electronics

MDPI

Article

An RG-FLAT-CRF Model for Named Entity Recognition of Chinese Electronic Clinical Records

Jiakang Li [1,2], Ruixia Liu [1], Changfang Chen [1], Shuwang Zhou [1,3], Xiaoyi Shang [1] and Yinglong Wang [1,*]

[1] Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China; lijiakang178@163.com (J.L.); liurx@sdas.org (R.L.); chenchangfang012@163.com (C.C.); zhoushw@sdas.org (S.Z.); shangxy@sdas.org (X.S.)
[2] Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China
[3] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
* Correspondence: wangylscsc@126.com

Abstract: The goal of Clinical Named Entity Recognition (CNER) is to identify clinical terms from medical records, which is of great importance for subsequent clinical research. Most of the current Chinese CNER models use a single set of features that do not consider the linguistic characteristics of the Chinese language, e.g., they do not use both word and character features, and they lack morphological information and specialized lexical information on Chinese characters in the medical field. We propose a RoBerta Glyce-Flat Lattice Transformer-CRF (RG-FLAT-CRF) model to address this problem. The model uses a convolutional neural network to discern the morphological information hidden in Chinese characters, and a pre-trained model to obtain vectors with medical features. The different vectors are stitched together to form a multi-feature vector. To use lexical information and avoid the problem of word separation errors, the model uses a lattice structure to add lexical information associated with each word, which can be used to avoid the problem of word separation errors. The RG-FLAT-CRF model scored 95.61%, 85.17%, and 91.2% for F1 on the CCKS 2017, 2019, and 2020 datasets, respectively. We used statistical tests to compare with other models. The results show that most *p*-values less than 0.05 are statistically significant.

Keywords: clinical named entity recognition; Chinese medical text; pre-trained model

Citation: Li, J.; Liu, R.; Chen, C.; Zhou, S.; Shang, X.; Wang, Y. An RG-FLAT-CRF Model for Named Entity Recognition of Chinese Electronic Clinical Records. *Electronics* **2022**, *11*, 1282. https://doi.org/10.3390/electronics11081282

Academic Editors: Agnieszka Konys and Agnieszka Nowak-Brzezińska

Received: 9 March 2022
Accepted: 15 April 2022
Published: 18 April 2022

1. Introduction

Informatization has penetrated all aspects of social life. In the medical field, more and more hospitals are building information systems to improve their service level and core competitiveness, effectively use limited medical resources, and provide patients with high-quality treatment. These information systems can not only improve doctors' efficiency but also enhance internal management, making information communication among departments more efficient and simplifying and standardizing the medical treatment process. Medical staff can be released from tedious and repetitive work, with extra time and energy being used to provide better patient services.

Existing medical systems have generated countless medical data, and if the data cannot be used effectively, it will be a waste of professional knowledge. As a medical record, Electronic Medical Record (EMR) has received great attention in scientific research [1] because it contains complete and detailed clinical information generated by patients during each visit. EMR refers to the digital information such as words, symbols, charts, graphics, data, images, and so on, generated by medical personnel using the information system of medical institutions in medical activities. EMR contains various information such as text and medical images. Medical images are mainly the results of laboratory tests of patients, such as CT and B-ultrasound. These medical images can currently be analyzed

using pattern recognition and machine learning methods, but EMR also contains much textual data. To make use of the text data, Natural Language Processing (NLP) technology is essential. Electronic medical records cover all patient information from admission to discharge, including admission time, symptoms, body parts, examination methods, medication, and other physical information [2]. Medical services may consider providing patients with the facility to submit inquiries in the form of comments [3].

EMR information extraction is to identify various medical entities from texts and establish relationships among them. The information extraction of EMR was first carried out on English medical records, and many achievements have been achieved, while domestic research on Chinese EMR is still in its infancy. Therefore, it is our top priority.

Named Entity Recognition (NER) is the foundation of text data mining and information processing. For entity recognition in the medical field, it refers to identifying entities such as symptoms, body parts, examinations, etc. Identifying this information and analyzing the relationship among different entity information plays an indispensable role in establishing a knowledge map in the medical field, building an auxiliary diagnosis model, and providing data support for clinical decision-making.

Early NER systems are mainly rule-based approaches. This method extracts the target entity through the preset rule template and has achieved certain results. Although for some uncommon fields, experts need to write rules, which is demanding, time-consuming, and limited, rule-based approaches are not outdated but are still an important complement to other approaches.

Feature-based Supervised Learning Approaches transform NER tasks into classification tasks or sequence labeling tasks. Conditional Random Fields (CRF) and Hidden Markov Models (HMM) [4] are two common algorithms.

With the rapid expansion of deep learning, Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) are applied to the CNER tasks [4]. Alam et al. [5] proposed a new framework based on association rule mining for prognostic factor identification in malignant mesothelioma. At present, the integration of LSTM and CRF is a common method. However, there are limitations. Transformer [6] proposes self-attention, enabling the LSTM networks to solve long-distance dependencies. Transformers gradually replaced LSTM as the mainstream feature extractor in NLP.

Unsupervised pre-trained models are suitable for general domains but not appropriate for the medical domain.

In addition, Chinese NER is related to word segmentation. Since Chinese entities are generally composed of words, word segmentation errors will lead to errors in Chinese NER. The character-based Chinese NER model cannot fully utilize the information of words. The Lattice-LSTM proposed by Zhang et al. [7] improves the accuracy of this task by adding dictionary information to the model. However, due to the complexity of the lattice structure, it does not support parallel computing. Li et al. [8] proposed a Flat Lattice Transformer (FLAT), which uses a flatten lattice structure and transformer to realize parallel processing. At the same time, FLAT uses the calculation method of relative position in the Transformer-XL model [9], and by adding additional position information in the Transformer structure, it solves the modeling of long text and captures ultra-long distance dependencies.

Also, unlike other languages, Chinese is a pictograph. Chinese characters contain rich semantic information. Many words with similar meanings are similar in composition and structure of Chinese characters, which is especially obvious in the medical field. The glyph information of Chinese characters is also of significant reference value. Glyce, proposed by Meng [10], can extract the glyph vectors of Chinese characters. It attempts to extract the semantics of Chinese characters from various ancient and modern Chinese characters and various writing styles, and the performance is improved.

To solve problems, we propose a RoBerta Glyce-Flat Lattice Transformer-CRF (RG-FLAT-CRF) model suitable for Chinese CNER tasks. First, the glyph vector is obtained by Glyce, the character vector and word vector are obtained by Word2vec [11], and the character vector obtained by RoBerta is spliced with the glyph vector and the word vector

obtained by Word2vec. At the same time, the Flat-lattice structure is used, word information is added, the head position code and tail position code are constructed for each character and vocabulary, and the relative position code is calculated. The concatenation of vectors and the corresponding position encoding are sent to a transformer to extract the context information of every Chinese character. Finally, we jointly decode the labels of the entire sentence using CRF. Our main contributions are as follows:

Contribution

In our contributions, we have:

1. A RoBerta Glyce-Flat Lattice Transformer-CRF model is proposed, which can make full use of the glyph information and language features of Chinese medical texts, has strong coding and text representation capabilities, and can accurately identify various types of Chinese electronic clinical Entity records.
2. According to the particularity of medical entities and the language characteristics of Chinese, a multi-feature fusion vector is constructed. The pre-trained model is used to obtain vectors that conform to medical characteristics. At the same time, to strengthen the semantic representation of medical entities, convolutional neural networks are used to extract the glyph features of medical entities, and different character vectors are spliced together to form a composite character vector.
3. Use of a lattice structure to add potential lexical information to each word to avoid word segmentation errors. The relative position vector in the improved transformer directly captures the dependencies between words and vocabulary, makes full use of case information, and can be implemented in parallel.

The rest of this article is organized as follows. Section 2 provides a brief review of related work of NER. The proposed model is presented in Section 3. The relevant content of the experiment is described in detail in Section 4. Finally, Section 5 gives the conclusions.

2. Related Work

We include the following studies: (1) How to enhance the semantic representation of Chinese word vectors. (2) Feature extraction networks more applicable to the Chinese language. (3) The characteristics and difficulties of named entity recognition in Chinese electronic medical records. (4) Related Evaluation Metrics [12]. We used multiple strings such as “Chinese electronic medical record named entity recognition”, “Chinese named entity recognition”, and “medical named entity recognition” to retrieve peer-reviewed articles using Multiple databases, including Scopus, ACM Digital Library, IEEE Xplore, ScienceDirect, SpringerLink, and Google Scholar [13].

This section primarily provides a brief introduction to rule-based and dictionary-based methods, machine learning-based methods, and deep learning-based methods. Then, the representation method of the word vector is introduced.

2.1. Rule-and-Dictionary-Based Clinical Named Entity Recognition

Nowadays, Rule-and-Dictionary-Based CNER is commonly used, and these methods benefit from the development of professional medical dictionaries. Researchers complete the NER task by pattern matching according to the belonging list in the dictionary. Friedman et al. [14] developed a clinical document processor that recognized medical information in the medical record and mapped this information into a structured representation containing medical terms. Fukuda et al. [15] proposed a method to identify the names of substances such as proteins from biological papers, using the characteristics of proper noun descriptions in the professional field, which eliminates the need to prepare a professional term dictionary in advance. Names can be extracted with precision, whether they are known or newly defined or are single or compound words.

The completeness and accuracy of the dictionary and the accuracy of the matching algorithm can determine the accuracy of such methods. Therefore, dictionary-based methods are more suitable for fields where proper nouns are fixed and updated infrequently. In the

biomedical field, there are problems such as the fast updating of proper nouns and different expressions of the same entity name. Experts need to spend much time and effort writing rules, and the cost is high. In addition, different rules are needed for different systems. They are of poor portability and are hard to reuse quickly.

2.2. Clinical Named Entity Recognition Based on Machine Learning

In the past, traditional machine learning based on CNER has been widely used, including HMM, CRF, Support Vector Machine (SVM) [16], Naive Bayesian Model (NBM) [17], etc. Settles [18] used combined feature sets with CRF in biomedical NER tasks. Tang [19] developed an SVM-based NER system for medical entities in the medical record. Roberts et al. [20] utilized SVM with a manually constructed dictionary to classify. Liu [21] evaluated the contribution of different features in the CRF-based CNER task.

Compared with the methods analyzed in Section 2.1, the method in Section 2.2 does not require the experimenter to master much language knowledge, thus saving time and effort. However, this type of method requires a lot of energy to design features. The effect of the model depends on the designed features. With deep learning modeling, the feature extraction problem in traditional machine learning can be addressed.

2.3. Deep-Learning-Based Clinical Named Entity Recognition

Recently, we have witnessed the great success of deep learning in the field of NLP, such as NER and event extraction tasks. Commonly used network models include Convolutional Neural Networks (CNN) [22], Recurrent Neural Networks (RNN) [23], and LSTM. Ma et al. [24] proposed the Bi-directional LSTM-CNNs-CRF model, character-level representations are extracted using CNN, Bi-directional LSTM (BiLSTM) is responsible for modeling the contextual information of each word. Xu et al. [25] combined bidirectional LSTM and CRF based, BiLSTM-CRF model can learn the information features of a given dataset and achieved a score of 0.8022 at NCBI, outperforming many widely used baseline methods. Yin et al. [26] used convolutional neural nets for Chinese character radical feature extraction and captured the correlation between characters using self-attentiveness. Kong et al. [27] proposed a Chinese medical named entity recognition based on a multi-layer CNN and attention mechanism, constructing a multi-layer CNN to extract short-term and long-term memories and using an attention mechanism to capture global information. However, the above deep neural network-based CNER methods cannot model the ambiguity of Chinese.

The BERT-BiLSTM-CRF model was proposed by Jiang et al. [28] to be applied to CNER. The semantic representation of words was enhanced with a BERT pre-trained language model, and the BiLSTM was to learn contextual information. Qin et al. [29] proposed a BERT-BiGRU-CRF model in the field of Chinese electronic medical records, which uses BERT to convert the electronic medical record text into low-dimensional vectors and BiGRU to obtain contextual features. Wu et al. [30] used a bi-directional LSTM model to learn a medical entity's partial head information using Roberta to learn medical features. Wang et al. [31] used information from medical encyclopedias as additional information to enhance the recognition of Chinese electronic medical record entities. However, these models do not fully consider the characteristics of medical domain data, and it is not very effective in medical entity extraction.

2.4. Research Status of Word Vector Representation Methods

If you want to reflect a word in a text and perform mathematical calculations, it must be done through word embedding. The bag-of-words model simply represents words without any semantic features. As the number of words increases, so does the dimension. Researchers propose a way to solve this problem using a pre-trained language model for word representation. Pre-training refers to obtaining a training model independent of subsequent tasks from a large-scale corpus using self-supervised learning. The model can be transferred to other tasks, thereby reducing the training burden of subsequent tasks. The Word2Vec model was proposed by Mikolov et al. to obtain vectors. The GloVe

algorithm was proposed by Pennington et al. [32]. In recent years, pre-trained models have received increasing attention. Since this type of model is a context-independent word vector trained by static pre-training technology, it cannot accurately model the polysemy of a word. Therefore, Peters et al. [33] proposed the ElMo algorithm. The bidirectional LSTM network structure was used for context encoding, which could effectively capture context information.

2.4.1. Models for BERT and Its Variants

Devlin et al. [34] proposed Bidirectional Encoder (BERT). The emergence of Bert opened a new era of research in the field of NLP. Then some improved pre-training models based on BERT, mainly including ERNIE [35], BERT-WWM [36], RoBerta [37], and XLNet [38]. The ERNIE model is pre-trained using massive corpora in multiple fields, including encyclopedias, news, forums, etc. BERT-WWM's improvement over BERT is to replace a complete word with a Mask label instead of a subword. The RoBerta model uses a dynamic mask mechanism for pre-training, cancels the NSP task, and expands the batch size. As an auto-regressive model, the XLNet model can expand the language model and increase the prediction of bidirectional words, the above predicting the next word and the following predicting the previous words.

2.4.2. Research on Chinese Characters

The structure of Chinese characters is different from that of English. Chinese characters are pictographs, and their glyphs also contain rich meanings. Therefore, many scholars have carried out characterization studies on the glyph features of Chinese characters. Sun [39] proposed to learn the radical features of Chinese. Wang et al. [40] proposed a Chinese character root and stroke-enhanced embedding method for learning Chinese character roots from the internal information of semantics and form. Wei [41] proposed a visual embedding method for semantic association among visual words, segmented the glyph, spliced the average embedding vectors corresponding to each sub-region, and converted it into a fixed-length vector for keyword detection. Su [42] used convolutional autoencoders to learn glyph features from images of traditional Chinese characters and introduced glyph features during training using the corpus. Meng [6] proposed the Glyce model. It tried to extract the semantics of Chinese characters from various ancient and modern Chinese characters and various writing styles, and the performance was improved.

These are the characteristics of Chinese, which improve CNER tasks. However, the current mainstream CNER methods cannot integrate the pre-trained model with the Chinese glyph information.

3. Proposed Method

In the NER task, the character sequence of the input text is represented by $X = (x_1, x_2, \ldots, x_n)$. The labels of the input text are represented by $Y = (y_1, y_2, \ldots, y_n)$. The goal of a NER system is to predict the correct sequence Y of labels for the text given the known sequence of characters X of the text. The RG-FLAT-CRF model proposed in this chapter consists of three parts; the embedding layer, the encoding layer, and the decoding layer. The overall structure is shown in Figure 1.

The model first matches the latent words related to the character in the input text and splices the character information and words information into the embedding layer. The embedding layer consists of three parts, and the character vector is spliced after processing by RoBerta, Glyce, and Word2vec. The word vector is obtained using Word2vec, head and tail position encoding are constructed for each character and word, and the relative position encoding is calculated. The concatenation of word vectors and the corresponding position encoding are input into the encoding layer, consisting of a Transformer neural network that captures deep features and encodes the input sequence. Finally, the output of the encoding layer is input to the decoding layer, which predicts the final label sequence.

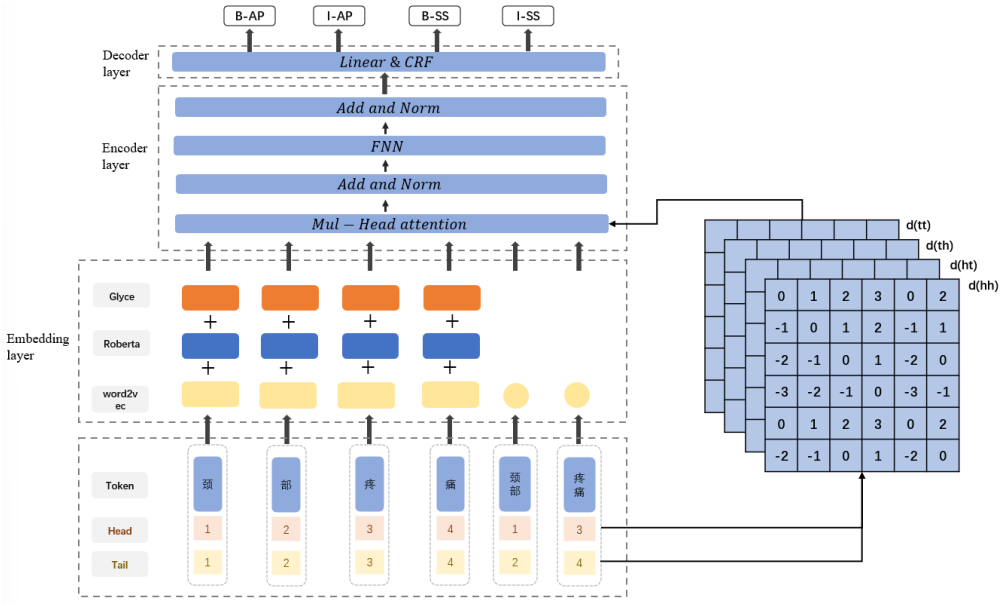

Figure 1. Model structure diagram of RG-FLAT-CRF.

This study uses NER to perform entity recognition on Chinese EMR. Specific steps are as follows:

(1) Electronic medical record data preprocessing, that is, the original electronic medical record text data set is processed, and the electronic medical record text set is represented as $J = (j_1, j_2, \ldots, j_n)$, where the i-th electronic medical record text is represented as j_i. The predefined entity category $C = (c_1, c_2, \ldots, c_m)$, is divided and annotated according to the character level, and the characters and predefined categories are separated by spaces when annotating.

(2) Establish a Chinese EMR text training dataset.

(3) Model training, that is, training the RGT-CRF model. Take the Chinese EMR test text set $J_{test} = (j_1, j_2, \ldots, j_N)$ as input and take the entity and its corresponding category pair as output: $\{\langle m_1, c_1\rangle, \langle m_2, c_2\rangle, \ldots, \langle m_p, c_p\rangle\}$. The entity m_i represents the entity that appears in the document, and b_i and e_i represent the start and end positions of m_i, respectively. There is no need to overlap between entities; that is, $e_i < b_i + 1$. C_{m_i} represents the predefined category of entity m_i, calculates the F1 score according to the precision and recall rate, and uses the F1 score as the comprehensive evaluation index of the model.

3.1. Embedding Layer

The embedding layer consists of three parts: RoBerta layer, Glyce layer, and Word2vec layer:

(1) RoBerta layer: the model adopts the better pre-training model RoBerta to capture the characteristics of medical text and converts each word of medical text into a low-dimensional vector form through RoBerta.

(2) Glyce layer: scan each word in the sentence to obtain the glyph vector corresponding to each word, and enhance the representation of the word.

(3) Word2vec layer: Using Word2vec, the vector representation of each word in the medical text and the vector representation of the latent words can be obtained to enrich the semantic representation.

The character vectors processed by RoBerta, Glyce, and Word2vec are spliced to obtain multi-feature word vectors, and then the character vectors and word vectors processed by Word2vec are spliced together.

3.1.1. RoBerta

Pretrained language models are often used in NER tasks to generate richer semantic representations. BERT and its variant RoBerta are widely used in research. We use RoBerta for text encoding instead of BERT. Compared with BERT, the model structure of RoBerta has not changed. They are all composed of 12 stacked transformers. Each layer has a hidden state of 768 dimensions. Each Transformer uses a 12-head self-attention mechanism. The only thing that has changed is the pre-training method. Dynamic masks and text encoding are adopted to remove the NSP task and use more data to train the model.

The vector is obtained through the RoBerta. The RoBerta structure is shown in Figure 2. The input text is $Z = \{Z_1, Z_2, \ldots, Z_x\}$. First, the sequence is vectorized. This part consists of token embedding, clause embedding, and position embedding. These three embedding layers are essentially equivalent to the static embedding layers, and the table lookup is performed by the embedding matrix. For the x-th token in the processed token sequence, the vector calculation is as follows:

$$e_x = W_t(E_{t_x}) + W_s(E_{s_x}) + W_p(E_x) \tag{1}$$

where W_t, W_s, W_p are the token embedding matrix, the clause embedding, and the position matrix.

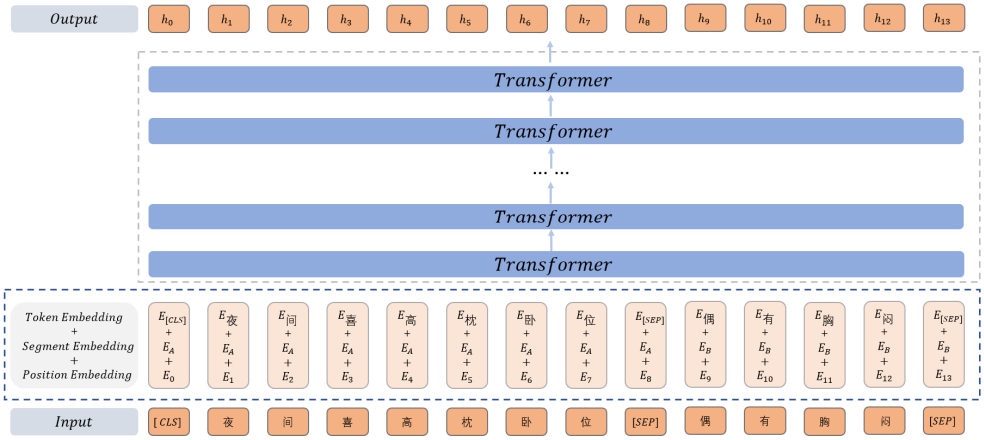

Figure 2. Structure diagram of RoBerta.

Token Embeddings represent the Embedding vector of each word. Segment Embeddings are used to distinguish different sentences before and after punctuation marks. Position Embeddings represent the embeddings of a word's position. The input feature of RoBerta is the sum of the above 3 embeddings. "[CLS]" is used as the starting symbol of the input, indicating that the feature can be used in the classification model. "[SEP]" indicates the clause symbol, which is used to cut off the clauses in the sentence.

The obtained vector is input into the stacked Transformer to extract features. The final output is the result of encoding the input sentence text. Finally, we obtained the sentence representation vector with the dependency information among words and words in the sentence text. The calculation is as follows:

$$H = Mul_{trans}(E) \tag{2}$$

where $Mul_{trans}(.)$ represents the stacked Transformer, outputting the text encoding of the entire sentence through the last layer H, which can be expressed as $H = h_0, h_1, \ldots, h_x$. Here h_x is the text representation vector to the xth token.

3.1.2. Glyce

Chinese characters are pictographs, and most Chinese characters are evolved from graphics. Chinese characters contain rich semantic information, especially in the medical

field. Most of the words for diseases have the same parts. Therefore, we believe that adding glyph information to word vectors can enhance the representation of characters.

Glyce used different versions of the writing method, as well as different writing to enhance the representation of the characters.

Glyce is different from traditional CNN. There are about 100,000 Chinese characters, but only a few thousand are commonly used. Compared with classification on the ImageNet dataset. There're few training examples for Chinese characters. Compared with the size of Imagenet images, Chinese images are usually smaller, with a size of 12 × 12. Thus according to the Chinese writing habits, a 2 × 2 Tianzi lattice structure is used. As shown in Figure 3, this structure can reflect the glyph information of Chinese, including components such as radicals, which is suitable for the extraction of glyph information.

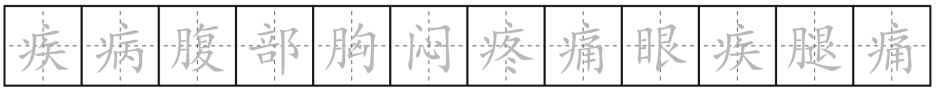

Figure 3. Schematic diagram of the Tianzi lattice.

The structure of Glyce Tianzi lattice-CNN is shown in Figure 4. The processing process is shown in Figure 5. To capture lower-level graph features, the input image approximation firstly passes through a convolutional layer with kernel size 5. In addition, the convolutional layer has to increase the number of feature channels to 1024. Then we apply a max-pooling layer with a pooling kernel of 4 × 4 to perform feature downsampling. After this, the resolution is reduced from 8 × 8 to 2 × 2. This 2 × 2 Tianzi lattice structure shows the glyph features of Chinese characters, and finally, we apply the group convolution operation to map the Tianzi lattice to the final output.

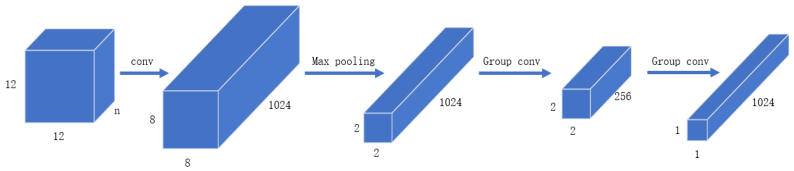

Figure 4. CNN structure diagram in Glyce.

layer	kernel	Output
input		n×12×12
Conv2d	5	1024×8×8
Relu		1024×8×8
Max pool	4	1024×2×2
8 group conv	1	256×2×2
16 group conv	2	1024×1×1

Figure 5. The Tianzi lattice—CNN structure.

For the input text $Z = \{Z_1, Z_2, \ldots, Z_x\}$, the glyph vector obtained by Glyce is $E_G = (e_{G0}, e_{G1}, \ldots, e_{Gx})$ as shown in Figure 6.

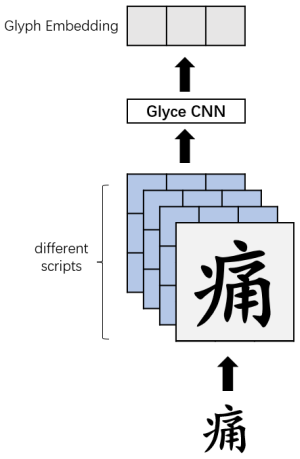

Figure 6. Glyce character embedding.

3.1.3. Word2vec

We use Word2vec to get word vectors, a typical representative of distributed representation. Compared with one-hot, Word2vec takes into account the relationships among words. In addition, Word2vec also optimizes the training efficiency of the model, so it is used more frequently.

3.2. Position Encoder

Chinese NER tasks are often considered sequence labeling tasks. By calculating the probability of each character corresponding to each entity type label, The label with the highest probability is used as the final identification result. There are usually two vectorization methods to vectorize Chinese characters into the model calculation: methods based on word vectors and methods based on character vectors.

The first task of the word vector-based model is to segment the text into the form of words. The improvement effect of word vectors on entities is significant. The word contains more semantic information, but if there is a false classification, it will affect the results of NER.

For instance, in Figure 7, this sentence can be divided into '济南人 (Jinan People)', '和 (and)', '山庄 (Mountain Villa)', and can also be divided into '济南 (Jinan People)', '人和山庄 (Renhe Mountain Villa)'. These two-word segmentation methods have a great impact on recognition.

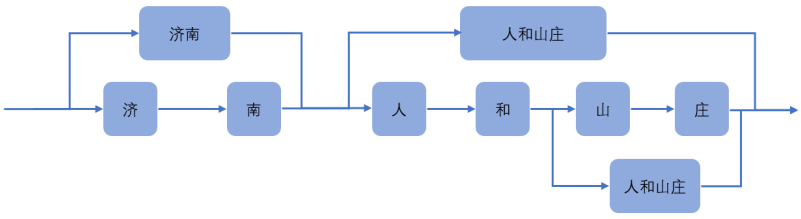

Figure 7. Structure diagram of Lattice.

Using character vector-based models avoids word segmentation error information but lacks lexical information. For example, '感冒 (cold)', separate the word '感 (feel)' and '冒 (emit)' represent different semantic information. '感 (feel)' means feeling, and '冒 (emit)' means to penetrate outward or rise upward. It is difficult to express the information of the word '感冒 (cold)' in medicine after '感 (feel)' and '冒 (emit)' are separated, which is especially obvious in the medical field.

To address the above problems, we adopted the FLAT-lattice structure, shown in Figure 8. This structure uses both character vectors and word vectors. Based on character vectors, the latent vocabulary of each character is matched, and the word vectors are added

to the model. This method utilizes the semantic relationship of words and avoids the phenomenon of word segmentation errors.

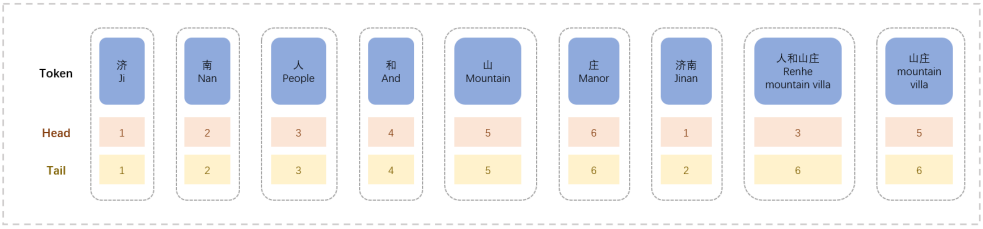

Figure 8. Structure diagram of Flat-lattice.

After using the dictionary to obtain lattice information from the string, it is flattened, and the structure is shown in Figure 8.

These flat lattices can also be defined as spans. A span comprises a token, a head, and a tail. A token is a word or character, and the head represents the starting position of the token in the original sequence, and the tail represents the ending position of the token in the original sequence. For characters, the head and tail are the same. For the matched words, head indicates the start position of the word in the sequence, and tail indicates the end position of the word in the sequence. The flat lattice can preserve the original structure of the lattice and, at the same time, preserve the word order information of the original sentence.

According to the Flat-lattice structure, there are three interrelationships, intersection, involvement, and separation. We use relative position encoding to encode the positional relationship among each span. Relative position encoding does not directly model the interaction relationship but obtains a dense vector by computing a set of head and tail changes. Not only the interrelationships among spans can be represented, but more detailed sequence relationships can be shown, such as the distance among words and characters. Let $tail_x$ and $tail_x$, $head_y$ and $tail_y$ denote the head and tail positions of s_x and s_y, respectively. Four kinds of relative distances can be used to represent the relative relationship between s_x and s_y. Their calculation formulas are as follows:

$$r_{xy}^{hh} = head_x - head_y \tag{3}$$

$$r_{xy}^{ht} = head_x - tail_y \tag{4}$$

$$r_{xy}^{th} = tail_x - head_y \tag{5}$$

$$r_{xy}^{tt} = tail_x - tail_y \tag{6}$$

where r_{xy}^{hh} stands for the distance from the head of s_x to the head of s_y, r_{xy}^{ht} is the distance from the head of s_x to the tail of s_y, r_{xy}^{th} represents the distance from the tail of s_x to the head of s_y, r_{xy}^{tt} is the distance from the tail of s_x to the tail of s_y. The final relative position encoding is a nonlinear transformation of the four distances, which can be calculated like:

$$L_{xy} = ReLU\left(W_l\left(P_{r_{xy}^{hh}} \bigoplus P_{r_{xy}^{ht}} \bigoplus P_{r_{xy}^{th}} \bigoplus P_{r_{xy}^{tt}}\right)\right) \tag{7}$$

among them, W_l is a learnable parameter, $\oplus$ represents the connection operator, and the calculation method of P_r refers to the calculation method of the transformer. The calculation is as shown in the equation:

$$P_r^{2k} = \sin \frac{r}{1000^{\frac{2k}{d_{model}}}} \tag{8}$$

$$P_r^{2k+1} = \cos \frac{r}{1000^{\frac{2k}{d_{model}}}} \tag{9}$$

3.3. Encoder

The encoding layer consists of Transformers, which aim to extract semantic and temporal features from the context automatically.

Before the transformer appeared, most NER used BiLSTM as the model's encoder. However, BiLSTM has some problems: (1) The sequential nature of the recurrent neural network represented by LSTM hinders the parallelization of training samples; (2) The problem of long-term dependence cannot be completely solved.

Transformer avoids recurrent model structure and uses attention mechanism for modeling. The structure is shown in Figure 9. We used its encoding part, which consists of two parts, a feedforward network and a multi-head self-attention layer, both of which have a residual network. Multi-head self-attention consists of stacked self-attentions, all accompanied by a "layer normalization" step.

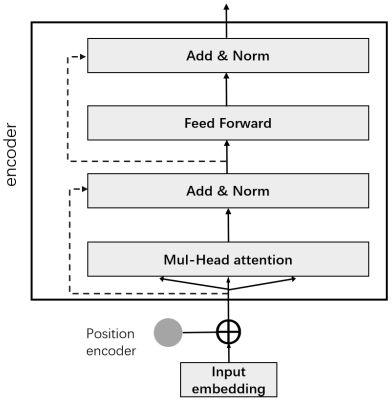

Figure 9. Structure diagram of Transformer.

When the encoder encodes this word, the self-attention mechanism can take other words in this sentence into consideration.

First, we send the vector output of the embedding layer and the corresponding relative position encoding to the encoding layer, using the encoding layer of the transformer. A Query vector, a Key vector, and a Value vector are created for each word by this self-attention mechanism. They are obtained through the vector multiplication by the three matrices we trained. Their calculation formula is as follows:

$$Q = Linear(X) = XW_q \tag{10}$$

$$K = Linear(X) = XW_k \tag{11}$$

$$V = Linear(X) = XW_k \tag{12}$$

The second step is to calculate the score, which will make the gradient more stable, and then it is divided by $\sqrt{d_{head}}$. The traditional Transformer model can capture contextual semantics by adding position information to the input, but there is a problem of sentence errors in the face of text segmentation input. Therefore, extra position information is added to the Transformer structure of the Transformer-XL model, and the absolute vector is converted into a relative vector. Solve the modeling of long text, capture ultra-long distance dependencies, and calculate the attention score vector among input vectors by the formula:

$$A^*_{x,y} = \frac{W_q^\intercal E_{s_x}^\intercal E_{s_y} W_{k,E} + W_q^\intercal E_{s_x}^\intercal L_{xy} W_{k,R} + u^\intercal E_{s_x} W_{k,E} + v^\intercal L_{xy} W_{k,R}}{\sqrt{d_{head}}} \tag{13}$$

where $W_q, W_{k,E}, W_{k,R}, u, v$ are learnable parameters,E_{s_x}, E_{s_y} are the embedded representations of s_x and s_y.

Then pass the result through softmax, which normalizes the scores for all words. For the weighted value vector, the output of the self-attention layer at that position is obtained, and the following is its formula:

$$Attention(A, V) = softmax(A)V \tag{14}$$

The multi-head attention mechanism consists of multiple self-attentions. Define multiple groups of different Q, K, and V, and let them focus on different contexts, respectively. The process of calculating Q, K, V is still the same, except that the matrix of linear transformation has changed from one set of (W_Q, W_K, W_K) to multiple sets of (W_Q, W_K, W_K).

For the input matrix X, each group of Q, K, V can get an output matrix Z. Concatenate the different matrices together and multiply with an additional matrix W_o.

The multi-head attention mechanism enhances the attention layer's performance in two aspects:

(1) It empowers the model with a closer focus on different locations.
(2) Multiple "representation subspaces" are given to the attention layer, and multi-head attention allows us to possess multiple sets of Q, K, and V matrices. After training, each group projects the output into a different representation subspace. The calculation formula is as (15):

$$MH_{att}(A, V) = Concat(head_1, \ldots, head_h)W_o \tag{15}$$

The resulting output is subjected to layer normalization and residual connections. The specific formula is as follows:

$$X_{MH_{att}} = X_{MH_{att}} + X \tag{16}$$

$$X_{MH_{att}} = LayerNorm(X_{MH_{att}}) \tag{17}$$

After the operation of Feedforward, the formulas are shown in equations:

$$X_{hidden} = Linear(ReLu(Linear(X_{attention}))) \tag{18}$$

$$X_{hidden} = X_{attention} + X_{hidden} \tag{19}$$

$$X_{hidden} = LayerNorm(X_{hidden}) \tag{20}$$

3.4. Decoder

The decoding layer consists of CRFs, whose purpose is to resolve the correlation between the output labels to obtain the globally optimal annotation sequence for the text.

For the input sequence $X = (x_1, x_2, \ldots, x_n)$, its predicted label is $Y = (y_1, y_2, \ldots, y_n)$. The score matrix P output by the encoding layer is n×k in size, n is the length of the input sequence, and q is the different types of labels defined. P_{i,y_i} represents the score of the ith character in the sentence on the y_i label. A state transition score matrix A represents the probability score of transition among different labels. $A_{y_i,y_{i-1}}$ represents the transition score from label y_i to label y_{i+1}. y_0, y_{n+1} represent the start tag and the end tag, respectively. Under the condition of the given sequence, the score $S(X, y)$ of the corresponding sequence tag is obtained. The functions can be described as follows:

$$SX, y = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \tag{21}$$

The predicted probability is $P(y|X)$. The calculation formula is shown in (22):

$$P(y|X) = \frac{e^{S(X,y)}}{\sum_{y' \in Y_X} e^{S(X,y')}} \tag{22}$$

The loss function, as shown in the formula:

$$-\log(P(y|X)) = \log \sum_{y' \in Y_X} e^{S(X,y')} - S(X,y) \tag{23}$$

In the last, we adopted the Viterbi algorithm to get the optimal path, that is, a more reasonable predicted label of the input sequence. The calculation formula is as follows (24):

$$y^* = arg_{y' \in Y_X} max\ S(X,y') \tag{24}$$

3.5. Time Complexity Analysis

We discuss the time complexity of the model.

$$O\left(n^2 \cdot d + n \cdot d^2 + \sum_{l=1}^{N} M_l^2 \cdot K_l^2 \cdot C_{l-1} \cdot C_l + n \cdot k^2\right)$$

where n is the sequence length and d is the dimension of embedding. n is the number of convolutional kernels the neural network has; l is the lth convolutional layer of the neural network; C is the number of output channels of the lth convolutional layer of the neural network; and for the lth convolutional layer, the number of input channels C_n is the number of output channels of the l-1st convolutional layer. k is the number of labels as

4. Experiment Design

This section presents the following aspects: the dataset used for the experiments, the labeling rules, the evaluation metrics, and an introduction to the comparative experimental model.

4.1. Dataset

Our proposed RG-FLAT-CRF model is validated with real datasets of three clinical NER tasks.

These three datasets are all from the CCKS competition dataset. The following is the introduction to these datasets.

CCKS-2017 data is adopted for the experiment. Since we did not participate in the competition, we only found some open-source data. The CCKS-CNER2017 dataset. Provides 300 electronic clinical record texts with 29,865 annotated instances (7816 sentences). It is annotated with five entity types: symptoms and signs, diseases and diagnosis, body parts, examinations and tests, and treatment. Table 1 lists its detailed statistics. The proportion of each part of the data is shown in Figure 11.

Table 1. Entity statistics of the three datasets.

Entity Type	CCKS2017	CCKS2019	CCKS2020	Total
symptoms and signs	7831	-	-	7831
diseases and diagnosis	721	5488	5628	11,837
body parts/anatomical parts	10,719	11,468	11,420	33,607
examinations and tests	9546	-	-	9546
treatment	1048	-	-	1048
examinations	-	1302	1626	2928
surgery	-	1182	1136	2318
tests	-	1678	2081	3759
drugs	-	2266	2814	5080
Total	29,865	23,384	24,705	77,954

CCKS-2019 contains 23,384 annotated instances (10,179 sentences). They are annotated with six entity types, namely diseases and diagnosis, examinations, tests, surgery, drugs,

and anatomical parts. The elaborated statistics are shown in Table 1. The proportion of each part of the data is shown in Figure 10.

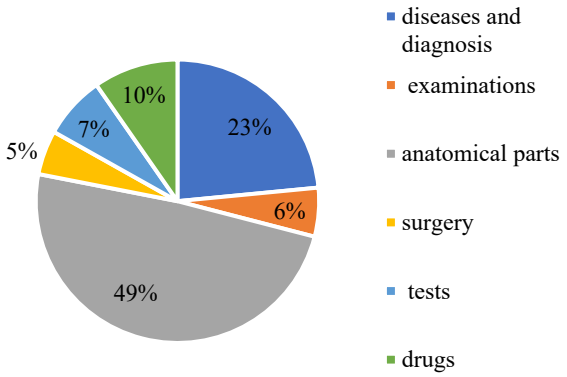

Figure 10. The proportion of medical entities on CCKS2019.

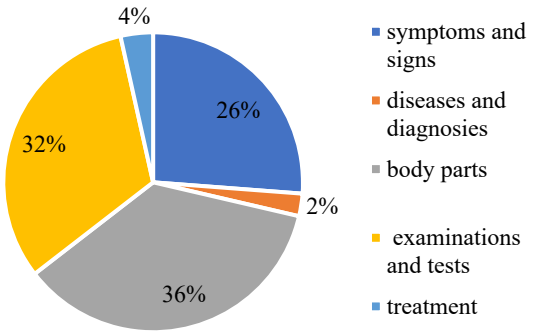

Figure 11. The proportion of medical entities on CCKS2017.

CCKS-2020 contains 24,341 annotated instances (13,308 sentences) with six entity types: diseases and diagnosis, examinations, tests, surgery, drugs, and anatomical parts. Table 1 shows the specific statistics. The proportion of each part of the data is shown in Figure 12.

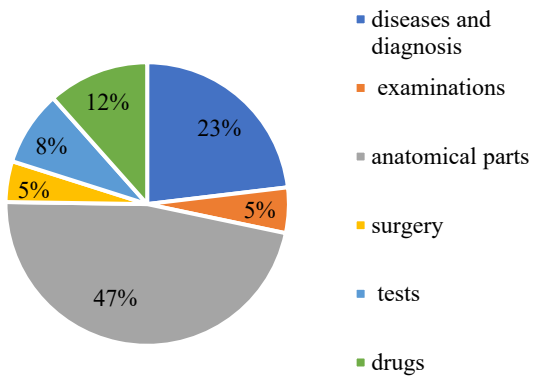

Figure 12. The proportion of medical entities on CCKS2020.

4.2. Labeling Rules

We adopt the BOI rule, where the entity's beginning is represented by B, I is the interior, and O stands for the other categories.

Annotation methods of five entity categories in CCKS2017: SS for symptoms and signs, DD for disease and diagnosis, AP for body parts, EE for inspection and examination, TM for treatment.

Annotation methods of six entity types in CCKS2019 and 2020: DD for disease and diagnosis, GEXA for examination, AP for the anatomical site, SU for surgery, EEXA for the test, and DR for the drug.

4.3. Evaluation Indicators

This paper uses the most common evaluation metrics in the NER field Precision, Recall, and F1 scores are used as the evaluation indicators of the model to evaluate the performance of the evaluation model comprehensively. TP is the number of positive samples predicted as positive samples, FN is the number of positive samples predicted as negative samples, and FP is the number of negative samples predicted as positive samples. They are widely used to evaluate classification and sequence annotation tasks [43].

Precision: The ratio of the number of recognized entities to the number of recognized entities is recorded as Precision, abbreviated as P. The calculation formula is Equation (25).

Recall: The percentage of correctly identified entities out of the number of entities in the sample. The calculation formula is Equation (26).

Both take values between 0 and 1, and the closer the value is to 1, the higher the precision or recall. Precision and recall are sometimes contradictory; a weighted harmonic mean that needs to be considered, and the F_1-score is a combination of the two. The higher the F1 score, the more robust the classification model is. The calculation formula is Equation (27).

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

$$F_1\text{-}score = \frac{2 \times Precision \times Recall}{Recall + Precision} \tag{27}$$

4.4. Experimental Parameters

The parameters of the RG-FLAT-CRF were tuned by Adam, and a hierarchical lr mechanism introduced. For the pre-trained RoBerta model, a learning rate of 3×10^{-5} is used, and for the other parts a learning rate of 2×10^{-4} is used. For the RG-FLAT-CRF model, the batch size used is 12. Details are shown in Table 2.

Table 2. Parameter settings.

Parameter	Value
lattice embedding	50
bigram embedding	50
glyce embedding	768
linear projection dim	768
dropout	0.1
Transformer head dim	20
Transformer head num	8

5. Results and Analysis

This part is divided into two parts: performance comparison with existing models, and ablation research.

5.1. Performance Comparison with Existing Models

To verify the effect of the RG-FLAT-CRF-model, the RGT-CRF model is compared to the existing state-of-the-art models. Evaluated on CCKS2017, CCKS2019, and CCKS2020 datasets, respectively. The comparison model is as follows:

(1) RoBerta: Liu et al. [37] improved the BERT model and proposed the RoBerta model. RoBerta performed better than BERT on NLP downstream tasks, and used RoBerta to enhance semantic representation and complete NER tasks.

(2) RoBerta-BiLSTM-CRF: Xu et al. [25] combined the bi-directional LSTM and CRF, which has become a classic model, and combined the RoBerta model with BiLSTM-CRF on this basis. Use RoBerta trained vectors and then use the BiLSTM-CRF model to extract entities.
(3) RoBerta-BiGRU-CRF: Qin et al. [29] proposed a BERT-BiGRU-CRF model in the field of Chinese electronic medical records, where the pre-trained model was replaced with an improved RoBerta.
(4) Ra-RC: Wu et al. [30] used RoBerta to obtain medical semantic features while using a bidirectional long short-term memory network to learn the radical features of Chinese characters.
(5) AR-CCNER: Yin et al. [26] used a convolutional neural network to extract radical features while using a self-attention mechanism to capture the dependencies between characters.
(6) ACNN: Kong et al. [27] used a multi-layer CNN structure to capture short-term and long-term contextual relations. CNN can also solve the problem that LSTM is difficult to exploit GPU parallelism, and the model uses an attention mechanism that can obtain global information.
(7) BE-Bi-CRF-JN: Wang et al. [31] cite additional medical knowledge information to correlate the original text in the named entity recognition task with its encyclopedic knowledge and enhance the ability of entity recognition by building a connection network.

Tables 3–5 show the precision, recall, and F1 results detailing various medical entities and all medical entities. From the comparison results of Table 6, the performance of the RGT-CRF model proposed in this chapter has achieved the best results on the three datasets, and the improvement on CCKS2017 is about 2~5%. The improvement is about 0.3~8% on CCKS2019 and about 3~9% on CCKS2020.

Table 3. Results of different models on CCKS2017.

Model	Evaluation Index	Entity Type					Comprehensive Value
		Symptoms and Signs	Diseases and Diagnosis	Body Parts	Examinations and Tests	Treatment	
RoBerta	P	96.95	74.21	88.09	95.59	76.39	91.99
	R	98.05	82.52	88.44	96.36	78.95	93.01
	F1	97.49	78.14	88.26	95.976	77.64	92.49
RoBerta-BiGRU-CRF	P	97.6	82.97	88.38	95.75	75.81	92.56
	R	97.99	81.82	88.69	96.58	77.99	93.09
	F1	97.79	82.39	88.53	96.16	76.88	92.82
RoBerta-BiLSTM-CRF	P	97.02	79.22	88.39	95.59	80.66	92.41
	R	98.39	85.31	90.65	96.53	81.81	94.11
	F1	97.7	82.15	89.5	96.05	81.23	93.25
Ra-RC	P	95	89.44	89.29	95.73	61.25	94.14
	R	98.11	89.17	90.79	97	69.01	92.39
	F1	96.53	89.31	90.03	96.36	64.9	93.26
ACNN	P	93.2	79.67	87.81	94.25	74.26	90.19
	R	97.92	79.39	84.41	95.9	75.7	90.78
	F1	95.5	79.52	86.08	95.07	74.97	90.49
AR-CCNER	P	96.53	74.07	89.38	94.78	82.91	92.27
	R	97.83	73.17	90.41	97.22	82.91	93.73
	F1	97.18	73.62	89.89	95.98	82.91	93
RGT-CRF	P	98.48	82.79	91.72	98.58	81.37	95.47
	R	98.66	85.31	90.16	97.2	82.99	95.76
	F1	98.56	84.03	90.93	97.88	82.17	95.61

Table 4. Results of different models on CCKS2019.

Model	Evaluation Index	Entity Type						Comprehensive Value
		Diseases and Diagnosis	Examinations	Anatomical Parts	Surgery	Tests	Drugs	
RoBerta	P	74.81	83.23	82.62	80.38	71.71	81.74	79.85
	R	75.27	83.48	84.14	78.4	63.79	82.08	79.85
	F1	75.03	83.35	83.37	79.37	67.51	81.9	79.85
RoBerta-BiGRU-CRF	P	67.15	65.47	82.94	76.1	65.58	64.62	72.95
	R	77.17	84.64	82.78	74.69	67.93	80.21	79.78
	F1	71.81	73.83	82.85	75.38	66.73	71.57	76.21
RoBerta-BiLSTM-CRF	P	75.76	84.97	83.06	78.26	69.73	77.49	79.7
	R	79.22	85.22	83.66	77.78	66.72	81.04	80.75
	F1	77.45	85.09	83.35	78.01	68.19	79.22	80.22
Ra-RC	P	78.74	81.23	82.67	79.61	74.28	93.27	83.31
	R	79	85.71	85.23	77.56	69.96	92.27	82.44
	F1	78.87	83.41	83.93	78.57	72.06	92.77	82.87
ACNN	P	75.84	85.37	88.2	76.27	68.23	92.34	83.07
	R	87.3	90.52	89.37	83.33	71.69	91.96	87.29
	F1	81.17	87.87	88.78	79.65	69.92	92.15	85.13
BE-Bi-CRF-JN	P	83.79	84.14	82.02	83.82	82.4	87.43	83.16
	R	83.66	89.71	87.55	90.48	89.8	85.11	86.67
	F1	83.73	86.83	84.7	87.02	85.94	86.25	84.88
RGT-CRF	P	78.3	88.84	85.58	84.23	78.5	88.44	85.36
	R	80.18	88.09	86.98	80.16	72.62	85.04	84.99
	F1	79.22	88.46	86.27	82.14	75.44	86.7	85.17

Table 5. Results of different models on CCKS2020.

Model	Evaluation Index	Entity Type						Comprehensive Value
		Diseases and Diagnosis	Examinations	Anatomical Parts	Surgery	Tests	Drugs	
RoBerta	P	85.28	78.05	90.03	88.03	69.28	87.77	86.68
	R	86.76	95.17	88.23	93.21	86.18	87.49	88.2
	F1	86.01	85.76	89.12	90.54	76.81	87.62	87.43
RoBerta-BiGRU-CRF	P	69.23	70.08	88.5	70.66	67.96	87.26	76.74
	R	84.65	91.5	89.4	91.4	86.59	87.92	88.09
	F1	76.16	79.37	88.94	79.7	76.15	87.58	82.02
RoBerta-BiLSTM-CRF	P	84.8	78.3	90.27	90.04	72.54	87.93	87.05
	R	84.8	92.56	88.57	94.12	83.74	87.27	87.67
	F1	84.8	84.83	89.41	92.03	77.73	87.59	87.35
BE-Bi-CRF-JN	P	81.88	81.25	94.64	83.7	76.82	94.64	82.52
	R	81.32	85.71	92.44	86.52	87.88	92.44	85.05
	F1	81.6	83.42	93.53	85.08	81.98	93.53	83.76
RGT-CRF	P	88.04	84.06	92.16	91.38	81.08	90.73	90.85
	R	87.33	94.05	91.81	92.4	88.96	90.35	91.57
	F1	87.68	88.77	91.98	91.88	84.83	90.53	91.2

The effect of ACNN is unstable in CCKS2017 and CCKS2019. Compared with other models, ACNN does not use BERT or an improved model based on BERT to enhance semantic representation, but multi-layer CNN and attention mechanisms play a certain positive role. From the three datasets, most of the models use BERT or an improved pre-training model based on BERT to enhance semantic representation and have achieved good experimental results. RoBerta-BiLSTM-CRF performs better than RoBerta-BiGRU-CRF on the three datasets. Although BiGRU has a simpler structure than BiLSTM, it is clear that BiLSTM is more suitable for Chinese electronic medical record NER. At the same time, these two models perform moderately well on the three datasets, as the feature extraction networks of the two models are variations of recurrent neural networks and cannot solve the long-range dependency problem. AR-CCNER and Ra-RC performed better on the

CCKS2017 and CCKS2019 datasets overall. Although AR-CCNER did not use a BERT-based pre-training model to enhance semantic representation, both AR-CCNER and Ra-RC were based on the characteristics of Chinese. BiLSTM and CNN are used to extract and use radical features, respectively, which utilize the glyph information of Chinese characters to a certain extent, but do not consider the information of learning the overall glyph structure of Chinese characters, and the model also lacks medical vocabulary information. BE-Bi-CRF-JN also achieved good results, proving that the use of external corpus in Chinese electronic medical records NER is effective. The above analysis shows that the RGT-CRF model is more suitable for Chinese electronic medical record named entity recognition electronic medical record recognition. This is mainly because the model adds glyph information while introducing lexical information based on words.

Table 6. Comparison of the results of different F1 of each model on different datasets.

Model	Dataset (F1)		
	CCKS2017	CCKS2019	CCKS2020
RoBerta	92.49	79.85	87.43
RoBerta-BiGRU-CRF	92.82	76.21	82.02
RoBerta-BiLSTM-CRF	93.25	80.22	87.35
Ra-RC	93.26	82.87	-
ACNN	90.49	85.13	-
AR-CCNER	93	-	-
BE-Bi-CRF-JN	-	84.88	83.76
RGT-CRF	95.61	85.17	91.2

From the perspective of entity type, the overall recognition effect of different medical entities is compared longitudinally. From Figures 13–15, it can be seen that the recognition results of different models on CCKS2017 show disease and diagnosis. Poor, because there are many long entities like '右股骨颈骨折髋关节股骨头表面置换术 (Right femoral neck fracture hip femoral head resurfacing)' in the two types of entities in the CCKS2017 dataset, and the boundaries of each entity cannot be clearly identified. The recognition results of different models on CCKS2019 and CCKS2020 show disease and diagnosis. The recognition results of these two types of entities are poor because the two types of entities in the CCKS2019 dataset and CCKS2020 dataset are similar to 'CA125', 'CEA'. Many entities coexist with English and numbers, such as 'CA199', which will also cause the model to fail to identify the boundaries of each entity.

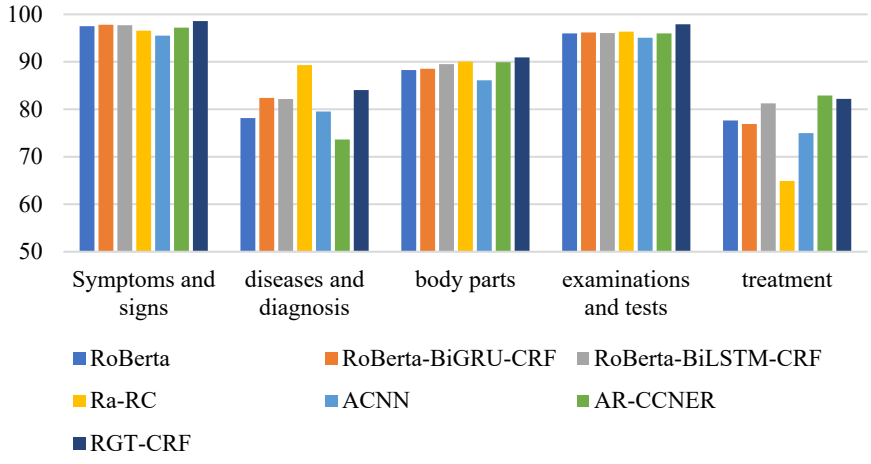

Figure 13. F1 values of different entities on CCKS2017 for different models.

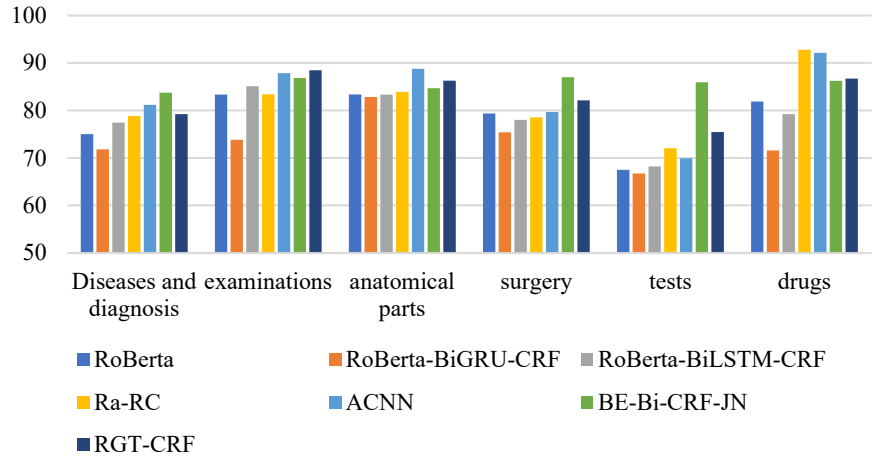

Figure 14. F1 values of different entities on CCKS2018 for different models.

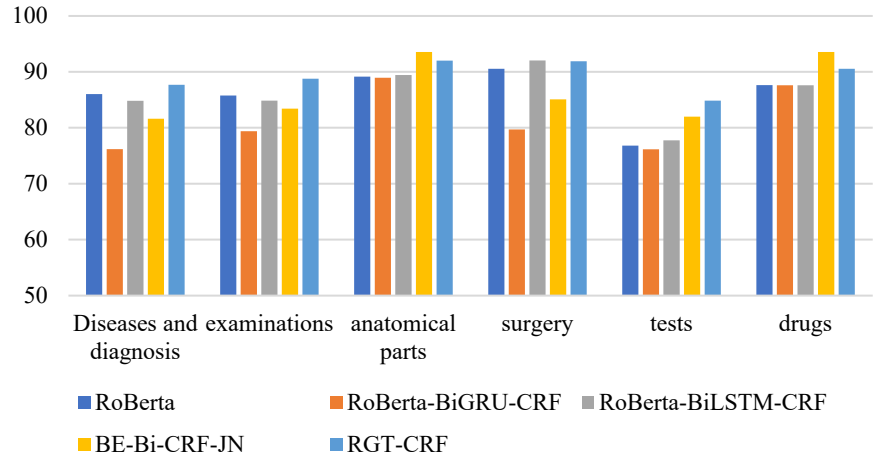

Figure 15. F1 values of different entities on CCKS2020 for different models.

To make the comparative results more convincing, a further hypothesis test was performed by calculating p-values using the t-test method, and *p*-values smaller than the significance level (usually 0.05) were considered statistically significant. Table 7 shows the statistical comparison of the proposed method with other methods. Most of the results are significant.

Table 7. Comparison results with different models on different datasets.

Model	CCKS2017	CCKS2019	CCKS2020
RoBera	0.0217	0.0019	0.0424
RoBerta-BiGRU-CRF	0.0382	0.0043	0.0049
RoBerta-BiLSTM-CRF	0.0029	0.0055	0.2025

5.2. Ablation Research

We design a set of ablation experiments to verify the contribution of each part to the model, where RGT-CRF-NG indicates that the model does not add glyph information. RGT-CRF-NF shows that the model does not add lexical information and its corresponding positional encoding. Finally, it is compared with RoBerta-BiLSTM-CRF and RGT-CRF on three datasets, and the results are shown in Table 8.

Table 8. Performance of different variants on three datasets.

Model	Dataset (F1)		
	CCKS2017	**CCKS2019**	**CCKS2020**
RoBerta-BiLSTM-CRF	93.25	80.22	87.35
RGT-CRF-NG	93.87	82.65	88.13
RGT-CRF-NF	94.92	84.03	89.83
RGT-CRF	95.61	85.17	91.2

The experimental results of RGT-CRF-NF and RGT-CRF-NG are better than the RoBerta-BiLSTM-CRF model regarding the three datasets, indicating that the glyph information and the use of lattice structure to add lexical information are effective for Chinese electronic medical record named entity recognition. The result of RGT-CRF-NG is slightly worse than that of RGT-CRF-NF, indicating that adding medical glyph information to the Chinese electronic medical record NER task is more effective than word information. This comparison can also be found in the above experiments using glyph information. Similarly, the final model with radical information is better than the model without radical information. This is because many Chinese characters in medical entities have the same glyph structure, so their meanings are also similar.

For example, '疼 (pain) ', '痛 (pain)', '病 (sick)', '腹 (belly)', '腰 (waist)', '肝 (liver)', '脾 (spleen)', '呕 (vomit)', '吐 (threw up)', '咳 (cough)', '嗽 (cough)', '胰 (pancreatic)', '肠 (intestinal)', '肿 (swell)', '胀 (swell)'. And this is very common in medical entities.

6. Conclusions

In this paper, an RG-FLAT-CRF model is proposed for Chinese CNER, which can learn the glyph features of medical fonts, and at the same time introduces word information to enhance word boundaries, and finally achieves good performance on three datasets. The RG-FLAT-CRF model obtains character vectors through RoBerta, Glyce, word2vec, and word vectors through word2vec. The word information is fused using the Flat-lattice structure and then encoded by the transformer network. In line with the output of the encoding layer, the label of each input character is predicted by the CRF layer. It addresses problems like word segmentation errors and lack of lexical information, given the characteristics of Chinese medical characters and the vector of multi-feature fusion. The final experimental results demonstrate that our proposed model outperformed the baseline models.

Several issues require further research. At this stage, deep learning requires a large amount of annotated data to train the model, as does our proposed model, but large-scale annotated data in the Chinese electronic medical record domain requires medical experts to annotate, which can be time-consuming. Therefore, our next research investigates how to perform named entity recognition on medical record texts with sparse data.

Author Contributions: J.L.: Conceptualization, Methodology, Software, Writing—original draft. Y.W.: Supervision, Project administration. R.L., C.C., and X.S.: Investigation, Writing—review & editing. S.Z.: Data curation, Resources. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program (Demonstration of R&D and Application of Integrated Science and Technology Service Platform for Central Plains Urban Agglomeration), grant number 2018YFB1404500.

Data Availability Statement: We used the CCKS open-source Chinese electronic medical record named entity recognition dataset and cite it in the article.

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

References

1. Chowdhury, S.; Dong, X.; Qian, L.; Li, X.; Guan, Y.; Yang, J.; Yu, Q. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. *BMC Bioinform.* **2018**, *19*, 75–84. [CrossRef] [PubMed]
2. Wang, Q.; Zhou, Y.; Ruan, T.; Gao, D.; Xia, Y.; He, P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *J. Biomed. Inform.* **2019**, *92*, 103133. [CrossRef] [PubMed]
3. Shaukat, K.; Shaukat, U. Comment extraction using declarative crowdsourcing (CoEx Deco). In Proceedings of the 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), Quetta, Pakistan, 11–12 April 2016; pp. 74–78.
4. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
5. Alam, T.M.; Shaukat, K.; Hameed, I.A.; Khan, W.A.; Sarwar, M.U.; Iqbal, F.; Luo, S. A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining. *Biomed. Signal Processing Control* **2021**, *68*, 102726. [CrossRef]
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
7. Zhang, Y.; Yang, J. Chinese NER using lattice LSTM. *arXiv* **2018**, arXiv:1805.02023.
8. Li, X.; Yan, H.; Qiu, X.; Huang, X. FLAT: Chinese NER using flat-lattice transformer. *arXiv* **2020**, arXiv:2004.11795.
9. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:1901.02860.
10. Meng, Y.; Wu, W.; Wang, F.; Li, X.; Nie, P.; Yin, F.; Li, M.; Han, Q.; Sun, X.; Li, J. Glyce: Glyph-vectors for chinese character representations. *arXiv* **2019**, arXiv:1901.10125.
11. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*; Curran Associates Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119.
12. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Xu, M. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access* **2020**, *8*, 222310–222354. [CrossRef]
13. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Chen, S.; Liu, D.; Li, J. Performance comparison and current challenges of using machine learning techniques in cybersecurity. *Energies* **2020**, *13*, 2509. [CrossRef]
14. Friedman, C.; Alderson, P.O.; Austin, J.H.; Cimino, J.J.; Johnson, S.B. A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.* **1994**, *1*, 161–174. [CrossRef] [PubMed]
15. Fukuda, K.; Tamura, A.; Tsunoda, T.; Takagi, T. Toward information extraction: Identifying protein names from biological papers. *Pac. Symp. Biocomput.* **1998**, *707*, 707–718.
16. McCallum, A.; Freitag, D.; Pereira, F.C. Maximum entropy Markov models for information extraction and segmentation. *ICML* **2000**, *17*, 591–598.
17. Možina, M.; Demšar, J.; Kattan, M.; Zupan, B. Nomograms for visualization of naïve Bayesian classifier. In *European Conference on Principles of Data Mining and Knowledge Discovery*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 337–348.
18. Settles, B. Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP), Geneva, Switzerland, 28–29 August 2004; pp. 107–110.
19. Tang, B.; Cao, H.; Wu, Y.; Jiang, M.; Xu, H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Med. Inform. Decis. Mak.* **2013**, *13*, S1. [CrossRef]
20. Roberts, K.; Shooshan, S.E.; Rodriguez, L.; Abhyankar, S.; Kilicoglu, H.; Demner-Fushman, D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. *J. Biomed. Inform.* **2015**, *58*, S111–S119. [CrossRef]
21. Liu, K.; Hu, Q.; Liu, J.; Xing, C. Named entity recognition in Chinese electronic medical records based on CRF. In Proceedings of the 2017 14th Web Information Systems and Applications Conference (WISA), Liuzhou, China, 11–12 November 2017; pp. 105–110.
22. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
23. Mikolov, T.; Karafiát, M.; Burget, L.; Cernocky, J.; Khudanpur, S. Recurrent neural network based language model. *Interspeech. Makuhari* **2010**, *2*, 1045–1048.
24. Ma, X.; Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv* **2016**, arXiv:1603.01354.
25. Xu, K.; Zhou, Z.; Hao, T.; Liu, W. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017; Springer: Cham, Swizerland, 2017; pp. 355–365.
26. Yin, M.; Mou, C.; Xiong, K.; Ren, J. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. *J. Biomed. Inform.* **2019**, *98*, 103289. [CrossRef]
27. Kong, J.; Zhang, L.; Jiang, M.; Liu, T. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition. *J. Biomed. Inform.* **2021**, *116*, 103737. [CrossRef]
28. Zhang, W.; Jiang, S.; Zhao, S.; Hou, K.; Liu, Y.; Zhang, L. A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition. In Proceedings of the 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA), Xiangtan, China, 26–27 October 2019; pp. 166–169.

29. Qin, Q.; Zhao, S.; Liu, C. A BERT-BiGRU-CRF Model for Entity Recognition of Chinese Electronic Medical Records. *Complexity* **2021**, *2021*, 6631837. [CrossRef]
30. Wu, Y.; Huang, J.; Xu, C.; Zheng, H.; Zhang, L.; Wan, J. Research on Named Entity Recognition of Electronic Medical Records Based on RoBERTa and Radical-Level Feature. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 2489754. [CrossRef]
31. Wang, Q.; Haihong, E. Bi-directional Joint Embedding of Encyclopedic Knowledge and Original Text for Chinese Medical Named Entity Recognition. In Proceedings of the 2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT), Sanya, China, 27–29 December 2021; pp. 304–309.
32. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
33. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
35. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; Wu, H. Ernie: Enhanced representation through knowledge integration. *arXiv* **2019**, arXiv:1904.09223.
36. 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]
37. 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.
38. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*; Curran Associates Inc.: Red Hook, NY, USA, 2019; p. 32.
39. Sun, Y.; Lin, L.; Yang, N.; Ji, Z.; Wang, X. Radical-enhanced chinese character embedding. In *International Conference on Neural Information Processing*; Springer: Cham, Switzerland, 2014; pp. 279–286.
40. Wang, S.; Zhou, W.; Zhou, Q. Radical and Stroke-Enhanced Chinese Word Embeddings Based on Neural Networks. *Neural Process. Lett.* **2020**, *52*, 1109–1121. [CrossRef]
41. Wei, H.; Zhang, H.; Gao, G. Word image representation based on visual embeddings and spatial constraints for keyword spotting on historical documents. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3616–3621.
42. Su, T.R.; Lee, H.Y. Learning chinese word representations from glyphs of characters. *arXiv* **2017**, arXiv:1708.04755.
43. Shaukat, K.; Luo, S.; Chen, S.; Liu, D. Cyber threat detection using machine learning techniques: A performance evaluation perspective. In Proceedings of the 2020 International Conference on Cyber Warfare and Security (ICCWS), Islamabad, Pakistan, 20–21 October 2020; pp. 1–6.