Bi-LSTM-CRF Network for Clinical Event Extraction With Medical Knowledge Features

SHUNLI ZHANG¹, YANCUI LI², SHIYONG LI¹, AND FANG YAN¹
¹Department of Information Engineering, Henan Institute of Science and Technology, Xinxiang, Henan 453003, China
²College of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan 453007, China
Corresponding author: Shunli Zhang (zhangshunli@hist.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502149 and Grant 22YJCZH091, and in part by the Science and Technology Project of Henan Province under Grant 222102210020.

ABSTRACT Extracting clinical event expressions and their types from clinical text is a fundamental task for many applications in clinical NLP. State-of-the-art systems need handcraft features and do not take into account the representation of the low-frequency words. To address these issues, a Bi-LSTM-CRF neural network architecture based on medical knowledge features is proposed. First, we employ convolutional neural networks (CNNs) to encode character-level information of a word and extract medical knowledge features from an open-source clinical knowledge system. Then, we concatenate character-level and word-level embedding and the medical knowledge features of words together, and feed them into bi-directional long short-term memory (Bi-LSTM) to build context information of each word. Finally, we jointly use a conditional random field (CRF) to decode labels for the whole sentence. We evaluate our model on two publicly available clinical datasets, namely THYME corpus and 2012 i2b2 dataset. Experimental results show that our model outperforms previous state-of-the-art systems with different methodologies, including machine learning-based methods, deep learning-based methods, and Bert-based methods.

INDEX TERMS Clinical text, entity recognition, deep learning, natural language processing.

I. INTRODUCTION
Recently, automatically extracting clinical event spans and their types from unstructured clinical narratives has received considerable research attention because of its important role on medical research or applications [1], [2]. In the clinical natural language processing (NLP) community, there are a number of evaluation tasks [3], [4], [5], [6] about clinical event identification. Clinical events refer to the phrases associated with a patient’s timeline including procedures, diseases, diagnoses, etc. The task belongs to the named entity recognition (NER) task in clinical natural language processing.

Researchers have explored many ways to extract event expressions in clinical domains, the more effective methods are rule-based, machine learning-based methods and deep learning-based methods. Rule-based methods [7], [8], [9] were the dominant methods in the early clinical entity recognition systems, which mainly relied on biomedical dictionaries and regular expressions to recognize entities. Along with the provision of publicly corpora by some share challenge organizations, lots of machine learning-based methods [10], [11], [12] were proposed, such as conditional random field (CRF), hidden Markov model (HMM), and support vector machine (SVM). Although these models have achieved the best results in many clinical tasks, they depend heavily on feature engineering. With the popularization of neural network in image processing and natural language processing [13], [14], [15], the methods based on deep learning such as recurrent neural networks (RNNs) [16], RNN-LSTM [17], and BERT-based [18], [19], [20] have been proposed in clinical information extraction tasks in recent years. Compared with traditional machine learning methods, these deep learning-based methods can automatically learn word context representations (word embeddings) from large unlabeled clinical text, reducing manual intervention. However, neither
RNN-based methods nor BERT-based methods can solve the identification errors caused by low-frequency words. Because of the existing plenty of non-uniform writing, medical terms and abbreviations in clinical notes, there are more low-frequency words in clinical notes than in general texts in other fields. Some previous studies [21], [22] have proved that medical domain knowledge can reduce this identification error.

To minimize artificial features in the process of recognition and reduce the recognition errors caused by low-frequency words, we propose a deep learning-based architecture for recognizing clinical event recognition, whose architecture is illustrated in Figure 2. It only requires pre-trained word embeddings on unlabeled datasets and clinical meaning features based on medical knowledge. As a result, our approach can be easily migrated to information extraction tasks in different fields and languages. Overall, firstly, the convolutional neural networks (CNNs) are employed to model the morphological structure of a word into its character-level word embedding; clinical meaning feature is extracted from an open-source clinical knowledge system. Subsequently, character-level and word-level word embeddings and the clinical meaning features of the whole sentence are concatenated and fed into the bi-directional LSTM (Bi-LSTM) to model context information. Ultimately, a CRF layer is utilized to jointly decode labels for the whole sentence. We evaluate our model on two public clinical datasets: THYME corpus and the 2012 i2b2 dataset. Evaluation results show that our model outperforms previous top systems and we are able to obtain state-of-the-art performance on the datasets only using domain knowledge as features. The contributions of our paper are as follows:

1. We propose a neural architecture to model the context information of each word in a sentence, which can automatically learn word context information and avoid recognition errors caused by manual features.

2. We introduce a domain knowledge feature to augment the semantic association of different clinical vocabularies with similar meanings, which can improve the recognition rate of low-frequency words.

3. We evaluate our method on two well-known clinical event recognition datasets. On each dataset, our method outperforms other state-of-the-art methods, including machine learning-based methods, deep learning-based methods, and BERT-based methods.

II. RELATED WORK

Numerous approaches have been proposed to identify clinical event expressions and their types, which can be divided into three categories: rule-based methods, machine learning-based methods and deep learning-based methods.

A. RULE-BASED METHODS

In the early studies, due to the lack of public datasets in the medical fields, the identification of clinical entities mainly adopted the rule-based methods, such as MedLEE [7], UMLS [23], cTAKES [9] etc. MedLEE was designed to discover clinical concepts in radiology reports, later broadened to other domains. In UMLS, medical knowledge is organized by concept meaning, synonymous terms are clustered together to form a concept and concepts are linked to other concepts by means of various types of relationships. The cTAKES is an open-source natural language processing system for clinical information extraction, which implements a terminology dictionary look-up algorithm for identifying clinical entities, mapping each entity to UMLS concept unique identifier (CUI).

These rule-based systems identified medical entities based on medical dictionaries and medical knowledge databases by designing regular expressions. they need to pay much attention on regular design and relies on task-specific knowledge. Although they are no longer the mainstream methods for medical entity recognition, they are often used as medical NLP tools to obtain lexical or semantic features.

B. MACHINE LEARNING-BASED METHODS

Compared with the rule-based methods, machine learning-based methods have proven to be more effective for clinical event recognition. In 2012 i2b2 challenges, CRF is the most popular method for clinical event extraction [3]. Xu et al. [24] employed a two-phase classifier based on CRFs, firstly recognize the event boundaries, and then classified the types of events. Tang et al. [25] also used a two-stage CRF pipeline model, the first stage recognized three types medical entities, and the second stage identified the remaining three types entities, the results of the first stage were used as features for the second stage. In Semeval-2016 Task12, Cohan et al. [11] trained a CRF model for detecting event expressions and supplement the CRF output spans with manual-crafted rules. Lee et al. [26] employed a HMM-SVM sequence tagger to identify the spans of event expressions as well as their types simultaneously, achieved the best F1 score in Semeval-2016 Task12, they used various features that have been successfully used for many entity recognition tasks in clinical domain such as lexical features, syntactic features, and discourse level features. Grouin and Moriceau [27] used CRF system and merged the tasks of spans and types identification together, they used lots of features in common entity recognition tasks and fifty Brown clusters as features in this system.

Though these machine learning-based methods have received have achieved satisfactory results, but a variety of feature engineering are time-consuming and costly.

C. DEEP LEARNING-BASED METHODS

Over the past few years, several Deep Learning-based Methods have been proposed in clinical domain [28], [29], [30]. In Semeval-2016 Task12 Challenge, Fries [16], Li and Huang [31] employed RNN to identify event expressions, but the result is lower than machine learning methods. Tourille et al. [32] constructed a Bi-LSTM neural Architecture for clinical temporal relation extraction. Liu et al. [17] employed two neural networks representing character-level
embedding to investigate the performance of LSTM on clinical entity recognition. Wu et al. [21] examined three ways to integrate medical knowledge into deep leaning methods; the results showed medical domain knowledge can improve the decent representation of low-frequency words.

Recently, with the emergence of BERT (Bidirectional Encoder Representation from Transformers), a popular pre-trained language representation model, some BERT-based methods have achieved new state-of-the-art performances in clinical tasks. RoBERTa-MIMIC integrated the four clinical BERT-based models into an open-source package for clinical concept extraction [19]. Bio_BERT was pre-trained on large-scale biomedical corpora, it largely outperforms BERT and previous best models in biomedical domain tasks, such as biomedical named entity recognition, biomedical relation extraction tasks [18]. UmlsBERT used the unified medical language system (UMLS) metathesaurus to augment contextual word embedding [20]. Though these BERT-based methods have achieved new state-of-the-art performances in clinical tasks, but they have two common disadvantages. one is that the computing cost of pre-training and fine tuning is huge, the other is that BERT-based methods used some algorithms to breaks words down into sub-word tokens, sometimes the results may not exactly conform to traditional medical term morphology (e.g., “chemotherapy” is broken into “che”, “mot”, “her”, “apy”, as opposed to having the prefix “chemo”), this reduces the accuracy of identification.

III. THE PROPOSED METHOD

We employ recurrent neural networks with long short-term memory to construct our architecture. Figure 1 illustrates the architecture of our neural network method. In particular, the architecture is composed of four layers: (1) word preprocessing layer, (2) input embedding layer, (3) Bi-LSTM layer, and (4) CRF layer. Specifically, word preprocessing layer split every document into sentences and then into words, the input embedding layer maps each word of a sentence to a vector representation sequence. The vector representation sequence of a sentence is fed into the Bi-LSTM layer, which outputs the sequence of probability vectors containing the probability of each label for each corresponding word. Finally, the CRF layer outputs the most likely sequence of predicted labels based on the sequence of probability vectors from the previous layer. All layers are learned jointly.

A. WORD PREPROCESSING LAYER

In the word preprocessing layer, we used the Stanford CoreNLP toolkit (version 3.4) [33] to split documents into sentences and then tokenized these sentences into words. We transformed all the words into their lowercase forms, replaced all the numbers with the letter ‘‘X’’ and remove all verb tenses.

B. INPUT EMBEDDING LAYER

The input embedding layer takes a word sequence of a sentence \(s = \{s_1, s_2, \ldots, s_n\}\) as input and outputs its vector representation sequence \(\{x_1, x_2, \ldots, x_n\}\). An overview of the word representation layer is presented in Figure 2. The input embedding of a word the concatenation of four embeddings: random embedding, character-level embedding, word-level embedding and mk-embedding. Random word embeddings are randomly initialized firstly, then tuned during training jointly.

1) CHARACTER-LEVEL EMBEDDING

Previous studies have shown the morphological features (like the prefix or suffix of a word) are useful for clinical event recognition. Following Ma et al. [34], convolutional neural networks (CNNs) are employed to model character-level information of a word.

Given a word \(s = \{a_1, a_2, \ldots, a_m\}\), \(a_i\) denotes its \(i\)-th character. We first map each character \(a_i\) to a char embedding \(\text{emb}(a_i)\). Given \(C\) represents the convolution window size, the vector representation \(r_i^{\text{chr}}\) of character \(a_i\) is the concatenation of the char embeddings of \(2 \times C + 1\) characters. For example, if \(C = 1\),

\[
r_i^{\text{chr}} = [\text{emb}(a_{i-1}), \text{emb}(a_i), \text{emb}(a_{i+1})]
\]

where ‘‘[]’’ represents the vector concatenation operation. Then the convolutional kernel of CNN needs \(m\) times of convolutions for all the characters in the word and for each convolution \(i\), the kernel output \(Z_i\) is calculated by

\[
Z_i = \tan W^1 r_i^{\text{chr}} + b^1
\]

where \(W^1\) and \(b^1\) are parameter matric and bias vector obtained by learning, and \(\tanh\) is the activation function. To get the character-level vector representation \(r_s^{\text{chr}}\) of the
problems\textsuperscript{37}. LSTM networks are designed to cope with these problems due to the gradient vanishing/exploding\textsuperscript{35},\textsuperscript{36}, but in practice it is limited to looking back only a few steps. In theory, it can make use of information in a sentence of any length, but in practice it is limited to looking back only a few steps due to the gradient vanishing/exploding problems \textsuperscript{35, 36}. LSTM networks are designed to cope with these problems \textsuperscript{37}.

In previous studies \textsuperscript{20, 21}, \textsuperscript{20, 21}. In this paper, the UMLS semantic network has proved to be effective in pre-processing words,\textsuperscript{20, 21}. UMLS semantic network has proved to be effective in pre-processing words. For example, the words ''lungs'' and ''pulmonary'' are linked to the common concept unique identifier (CUI) CUI: C0024109, and the words ''chest'' and ''cavity'' have the same semantic type ''Body Location or Region''. The semantic type feature from the UMLS semantic network has proved to be effective in previous studies \textsuperscript{20, 21}. In this paper, the UMLS semantic type of a word is obtained by the open-source system cTAKES\textsuperscript{9}, the words that are not identified in UMLS are set to a zero-filled vector.

3) MK-EMBEDDING

The mk-embedding is the semantic types defined in UMLS. In UMLS, medical terms sharing a similar meaning are mapped to the same concept, and concepts are organized according to their semantic types. For example, the words “lungs” and “pulmonary” are linked to the common concept unique identifier (CUI) CUI: C0024109, and the words “chest” and “cavity” have the same semantic type “Body Location or Region”. The semantic type feature from the UMLS semantic network has proved to be effective in previous studies \textsuperscript{20, 21}. In this paper, the UMLS semantic type of a word is obtained by the open-source system cTAKES\textsuperscript{9}, the words that are not identified in UMLS are set to a zero-filled vector.

C. BI-LSTM LAYER

Recurrent Neural Network (RNN) is a neural network architecture designed to model the sequential information. In theory, it can make use of information in a sentence of any length, but in practice it is limited to looking back only a few steps due to the gradient vanishing/exploding problems \textsuperscript{35, 36}, and LSTM networks are designed to cope with these problems \textsuperscript{37}.

LSTM networks are same as RNN except that the hidden layer units are replaced by purpose-built memory cells called LSTM cell\textsuperscript{38}. The LSTM cell is composed of multiple gates which control the proportions of information to forget and pass on to the next time step. More specifically, given a sequence of vectors $x_1, x_2, \ldots, x_n$, at each step $t \in 1, 2, \ldots, n$, each LSTM unit takes $\{x_t, h_{t-1}, c_{t-1}\}$ as input and produces the hidden state $h_t$ and the memory cell $c_t$ based on the following formulas:

$$f_t = \sigma(W_f \cdot [h_{t-1} \oplus x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1} \oplus x_t] + b_i)$$
$$c_t = \tanh(W_c \cdot [h_{t-1} \oplus x_t] + b_c)$$
$$o_t = \sigma(W_o \cdot [h_{t-1} \oplus x_t] + b_o)$$
$$h_t = o_t \times \tanh(c_t)$$

(4)

For a standard LSTMs, its disadvantage is that cannot make use of the future information of a sentence, only knowing about the past context meanings. Clinical event expression recognition is a sequence labeling task; it is beneficial to have access to the context information (both previous and latter) in a sentence. Bidirectional LSTM (Bi-LSTM) is an effective solution to solve this problem. The main idea is to feed each sequence into forwards and backwards LSTMs respectively, the results of two separate LSTMs (the two hidden layers) are concatenated to form the final output.

Figure 1 shows an example of bidirectional LSTMs with a clinical sentence. For a given sentence $s = \{s_1, s_2, \ldots, s_n\}$ in the clinical text containing $n$ words, the Bi-LSTM layer takes the vector sequence $(x_1, x_2, \ldots, x_n)$ as input. $i_t$ is the concatenation of the word-level, character-level embedding, random embedding, and mk-embedding of $x_t$ given by

$$x_t = [\text{emb}(s_t), r_{s_t}^{chr}, \text{rand}(s_t), \text{mk}(s_t)]$$

(5)

Based on $x = (x_1, x_2, \ldots, x_n)$, we compute two representations: the forward LSTM calculates the forward hidden states $h_t^{(f)} = \{h_1^{(f)}, h_2^{(f)}, \ldots, h_n^{(f)}\}$ and the backward LSTM calculates the backward hidden states $h_t^{(r)} = \{h_1^{(r)}, h_2^{(r)}, \ldots, h_n^{(r)}\}$. The sequence $\tilde{h_t} = \{\tilde{h_1}, \tilde{h_2}, \ldots, \tilde{h_n}\}$ of the Bi-LSTM layer is obtained by concatenating the hidden states of the forward and backward LSTMs. This procedure can be formulated as

$$\tilde{h_t} = \tanh(W_h[h_t^{(f)} \oplus h_t^{(r)}] + b_h)$$

(6)

D. CRF LAYER

CRFs, a class of undirected probabilistic graphical model Lafferty et al.\textsuperscript{39}, it has been demonstrated to be very effective in extracting entities in clinical domains. Despite the Bi-LSTM model success in some tasks like POS tagging, its independent classification labels are limited. It is beneficial to consider the correlation between adjacent labels in sequence labeling tasks\textsuperscript{40}. Therefore, in Figure 1, we construct our model by feeding the output vectors of Bi-LSTM into a CRF layer.

For an input sentence: $x = (x_1, x_2, \ldots, x_n)$, We denote $P$ as the matrix of scores output by the Bi-LSTM layer. The size
of \( P \) is \( n^k \), where \( k \) is the number of different labels, and \( P_{ij} \) corresponds to the score of the \( j \)-th label of the \( i \)-th word in a sentence. For a sequence of predictions \( y = (y_1, y_2, \ldots, y_n) \), its score can be formulated as follow:

\[
S(x, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} p_{i, y_i}
\]

(7)

where \( A \) is a matrix of transition scores, \( A_{ij} \) represents the score of a transition from the label \( i \) to the label \( j \). \( y_0 \) and \( y_{n+1} \) are the start and end labels of a sentence, which we add to the set of possible labels. We predict the output sequence that obtains the maximum score given by:

\[
y^* = \arg \max(S(x, y))(\hat{y} \in f(x))
\]

(8)

We encoded output labels in IOB2 format [41]. The label B-EVIDENTIAL represents a token which is the beginning of an EVIDENTIAL-type event entity, and I-N/A represents a token which is inside an N/A-type event entity. Through this approach, the CRF model predicts the span and the type of an entity simultaneously.

IV. EXPERIMENTAL RESULTS

In this section, we make a comprehensive evaluation of the proposed method. In particular, we compare our method with different state-of-the-art methods to show the efficiency of our proposed model on two public clinical event recognition tasks. In addition, we provide an ablation analysis to examine the effectiveness of each part in our method. We also conduct a comparison of different categories, computation cost. Finally, we conduct the case study to further validate the merits of our method in an intuitive way.

A. EXPERIMENTAL SETTINGS

1) DATASETS

We evaluate the proposed method on two widely used public clinical datasets: The THYME corpus [42] and 2012 i2b2 dataset [3]. The THYME corpus contains 192 de-identified clinical and pathology notes of cancer patients collected from the Mayo Clinic Center, which was released several times in Clinical TempEval challenges [4], [5], [43]. The training, development, and test set include 293, 147, and 152 documents, which contain 38890, 20974, and 18990 event expressions respectively. The corpus includes three types of clinical event entities, namely ASPECTUAL, EVIDENTIAL and N/A. The 2012 i2b2 dataset contains 310 discharge records from Partners Healthcare and the Beth Israel Deaconess Medical Center, which was released by 2012 i2b2 challenge. The training and test set include 190 and 120 records, which contain 16619 and 13593 event expressions respectively. The dataset includes six types of clinical event entities, namely PROBLEM, TREATMENT, TEST, OCCURRENCE, EVIDENTIAL, CLINICAL_DEPT. Each dataset was split into three parts: training, development, and test sets. For the THYME corpus, we used the official splits provided by the authors of the datasets. Because there was no official development set in 2012 i2b2, we split the dataset followed by previous work [17].

2) METRIC

We evaluate the performances of all models by Precision (P), Recall (R) and F1-score (F1) in percentage. Precision refers to the proportion of true positive examples to the examples predicted by the system, and Recall refers to the proportion of true positive examples to the examples annotated by the humans. If we only consider the Precision or Recall rate, we cannot accurately evaluate the quality of a model, so we use the F1 score to be compatible with the accuracy and recall rate. The F1 metric is defined as follows:

\[
F1 = \frac{2 \times P \times R}{P + R}
\]

(9)

Because random initialization of the networks weights and shuffling of batches during training would affect the final results of the model, we run the training and evaluation procedure three times on each dataset, and take the average scores of the three experiments as our final experimental results.

3) BASELINES

To further demonstrate the effectiveness of our model, we compare our method with previous 8 state-of-the-art methods for clinical event mention recognition namely UTHEalth [26], UTA-5 [31], Char-LSTM [17], Bio_BERT [18], RoBERTa-MIMIC [19], UMLSBERT [20], CRFs_X [24], and BERT-Large [44]. Machine learning-based methods include UTHEalth and CRFs_X, which got the best scores on THYME corpus and i2b2 2012 dataset respectively. Deep learning methods include UTA-5 and Char-LSTM, which got the highest scores among deep learning methods on THYME corpus and i2b2 2012 dataset respectively. Our comparison methods also include four state-of-the-art BERT-based models (Bio_BERT, RoBERTa-MIMIC, UMLSBERT, and BERT-Large). BERT-Large got the best scores on i2b2 2012 event recognition among the BERT-based methods. UMLSBERT is a BERT-based method, which integrates domain knowledge during the pre-training stage.

B. TRAINING AND PARAMETER SETTINGS

We exploit gradient descent algorithm to train our model, and use AdaGrad algorithm following previous work [45]. Derivatives are calculated from standard back propagation [46]. We adopt the dropout method to prevent our neural networks from overfitting following [47]. We apply the dropout method on both the input and output vectors of Bi-LSTM layer, and the dropout rate is fixed at 0.4 for THYME corpus and 0.3 for 2012 i2b2. We achieved obvious improvements in model performance after using dropout, and the results of the dropout will be illustrated in Table 5.

Parameters are tuned based on the development set. Some of the values are chosen empirically following previous work [13], [16]. The initial AdaGrad learning rate is set as 0.03 and regularization parameter is set as \( 10^{-8} \).
The weight matrices $W$, bias vectors $b$ are randomly initialized in the range $(-0.01, 0.01)$ with a uniform distribution. Table 1 shows the chosen values of hyper-parameters in our models. We tune the hyper-parameters on the development sets by random search, and stop the training when we find the best scores in the development set.

Character embeddings are initialized with uniform samples from range $[-\sqrt{3\text{dim}}, +\sqrt{3\text{dim}}]$, where dim is 30. Word embeddings derived from unlabeled text have been showed significant improvement on entity recognition in biomedical text [19], [20]. We use publicly available pubmed embeddings [48] which are trained in biomedical literature containing over 5 billion words in publication abstracts and full texts.

### C. EVALUATION ON THYME CORPUS

Table 2 illustrates the results of our method for event expressions recognition on THYME corpus, comparing with six previous state-of-the-art systems. Encourage by previous study, we recognize the event and their types simultaneously.

The results of UTHealth [26] and UTA-5 [31] come from their published papers. Char-LSTM [17] got state-of-the-art results on i2b2 2012 among deep machine methods, Bio_BERT [18], RoBERTa-MIMIC [19], and UMLSBERT [20] are BERT-based methods, which received state-of-the-art results on i2b2 2012. Because these methods have no published results on THYME corpus, we utilized the provided parameter settings by the authors to run their released source code on THYME and output the highest results, these results are marked with ‘*’. We run three BERT-based models [18], [19], [20] with the best parameters their authors provided separately; their grades are all lower than ours. These may be due to these BERT-based models have too many layers and complex architectures; fine tune the entire language models on a task-specific architecture are a sophisticated work.

To further evaluate the performance of our method, we plotted ROC (receiver operating characteristic) and PR (Precision/Recall) curves for Char_LSTM [17], UMLSBERT [20] and ours model by changing the confidence threshold. The results are shown in Figure 3 and Figure 4.

From Figure 3, we can see that the AUC (area under the ROC curve) of our method is bigger than the other two models. In Figure 4, the curve of our model almost completely encloses the PR curves of the other two methods. These demonstrate our method is more effective than other models in performance.

### D. EVALUATION ON i2b2 2012 DATASET

Table 3 shows the comparison of the previous best six methods with our proposed methods on i2b2 2012 dataset. Because most of the best systems have not release the results of event type recognition, we only compare the recognition results of event expression with relevant systems. Some methods did not publish their Precision and Recall values in their papers, these results are replaced with ‘-’. We adopt paired two-tailed Student’s t-tests with $p < 0.05$ between our method and the other methods.

Comparing with machine learning-based method UTHealth [26], our method achieves better P, R, and F1 value. [26] achieved the best result in Semeval-2016 Task12 competition. There is little difference between their scores and ours, but huge features in the clinical and natural language processing domain were employed in their method. UTA-5 [31] used RNN to predict event spans attaching each word with their part-of-speech tag and shape information as extra features, their results are much lower than ours. Our method achieves significant improvements over the Char-LSTM method [17]. Our method is similar to theirs, there are only two differences: (1) we map the concept unique identifier in UMLS as input features; (2) we utilize a CNN model to represent the character-level information, but they used LSTM. These results demonstrate the CNN model and the medical knowledge features we used have more effective than theirs.

We run three BERT-based models [18], [19], [20] with the best parameters their authors provided separately; their grades are all lower than ours. These may be due to these BERT-based models have too many layers and complex architectures; fine tune the entire language models on a task-specific architecture are a sophisticated work.

To further evaluate the performance of our method, we plotted ROC (receiver operating characteristic) and PR (Precision/Recall) curves for Char_LSTM [17], UMLSBERT [20] and ours model by changing the confidence threshold. The results are shown in Figure 3 and Figure 4.

From Figure 3, we can see that the AUC (area under the ROC curve) of our method is bigger than the other two models. In Figure 4, the curve of our model almost completely encloses the PR curves of the other two methods. These demonstrate our method is more effective than other models in performance.
while our model only added the feature provided by UMLS, slightly improves CRFs_X by 0.21%, achieving superior performance. Similar to the comparisons of THYME corpus, our method achieves significant improvements over the Char-LSTM [17]. We have to point out that their result (92.28) is a partial-matching F1 value, and ours partial-matching F1 is higher than theirs (93.29).

Compared with these BERT-based models, our model achieves higher Precision, Recall and F1 scores, obtaining 13.5% improvement on F1 over the best model reported results by Si et al. [43]. The algorithm we use domain knowledge as features is similar to UMLSBERT [20], Our F1 value is 15.69% higher than theirs, it demonstrates our method is more effective.

### E. ABLATION ANALYSIS

In order to quantify the role of each layer of our deep neural network model, we run 4 different experiments by ablation test for each task; all these models for each task are run using the same hyper-parameters as shown in previous sections. The results of the ablation tests are shown in Table 4. Removing CNN layer, the recall and the precision decreased obviously, this demonstrates that the character-level representation is effective for our tasks. Removing mk-embeddings, the recall decreased from 90.44 to 89.70, this demonstrates that the mk-embeddings are effective features, and the domain knowledge can improve the recognition rate of low frequency words. Removing LSTM (Backward) layer, the performance of model was obviously decreased; it proves the Bi-LSTM is necessary. Eliminating the CRF layer, the performance was obviously decreased. This demonstrates that CRF layer by joint decoding label sequence can effectively improve the final performance of neural network models.

### F. EFFECT OF DROPOUT

Table 5 compares the results of various dropout values for each dataset. All other hyper parameters remain the same as our best model. We choose 0.3 as dropout for THYME, and 0.4 for i2b2 2012. It demonstrates the effectiveness of dropout in reducing overfitting.

### G. COMPUTATION COST

We evaluated the computation cost of our model with Char-LSTM [17] and UMLSBERT [20], which achieved
TABLE 6. Comparison of the computation cost.

| Methods       | Run Times(s) |
|---------------|--------------|
| UMLSBERT      | 1207         |
| Char-LSTM     | 935          |
| Ours          | 926          |

TABLE 7. Performances of each type on i2b2 2012.

| Type              | P     | R     | F1   |
|-------------------|-------|-------|------|
| PROBLEM           | 83.12 | 85.23 | 84.16|
| TREATMENT         | 86.01 | 84.73 | 85.37|
| TEST              | 93.53 | 91.66 | 92.59|
| EVIDENTIAL        | 79.26 | 78.37 | 78.81|
| CLINICAL_DEPT     | 82.28 | 77.57 | 79.86|
| OCCURRENCE        | 75.65 | 72.85 | 74.22|
| Overall           | 88.86 | 85.72 | 87.26|

TABLE 8. Performances of each type on on THYME.

| Type              | P     | R     | F1   |
|-------------------|-------|-------|------|
| ASPECTUAL         | 68.31 | 56.82 | 63.52|
| EVIDENTIAL        | 74.86 | 91.59 | 81.46|
| N/A               | 90.81 | 87.05 | 89.07|
| Overall           | 89.51 | 86.94 | 88.21|

H. DISCUSSION

In order to investigate on each type of clinical entity how our model performs, we list the performance of our method on each type of clinical entity under “strict” criterion in Table 7 and Table 8. Our model performs well on some types, such as “TEST” and “TREATMENT” on i2b2 2012, “N/A” on THYME corpus. However, the F1 value of some entity categories is very low, such as “OCCURRENCE” and “CLINICAL_DEPT” on i2b2 2012, “ASPECTUAL” on THYME corpus. The recognition performance of different types varies greatly, which may be mainly caused by the imbalance between the samples in the training dataset, such as in THYME training dataset, ASPECTUAL type samples only accounts for 1.4%, but N/A type samples accounts for 92.8%.

We used case study to further verify the advantages of our model in recognizing low-frequency words. We compared our method with Char-LSTM [17] and UMLSBERT [20] in the clinical event recognition task. We randomly choose two sentences from THYME corpus test set as examples, the detailed results are shown in Table 9. Column 1 represents sentences containing new event spans, which are not existing in training set, and column 2 represents gold mentions in the sentences. Column 3, column 4, and column 5 represent the predicted results by our model, Char-LSTM, UMLSBERT respectively.

From Table 9, we can see that Char-LSTM model cannot recognize the words “CHF” and “IDA”, which are all abbreviations of medical terms that do not appear in the training dataset. While our method (Bi-LSTM-CRF+MK) and UMLSBERT recognized such entities. Because our method employed the CUI from UMLS as medical knowledge features (mk-embedding) followed by previous works [20], [21]. Here, “CHF” is an acronym for “congestive heart failure”, they share a similar meaning with word “cardiac” and thus can be mapped to the same CUI, and “IDA” is an acronym for “iron-deficiency anemia”. Although the words “IDA” and “CHF” did not appear in the training set, because their synonyms appeared in the training set, they were correctly recognized. From the above examples, we can confirm that our method can learn the association of different clinical synonyms and improve the recognition rate of low-frequency words.

Although our method has achieved good performance, it also has some limitations: 1) the proposed method is also applicable to entity recognition in Chinese text, but we do not
V. CONCLUSION

In this paper, we proposed a neural network architecture for clinical event extraction. We first employed CNN to represent the morphological features of a word and medical domain language to augment the association of different clinical terms, the morphological features of a word and medical domain lan-

guage to augment the association of different clinical terms, and then we used Bi-LSTM to learn the contextual semantics of words automatically. Finally, we utilized a CRF layer to output labels for the whole sentence. The experiments on two public clinical datasets show that our method has good performance in F1 scores, and the medical knowledge features we employed is effective in recognizing low-frequency word.

REFERENCES

[1] R. Sivarathinamohan, S. Sujatha, and P. Biswas, “Envisioning the potential of natural language processing (NLP) in health care management,” in Proc. 7th Int. Conf. Res. Innov. Amid Global Pandemic (I2C), Feb. 2021, pp. 189–193.

[2] E. H. Houssein, R. E. Mohamed, and A. A. Ali, “Machine learning techniques for biomedical natural language processing: A comprehensive review,” IEEE Access, vol. 9, pp. 140628–140653, 2021.

[3] W. Sun, A. Rumshisky, and O. Uzuner, “Evaluating temporal relations in clinical text: 2012 I2B2 challenge,” J. Amer. Med. Inform. Assoc., vol. 20, no. 5, pp. 806–813, Sep. 2013.

[4] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, and M. Verhagen, “SemEval-2016 task 12: Clinical temporal,” in Proc. 10th Int. Workshop Semantic Eval. (SemEval), San Diego, CA, USA, 2016, pp. 1052–1062.

[5] S. Bethard, G. Savova, M. Palmer, and J. Pustejovsky, “SemEval-2017 task 12: Clinical temporal,” in Proc. 11th Int. Workshop Semantic Eval. (SemEval), Vancouver, BC, Canada, 2017, pp. 565–572.

[6] S. Henry, K. Buchan, M. Filannino, A. Stubbis, and O. Uzuner, “2018 N2C2 shared task on adverse drug events and medication extraction in electronic health records,” J. Amer. Med. Inform. Assoc., vol. 27, no. 1, pp. 3–12, Jan. 2020.

[7] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson, “A general natural-language text processor for clinical radiology.” J. Amer. Med. Inform. Assoc., vol. 1, no. 2, pp. 161–174, Mar. 1994.

[8] A. R. Aeronon and F.-M. Lang, “An overview of MetaMap: Historical perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[9] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. S. Anghavan, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical text analysis perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 457–470, Mar. 2020.

[10] Y. Xu, Y. Wang, T. Liu, J. Tsujii, and E. I.-C. Chang, “An end-to-end system to identify temporal relation in discharge summaries: 2012 I2B2 challenge,” J. Amer. Med. Inform. Assoc., vol. 20, no. 5, pp. 849–858, Sep. 2013.

[11] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, and H. Xu, “A hybrid system for temporal information extraction from clinical text,” J. Amer. Med. Inform. Assoc., vol. 20, no. 5, pp. 828–835, Sep. 2013.

[12] C. Grouin and V. Moriceau, “LIMSI at SemEval-2016 task 12: Machine-learning and temporal information to identify clinical events and time expressions,” in Proc. 10th Int. Workshop Semantic Eval. (SemEval), San Diego, CA, USA, 2016, pp. 1225–1230.

[13] C. Grouin and V. Moriceau, “LIMSI at SemEval-2016 task 12: Machine-learning and temporal information to identify clinical events and time expressions,” in Proc. 10th Int. Workshop Semantic Eval. (SemEval), San Diego, CA, USA, 2016, pp. 1225–1230.

[14] L. NqK and B. P. Nguyen, “Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles,” Proteomics, vol. 24, no. 6, pp. 1062–1071, Apr. 2017.

[15] K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical text analysis perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[16] C. Manning, M. Surdeanu, and P. Biswas, “Envisioning the potential of natural language processing (NLP) in health care management,” in Proc. 7th Int. Conf. Res. Innov. Amid Global Pandemic (I2C), Feb. 2021, pp. 189–193.

[17] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: Historical perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[18] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. S. Anghavan, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical text analysis perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[19] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. S. Anghavan, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical text analysis perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[20] A. R. Aronson and F.-M. Lang, “An overview of MetaMap: Historical perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, 2010.

[21] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. S. Anghavan, K. C. Kipper-Schuler, and C. G. Chute, “Mayo clinical text analysis perspective and recent advances,” J. Amer. Med. Inform. Assoc., vol. 17, no. 3, pp. 229–236, 2010.
include natural language processing, big data application, and computer science. She is currently a Lecturer with the Department of Information Engineering, Henan Normal University, Xinxiang, China. Her research interests include natural language processing, biomedical, and intelligent information processing.

Dr. Li is a Senior Member of CCF of China Computer Association.

SHIYONG LI received the B.S. degree from the Department of Computer Science, Henan University of Economics and Law, Zhengzhou, China, in 2004, and the M.S. degree from the Department of Computer and Information, Guangxi University, Nanning, China, in 2007.

He is currently a Lecturer with the Department of Information Engineering, Henan Institute of Science and Technology, Xinxiang, China. His research interests include data mining and machine learning.

SHUNLI ZHANG received the B.S. and M.S. degrees from the Department of Information Engineering, Zhengzhou University, Zhengzhou, China, in 2002 and 2006, respectively.

From 2006 to 2007, she was a Visiting Scholar with the School of Computer, Wuhan University, Wuhan, China. During this period, she has participated in the Challenge of SemEval-2017 Task 12 and completed research projects “automatic extraction of time expressions in clinical texts.” She is currently a Lecturer with the Department of Information Engineering, Henan Institute of Science and Technology, Xinxiang, China. Her research interests include natural language processing, big data application, and computer education.