one so that the boundary of the entity can be better distinguished. Where B-Entity denotes the first character in an entity, I-Entity denotes the character in the middle of an entity, E-Entity denotes the last character in an entity, and S-Entity denotes an entity composed of a single character, and O denotes a non-identified entity. Examples of sentence labels are shown in Table 3.

C. EVALUATION INDICATORS

In this paper, the precision rate P(Precision), recall rate R (Recall), and F_1 score are used as evaluation indexes. The calculation formula is shown in Equations (11)-(13).

$$P = \frac{T_P}{T_P + F_P} \times 100\%. \tag{11}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\%. \tag{12}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\%. \tag{13}$$

where:T_P is the number of correctly identified entities, F_P is the number of incorrectly identified entities, and F_N is the number of unidentified entities.

D. EXPERIMENTAL ENVIRONMENT

The specific environment of all experiments of the Chinese named entity identification study for the flow of information of patients throughout the COVID-19 epidemic is shown in Table 4.

TABLE 4. Experimental environment.

Environment	Configuration
OS	Window10
CPU	AMD Ryzen 7 4800H
GPU	NVIDIA GeForce GTX 1660 Ti
RAM	16GB
TensorFlow	1.14
Python	3.6

E. EXPERIMENTAL PARAMETERS

In the epidemiological investigation information of the COVID-19 dataset, the longest sentence consists of 195 characters, and the average sentence length is about 170 characters, so the maximum sentence length (max_len)is set to 200 characters. After the hyperparameter ablation experiment, other relevant settings are shown in Table 5.

TABLE 5. Experimental parameters.

Parameters	Parameter Value
Number of transformer layers	12
max_len	200
BERT-embedding-size	768
batch_size	8
Optimizer	Adam
LSTM_dim	128
learning rate	1e-4
Number of convolution blocks	3
Convolution kernel filter	1,1,2
dropout	0.4
clip	5

V. RESULTS AND ANALYSIS

A. COMPARATIVE EXPERIMENTAL DESIGN

The following sets of comparison experiments are set up in the Chinese named entity recognition experiments for the epidemiological investigation of information on COVID-19.

(1) Word2Vec-BiLSTM-CRF [12]: word2Vec generates character vectors, BiLSTM neural network extracts semantic features, and the CRF inference layer classifies different entities.

(2) Word2Vec-IDCN-CRF [13]: word2Vec generates character vectors, IDCNN neural network extracts semantic features, and the CRF inference layer classifies different entities.

(3) BERT-BiLSTM-CRF [20]: The BERT pre-trained model generates character vectors, the BiLSTM neural network extracts semantic features, and the CRF inference layer classifies different entities.

(4) BERT-IDCNN-CRF [24]: The BERT pre-trained model generates character vectors, the IDCNN neural network extracts semantic features, and the CRF inference layer classifies different entities.

(5) BERT-BiLSTM-Attention-CRF [22]: The BERT pre-trained model generates character vectors, the IDCNN neural network extracts semantic features, the Attention mechanism layer reinforces the extracted semantic features, and the CRF inference layer classifies different entities.

(6) BERT-Transformer-BiLSTM-CRF [23]: The BERT pre-trained model generates character vectors, the Transformer encoding area constructs contextual long-range semantic features of text, the BiLSTM neural network extracts semantic features, and the CRF inference layer classifies different entities.

(7) BERT-BiLSTM-IDCNN-CRF (Serial) [16]: The BERT pre-trained model generates character vectors, the BiLSTM and IDCNN neural networks extract semantic features and then fuse the features serially, and the CRF inference layer classifies different entities.

(8) BERT-BiLSTM-IDCNN-CRF (Parallel) [25]: The BERT pre-training model generates character vectors, the BiLSTM and IDCNN neural networks extract semantic features and then fuse the features in parallel, and the CRF inference layer classifies different entities.

(9) BBIEC: BERT pre-trained model generates character vectors, BiLSTM, and improved IDCNN (replacing the relu activation function with elu activation function in IDCNN) neural network extracts semantic features and later fuses the features in parallel, and CRF inference layer classifies different entities.

B. EXPERIMENTAL PROCEDURE

The epidemiological investigation of information on COVID-19 origin data is collected and preprocessed (including data de-duplication, screening of unqualified data, etc.) to construct the unannotated new coronavirus text dataset, and then the unlabeled dataset is divided to segment the text into the character-level corpus. And each character is an entity labeled according to the rules in section IV.A.1) (Dataset labeling), with 37 labeling types, using spaces between characters and labels for segmentation and line breaks for each patient's information to construct the labeled epidemiological investigation of information on COVID-19 dataset. This dataset was then fed into the four comparison models and the BBIEC model for prediction. The experimental flow is shown in Figure 9.

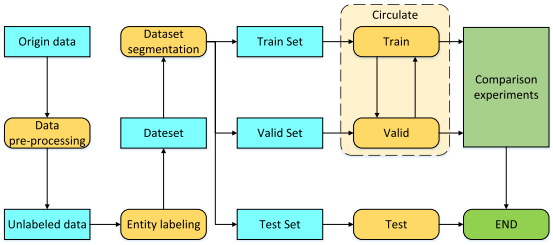


FIGURE 9. Experimental procedure.

C. COMPARISON EXPERIMENTS RESULTS

According to the process described in Section 4.4 (Experimental Procedure), comparative experiments were conducted on three evaluation metrics: accuracy, recall, and F1 value. The results of the experiments are shown in Table 6.

TABLE 6. Comparison of experimental results.

Model	P	R	F_1
Word2Vec-IDCNN-CRF	0.8579	0.7548	0.7745
Word2Vec-BiLSTM-CRF	0.8674	0.8291	0.8452
BERT-IDCNN-CRF	0.9160	0.9230	0.9182
BERT-BiLSTM-CRF	0.9103	0.9276	0.9229
BERT-BiLSTM-Attention-CRF	0.9201	0.9083	0.9175
BERT-Transformer-BiLSTM-CRF	0.9288	0.9362	0.9307
BERT-BiLSTM-IDCNN-CRF(Serial)	0.9439	0.9232	0.9312
BERT-BiLSTM-IDCNN-CRF(Parallel)	0.9509	0.9455	0.9468
BBIEC	0.9492	0.9561	0.9521

As shown in Table 6. Without the BERT pre-training model, the BiLSTM-CRF model has improved all the metrics compared with the IDCNN-CRF model because the BiLSTM has stronger global context extraction ability compared with IDCNN, but also takes more time; after adding the BERT pre-training to the BiLSTM-CRF and ICDNN-CRF models, the metrics of both models have significantly improved. This is because BERT adds location coding and a multi-headed self-attention mechanism. Hence, it has a stronger semantic recognition ability and can make the downstream model perform better, but it does not change the fundamental strengths and weaknesses of the downstream model, so the evaluation index of the BERT-BiLSTM-CRF model is still higher than that of the BERT-IDCNN-CRF; and the advantage of IDCNN over BiLSTM is that it has a stronger extraction ability for local semantic features, which is determined by the principle of IDCNN; inserting the Attention mechanism into the BERT-BiLSTM-CRF model emphasizes extracting local semantic features, decreasing the model's extraction ability; inserting

a Transformer encoder to the BERT-BiLSTM-CRF model improves the ability of the model to construct contextual semantic vectors, which improves the model's extraction ability. Serially inserting the IDCNN model based on the BERT-BiLSTM-CRF model improves the extraction ability of local features; based on the BERT-BiLSTM-IDCNN-CRF model, the BiLSTM and IDCNN neural networks are connected in parallel with dual channels to improve in the extraction ability. The BBIEC surface model proposed in this paper is based on the dual channel parallel model and improves the IDCNN model in it by improving the activation function from relu to elu, and the three evaluation indexes reach 0.9492, 0.9561, and 0.9521, respectively, which is better than the BERT-BiLSTM-IDCNN-CRF (dual channel) model with the best results in the comparison experiments in terms of recall and channel) The best model in the comparison experiments has significantly improved recall and F1 value. The reason is that elu has most of the advantages of relu and does not have the Dead relu problem of the relu activation function; the elu function makes the gradient closer to the unit gradient by reducing the effect of bias shift, and the mean value of output is also closer to 0. However, the BBIEC model has decreased in the accuracy rate because the elu activation function contains power operations in its calculation, which is computationally intensive. The elu activation function introduces the case of wanting $x < 0$, which activates some neurons, resulting in a larger FP value in the accuracy formula, leading to a decrease in the accuracy rate.

D. SENTENCE-LEVEL RECOGNITION RESULTS

The effect of entity recognition at the sentence level of the BBIEC model is shown in Table 7.

TABLE 7. Sentence-level recognition results.

Model	Accuracy
Word2Vec-IDCNN-CRF	0.4465
Word2Vec-BiLSTM-CRF	0.4606
BERT-IDCNN-CRF	0.6225
BERT-BiLSTM-CRF	0.6512
BERT-BiLSTM-Attention-CRF	0.6316
BERT-Transformer-BiLSTM-CRF	0.6581
BERT-BiLSTM-IDCNN-CRF(Serial)	0.6718
BERT-BiLSTM-IDCNN-CRF(Parallel)	0.6870
BBIEC	0.7141

As shown in Table 7. Without using the BERT pre-training model, the whole-sentence recognition accuracy of the IDCNN-CRF model and BiLSTM-CRF model is less than 0.5; after using the BERT pre-training model, the two models have a greater improvement in whole-sentence recognition accuracy, which is because the BERT pre-training model brings stronger semantic extraction ability to the model; after using the attention mechanism, over-emphasis on local features, instead, is a decrease in extraction ability;stronger semantic extraction capability is obtained after using Transformer encoder; after the two neural networks are connected, The two-channel parallel connection method has better recognition effect than the serial connection method, which is because, in the serial connection method, the downstream neural network is vulnerable to the error of the upstream neural network, resulting in the poor recognition effect of the serial connection model; the BBIEC model proposed in this paper improves the activation function used in the IDCNN network, which makes some dormant neurons in the IDCNN activate and improves the overall model of semantic extraction ability.

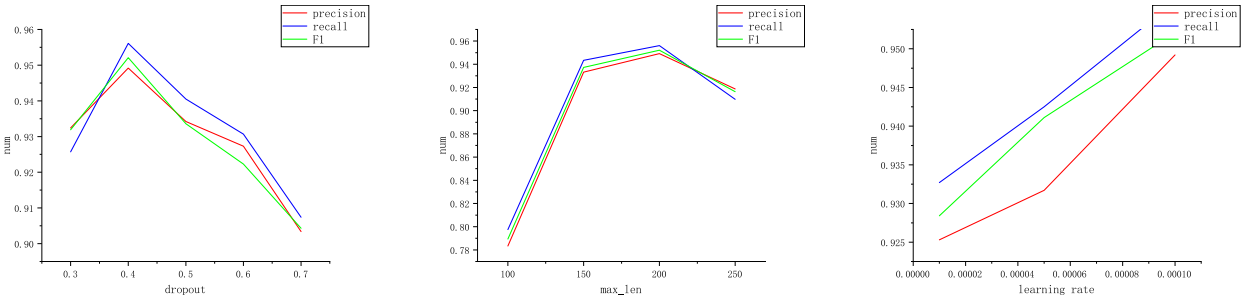
E. ENTITY-LEVEL RECOGNITION RESULTS

In some relevant application areas of the epidemiological investigation of information on COVID-19, entity-level evaluation metrics are required, and the evaluation metrics for each entity of the BBIEC model are shown in Table 8.

TABLE 8. Entity-level recognition results.

Entity	P	R	F_1
B-PER	0.8000	0.8000	0.8000
I-PER	0.7143	0.7143	0.7143
E-PER	0.8108	0.8571	0.8333
B-LOC	0.9566	0.9622	0.9594
I-LOC	0.9568	0.9526	0.9547
E-LOC	0.9206	0.9206	0.9206
S-LOC	1.0000	0.8462	0.9167
B-ORG	0.8824	1.0000	0.9375
I-ORG	0.9167	0.9167	0.9167
E-ORG	0.8824	1.0000	0.9375
B-VEH	0.9838	0.9492	0.9662
I-VEH	0.9429	0.9900	0.9659
E-VEH	0.9675	0.9297	0.9482
B-TEL	1.0000	1.0000	1.0000
I-TEL	1.0000	1.0000	1.0000
E-TEL	1.0000	1.0000	1.0000
B-TEMP	1.0000	1.0000	1.0000
I-TEMP	1.0000	1.0000	1.0000
E-TEMP	1.0000	1.0000	1.0000
B-NUM	0.9718	0.9857	0.9787
I-NUM	0.9868	0.9912	0.9890
E-NUM	0.9437	0.9571	0.9504

According to the numerical analysis of the indicators, we can get: (1) the three evaluation indicators are 1.0000 have TEL and TEMP entities because the format of these two entities is very fixed, indicating the phone number and the patient's body temperature, and the model is easier to achieve the result of successful recognition of all of them; (2) the three evaluation indicators are more than 0.9000 have four entities: VEH, NUM, TIME and DATE respectively entities; because these four types of entities all have strong regularity and are mostly represented by English, numeric and Chinese characters, the recognition difficulty is small and the indicators are high. (3) There are three entities whose remaining



(a) Effect of dropout on model performance (b) Effect of max_len on model performance (c) Effect of lr on model performance

FIGURE 10. Effect of different hyperparameters on model performance.

three indicators do not reach 0.9000: PER, LOC, and ORG, respectively, and the reasons are analyzed as follows.

Analysis by entity type shows that (1) PER entity: the recognition effect of I-PER is worse than that of the boundary entities B-PER and E-PER because, in the epidemiological investigation of information on COVID-19 corpus, the patient's name is treated as ''last name'' plus ''a,'' or the patients are numbered and their real names are not published, so the training set of I-PER entities is small and cannot learn the semantic features effectively, resulting in low metrics. (2) LOC entities: The epidemiological investigation information of the COVID-19 corpus is published by the relevant institutions in each region, which contains many single character place name abbreviations, i.e., S-LOC entities, and these individual character entities do not have accurate upper and lower boundaries and cannot be corrected for errors by the CRF classifier, so the metrics are relatively low compared to other entities. (3) ORG entities: In natural language expressions, LOC entities and ORG entities have a great connection, and some words can be represented as both LOC entities and ORG entities, and the complex semantic understanding leads to a low accuracy rate.

F. EFFECT OF HYPERPARAMETERS ON MODEL PERFORMANCE

In this paper, we conduct experiments on the effects of three model hyperparameters, dropout rate (dropout), maximum sentence length (max_len), and learning rate (lr), on model performance, and the results are shown in Figure 10.

As shown in Figure 7(a), using dropout to prevent the overfitting phenomenon, it can be concluded that the best fit is achieved when dropout is 0.4; when dropout=0.5, the three indicators are as close as when dropout=0.4, which proves that its fitting ability is more excellent; when dropout=0.3, all indicators decrease, which is due to dropout rate is low, leading to model overfitting and lower verification accuracy; when dropout>0.5, the indicators decrease significantly, which is due to the loss of too much semantic information due to too high dropout, leading to the decline of model performance; as shown in Figure 7(b), max_len takes 200 as the best choice, and max_len=100 intercepts a part of the sentences, and the distribution of epidemiological investigation of information on COVID-19 corpus entities is more dense, which will lose a large amount of semantic information, resulting in three indicators far below the optimal choice; when max_len=250, because the longest value of the sentence is 195 and the average length is about 170, filling in too much useless information will lead to low prediction results; when max_len=150, the truncated part is less and can still contain most of the semantics, so it is close to the optimal index; as shown in Figure 7(c), when lr=1e-4, the three indexes are the highest; when lr>5e-4, lr is too large, which leads to the model cannot converge and the indexes are 0; there is a graph that shows the overall change trend incrementally.

VI. CONCLUSION

In this paper, we address the problems in the field of epidemiological investigation of information on COVID-19, design the entity definition of the epidemiological investigation of information on COVID-19 named entity recognition dataset by analyzing the original corpus of epidemiological investigation of information on COVID-19, and select the appropriate annotation system to build the epidemiological investigation of information on COVID-19 dataset, and propose the BBIEC model with dual neural network serial, which reduces the labor cost and time cost in the annotation process.

The model can fully learn both global context and local features and improve the entity recognition by fusing both features. The model achieves better results in the epidemiological investigation information of the COVID-19 dataset and multiple entity levels. Three metrics reach 0.9492, 0.9561, and 0.9521, respectively.

The recognition of Chinese named entities of epidemiological investigation of information on COVID-19 achieved in this paper lays the foundation for the construction of a knowledge graph in the field of epidemiological investigation of information on COVID-19, which is a good aid to better control the epidemiological investigation of COVID-19. The next step in the research direction is to extract the relationship between the epidemiological investigation of information on COVID-19 entities.

REFERENCES

[1] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106779.

[2] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," 2019, *arXiv:1910.11470*.

[3] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: A systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–43, Aug. 2018.

[4] P. Sun, X. Yang, X. Zhao, and Z. Wang, "An overview of named entity recognition," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 273–278, doi: 10.1109/IALP.2018.8629225.

[5] B. Babych and A. Hartley, "Improving machine translation quality with automatic named entity recognition," in *Proc. EAMT*, 2003, pp. 1–8.

[6] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1441–1451.

[7] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the web: An experimental study," *Artif. Intell.*, vol. 165, no. 1, pp. 91–134, 2005.

[8] Z. Ju, J. Wang, and F. Zhu, "Named entity recognition from biomedical text using SVM," in *Proc. 5th Int. Conf. Bioinf. Biomed. Eng.*, May 2011, pp. 1–4, doi: 10.1109/icbbe.2011.5779984.

[9] A. Ekbal, S. Saha, U. K. Sikdar, and M. Hasanuzzaman, "A genetic approach for biomedical named entity recognition," in *Proc. 22nd IEEE Int. Conf. Tools Artif. Intell.*, Oct. 2010, pp. 354–355, doi: 10.1109/ICTAI.2010.125.

[10] X. Niu and Z. Jiang, "Comparative study of HMM and MEM in named entity recognition," in *Proc. Int. Conf. Comput. Sci. Service Syst. (CSSS)*, Jun. 2011, pp. 1150–1153, doi: 10.1109/CSSS.2011.5974861.

[11] H. Hu and H. Zhang, "Chinese named entity recognition with CRFs: Two levels," in *Proc. Int. Conf. Comput. Intell. Secur.*, Dec. 2008, pp. 1–6, doi: 10.1109/CIS.2008.72.

[12] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[13] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and accurate entity recognition with iterated dilated convolutions," 2017, *arXiv:1702.02098*.

[14] Q. Y. Yang, J. Jiang, and X. Y. Feng, "Named entity recognition of power substation knowledge based on transformer-BiLSTM-CRF network," in *Proc. Int. Conf. Smart Grids Energy Syst. (SGES)*, Nov. 2020, pp. 952–956, doi: 10.1109/SGES51519.2020.00174.

[15] Y. An, X. Xia, X. Chen, F.-X. Wu, and J. Wang, "Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF," *Artif. Intell. Med.*, vol. 127, May 2022, Art. no. 102282.

[16] X. Jiang, J. X. Ma, and H. Yuan, "Named entity recognition in the field of ecological management technology based on BILSTM-IDCNN-CRF model," *Comput. Appl. Softw.*, vol. 38, no. 3, pp. 134–141, Mar. 2021.

[17] X. Li, Y. Meng, X. Sun, Q. Han, A. Yuan, and J. Li, "Is word segmentation necessary for deep learning of Chinese representations," 2019, *arXiv:1905.05526*.

[18] Z. Liu, C. Zhu, and T. Zhao, "Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words?" in *Proc. Int. Conf. Intell. Comput. (ICIC)*. Berlin, Germany: Springer, vol. 6216, 2010, pp. 634–640, doi: 10.1007/978-3-642-14932-0_78.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[20] Y. Li, X. Xie, D. Chen, A. Wen, A. Li, and Z. Xie, "Thyroid discharge summary NER based on BERT-BiLSTM-CRF model," in *Proc. 3rd Int. Conf. Electron. Commun. Artif. Intell. (IWECAI)*, Jan. 2022, pp. 43–46, doi: 10.1109/IWECAI55315.2022.00016.

[21] H. L. Tang *et al.*, "BERT-LCRF named entity recognition method oriented clock domain," *Comput. Eng. Appl.*, vol. 4, no. 5, pp. 1–11, Jun. 2021. [Online]. Available: http://kns.cnki.net/kcms/detail/11.2127.TP.20210622.1045.016.html

[22] X. Gao and Q. Li, "Named entity recognition in material field based on BERT-BILSTM-attention-CRF," in *Proc. IEEE Conf. Telecommun., Opt. Comput. Sci. (TOCS)*, Dec. 2021, pp. 955–958, doi: 10.1109/TOCS53301.2021.9688665.

[23] Y. Gan, R. Yang, C. Zhang, and D. Jia, "Chinese named entity recognition based on BERT-transformer-BiLSTM-CRF model," in *Proc. 7th Int. Symp. Syst. Softw. Rel. (ISSSR)*, Sep. 2021, pp. 109–118, doi: 10.1109/ISSSR53171.2021.00029.

[24] N. Li, H. M. Guan, P. Yang, and "Chinese named entity recognition method based on BERT-IDCNN-CRF," *J. Shandong Univ. Sci. Ed.*, vol. 55, no. 1, pp. 102–109, 2020.

[25] X. Wu, T. Zhang, S. Yuan, and Y. Yan, "One improved model of named entity recognition by combining BERT and BiLSTM-CNN for domain of Chinese railway construction," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2022, pp. 728–732, doi: 10.1109/ICSP54964.2022.9778794.

[26] Y. Chang, L. Kong, K. Jia, and Q. Meng, "Chinese named entity recognition method based on BERT," in *Proc. IEEE Int. Conf. Data Sci. Comput. Appl. (ICDSCA)*, Oct. 2021, pp. 294–299, doi: 10.1109/ICDSCA53499.2021.9650256.

[27] A. Vaswani, N. Shazeer, and N. Parmar, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2017, pp. 5998–6008.

[28] J. L. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[31] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines Vinod Nair," in *Proc. Int. Conf. Mach. Learning (ICML)*, 2010, pp. 807–814.

[33] X. T. Guan, R. Z. He, and Z. R. Li, "The COVID-19 diagnosed patient trajectory structured data," in *Beijing Advanced Innovation Center for Big Data and Brain Computing*, Z. Hu and C. Zheng, Eds. Beijing, China: Beihang Univ., Mar. 2020. [Online]. Available: https://github.com/BDBC-KG-NLP/COVID-19-tracker

[34] Y. Wang, Y. Wang, Y. Chen, and Q. Qin, "Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures," *J. Med. Virol.*, vol. 92, no. 6, pp. 568–576, Jun. 2020.

[35] C. L. Yang *et al.*, "Research on COVID-19 text entity relation extraction and dataset construction methods," *Comput. Eng. Appl.*, pp. 1–9, Jun. 2022. [Online]. Available: http://kns.cnki.net/kcms/detail/11.2127.tp.20220622.1100.010.html

[36] H. M. Duan, J. Song, G. W. Xu, G. X. Hu, and S. W. Yu, "The development of a large-scale tagged Chinese corpus and its applications," *Appl. Linguistics*, vol. 2, no. 11-2888/H, pp. 72–77, May 2000, doi: 10.16499/j.cnki.1003-5397.2000.02.013.

[37] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022.

CHONGLUO YANG was born in Heilongjiang, China, in 1996. He received the bachelor's degree from the Harbin Institute of Technology, in 2018. He is currently pursuing the master's degree with the School of Information and Electrical Engineering, Hebei University of Engineering. His research interests include natural language processing and deep learning.

LONG SHENG was born in Hebei, China, in 1982. He received the bachelor's degree from Central South University, in 2004, and the Ph.D. degree from the University of Electronic Science and Technology, in 2012. He worked at the University of Information and Electrical Engineering, Hebei University of Engineering. He has published nine papers. His research interests include natural language processing and machine learning.

ZHONGCHENG WEI was born in Hebei, China. He received the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2016. He is currently an Assistant Professor with the School of Information and Electrical Engineering, Hebei University of Engineering, and works as a member of the Hebei Key Laboratory of Security and Protection Information Sensing and Processing. He has coauthored 20 publications, held eight China national invention patents and seven software copyrights. He has presided over or participated in more than ten projects on scientific research. His research interests include the Internet of Things, wireless communication, big data, and artificial intelligence.

WEI WANG received the Ph.D. degree in control science and engineering from the School of Information Engineering, University of Science and Technology Beijing, in 2012. He is currently an Associate Professor with the Hebei University of Engineering. His research interests include the public safety Internet of Things and implicit human–computer interaction learning.