applied sciences

MDPI

Article

Medical Named Entity Recognition Fusing Part-of-Speech and Stroke Features

Fen Yi [1], Hong Liu [1], You Wang [2], Sheng Wu [3], Cheng Sun [4], Peng Feng [3] and Jin Zhang [1,3,*]

[1] College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China
[2] The State Key Laboratory of Industrial Control Technology, Institute of Cyber Systems and Control, Zhejiang University, Hangzhou 310027, China
[3] School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China
[4] School of Mathematics and Statistics, Hunan Normal University, Changsha 410081, China
* Correspondence: mail_zhangjin@163.com

Abstract: It is highly significant from a research standpoint and a valuable practice to identify diseases, symptoms, drugs, examinations, and other medical entities in medical text data to support knowledge maps, question and answer systems, and other downstream tasks that can provide the public with knowledgeable answers. However, when contrasted with other languages like English, Chinese words lack a distinct dividing line, and medical entities have problems such as long length and multiple entity types nesting. Therefore, to address these issues, this study suggests a medical named entity recognition (NER) approach that combines part-of-speech and stroke features. First, the text is fed into the BERT pre-training model to get the semantic representation of the text, while the part-of-speech feature vector is obtained using the part-of-speech dictionary, and the stroke feature of the text is extracted through a convolution neural network (CNN). The word vector is then joined with the part-of-speech and stroke feature vectors, respectively, and input into the BiLSTM and CRF layer for training. Additionally, to balance the disparity in data volume across several types of entities, the class-weighted loss function is included in the loss function. According to the experimental findings, our model's F1 score on the CCKS2019 dataset reaches 78.65%, and the recognition performance exceeds many existing algorithms.

Keywords: entity recognition; BERT; BiLSTM; multiple features; CRF

Citation: Yi, F.; Liu, H.; Wang, Y.; Wu, S.; Sun, C.; Feng, P.; Zhang, J. Medical Named Entity Recognition Fusing Part-of-Speech and Stroke Features. *Appl. Sci.* **2023**, *13*, 8913. https://doi.org/10.3390/app13158913

Academic Editor: Giacomo Fiumara

Received: 30 May 2023
Revised: 21 July 2023
Accepted: 28 July 2023
Published: 2 August 2023

1. Introduction

With the rise of "Internet + Medical", major online medical service platforms bring convenience to people and create a significant volume of text data containing rich medical knowledge. Most of these text data are in an unstructured form, in which it is difficult to intuitively reflect information. Processing and mining these textual data can help professionals quickly and accurately obtain information to assist in diagnostic decision-making, save time and cost, and better provide treatment for patients.

The aim of named entity recognition (NER) is to find entities in a context that have specific meanings. NER extracts and classifies the names of people, places, institutions, biological proteins, drugs, and other entities in unstructured text to serve downstream tasks including text mining, information extraction, intelligent question answering, machine translation, and so on [1]. Numerous scholars have studied entity recognition in a variety of domains, including chemistry [2,3], politics and law [4,5], social media [6,7], and geography [8], thanks to the ongoing advancement of NER technology. Luo et al. [2] presented an Att-BiLSTM-CRF model for chemistry NER, which applies attention to the document to determine the similarity between various words in the document and assigning weights. Liu et al. [4] introduced a legal judgment named the entity recognition approach based on small-scale tag data, which uses bootstrapped mode to learn how to

label data from small-scale manually labeled data, and then embed words with the Chinese model pre-trained by BERT. Aguilar et al. [6] introduced a multitasking approach applied to social media data, which combines named entity segmentation with fine-grained network element classification main tasks. Within the biomedical field, Zhu et al. [9] suggested an end-to-end deep learning method called GRAM-CNN, which utilizes local context embedded by n-gram characters and words via convolutional neural networks without requiring any specific knowledge or feature engineering. Yoon et al. [10] suggested a model known as CollaboNet that combines many NER models. The target model can obtain information from other cooperation models thanks to the connections between models trained on various datasets, which lowers the number of false positives and incorrectly categorized entities. Unlike languages such as English, Chinese lacks a distinct demarcation between words, and the grammatical structure of Chinese is complex and diverse. Chinese medical entities usually contain a large number of professional words and terms, making it challenging to understand and analyze these complex language structures for the model. And the technical terms involved in the medical field may have different forms and variants such as synonyms, abbreviations, spelling mistakes, etc. Additionally, The length of medical entities, such as "根治性左半肠切除术 (radical left hemicolectomy)", is excessively long. And the nested nature of various entity types, for instance, the imaging examination entity "胃肠道碘水造影 (gastrointestinal tract iodine hydrography)" nested in the symptom entity "胃肠道 (gastrointestinal tract)". These problems make it more difficult for Chinese medical entity recognition.

Based on the aforementioned issues, this research suggests a named entity recognition model fusing part-of-speech and stroke features for medical texts. Firstly, in order to discriminate words with distinct meanings in the text, word parts-of-speech and stroke information are incorporated as multiple characteristics in low-dimensional word vectors acquired by the BERT pre-training model, improving the border recognition result, in which stroke features are extracted by Convolution Neural Network (CNN). Following that, word vectors are spliced with the parts-of-speech and stroke vectors, respectively, to be input into the bidirectional short-term memory network (BiLSTM) for feature extraction. At last, the retrieved context characteristics are combined and fed into a conditional random field (CRF) and predicted label results are generated using constraints discovered from the CRF. At the same time, the class-weighted loss function, which, along with the CRF loss function, makes up the model's overall loss function, is added to balance out the disparity in entity counts among different categories. The experimental results show that the model, which outperforms other models currently in use, achieved favorable outcomes on the CCKS2019 dataset.

2. Related Work

2.1. NER Approaches

Dictionary and rule-based, traditional machine learning, and deep learning approaches are the main NER approaches.

Dictionary and rule-based approaches count on linguists to manually create unique rule templates or specialized dictionaries depending on the properties of data sets [11–13]. This rule-based approach has various drawbacks, including a high level of human involvement required, challenges in extending it to new entity types or datasets due to the specificity of rules to a given domain, the absence of comprehensive dictionaries, and limited portability.

According to the machine learning approach, NER is viewed as a sequential labeling issue, and the labeling model is learned from a huge corpus by a statistical machine learning algorithm, to label sentences. Currently, the traditional machine learning approaches used in NER mainly include CRF [14], Support Vector Machine (SVM) [15], Hidden Markov Model (HMM) [16,17], Decision trees [18], etc.

Among the deep learning-based methods, an NER technique based on neural networks was first described by Collobert et al. [19] and offers an integrated neural network

design for handling numerous tasks related to NLP. Huang et al. [20] proposed using the BiLSTM-CRF model for sequence annotation in 2015. BiLSTM can effectively employ input features for the past and future, while the CRF layer's transition matrix can learn a variety of limitations to improve the accuracy of its forecast outcomes. Strubell et al. [21] suggested a network IDCNN with better prediction ability than CNN in 2017. IDCNN used expansive convolution to add holes in the convolution kernel, which expanded the receptive field of the model and ensured accuracy while increasing speed. Google introduced the unsupervised pre-training language model BERT [22] in 2018. BERT adopted the Decoder structure of Transformer [23] and trained with large-scale unlabeled corpus, which showed strong semantic information extraction ability. Later, a series of improved models of BERT appeared, such as RoBERTa [24] and BERT-WWM [25]. Deep learning-based NER methods have gained popularity in recent times. Wang et al. [26] suggested a multi-tasking learning structure that uses training data from various sorts of entities in tandem to improve each type of entity's performance. Cho et al. [27] created a model according to the BiLSTM-CRF model, which enhanced CNN and BiLSTM by combining two distinct character-level representations collected from convolution. Chang et al. [28] put forward an NER approach based on BERT, which is used for pre-training and simultaneously uses BiLSTM and IDCNN to extract features.

2.2. NER Approaches Based on Deep Learning

The embedding layer, encoding layer, and decoding layer are the three components of deep learning-based NER methods. The embedding layer learns the semantic and grammatical information of the input text and displays them as a vector. The encoding layer performs feature extraction according to the contextual data incoming from the embedding layer. The decoding layer predicts the label corresponding to each word in the text according to the encoding layer's results.

For the embedding layer, the full input Chinese text must be segmented. Word-based and character-based segmentation processing techniques are the two main categories. With the word-based approach, the text is divided into words, and the pertinent information in the words is used to extract contextual characteristics. However, issues will unavoidably arise throughout the word segmentation phase, and the mistakes created during this process will be conveyed step by step in the subsequent recognition process, impacting the final recognition result. The character-based model uses the character as its input, does not require splitting words in the text, avoids the error introduced by word separation, and frequently outperforms the word-based model, though words still have more detailed boundary and semantic information.

In the encoding layer, CNN, Recurrent Neural Network (RNN), and Transformer networks are frequently employed. The CNN-based model concentrates on extracting local features and can be calculated in parallel to shorten training time and save time and cost. CNN, however, is unable to record distant characteristics and ignores contextual data. Yao et al. [29] introduced a CNN-based biomedical NER method with a multi-layer structure that uses neural networks to produce a significant amount of possible feature data embodied in word vectors. Data from linear sequences are routinely processed using a model based on RNN. An RNN variation called BiLSTM is capable of using sequential data in both forward and backward directions to learn the connection between distant dependencies. The computation of the current element, however, depends on the calculation outcome of the preceding element owing to the sequential sequence; hence, simultaneous processing is not possible. Lin et al. [30] introduced a multi-channel neural structure for social media messaging, which integrated whole-word representations with multi-channel data and conditional random fields into BiLSTM without the need for any extra hand-made features. The model based on Transformer can make use of attention mechanisms to better capture long-distance dependencies and parallel computation; however, it is unable to capture information such as direction and position. Yu et al. [31] propose a multimodal NER approach using a Transformer as a foundation, and a multimodal interaction module may

be used to gain the verbal description of picture perception as well as the visual depiction of verbal perception.

CRF and Multi-Layer Perceptron (MLP) are commonly used in decoding layers. The Multi-Layer Perceptron views sequence labeling as a multi-classification issue as the decoding layer, and each word's prediction simply makes inferences based on context-related information rather than considering the predicted label outcomes of neighboring words. As the decoding layer, CRF considers the whole sequence, learns the relationships between nearby tags during training, and incorporates restrictions into the prediction to provide the ideal tag sequence.

3. Methods

The model that is given in this paper employs a BERT pre-training model in the embedding layer, BiLSTM as the encoder and CRF as the decoder. Figure 1 displays the main layout of the construction. A word vector to represent the content is produced by the BERT pre-training model in the embedding layer, and this word vector is later combined with the parts-of-speech feature vector and the stroke feature vector and fed into the BiLSTM layer, respectively, to produce text features using contextual information. After restrictions, the combined feature information is entered into the CRF layer to produce the predictive text labels.

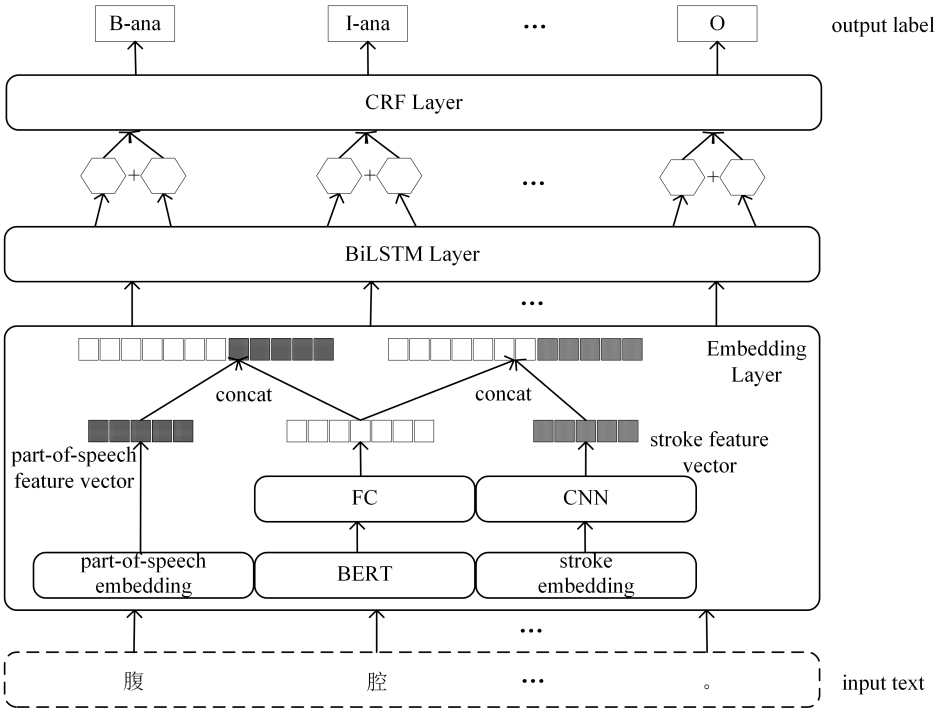

Figure 1. This is the model structure. The "腹腔 (abdominal cavity)..." is the input Chinese text. And the "Embedding Layer" shows the process that a single word enters the embedding layer.

3.1. Embedding Layer

The entered content is mapped into a low-dimensional space in units of words in the embedding layer, and the words are represented by vectors, which can well reflect the relationship between words. Words with similar contexts usually have similar semantic and grammatical attributes and in this low-dimensional space, are frequently near to each other. In this study, a character-based approach is adopted to process the input text. Additionally, to utilize the word's semantic data to its fullest extent and distinguish different semantic elements, and then enhance the recognition effect of the boundary, this work introduces the part-of-speech and stroke characteristics as multiple features and uses the BERT pre-training model to produce the word vector.

Assuming the model receives the following sentence as input: $X = (x_1, x_2, \ldots, x_n)$, after entering the embedding layer, the word vector $W = (w_1, w_2, \ldots, w_n)$ is acquired through the BERT model while the part-of-speech feature vector $C = (c_1, c_2, \ldots, c_n)$ is obtained through the part-of-speech dictionary, and the stroke feature vector $B = (b_1, b_2, \ldots, b_n)$ is extracted using CNN. The word vector was spliced with the part-of-speech feature vector and stroke feature vector, respectively, to get $E_1 = [W; C]$ and $E_2 = [W; B]$, and then input E_1 and E_2 into the BiLSTM layer, respectively.

3.1.1. BERT Model

The multi-layer Transformer framework used by the BERT model is formed by stacking encoders of multiple transformers. There is a Self-attention mechanism in the encoder, which will use the token of its context when encoding a token. The three embeddings that make up the text data input into the BERT pre-training model are token embedding, segment embedding, and position embedding. Token embedding represents the current word ID, segment embedding shows the sentence ID in which the word at hand is situated, and position embedding displays the position ID of the current word. Figure 2 depicts the BERT model's general structure. In this experiment, BERT-base, which has 12 stacked encoder layers, is employed. Each encoder employs 12 heads of attention, and the bottom encoder's input is the top encoder's output. There are 768 hidden nodes in the encoder's feedforward network.

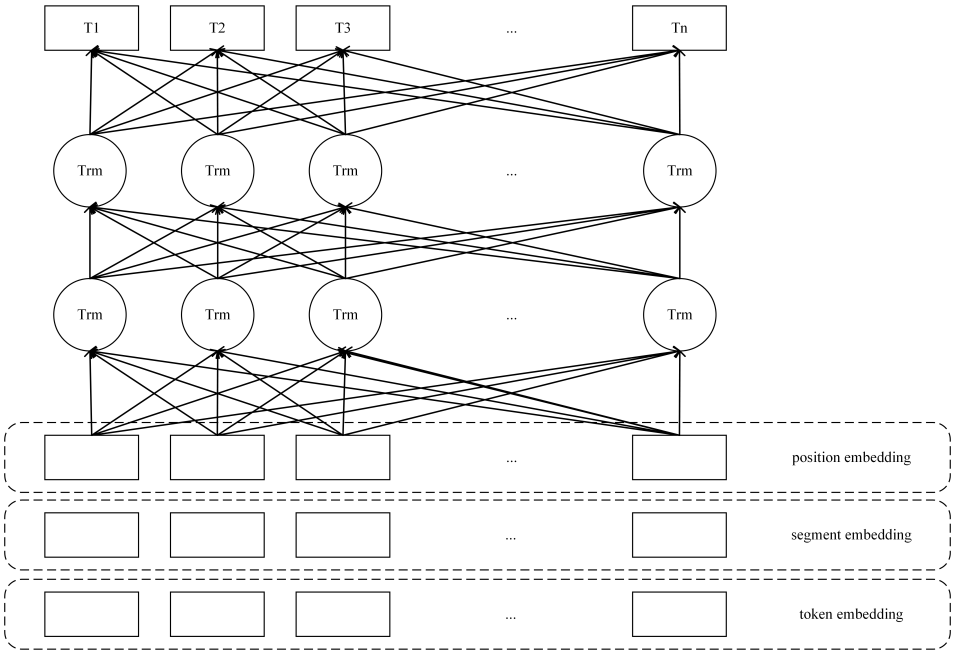

Figure 2. Structure of BERT model.

3.1.2. Part-of-Speech Embedding

As a part of the word attribute, adding part-of-speech features in the embedding layer can more accurately convey the meaning of the word. For example, diseases and drugs are mostly nouns and symptoms are mostly verbs. The model developed in this study employs the posseg package in the jieba library to segment the input text into parts of speech. A total of 23 parts of speech and a unique part-of-speech marker O are used to mark the text, except for a few parts of speech that are not frequently used. After the part-of-speech of the text is entered into the embedding layer, the part-of-speech vector of the input text is acquired by looking up the part-of-speech dictionary. The part-of-speech vector is initially initialized randomly.

3.1.3. Stroke Embedding

It is possible to extract stroke features using both BiLSTM and CNN. CNN gathers information by grabbing specific word strokes, whereas BiLSTM collects information following the entire word's stroke. In this study, BiLSTM and CNN were respectively used for feature extraction of stroke information during the experiment. The results show that the extraction result of CNN is superior to that of BiLSTM (see the ablation experiment in Section 4.4 for the results). As a result, CNN is used in the model suggested in this study to extract the text's stroke characteristics.

First, we use the pywubi package to get the strokes of the input text and map the strokes to a low-dimensional vector. The vector matrix is then subjected to the application of convolution kernels of various sizes to extract relevant information between neighboring strokes. To capture various characteristics during the convolution process, two kernels of convolution of differing sizes are utilized in this work. Finally, in order to extract stroke features, the maximum pooling method is used on all feature mappings, and the resultant features are connected to form the stroke feature vector of each text. Figure 3 depicts the structure described above.

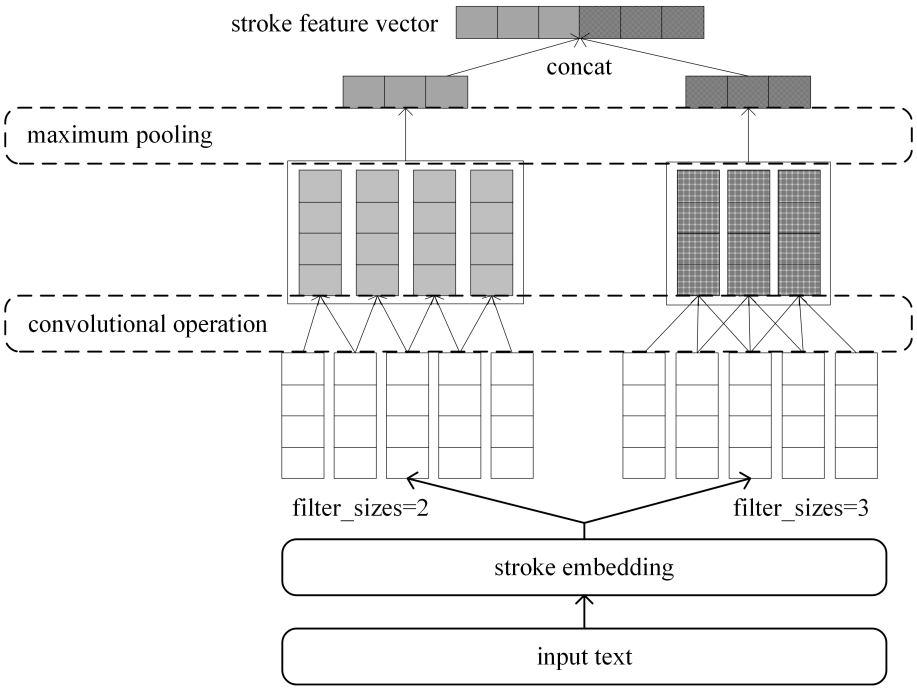

Figure 3. Structure of CNN extraction stroke feature. The vectors obtained from stroke embedding are extracted through two CNN kernels of different sizes and then concatenated together.

3.2. BiLSTM Layer

Long Short-Term Memory (LSTM), which employs memory units with adaptive gating mechanisms to handle information selectively, solves the challenge of gradient vanishing and explosion in RNN structures. Three different gates are present in an LSTM cell structure: an input gate for selecting data that can be input, a forgetting gate for selecting data that should be forgotten, and an output gate for selecting data that will be output. The LSTM hidden state's precise calculation algorithm at time t is as follows:

$$I_t = \sigma(X_t W_{xi} + H_{t-1} W_{hi} + b_i) \tag{1}$$

$$F_t = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right) \tag{2}$$

$$O_t = \sigma(X_t W_{xo} + H_{t-1} W_{ho} + b_o) \tag{3}$$

$$\widetilde{C}_t = \tanh(X_t W_{xc} + H_{t-1} W_{hc} + b_c) \tag{4}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \widetilde{C}_t \tag{5}$$

$$H_t = O_t \odot \tanh(C_t) \tag{6}$$

Both the forward and backward LSTM components make up BiLSTM. It obtains semantic information from both directions through forward and backward input, and can effectively extract the features of context.

3.3. CRF Layer

There will be some issues if the content output of the BiLSTM layer is classed directly because the dependency between labels is not considered. The drug entity, for instance, can only begin with "B-dru" rather than "I-dru" and the disease and diagnosis label "B-dis" can only be followed by "I-dis" rather than "O" or "I-dru". The role of conditional random fields is to provide some restrictions to the final tag prediction to guarantee the output tag's validity.

The sequence $S = (s_1, s_2, \ldots, s_n)$ is sent as input to the CRF layer, and then CRF outputs the most probable sequence $T = (t_1, t_1, \ldots, t_n)$. The true path's score and the sum of all pathways' scores make up the CRF loss function. The specific formula is shown in Formula (7), where $S_{RealPath}$ is the real path score, and $S_i (i \in n, n$ is the overall quantity of paths) is the path score function.

$$LossFunction = \log(e^{s_1} + e^{s_2} + \cdots + e^{s_n}) - S_{RealPath} \tag{7}$$

The path score function is divided into two parts: emission score and transfer score. The output from the previous BiLSTM layer provides the emission score, indicating the likelihood of the sequence and its corresponding tags. The transfer score is derived from the transfer matrix in the CRF layer, representing the likelihood of transitioning between labels. The transition matrix is continuously updated during training, and the optimal sequence of tags can be determined by learning the relationships between tags and constraints.

4. Analysis and Results from Experiments

4.1. Experimental Dataset

This experiment uses the CCKS2019 dataset, which is the assessment task dataset made available by the China Conference on Knowledge Graph and Semantic Computing in 2019. This dataset is the real desensitized electronic medical record data, which was manually edited by Yiduyun Medicine, and includes six entity types: disease and diagnosis, laboratory test, image examination, drug, operation, and anatomical site. In this experiment, the number of sentences in the CCKS2019 dataset is divided into a train set and a dev set according to the ratio of 7:3. Table 1 displays the statistics of the dataset's entities.

Table 1. Statistics of the dataset.

Type	Train	Dev
disease and diagnosis	2924	1288
laboratory test	799	396
image examination	661	308
drug	1250	572
operation	709	320
anatomical site	5804	2622
total	12,147	5506

4.2. Improvement of Loss Function

Because there are differences in the number of entities in the experimental data, the imbalance in the number of sample categories may lead to the loss being dominated by categories with large data during the training process. This can result in the model failing to effectively train the features of categories with fewer samples. To address this issue, in addition to the CRF loss function, this study introduces the class-weighted loss function. Both loss functions are combined to create the total loss function. By assigning various weights to the sample losses of different categories and adjusting the weights based on the number of entities in each category, the class-weighted loss function balances the differences in data volume across multiple categories. The category with the fewest samples is given the greatest weight, while the category with the largest number of samples is given the least weight.

4.3. Experimental Setup

In this experiment, the training framework used is Pytorch, and the optimizer used is Adam, and the main hyperparameter settings in this paper are shown in Table 2.

Table 2. Hyperparameters of the model.

Hyperparameter	Value
BiLSTM hidden size	128
Embedding size	768
Learning rate	3×10^{-5}
Max sequence length	150
Dropout	0.5
Batch size	16
Epoch	30

4.4. Results

In this experiment, the evaluation criteria are F1 score (F1), precision (P), and recall (R). On the CCKS2019 dataset, the following models were used to compare:

- IDCNN-CRF, the encoding layer uses the expansion convolution network. Unlike the traditional convolution network, the expansion convolution expands the receptive field of the model by adding holes in the convolution kernel and obtains more context by using fewer convolution layers.
- BiLSTM-CRF, a popular model for handling NER tasks, takes the word vector representing the input text from the embedding layer, feeds it into BiLSTM to obtain characteristics, and then outputs the expected tag results following CRF.
- BERT-CRF, the encoding layer using the BERT model.
- BERT + BC, acquiring word vectors by pre-training the BERT model, followed by BiLSTM-CRF for the NER task.
- BERT-WWM + BC obtains word vectors by pre-training the BERT-WWM model. Unlike BERT, in the initial pre-training phase, BERT-WWM modifies the training sample generation approach and increases the whole word mask.
- RoBERTa-WWM-ext + BC, using the RoBERTa-WWM-ext model in the embedding layer, RoBERTa has enhanced the training tasks and data generation methods over BERT.

Table 3 displays the outcomes of the experiment.

Table 3. Experimental result.

Model	P/%	R/%	F1/%
IDCNN-CRF	72.62	73.88	73.24
BiLSTM-CRF	73.43	74.32	73.87
BERT-CRF	73.90	80.43	77.03
BERT + BC	75.58	79.01	77.26
BERT-WWM + BC	76.46	79.89	78.14
RoBERTa-WWM-ext + BC	75.16	79.89	77.45
Our Model	77.49	79.84	78.65

It can be seen from the experimental outcomes that the model suggested in this study has F1 score, precision, recall that are 78.65%, 77.49%, and 79.84%, respectively, on the CCKS2019 dataset, among which the F1 score and precision reach the maximum in all comparison models. The results of the first three experiments reveal that the BERT model outperforms IDCNN and BiLSTM in terms of recognition effect when CRF is utilized as the decoder. By first masking part of the sentence's terms and then letting the model predict the unmasked words, BERT improved its capacity for semantic processing and did well on the NER challenge. The recognition effect is improved by adding a BiLSTM layer between BERT and CRF, as can be seen from a comparison of the third model's and fourth model's results. This is because the forward and backward LSTM layers reinforce the relationship before and after the text sequence, which enhances the recognition effect. The model described in this study builds on the BERT pre-training model by adding category weighted loss function, part-of-speech information, and stroke information. All three assessment indexes have advanced in comparison to the BERT + BC model, with the F1 score rising by 1.39%. Although the recall rate of our model is slightly lower than that of the RoBERTa-WWM-ext + BC model and the BERT-WWM + BC model, its precision is higher. On the whole, the F1 score of the comprehensive index is increased by 0.51% and 1.20%, respectively, which shows that embedding word part-of-speech information and stroke information as multiple features is helpful for the model to achieve a better recognition effect.

To further support the reliability of each element of the model outlined in this study, the following comparative tests were conducted: (1) by adding the class-weighted loss function; (2) by adding part-of-speech features; (3) by adding stroke features extracted by BiLSTM; (4) by adding stroke features extracted by CNN. In Table 4, the experimental findings are displayed.

Table 4. The results of the Ablation experiment.

Model	P/%	R/%	F1/%
Baseline	75.58	79.01	77.26
Baseline + Loss	75.95	79.84	77.85
Baseline + Part-of-speech embedding	76.37	79.55	77.93
Baseline + Stroke embedding(BiLSTM)	75.03	79.94	77.41
Baseline + Stroke embedding(CNN)	75.83	79.80	77.76
Baseline + all	77.49	79.84	78.65

When the loss function improves, the F1 score increases by 0.59% in comparison to the Baseline model. And the F1 score rises by 0.67% when the part-of-speech element is included. When adding stroke features, BiLSTM and CNN are used to extract stroke features, the F1 score is enhanced by 0.15% using BiLSTM and by 0.50% using CNN. Each component of the model described in this research is found to be effective. It can be seen from the results that CNN is more suitable for stroke feature extraction; thus, CNN is used in our model to extract stroke features.

5. Conclusions

In this study, an NER method fusing part-of-speech and stroke features is designed for medical text. In the embedding layer, the part-of-speech data and stroke data of words are embedded as multiple features to distinguish words with different semantics, and the features of stroke information are extracted by CNN using two convolution kernels with different sizes. After that, the part-of-speech and stroke feature vectors are spliced with the word vector, respectively, and input into the BiLSTM and CRF layer for training, wherein the word vectors are obtained by the BERT pre-training model. From the experimental data, it can be seen that using the BERT pre-training model to obtain word vectors makes the entity recognition effect better because BERT is a depth model with high semantic information extraction capabilities that was trained using a vast amount of unlabeled data. Moreover, using the BiLSTM layer can extract features better than not using it, which makes the results better. The embedding layer uses BERT pre-training models, the encoder uses BiLSTM, and the decoder uses CRF, which shows better performance than other combinations. The part-of-speech and stroke features used in the model suggested in this paper, as part of the word features, make it simpler to distinguish between words with different semantics while also making it easier to identify similar words. This makes it possible to more accurately identify the boundaries of long medical entities and improves the identification effect of medical entities, providing better services for subsequent tasks such as building medical knowledge maps. Additionally, the difference in data volume between various categories of samples is balanced by adding a class-weighted loss function.

Our research is limited in its ability to process long sentences since the BERT processing sequence is limited to a certain length. Additionally, because our model processes input text using a character-based approach, text features lack word information. In further studies, we will strengthen our model, to address these issues and investigate the impact of additional semantic elements and labeling strategies on boundary word recognition.

Author Contributions: Conceptualization, F.Y. and H.L.; methodology, F.Y.; validation, Y.W., S.W., C.S. and P.F.; investigation, J.Z.; writing—original draft, F.Y.; writing—review and editing, F.Y. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Hunan Province (2021JJ30456, 2021JJ30734), the Open Research Project of the State Key Laboratory of Industrial Control Technology (No. ICT2022B60), and the National Defense Science and Technology Key Laboratory Fund Project (2021-KJWPDL-17), the National Natural Science Foundation of China (61972055).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
2. Luo, L.; Yang, Z.; Yang, P.; Zhang, Y.; Wang, L.; Lin, H.; Wang, J. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics* **2018**, *34*, 1381–1388. [CrossRef] [PubMed]
3. Tarasova, O.A.; Rudik, A.V.; Biziukova, N.Y.; Filimonov, D.A.; Poroikov, V.V. Chemical named entity recognition in the texts of scientific publications using the naive Bayes classifier approach. *J. Cheminform.* **2022**, *14*, 55. [CrossRef] [PubMed]
4. Liu, J.; Ye, L.; Zhang, H.; Guo, X. Named entity recognition of legal judgment based on small-scale labeled data. In Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies, Guangzhou, China, 4–6 December 2020; pp. 549–555. [CrossRef]
5. Donnelly, J.; Roegiest, A. The Utility of Context When Extracting Entities from Legal Documents. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, New York, NY, USA, 19–23 October 2020; pp. 2397–2404. [CrossRef]
6. Aguilar, G.; Maharjan, S.; López-Monroy, A.P.; Solorio, T. A multi-task approach for named entity recognition in social media data. *arXiv* **2019**, arXiv:1906.04135. [CrossRef]

7. Ruokolainen, T.; Kauppinen, P.; Silfverberg, M.; Lindén, K. A Finnish news corpus for named entity recognition. *J. Lang. Resour. Eval.* **2020**, *54*, 247–272. [CrossRef]
8. Gaio, M.; Moncla, L. Extended named entity recognition using finite-state transducers: An application to place names. In Proceedings of the Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2017), Nice, France, 19–23 March 2017; pp. 15–20.
9. Zhu, Q.; Li, X.; Conesa, A.; Pereira, C. GRAM-CNN: A Deep Learning Approach with Local Context for Named Entity Recognition in Biomedical Text. *Bioinformatics* **2018**, *34*, 1547–1554. [CrossRef]
10. Yoon, W.; So, C.H.; Lee, J.; Kang, J. CollaboNet: Collaboration of Deep Neural Networks for Biomedical Named Entity Recognition. *Bioinformatics* **2019**, *20* (Suppl. S10), 249. [CrossRef]
11. Popovski, G.; Kochev, S.; Seljak, B.; Eftimov, T. FoodIE: A Rule-Based Named-Entity Recognition Method for Food Information Extraction. In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, Prague, Czech Republic, 19–21 February 2019; pp. 915–922. [CrossRef]
12. Gabbard, R.; DeYoung, J.; Lignos, C.; Freedman, M.; Weischedel, R. Combining Rule-Based and Statistical Mechanisms for Low-Resource Named Entity Recognition. *Mach. Transl.* **2017**, *32*, 31–43. [CrossRef]
13. Gorinski, P.J.; Wu, H.; Grover, C.; Tobin, R.; Talbot, C.; Whalley, H.; Sudlow, C.; Whiteley, W.; Alex, B. Named Entity Recognition for Electronic Health Records: A Comparison of Rule-Based and Machine Learning Approaches. *arXiv* **2017**, arXiv:1903.03985. [CrossRef]
14. Patil, N.; Patil, A.; Pawar, B.V. Named Entity Recognition Using Conditional Random Fields. *Procedia Comput. Sci.* **2020**, *167*, 1181–1188. [CrossRef]
15. Suthaharan, S. Support Vector Machine. *Mach. Learn. Model. Algorithms Big Data Classif.* **2016**, *36*, 207–235. [CrossRef]
16. Morwal, S.; Jahan, N.; Chopra, D. Named Entity Recognition Using Hidden Markov Model (HMM). *Int. J. Nat. Lang. Comput.* **2012**, *1*, 15–23. [CrossRef]
17. Setiyoaji, A.; Muflikhah, L.; Fauzi, M. Named entity recognition menggunakan hidden markov model dan algoritma viterbi pada teks tanaman obat. *J. Pengemb. Teknol. Inf. Dan Ilmu Komput. e-ISSN* **2017**, *2548*, 964X.
18. Szarvas, G.; Farkas, R.; Kocsor, A. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In Proceedings of the Discovery Science, Berlin/Heidelberg, Germany, 7–10 October 2006; pp. 267–278. [CrossRef]
19. 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.
20. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991. [CrossRef]
21. Strubell, E.; Verga, P.; Belanger, D.; McCallum, A. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. *arXiv* **2017**, arXiv:1702.02098. [CrossRef]
22. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
24. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
25. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-Training with Whole Word Masking for Chinese BERT. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
26. Wang, X.; Zhang, Y.; Ren, X.; Zhang, Y.; Zitnik, M.; Shang, J.; Langlotz, C.; Han, J. Cross-Type Biomedical Named Entity Recognition with Deep Multi-Task Learning. *Bioinformatics* **2019**, *35*, 1745–1752. [CrossRef] [PubMed]
27. Cho, M.; Ha, J.; Park, C.; Park, S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *J. Biomed. Inform.* **2020**, *103*, 103381. [CrossRef] [PubMed]
28. Chang, Y.; Kong, L.; Jia, K.; Meng, Q. Chinese Named Entity Recognition Method Based on BERT. In Proceedings of the 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 29–31 October 2021; pp. 294–299. [CrossRef]
29. Yao, L.; Liu, H.; Liu, Y.; Li, X.; Anwar, M.W. Biomedical Named Entity Recognition Based on Deep Neutral Network. *Int. J. Hybrid Inf. Technol.* **2015**, *8*, 279–288. [CrossRef]
30. Lin, B.Y.; Xu, F.; Luo, Z.; Zhu, K. Multi-Channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media. In Proceedings of the 3rd Workshop on Noisy User-Generated Text, Copenhagen, Denmark, 7 September 2017; pp. 160–165. [CrossRef]
31. Yu, J.; Jiang, J.; Yang, L.; Xia, R. Improving Multimodal Named Entity Recognition via Entity Span Detection with Unified Multimodal Transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3342–3352. [CrossRef]