Received 17 April 2024, accepted 23 May 2024, date of publication 27 May 2024, date of current version 20 June 2024.
Digital Object Identifier 10.1109/ACCESS.2024.3405997

RESEARCH ARTICLE

A Named Entity Recognition Method Based on Knowledge Distillation and Efficient GlobalPointer for Chinese Medical Texts

ZHENGWEI ZHAI[1], RONGLI FAN[1], JIE HUANG[1], NEAL N. XIONG[2], (Senior Member, IEEE), LIJUAN ZHANG[1], JIAN WAN[1], AND LEI ZHANG[1]
[1]School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
[2]Department of Computer Science and Mathematics, Sul Ross State University, Alpine, TX 79830, USA

Corresponding author: Lijuan Zhang (zhanglijuan931026@foxmail.com)

This work was supported in part by the Key Research and Development Program of Zhejiang Province under Grant 2021C03145, in part by the National Natural Science Youth Science Foundation Project under Grant 62201508, and in part by Zhejiang Provincial Natural Science Foundation Youth Fund Project under Grant LQ23F010004.

ABSTRACT The task of named entity recognition has been widely used in medical text analysis, but there is still the problem of poor transfer ability in practical applications. This work proposes a novel named entity recognition method based on a proposed knowledge distillation framework and Efficient GlobalPointer for Chinese biomedical and clinical data. Specifically, our study leverages the Efficient GlobalPointer to address the issue of entity nesting and introduces a context shield window to mitigate interference from redundant information. Furthermore, the model's generalization ability is enhanced through a novel knowledge distillation framework. The proposed knowledge distillation framework solves the problem of independent feature learning process in feature distillation by using linkage mechanism. The recognition accuracy is improved by the proposed knowledge distillation method while keeping the model complexity low, so that our method can meet the inference speed requirement in real applications while ensuring a certain recognition accuracy. Our method achieves excellent experimental results on three publicly available Chinese datasets, where the comprehensive evaluation metric F1 exceeds the best results achieved by existing methods.

INDEX TERMS Context shield window, efficient GlobalPointer, knowledge distillation, medical data analysis, named entity recognition.

I. INTRODUCTION

Biomedical informatics has been applied in many fields of biomedical and clinical data analysis [1], where the Named Entity Recognition (NER) technology is widely used in medical data analysis to classify the named entities mentioned in sentences into some predefined categories [2], such as diseases, symptoms, etc. Furthermore, NER is a basic task in Natural Language Processing (NLP) and is usually used as an information extraction tool embedded in many downstream tasks such as Clinical decision making and medical question answering system. Therefore, NER models are critical to biomedical and clinical text analysis, providing deeper text understanding and analysis capabilities to provide important support for intelligent healthcare applications.

In view of its importance, scholars' research on NER depth mainly focuses on two directions [3]. One effective direction is to improve NER performance by using more complex model architectures, such as BERT [4], RoBERTa [5], GPT3 [6], and other pre-trained models based on the Transformer architecture [7]. These models understand and pay attention to the importance of different components in

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

the sentence through the attention mechanism and analyze the content of the context, which is crucial for NER tasks. However, an important flaw in this direction is that it has high environmental requirements, and the high computational cost limits its scalability and adaptability for different medical scenarios.

Research in the other direction is mainly to adapt to different application scenarios and overcome the challenges of limited computing resources. This direction generally hopes to simplify these large models into lighter counterparts, preferably without any significant performance degradation [8]. To solve this problem, Knowledge Distillation (KD), as one of the most common strategies, is considered a very promising technology based on its good performance and excellent model compression capabilities. KD often includes a pre-trained teacher model. By incorporating the predictions or intermediate features of the teacher model as soft labels into the training of the student model, the student model learns the knowledge bias embedded in the teacher model. This helps avoid learning absolute knowledge from hard labeling alone. By transferring knowledge from the large teacher model to the compact student model, higher recognition accuracy is achieved without changing the complexity of the student model.

Inspired by the Ambiguity-aware Robust Teacher-Knowledge Distillation (ART-KD) [9] in the field of image recognition, we propose a simplified but effective NER method for knowledge distillation in Chinese medical text analysis applied in the field of natural language processing. The core of our proposed KD framework is to solve the problem of independent feature learning process in traditional feature distillation and overcome the limitation of predicting ambiguous samples by a single teacher model in NER knowledge distillation by proposing a linkage mechanism. This framework can provide refined teacher knowledge to the student network through the linkage mechanism, thus improving the distillation effect and recognition accuracy.

In addition, our study first introduces the context shield window to NER task for optimizing the NER models: Efficient GlobalPointer [10]. By limiting the scope of entity recognition, the context shield window not only solves the problem of redundant information interference caused by the sparseness of the Efficient GlobalPointer model, but also has the effect of filtering out incorrect entities for too long.

In summary, the main work of the method proposed in this paper is summarized as follows:

(1) By proposing a novel knowledge distillation framework and applying to the Efficient GlobalPointer, we solve the problem of independent teacher and student models during feature distillation, which achieves excellent recognition accuracy with lower complexity.
(2) By introducing context shield window, the problem of redundant information interference caused by sparse Efficient GlobalPointer in NER is solved, and the discrimination of nested and flat entities in NER is realized.
(3) The effectiveness of the NER method proposed in this paper is verified by experiments on three Chinese public datasets: CLUENER in the general field and CMeEE and CMeEE-V2 in the medical field. Experiments have proven that the proposed method has significant effects in reducing parameters and improving recognition accuracy.

The rest of this paper is structured as follows: Firstly, Section II presents some related works used in our study and Section III describes the proposed KD method and NER model in detail. Next, Section IV shows the comparative experiments with some mainstream models and ablation studies on three datasets. Finally, Section V gives the conclusions and discusses future research directions.

II. RELATED WORK

This paper will describe the related work from two aspects: NER and knowledge distillation.

Earlier researches on NER task are mainly based on rule template [11], [12], [13], [14] and machine learning [15], [16], [17], [18]. For example, Hanisch et al. [19] and Quimbaya et al. [20] adopted the rule-based NER method to realize the identification of proteins, genes and medical entities in electronic health records, respectively. Feng et al. [21] used machine learning methods to achieve excellent results in identifying bridge entities by combining the HMM [22] model with lexical features and proprietary rules of the bridge domain. With the development of pre-trained models, deep learning-based methods have gradually become the mainstream [23]. Devlin et al. [4] proposed a method of using pre-trained language models to fine-tune NER tasks. In paper [24], a method is proposed to integrate the dictionary features into the bi-directional long short term memory with conditional random field (BiLSTM-CRF) model. Yoon et al. [25] proposed a multi-model aggregation framework to train multiple single-task models separately and then map them to the same BiLSTM-CRF model through transformation matrix. Frei and Kramer [26] proposed a NER method applied to German clinical text mining. Sun et al. [27] combined pre-trained model BERT with BiLSTM-CRF to improve the performance of NER tasks in healthcare and other fields. In the Chinese Field, Yang et al. proposed a NER method for Chinese medical texts using BiLSTM-CRF structure. Gao et al. [28] improved medical domain recognition by using pre-trained models. Tarcar et al. [29] also adopted language model pretraining for the task of Healthcare NER. Paper [30] added a large number of biomedical corpora into the training of the pre-trained language model, effectively improving the model's performance on biomedical and clinical data analysis. Su et al. [10] proposed GlobalPointer that is not inferior to CRF and recognizes nested and flat entities indiscriminately, but this GlobalPointer suffers from redundant information interference.

Knowledge Distillation (KD) has excellent applications in fields such as computer vision. The student-teacher learning model has become one of the hot spots in the research

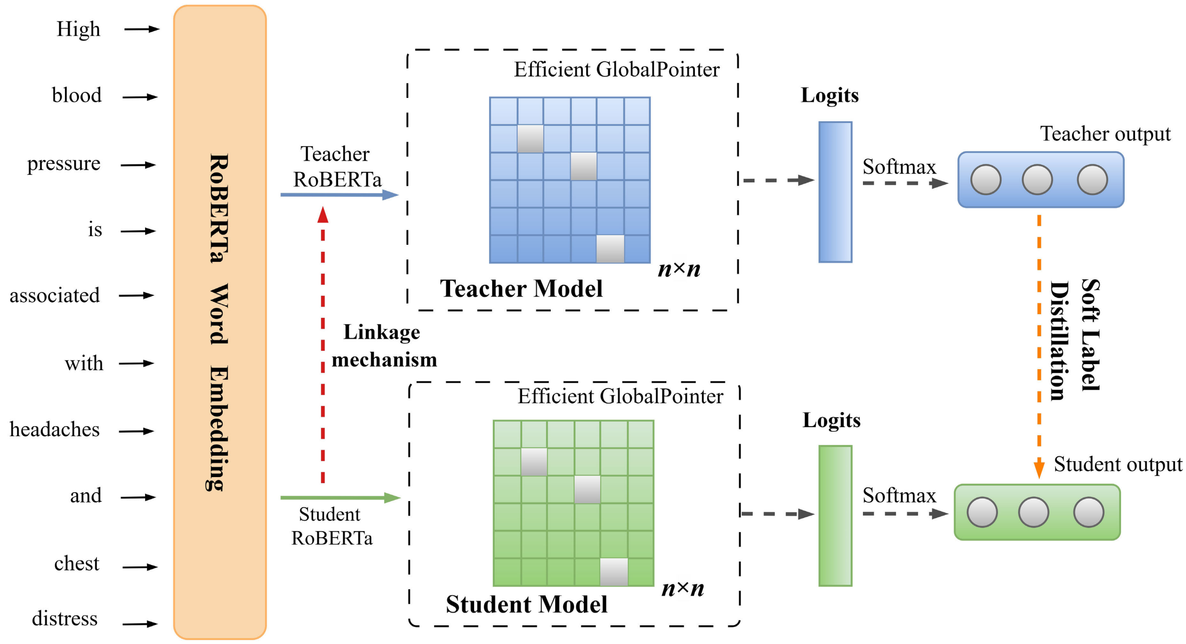

FIGURE 1. The overall structure of the proposed NER method.

field. Hinton et al. [31] first proposed a basic knowledge distillation method: by guiding the student model to imitate the teacher model and generate soft labels, the logarithm of the teacher model output is converted into the target of the student model to improve performance. Chen et al. [32] took advantage of the characteristics of generative adversarial network to randomly generate training data and propose a method combined with knowledge distillation. In the case of no data, it achieved certain results in image classification task. In the field of natural language processing, Jiao et al. [33] proposed a knowledge distillation method to learn the hidden vectors in the middle layer of Transformer structure, which realizes effective compression of Bert model. Doğan et al. [34] proposed a method to simultaneously learn the output of the feedforward neural network layer in the middle layer of Transformer structure and the output of the attention matrix of the multi-head self-attention mechanism layer, which further improves the distillation effect. Cho et al. [9] proposed a self-knowledge distillation framework applied in fine-grained visual recognition, which makes full use of the teacher's information and performs pruning and other operations to obtain refined knowledge for distillation.

III. METHODOLOGY

The knowledge distillation framework proposed in our study is shown in Fig. 1. This method combines the feature input of two pre-trained models as the input part of the teacher network by using the linkage mechanism, so as to improve the credibility of the teacher network knowledge more fully. After the teacher network processes the input features, the student network's input and output features in the subsequent soft label distillation are more similar to the teacher's features, which can also make the learning process smoother and more consistent.

In the selection of teacher model and student model, our method adopts Efficient GlobalPointer with different pre-trained weights as teacher model and student model respectively. Based on this model, a context shield window is first introduced to solve the problem of redundant information interfering with entity recognition in Efficient GlobalPointer, and the optimized Efficient GlobalPointer is shown in Fig. 2. In this section, the proposed NER knowledge distillation framework and the modules included in the NER structure will be introduced in detail: text embedding, Efficient GlobalPointer and the context shield window.

A. KNOWLEDGE DISTILLATION FRAMEWORK

When using knowledge distillation methods, especially when using uncertain information as knowledge of the teacher model, model performance degradation often occurs. Furthermore, traditional feature learning in knowledge distillation methods usually treat features as independent learning tasks and input them into the neural network for training. However, this independent learning approach can lead to inconsistencies and incompleteness among features, thereby limiting the representation and generalization ability of the model.

In order to solve these problems, a novel knowledge distillation method is introduced by proposing a linkage mechanism to link two features as two parts of inputs for the teacher model. Specifically, the input feature of the teacher network is processed so that it contains the knowledge of

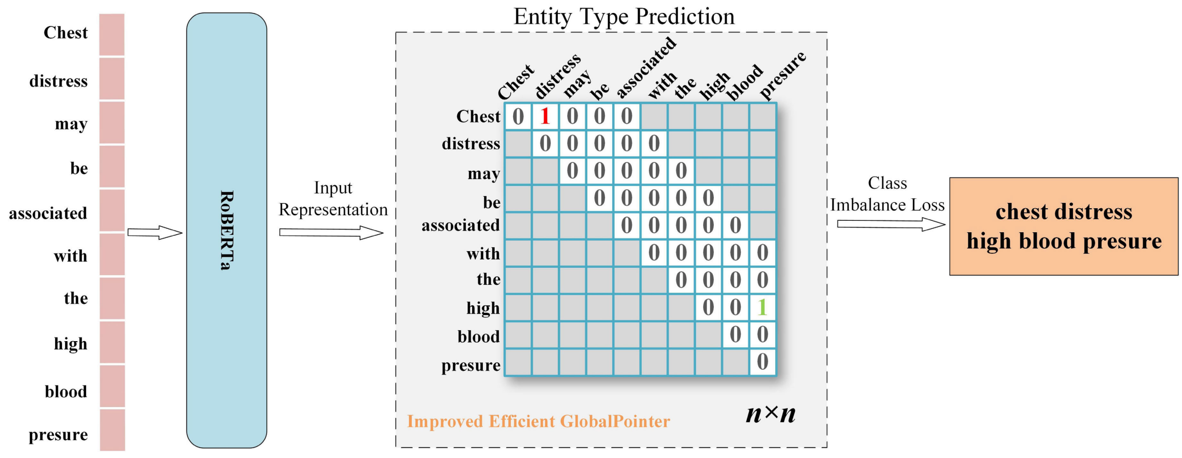

FIGURE 2. The improved efficient GlobalPointer structure.

both the teacher and the student, and the output feature of student network is more similar to the teacher feature after distillation. This achieves mutual correlation and consistency between the two features.

The introduction of this linkage mechanism brings many benefits. Firstly, it enables smooth transition and integration between features. By guiding the student network to imitate the feature representation of the teacher network as much as possible, the student network can learn the high-level representation of the teacher network, thereby improving the expression ability of the model. Secondly, this method can also enhance the stability and robustness of the model. By using common information between multiple features for training, even if a feature is abnormal or wrong, other features can provide supervision signals and reduce the risk of the model being affected by abnormal features.

As shown in Fig. 1, the text is first loaded into the teacher and student models, each embedded in distinct Roberta pre-trained models (indicated by the blue and green lines). Secondly, the two embeddings are fed into Efficient GlobalPointer as inputs of the teacher model and the student model, respectively, during which feature information from the student model is introduced into the teacher model via the linkage mechanism (indicated by the red line). This linkage mechanism makes the knowledge of the teacher model integrate some features of the student model, which enhances the performance and stability of the model and promotes the consistency and integration between features. Then, the student output and the teacher output are obtained by extracting the Efficient GlobalPointer. Finally, the student model learns the teacher's knowledge generated by the joint input through soft label distillation. This enables the student network to acquire a more accurate feature representation from the teacher network.

In the NLP field, most knowledge distillation methods use the Mean Square Error (MSE) loss function L_{MSE} as shown in Eq. (1), which represents the mean sum of squares of the difference between the predicted value $f(x)$ and the target value y. In our study, the KL divergence loss function D_{KL} shown in Eq. (2) is used, which represents the KL divergence of p to q, where $p(x)$ and $q(x)$ are two probability distributions of random variable x. The reason for using KL divergence loss function D_{KL} is that MSE loss function L_{MSE} is easy to fall into local optimal solution prematurely in the training optimization process, and the optimization process using KL divergence as loss is convex, which can better make the model converge to the global optimal.

$$L_{MSE} = \frac{\sum_{i=1}^{n}(f(x)-y)^2}{n} \tag{1}$$

$$\begin{aligned} D_{KL}(p\|q) &= E_{p(x)} \log \frac{p(x)}{q(x)} \\ &= \sum_{i=1}^{N} p(x_i) \cdot (\log p(x_i) - \log q(x_i)) \end{aligned} \tag{2}$$

Common soft label knowledge distillation using KL loss usually uses the loss function $\mathcal{L}_{KD}$ shown in Eq. (3). The output of both the student model and the teacher model is input to $D_{KL}(p\|q)$ after adjusting the softmax function and distillation temperature K.

$$\begin{aligned} &\mathcal{L}_{KD}(\boldsymbol{x}; \theta_s, \theta_t, K) \\ &\quad = D_{KL}\left(\operatorname{softmax}\left(\frac{f_s(\boldsymbol{x}; \theta_s)}{K}\right) \| \operatorname{softmax}\left(\frac{f_t(\boldsymbol{x}; \theta_t)}{K}\right)\right) \end{aligned} \tag{3}$$

where f_s and f_t represent the student model and teacher model respectively, θ_s and θ_t are the parameters of the student model and teacher model respectively, and K is the parameter distillation temperature. But such incorrect knowledge generated by the model can lead to interference in the teacher's model. The difference between traditional KD method and our proposed KD method is that the teacher model uses an intermediate mapping of student input features through the linking mechanism, so that the recognition of teacher model includes not only the trained weight of the teacher, but also a part of the student knowledge. And the Teacher output $\hat{p}_{\text{joint}}$ obtains under the joint action of the two parts of knowledge. Therefore, the logits output by the teacher model are no longer independent $f_t(\boldsymbol{x};\theta_t)$ but jointly generated $\hat{p}_{\text{joint}}$. Specifically, the $\hat{p}_{\text{joint}}$ is obtained from the Eq. (4) and the distillation loss $\mathcal{L}_{KD}^{e}$ is shown in Eq. (5).

$$\hat{p}_{\text{joint}} = f_t'(x_s, x_t; \theta_t) \tag{4}$$

$$\mathcal{L}_{KD}^{e}(\boldsymbol{x};\theta_s,\theta_t,K) = D_{KL}\left(\text{softmax}\left(\frac{f_s(\boldsymbol{x};\theta_s)}{K}\right) \| \text{softmax}\left(\frac{\hat{p}_{\text{joint}}}{K}\right)\right) \tag{5}$$

where $\hat{p}_{\text{joint}}$ is the joint distribution generated by linkage mechanism, which is jointly generated by the knowledge x_s of the original teacher model and the intermediate feature x_s of the student model. In this way, the redefined teacher knowledge will be high-confidence knowledge rather than independent information. Subsequently, the predictions from both models will be combined for the final soft label distillation.

With the implementation of the distillation framework, the optimization goal of the whole NER task with $\mathcal{L}_{\text{Total}}$ will be the joint realization of the soft label distillation loss $\mathcal{L}_{KD}^{e}$ and the NER task loss $\mathcal{L}_{NER}$ as shown in the Eq. (6).

$$\mathcal{L}_{\text{Total}}(\boldsymbol{x},y;\theta_s,\theta_t,K) = \mathcal{L}_{NER}(\boldsymbol{x},y;\theta_s) + \alpha\cdot\mathcal{L}_{KD}^{e}(\boldsymbol{x};\theta_s,\theta_t,K) \tag{6}$$

where α is hyperparameter, which accords to datasets; $\mathcal{L}_{NER}$ is the loss function of Efficient GlobalPointer and is described in detail in Section III-E.

B. TEXT EMBEDDING

This method uses RoBERTa [5] pre-trained model for text embedding. RoBERTa (Robustly Optimized BERT approach) model is a pre-trained language model based on Transformer architecture and developed by Facebook AI. It optimizes BERT to boost performance with larger amounts of data and longer training durations. In the field of Chinese language processing, the RoBERTa model is a Chinese pre-trained language model that has been fine-tuned through pre-training on a Chinese corpus. It inherits the advantages and characteristics of the RoBERTa model and demonstrates exceptional performance in various Chinese NLP tasks.

The difference in performing knowledge distillation is that the teacher model of our NER method uses the roberta_zh_large_pytorch pre-trained model with a hidden layer size of 1024, while the student model uses the hfl/chinese-roberta-wwm-ext pre-trained model with a hidden layer size of 768 [35]. Both pre-trained models are essentially a stack of 12 encoder layers under the transformer architecture. However, roberta_zh_large_pytorch's training data and parameters are more extensive and complex, resulting in superior training outcomes. Consequently, it can be employed as a teacher model to achieve enhanced entity recognition performance. In contrast, the hfl/chinese-roberta-wwm-ext model has fewer parameters and offers greater portability, making it easier to adapt to various application scenarios as a student model.

C. EFFICIENT GLOBALPOINTER

The task of NER is primarily to identify and classify named entities with a specific meaning from given text, such as names of people, places, organizations, etc. Specifically, NER aims to identify the starting and ending positions of an entity, with the entity's content located between these two positions. One of the challenges in the NER task is dealing with nested entities. For example, the disease entity ''lung cancer'' also has the organ entity ''lung'' nested in it. The model GlobalPointer [10] can identify nested and flat entities without distinction, achieving excellent recognition performance. The main principle of the GlobalPointer is to identify the starting and ending positions of each possible entity in the sentence through a matrix that is the length of the sentence multiplied by the length of the sentence.

Specifically, it is first necessary to set the maximum sentence length and truncate when the sentence length exceeds this value. Secondly, let a sentence of length n be encoded by the RoBERTa model to obtain the sequence $[\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n]$. Then, the sequence $[\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n]$ is processed by the feedforward layers of the Eq. (7) and Eq. (8), and the matrix $s_\alpha(i,j)$ of size n is obtained from the input Eq. (9). When it is necessary to identify the entities of the sentence of length n, and the number of entity types to be determined is K, then there are K matrices of size n used to identify the entities. In terms of type α, there are $n(n+1)/2$ candidate entities to exhaustively enumerate all possible entities in the matrix $s_\alpha(i,j)$. The role of the GlobalPointer for each entity type is to compare the $n(n+1)/2$ candidates and select the correct number of entities.

$$\boldsymbol{q}_{i,\alpha} = \boldsymbol{W}_{q,\alpha}\boldsymbol{c}_i + \boldsymbol{b}_{q,\alpha} \tag{7}$$

$$\boldsymbol{k}_{j,\alpha} = \boldsymbol{W}_{k,\alpha}\boldsymbol{c}_j + \boldsymbol{b}_{k,\alpha} \tag{8}$$

$$s_\alpha(i,j) = \boldsymbol{q}_{i,\alpha}^{\top}\boldsymbol{k}_{j,\alpha} \tag{9}$$

where $\boldsymbol{q}_{i,\alpha}$ and $\boldsymbol{k}_{i,\alpha}$ are vector representations used to identify the entity type α. And $\boldsymbol{q}_{i,\alpha}$ transposed to multiply $\boldsymbol{k}_{j,\alpha}$ is the scoring function $s_\alpha(i,j)$ of entity type α.

Another advantage of GlobalPointer is the utilization of Rotary Position Embedding (RoPE) [36] to make better use of location information. The key distinction between RoPE and traditional relative position coding [37], [38], [39] is that traditional position coding identifies relative positions after truncation, whereas RoPE has more superiority in extrapolation, enabling it to effectively handle longer text sequences. RoPE representation in two dimensions by complex number can be obtained $R_f(\boldsymbol{q},m) = \|q\| = \boldsymbol{q}$, its essence indicates that the transformation of RoPE corresponds to the rotation of the vector, so it is called rotational position coding. In two-dimensional case, it is expressed in matrix form as follows:

$$f(\boldsymbol{q}, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} q_0 \\ q_1 \end{pmatrix} \tag{10}$$

where $f(\boldsymbol{q}, m)$ is the RoTE encoding function. According to the principle that the inner product satisfies linear superposition, even dimensional cases can be realized by two-dimensional concatenation. And the rotation encoding matrix applied to the GlobalPointer is as follows:

$$\begin{aligned} s_\alpha(i,j) &= \left(\mathcal{R}_i \boldsymbol{q}_{i,\alpha}\right)^\top \left(\mathcal{R}_j \boldsymbol{k}_{j,\alpha}\right) \\ &= \boldsymbol{q}_{i,\alpha}^\top \mathcal{R}_i^\top \mathcal{R}_j \boldsymbol{k}_{j,\alpha} = \boldsymbol{q}_{i,\alpha}^\top \mathcal{R}_{j-i} \boldsymbol{k}_{j,\alpha} \end{aligned} \tag{11}$$

where $\mathcal{R}_i$ is the rotational position coding matrix, so that the relative position information obtained by rotational position coding can be injected into the scoring function $s_\alpha(i,j)$ through the relation $\mathcal{R}_i^\top \mathcal{R}_j = \mathcal{R}_{j-i}$.

Efficient GlobalPointer is proposed to optimize GlobalPointer by addressing issues such as parameter redundancy and poor utilization. Efficient GlobalPointer can improve the effect of entity recognition while reducing the number of parameters. The principle behind this is that since the scoring function $s_\alpha(i,j)$ for each entity type α to be extracted exhibits significant similarity, there's no need to establish a separate scoring function for each entity type α, but the entity identification process can be decomposed into two steps. In other words, entities of all entity types can be directly obtained with a scoring matrix, and the classification process of the second step to determine the entity type is realized by feature concatenation and dense layer. Thus, the improved scoring function $s_\alpha(i,j)$ is implemented as follows:

$$s_\alpha(i,j) = \left(\boldsymbol{W}_q \boldsymbol{h}_i\right)^\top \left(\boldsymbol{W}_k \boldsymbol{h}_j\right) + \boldsymbol{w}_\alpha^\top \left[\boldsymbol{h}_i; \boldsymbol{h}_j\right] \tag{12}$$

where $\left(\boldsymbol{W}_q \boldsymbol{h}_i\right)^\top \left(\boldsymbol{W}_k \boldsymbol{h}_j\right)$ represents the process of extracting the entity span, and $\boldsymbol{w}_\alpha^\top \left[\boldsymbol{h}_i; \boldsymbol{h}_j\right]$ represents the process of classifying the entity type. Furthermore, if $\boldsymbol{q}_i$ and $\boldsymbol{k}_i$ represent $\boldsymbol{W}_q \boldsymbol{h}_i$ and $\boldsymbol{W}_k \boldsymbol{h}_j$ respectively, then $[\boldsymbol{q}_i; \boldsymbol{k}_i]$ replaces $\boldsymbol{h}_i$ to further reduce the number of parameters. Therefore, the final scoring function $s_\alpha(i,j)$ is shown as follows:

$$s_\alpha(i,j) = \boldsymbol{q}_i^\top \boldsymbol{k}_j + \boldsymbol{w}_\alpha^\top \left[\boldsymbol{q}_i; \boldsymbol{k}_i; \boldsymbol{q}_j; \boldsymbol{k}_j\right] \tag{13}$$

D. CONTEXT SHIELD WINDOW

Both the GlobalPointer and the Efficient GlobalPointer are marked based on the matrix of sentence length multiplied by sentence length. Due to the long sentence length in the actual scene, the entities that need to be identified are mostly short entities within a few words. This leads to too much redundant information in the entity recognition matrix, which does not produce many positive effects and instead interferes with entity recognition.

To solve this problem, the context shield window in joint extraction is introduced to our NER method. The context shield window further reduces the participation of redundant parameters in entity recognition by limiting the scope of extraction, thus achieving the effect of shielding interference.

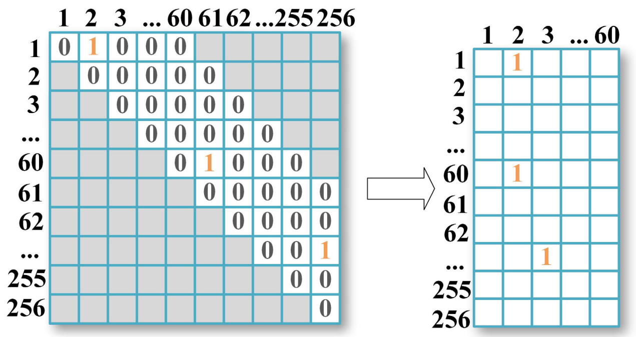

FIGURE 3. The detial of context shield window.

The context shield window can also play a role in filtering the recognized error-long entities to improve the recognition accuracy. For example, when context shield windows are added to the sentences depicted in Fig. 2, the Efficient GlobalPointer only needs to recognize the two entities of ''high blood pressure'' and ''chest distress'' from the sentence of half length by compressing the extraction range. The specific implementation of the context shield window is shown in Fig. 3.

E. CLASS IMBALANCE LOSS

When identifying the type of α after the scoring function $s_\alpha(i,j)$ is determined, the conventional way of thinking to determine the type of all candidate entities by $n(n+1)/2$ binary categories, but the actual situation is that the sentence length of n is too large, resulting in a large number of candidate entities in the actual scenario. The actual number of right entities is only single digits, causing a serious class imbalance problem. Therefore, the multi-label classification loss [40] function $\mathcal{L}_{NER}$ adopted in our method is as shown in the Eq. (14).

$$\begin{aligned} &\mathcal{L}_{NER} \\ &= \log\left(1 + \sum_{(i,j)\in P_\alpha} e^{-s_\alpha(i,j)}\right) + \log\left(1 + \sum_{(i,j)\in Q_\alpha} e^{s_\alpha(i,j)}\right) \end{aligned} \tag{14}$$

where P_α is the starting and ending location set of all entity for type α, Q_α is the negative sample set, i.e., the starting and ending location set of non-entities or entity types that are not α, $s_\alpha(i,j)$ is the scoring function of span $s[i:j]$ for entity type α. Since the entity's start position doesn't appear after the end position, only the combination of $i < j$ needs to be considered. After adding the context shield window of size w, that is:

$$\Omega = \{(i,j) \mid 1 \le i \le j \le w\} \tag{15}$$

$$Q_\alpha = \Omega - P_\alpha \tag{16}$$

IV. PERFORMANCE ANALYSIS

In order to demonstrate the effectiveness of the NER method with proposed KD framework, this section mainly conducts

TABLE 1. The statistical information of CLUENER, CMeEE, CMeEE-V2.

datasets	Train	Test	Sentence length	Entity Type
CLUENER	10748	1343	37.38	10
CMeEE	15000	5000	54.15	9
CMeEE-V2	15000	5000	54.15	9

comparative experiments and analysis. Firstly, the three datasets used in the experiments are introduced; Secondly, the evaluation indicators, experimental parameters and experimental environment are described in detail; Then, several comparative models are introduced and the experimental results of the comparative models are analyzed; Next, the effectiveness of each part in our method are proved through ablation experiments; Finally, the effect of distillation model compression is verified.

A. DATASETS

The experiment is mainly conducted on three Chinese domain public datasets, including a Chinese general domain dataset CLUENER [41] and two Chinese medical domain datasets CMeEE and CMeEE-V2 [42].

- The CLUENER dataset is selected and integrated based on the text classification dataset THUCNEWS released by Tsinghua University. By selecting part of the data for annotation and cleaning, a high-quality fine-grained NER dataset is finally obtained. This dataset has a total of 10 entity categories and more than 10,000 pieces of data. Each data is relatively evenly distributed in 10 different categories.
- CMeEE stands for Chinese Medical Entity Extraction dataset. And the data mainly come from medical textbooks or clinical texts. The data quality index IAA F1 is 0.853, which was obtained by dozens of annotators. The dataset has a total of more than 20,000 pieces of data, with 9 entity types, including diseases, clinical manifestations, drugs, etc.
- CMeEE-V2 is a Chinese medical dataset optimized and improved based on CMeEE. It solves some of the problems of annotation errors in CMeEE, improves the quality of the corpus, and changes the labeling method of entities.

Table 1 provides statistics on the division of the training and test sets for each of the three datasets. For better comparison, the division criteria of all training sets and test sets are the same as those of the comparison models.

B. EXPERIMENTAL SETUP

The proposed NER method is implemented based on pytorch framework, and the parameter optimizer is Adam optimizer. Some of the hyperparameters in our NER method are as follows: Specifically, the learning rate is set to 2e-5 on three datsets, the batch size is set to 64 on CLUENER and 16 on both CMeEE and CMEEE-V2 datasets.

TABLE 2. Experimental environment configurations.

Item	Environment
Operating system	Ubuntu 20.04.4 LTS
CPU	Silver 4114T CPU @ 2.20GHz
GPU	NVIDIA GeForce RTX 3090
Memory	128G
Python version	3.8.1
Pytorch version	1.8.0
Transformers version	4.17.0

In distillation, the pre-trained model of the teacher model is roberta_zh_large_pytorch with hidden size of 1024, and the pre-trained model of the student model is hfl/chinese-roberta-wwm-ext with hidden size of 768. The hyperparameter distillation temperature K for all datasets defaults to 1, and the loss coefficient is set to 5000, 6000, and 7000 for CLUNER, CMeEE, and CMEEE-V2 datasets, respectively.

In order to uniformly measure the effect of entity recognition, the extraction effect is evaluated using precision (P), recall (R) and F1 as the evaluation functions of entity recognition effect, as follows:

$$P = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} \tag{17}$$

$$R = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \tag{18}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{19}$$

where **TP** (true positive) indicates the number of correctly identified positive samples. **FP**(false positive) indicates the number of samples that incorrectly classified negative samples as positive; **FN** (false negative) indicates the number of samples that the model incorrectly classifies as negative. F1 is a combined measure of accuracy P and recall R.

The model complexity statistics are represented by FLOPs and Params. FLOPs (Floating Point Operations) is a measure of the computational complexity of a model and represents the number of floating point operations performed during the model inference process. The higher the FLOPs, the greater the computational effort of the model. Params are the number of parameters in the model, which refers to the total number of adjustable parameters in the model that need to be trained or learned. Params is the number of parameters of a model, including learnable model parameters such as weight matrices and bias terms. The larger the Params, the larger the computational cost. FLOPs and Params are both important indicators for evaluating model complexity and computing resource consumption, which can help to understand the computational performance and scale of the model.

The detailed description of the experiment implementation is provided in Table 2, which mainly includes equipment information and environmental information.

C. RESULTS AND ANALYSIS

In this section, several comparison methods are introduced first, and then the proposed NER method is compared and analyzed with these comparison methods. Then, the effectiveness of the proposed distillation method and the context shield window is further demonstrated by ablation experiments. Finally, through parameter analysis, we prove that the proposed KD method can reduce the parameters and improve the recognition effect.

1) EXPERIMENTAL COMPARISON AND DISCUSSION

In order to prove the effectiveness of the entity recognition method, several mainstream methods compared with our NER method are summarized as follows:

- LSTM+CRF: This method is a NER method combining LSTM and CRF, which can effectively process sequence data and carry out label prediction.
- BERT-CRF [43]: A mainstream named entity recognition method based on sequence annotation, which mainly uses BERT pre-trained model and additional Conditional Random Field (CRF) to constrain the sequence to realize entity recognition.
- PFN [44]: The model defines tasks as form-filling problems and then extracts them based on the Parition Filter Network. PFN divides the feature encoding into two steps: partitioning and filtering.
- TsERL [45]: A model applied for Chinese medical NER enhances the performance of Chinese medical entity recognition by making full use of lexical and radical information through two stages.
- BERT-Biaffine [46]: A model that defines NER as tasks that determine the start and end indexes, which solves the Nested NER using the biaffine mechanism.
- GlobalPointer [10]: Based on the idea of GlobalPointer, it realizes the identification of nested and flat entities in a unified way, and also uses rotary position coding (RoPE) to make full use of position information.
- FFBLEG [47]: The model embeds pinyin information through feature fusion and uses bidirectional lattice embedding graph to realize Chinese nested and flat NER task.
- Efficient GlobalPointer [10]: Efficient GlobalPointer is realized by splitting entities into two steps of identification and classification on the basis of GlobalPointer, which improves the utilization of parameters to a certain extent.

In order to more comprehensively verify the effectiveness of the proposed NER method, comparative experiments are carried out on the CLUENER dataset and CMeEE dataset respectively, and compared with the above methods, the results shown in table 3 are finally obtained. The evaluation indicators covered mainly include P, R and F1. The metrics obtained from our proposed method are the mean and confidence intervals (0.95) from the results of ten different training sessions.

As shown in table 3, the bolded results are the best values. On the CLUENER dataset, our NER method is 1.33%, 0.81% and 1.08% higher than the best comparison model Efficient GlobalPointer in the comparison method on P, R and F1, respectively. On the CMeEE dataset, in addition to the lower P indicator, R improved significantly by 4.57% compared to the best-performing GlobalPointer, and F1 improved by 2.01% compared to the best-performing Efficient GlobalPointer. It is worth mentioning that the pre-trained model we used has about 72% fewer parameters and floating points than the best-performing Efficient GlobalPointer.

2) ABLATION EXPERIMENTS

Ablation experiments are performed on CLUENER, CMeEE, and CMEEE-V2 datasets to demonstrate and analyze the effectiveness of the proposed knowledge distillation framework and the context shield window, and to verify their contribution to the overall improvement. The influence of each point on the whole NER method is verified by removing one part of the knowledge distillation and context shield window one by one during each experiment.

As can be seen from Table 4, after removing the knowledge distillation (KD) part proposed by this method, the indicator F1 decreases by 0.81%, 1.62% and 1.63% on CLUENER, CMeEE and CMEEe-V2 datasets, respectively. When the Context Shield Window (CSW) is removed, the indicator F1 dropped by 0.16%, 0.78%, and 0.97% on three datasets, respectively. The effectiveness of the distillation part and the context shield window in this method is proved by the ablation experiments. That is, distillation operation can improve the extraction effect of model knowledge, and context shield window can improve the performance of entity recognition by limiting the extraction range.

3) COMPARISON OF MODEL COMPRESSION EFFECT

In Table 5, the realization of Efficient GlobalPointer knowledge distillation Method is described in detail, and the comparison results between traditional KD distillation and the knowledge distillation framework adopted by Our KD Method are given. The teacher network selects roberta_zh_large_pytorch, a pre-trained model with a larger number of participants, so the entity recognition effect is better. The student network choose hfl/chinese-roberta-wwm-ext, a lighter pre-trained model. Therefore, the performance indicator F1 is weaker than the teacher network. When the traditional KD method is used, there is a certain improvement in the three datasets compared with the teacher network and the student network in indicator F1. When using the proposed distillation framework, the indicator F1 can be further improved by 0.39%, 0.47% and 0.97% on the basis of traditional knowledge distillation framework. The overall results are 1.09%, 1.92% and 2.32% higher than student model, which uses the same pre-trained RoBERTa weights. It is also proved that this distillation method can solve the problem that the feature learning process in traditional knowledge distillation is independent of each other and it is

TABLE 3. Comparison results with different methods on the CLUENER and CMeEE dataset.

Methods	CLUENER			CMeEE		
	Precision	Recall	F1	Precision	Recall	F1
LSTM+CRF	0.7106	0.6897	0.7000	0.5487	0.4902	0.5178
BERT-CRF [43]	-	-	0.7870	0.5834	0.6408	0.6107
PFN [44]	-	-	0.7929	-	-	0.6368
TsERL [45]	-	-	-	0.6182	0.6478	0.6327
BERT-Biaffine [46]	-	-	-	0.6417	0.6129	0.6229
GlobalPointer [10]	0.7874	0.8059	0.7966	0.6517	0.6474	0.6442
FFBLEG [47]	-	-	-	0.6470	0.6492	0.6481
Efficient GlobalPointer [10]	0.7966	0.8109	0.8035	**0.6645**	0.6469	0.6504
Ours	**0.8099**	**0.8190**	**0.8143**	0.6483	**0.6931**	**0.6705**

TABLE 4. Results of ablation experiments on the CLUENER, CMeEE and CMeEE-V2 dataset.

Method	CLUENER	CMeEE	CMeEE-V2
	F1	F1	F1
Ours	0.8143	0.6705	0.7585
-KD	0.8062	0.6543	0.7422
-CSW	0.8127	0.6627	0.7488

TABLE 5. Comparative experiments of knowledge distillation on CLUENER, CMeEE and CMEEE-V2 datasets.

Method	CLUENER	CMeEE	CMeEE-V2
	F1	F1	F1
Teacher Model	0.8062	0.6543	0.7422
Student Model	0.8034	0.6513	0.7353
Traditional KD	0.8104	0.6658	0.7488
Our KD Method	0.8143	0.6705	0.7585

difficult to ensure that each feature of the output of the student network matches the corresponding feature of the teacher network by using the linkage mechanism.

In this study, FLOPs and Params are counted for several main models involved in the entire experiment. The specific results are shown in Table 6. First of all, from the comparison of table 3 and table 6, it can be seen that the indicator F1 of the efficient GlobalPointer is improved by 0.69% and 0.62% respectively compared with the global pointer model on the two datasets, while the number of FLOPs and Params are reduced by 0.25G and 1.03M respectively. If the student model uses the lightweight pre-trained model hfl/chinese-roberta-wwm-ext, although the computational complexity indicators FLOPs and Params are 71.84% and 71.74% respectively less than the teacher model, the performance indicator F1 will decrease. It can be seen from Table 5 that when our KD framework is used, the performance

TABLE 6. Statistics of floating point operations and parameter quantities of the model.

Models	FLOPs	Params
GlobalPointer	71.60 G	304.54 M
Efficient GlobalPointer	71.35 G	303.51 M
Teacher Model	71.35 G	303.51 M
Student Model	20.09 G	85.76 M

indicator F1 of the student model can be significantly improved by 1.09%, 1.92% and 2.32% while maintaining the model calculation complexity. These experiments show that our proposed model can achieve a further performance improvement of up to 2.32% when using only a pre-trained model with approximately 72% fewer parameters and floating-point numbers. These analyses further prove that the proposed rectification method has the advantages of both compression model and performance improvement.

D. ENGINEERING APPLICATIONS

The medical field contains a large amount of medical literature, medical record data, and drug instructions. These texts contain a large number of professional terms, medical entities, and concepts. Manual processing of these data is very time-consuming and labor-intensive, and is prone to omissions and errors. Therefore, using medical big data analysis technology for medical named entity recognition can greatly improve the organization and computability of medical information and facilitate the efficiency and accuracy of data processing, thereby promoting the accumulation of medical knowledge and the development of clinical research.

Specifically, the massive electronic medical records from hospitals and medical data queried online are sent to the Storage and Executive layer to be cleaned and organized by the NER method proposed in this paper. In the process of medical data analysis, the system realizes nested entity and flat entity recognition of medical texts by using the Efficient GlobalPointer model after knowledge distillation, and then

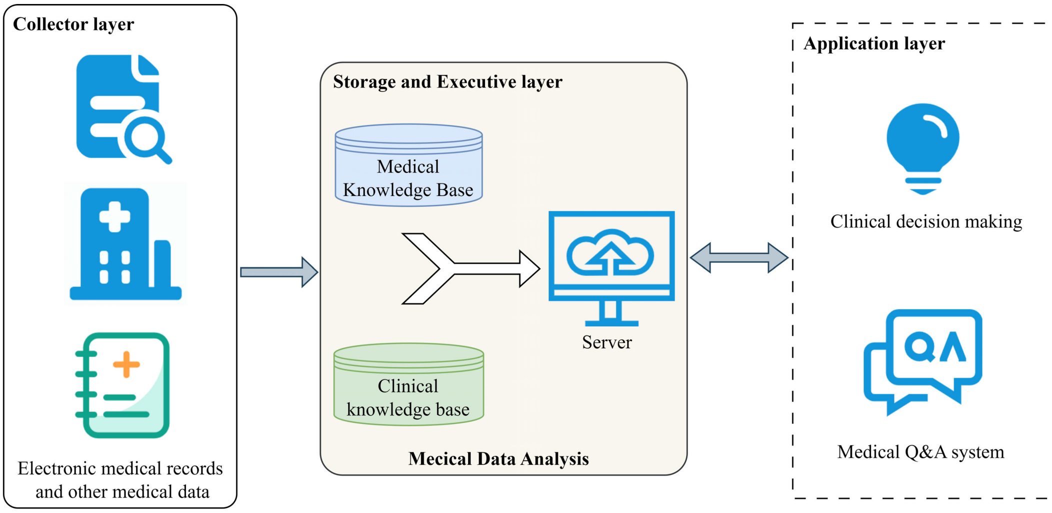

FIGURE 4. **The engineering application architecture.**

stores clinical data and medical data in the corresponding knowledge base respectively. Among them, the proposed knowledge distillation method improves the generalization ability of the model to adapt to different medical application scenarios, and the Efficient GlobalPointer reduces the interference of redundant information by using context shield window. The whole operation improves the quality of medical texts and reduces the errors and noise, which provides a reliable data basis for subsequent analysis work. At the same time, it can also provide support for clinical decision making, medical question and answer system and other applications, which is of great significance for realizing automated data analysis and model construction. The engineering application architecture is shown in Fig. 4.

V. CONCLUSION

In this paper, we propose a novel NER method using our proposed KD framework for biomedical and clinical text analysis and management. The proposed KD framework introduces a linkage mechanism to link two features as input and output components of the teacher model and solve the problem that feature learning processes are independent of each other in distillation. Besides, our NER method introduces the context shield window to Efficient GlobalPointer to solve the problem of redundant information interference. Finally, through the experimental comparison with comparative methods in three datasets, the superiority of our proposed method in both performance and scalability is proved.

The future goal of our research is to combine the knowledge distillation method with transfer learning to further improve the generalization ability and performance of small models. This enhancement can better adapt to more demanding medical scenarios and improve the stability of named entity recognition model.

REFERENCES

[1] Y. Zhang, J. Hong, and S. Chen, "Medical big data and artificial intelligence for healthcare," *Appl. Sci.*, vol. 13, no. 6, p. 3745, Mar. 2023.

[2] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *Proc. 16th Conf. Comput. Linguistics*, 1996.

[3] X. Zhou, X. Zhang, C. Tao, J. Chen, B. Xu, W. Wang, and J. Xiao, "Multi-grained knowledge distillation for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 5704–5716.

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

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[6] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[8] J. S. McCarley, R. Chakravarti, and A. Sil, "Structured pruning of a BERT-based question answering model," 2019, *arXiv:1910.06360*.

[9] Y. Cho, G. Ham, J.-H. Lee, and D. Kim, "Ambiguity-aware robust teacher (ART): Enhanced self-knowledge distillation framework with pruned teacher network," *Pattern Recognit.*, vol. 140, Aug. 2023, Art. no. 109541.

[10] J. Su, A. Murtadha, S. Pan, J. Hou, J. Sun, W. Huang, B. Wen, and Y. Liu, "Global pointer: Novel efficient span-based approach for named entity recognition," 2022, *arXiv:2208.03054*.

[11] P. Dino, S. Kumar, A. Ali, and H. Raj, "Bio-NER: Biomedical named entity recognition using rule-based and statistical learners," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 12, 2017.

[12] L. F. Rau, "Extracting company names from text," in *Proc. 7th IEEE Conf. Artif. Intell. Appl.*, 1991, pp. 29–30.

[13] X. Wang, Y. Zhang, Q. Li, C. H. Wu, and J. Han, "PENNER: Pattern-enhanced nested named entity recognition in biomedical literature," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 540–547.

[14] X. Li and D. Roth, "Learning question classifiers: The role of semantic information," *Natural Lang. Eng.*, vol. 12, no. 3, pp. 229–249, Sep. 2006.

[15] K.-J. Lee, Y.-S. Hwang, S. Kim, and H.-C. Rim, "Biomedical named entity recognition using two-phase model based on SVMs," *J. Biomed. Informat.*, vol. 37, no. 6, pp. 436–447, Dec. 2004.

[16] R. Leaman, C.-H. Wei, C. Zou, and Z. Lu, "Mining chemical patents with an ensemble of open systems," *Database*, vol. 2016, 2016, Art. no. baw065.

[17] N. Ponomareva, P. Rosso, F. Pla, and A. Molina, "Conditional random fields vs. hidden Markov models in a biomedical named entity recognition task," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, 2007, pp. 479–483.
[18] Y. Zhang, Z. Xu, and T. Zhang, "Fusion of multiple features for Chinese named entity recognition based on CRF model," in *Proc. Asia Inf. Retr. Symp.* Springer, 2008, pp. 95–106.
[19] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "ProMiner: Rule-based protein and gene entity recognition," *BMC Bioinf.*, vol. 6, no. S1, pp. 1–9, May 2005.
[20] A. P. Quimbaya, A. S. Múnera, R. A. G. Rivera, J. C. D. Rodríguez, O. M. M. Velandia, A. A. G. Peña, and C. Labbé, "Named entity recognition over electronic health records through a combined dictionary-based approach," *Proc. Comput. Sci.*, vol. 100, pp. 55–61, Jan. 2016.
[21] J. Feng, Z. Li, and D. Zhang, "Bridge detection text named entity recognition based on hidden Markov model," *Traffic World*, vol. 8, pp. 32–33, Jan. 2020.
[22] A. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
[23] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Oct. 2020.
[24] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for BERT model compression," 2019, *arXiv:1908.09355*.
[25] W. Yoon, C. H. So, J. Lee, and J. Kang, "CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition," *BMC Bioinf.*, vol. 20, no. S10, pp. 55–65, May 2019.
[26] J. Frei and F. Kramer, "GERNERMED: An open German medical NER model," *Softw. Impacts*, vol. 11, Feb. 2022, Art. no. 100212.
[27] M. Sun, Z. Guo, and X. Deng, "Intelligent BERT-BiLSTM-CRF based legal case entity recognition method," in *Proc. ACM Turing Award Celebration Conf.*, Jul. 2021, pp. 186–191.
[28] W. Gao, X. Zheng, and S. Zhao, "Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF," *J. Phys., Conf. Ser.*, vol. 1848, no. 1, Apr. 2021, Art. no. 012083.
[29] A. K. Tarcar, A. Tiwari, V. N. Dhaimodker, P. Rebelo, R. Desai, and D. Rao, "Healthcare NER models using language model pretraining," 2019, *arXiv:1910.11241*.
[30] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[32] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3513–3521.
[33] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," 2019, *arXiv:1909.10351*.
[34] R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization," *J. Biomed. Informat.*, vol. 47, pp. 1–10, Feb. 2014.
[35] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting pre-trained models for Chinese natural language processing," 2020, *arXiv:2004.13922*.
[36] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," 2021, *arXiv:2104.09864*.
[37] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1243–1252.
[38] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv:1803.02155*.
[39] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.
[40] J. Su, M. Zhu, A. Murtadha, S. Pan, B. Wen, and Y. Liu, "ZLPR: A novel loss for multi-label classification," 2022, *arXiv:2208.02955*.
[41] L. Xu, Y. Tong, Q. Dong, Y. Liao, C. Yu, Y. Tian, W. Liu, L. Li, C. Liu, and X. Zhang, "CLUENER2020: Fine-grained named entity recognition dataset and benchmark for Chinese," 2020, *arXiv:2001.04351*.
[42] N. Zhang et al., "CBLUE: A Chinese biomedical language understanding evaluation benchmark," 2021, *arXiv:2106.08087*.
[43] S. Hu, H. Zhang, X. Hu, and J. Du, "Chinese named entity recognition based on BERT-CRF model," in *Proc. IEEE/ACIS 22nd Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2022, pp. 105–108.
[44] Z. Yan, C. Zhang, J. Fu, Q. Zhang, and Z. Wei, "A partition filter network for joint entity and relation extraction," 2021, *arXiv:2108.12202*.
[45] D. Yang, H. Yang, and B. Wu, "TsERL: Two-stage enhancement of radical and lexicon for Chinese medical named entity recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2022, pp. 2719–2726.
[46] J. Yu, B. Bohnet, and M. Poesio, "Named entity recognition as dependency parsing," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6470–6476.
[47] Q. Cong, Z. Feng, G. Rao, and L. Zhang, "Chinese medical nested named entity recognition model based on feature fusion and bidirectional lattice embedding graph," in *Proc. Int. Conf. Database Syst. Adv. Appl.* Cham, Switzerland: Springer, 2023, pp. 314–324.

ZHENGWEI ZHAI is currently pursuing the degree with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology. His research interests include knowledge graphs and information extraction.

RONGLI FAN received the Ph.D. degree in genetics from Zhejiang University, China. She is currently a Lecturer with Zhejiang University of Science and Technology, China. Her research interests include bioinformatics, biomarker discovery, and gene editing.

JIE HUANG received the Ph.D. degree from Zhejiang University, Hangzhou, China, in 2008. He is currently an Associate Professor with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology. He is the author of two books, more than 20 articles, and more than ten patents. His research interests include cloud computing, blockchain, and bioinformatics.

NEAL N. XIONG (Senior Member, IEEE) received the Ph.D. degree in sensor system engineering from Wuhan University, in 2007, and the Ph.D. degree in dependable communication networks from Japan Advanced Institute of Science and Technology, in 2008. Before he attended Sul Ross State University, he worked with Georgia State University, Northeastern State University, and Colorado Technical University (a Full Professor for about five years) for about ten years. He is currently an Associate Professor (four-year credits) with the Department of Computer Science and Mathematics, Sul Ross State University, Alpine, TX, USA. His research interests include cloud computing, security and dependability, parallel and distributed computing, networks, and optimization theory. He has received the Best Paper Award from the 10th IEEE International Conference on High-Performance Computing and Communications (HPCC-08) and the Best Student Paper Award from the 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009). He has been the general chair, the program chair, the publicity chair, a program committee member, and an organizing committee member of more than 100 international conferences. He has been a Reviewer of about 100 international journals, including IEEE Journal on Selected Areas in Communications, IEEE Transactions on Systems, Man, and Cybernetics: Systems (Park: A/B/C), IEEE Transactions on Communications, IEEE Transactions on Mobile Computing, IEEE Transactions on Parallel and Distributed Systems. He is serving as the Editor-in-Chief, an associate editor, or an editor member for more than ten international journals.

LIJUAN ZHANG received the B.S. degree in applied mathematics from Inner Mongolia University, Inner Mongolia, China, in 2015, and the M.S. and Ph.D. degrees in communication and information systems from Dalian University of Technology, Dalian, China, in 2017 and 2021, respectively. Since 2021, she has been a Lecturer with the Department of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China. Her current research interests include machine learning, signal detection, and biomedicine.

JIAN WAN received the Ph.D. degree in computer application technology from Zhejiang University, Zhejiang, China, in 1996. He is currently a Professor with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology. His research interests include bioinformatics, service computing, and cloud computing.

LEI ZHANG received the Ph.D. degree in microbiology from the College of Life Sciences, Zhejiang University, China. She is currently a Professor with Zhejiang University of Science and Technology, China.