GCN with External Knowledge for Clinical Event Detection

Dan Liu, Zhichang Zhang†, Hui Peng, Ruirui Han
College of Computer Science and Engineering, Northwest Normal University, China
3394055228@qq.com, zzc@nwnu.edu.cn
penghuipwld@163.com, 1090400220@qq.com

Abstract

In recent years, with the development of deep learning and the increasing demand for medical information acquisition in medical information technology applications such as clinical decision support, Clinical Event Detection has been widely studied as its subtask. However, directly applying advances in deep learning to Clinical Event Detection tasks often produces undesirable results. This paper proposes a multi-granularity information fusion encoder-decoder framework that introduces external knowledge. First, the word embedding generated by the pre-trained biomedical language representation model (BioBERT) and the character embedding generated by the Convolutional Neural Network are spliced. And then perform Part-of-Speech attention coding for character-level embedding, perform semantic Graph Convolutional Network coding for the spliced character-word embedding. Finally, the information of these three parts is fused as Conditional Random Field input to generate the sequence label of the word. The experimental results on the 2012 i2b2 data set show that the model in this paper is superior to other existing models. In addition, the model in this paper alleviates the problem that "occurrence" event type seem more difficult to detect than other event types.

1 Introduction

Electronic medical records are an inevitable product of medical information. The use of natural language processing (NLP) technology to effectively detect clinical events in electronic medical records becomes crucial as the number of electronic medical records rapidly grows. It has been widely studied because of its potential help in constructing clinical event lines, medical Q & A, assisted diagnosis and other tasks. The task of Clinical Event Detection (CED) is to identify the boundary of the event in the electronic medical record and determine its type. The event detection to identify the boundary and determine type is usually considered as a sequence labeling task.

The emergence of deep learning has greatly improved the performance of the sequence labeling task model. Bidirectional long short-term memory network (BiLSTM) is widely employed in sequence labeling tasks owing to its high power to learn the contextual representation of words. Huang et al. (2015) was the first to apply the bidirectional long short-term memory (BiLSTM) and the conditional random field (CRF) to sequence labeling tasks. But BiLSTM needs to be processed sequentially over time, it cannot be calculated in parallel. Instead, the Transformer not only advantage in modeling the long-range context, but also fully make use of the concurrence power of GPUs. Yan et al. (2019) found that due to its position coding problem, the performance of Transformer in sequence labeling tasks is not as good as in other NLP tasks, and solves the position coding problem. However, due to the lack and particularity of clinical data, the performance of directly applying these technological advances to CED tasks is not ideal. For this reason, we need to use a large amount of medical information for tokens representation. In the past, Word2Vec (Mikolov et al., 2013) or Glove (Pennington et al., 2014) was used to train a large number of unlabeled clinical texts to generate word embedding. There are two problems with this method. On the one hand, it is difficult to obtain a large amount of clinical data. On the other hand, the

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, Augest 13 - 15, 2021.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

obtained data is not standardized, many mistakes, Various representations. Lee et al. (2020) proposed a pre-trained biomedical language representation model for biomedical text mining (BioBERT) in 2019, whose performance on various biomedical text mining tasks largely surpassed BERT and previous advanced models. We use BioBERT for word embedding to solve the problem of poor model recognition performance caused by a large number of obscure professional terms in the medical field. In addition, the character-level features of words may show word features, for example, the beginning of "un" generally indicates negative characteristics. Therefore, adding character-level encoding will also have an impact on improving the performance of the model, and can solve the problem of Out Of Vocabulary(OOV). Lample et al. (2016), Ma et al. (2016) and Liu et al. (2018) have added character-level coding to the model of the English NER task and proved its effectiveness. This article uses Convolutional Neural Network (CNN) as a character-level encoder for words (Chiu and Nichols, 2016).

After the Informatics Integrated Biology and Bedside Information (i2b2) sharing task was proposed in 2012 (Grouin et al., 2013), we found that in order to better enable the event extraction task to better serve the later tasks such as disease diagnosis, the event types of the shared task also include "occurrence" type. However, the experimental results of the organizations that participated in the challenge in the past show that the "occurrence" type is more difficult to predict than other types of events, and it is not due to the small amount of data in this type. We analyzed the reasons and found a solution. (1) Most of the events in this type are nouns or verbs. We use the Part-of-Speech generated by Stanford CoreNLP tools as attention to help them identify. (2) The past practice always solves the problem of poor recognition based on the particularity of the medical field, while ignoring the language commonality between the medical field and the general field, and incidents of "occurrence" type is more inclined to event recognition in the general field. For this reason, this article introduces external knowledge to alleviates the problem that occurrence event type seem more difficult to detect than other event types. At the same time, these external knowledge are helpful to the recognition of the event span and improve the overall performance of the model. In general, the contributions of this paper are as follows:

- The pre-trained BioBERT language model is used for word-level coding, which effectively solves the problem of the lack and particularity of clinical data.
- The Graph Convolutional Network(GCN) that introduces external knowledge alleviates the problem that occurrence event type seem more difficult to detect than other event types.
- The experimental results show that the model in this paper is better than the previous optimal model.

2 Related Work

Constructing the clinical timeline is crucial to the patient diagnosis and treatment. The 2012 Informatics for Integrating Biology and the Bedside (i2b2) shared task (Grouin et al., 2013) was the identification and linking of mentions of temporal expressions (TEs) (eg, dates, times, durations, and frequencies) and clinically relevant events (eg, patients problems, tests, treatments) in narratives. For this task, previous research on deep learning methods is mainly based on recurrent neural networks (RNN) (Fries, 2016; Cheng and Miyao, 2017), convolutional neural networks (CNN) (Dligach et al., 2017; Li and Huang, 2016), BiLSTM (Tourille et al., 2017; Lin et al., 2018) and Attention (Zhao et al., 2019) methods.

2.1 Clinical Event Detection

As a subtask of constructing the clinical timeline, Clinical Event Detection(CED) methods are mainly divided into the following two categories. (1) Method based on rules and machine learning: The method based on rules mainly summarizes relevant classification rules based on the experience and knowledge of knowledge engineers or domain experts, and then constructs corresponding rule templates as classifiers. Traditional machine learning methods are based on feature engineering. After 2012 i2b2 challenge task is proposed, The team involved in the task has adopted many different methods: rule-based , support vector machine (SVM) (Cortes et al., 1995) conditional random field (CRF) (Lafferty et al., 2001), Markov Logic and some combination of these methods, the best performance is the CRF-based model

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, Augest 13 - 15, 2021.

proposed by Beihang University, Microsoft Research Asia, Beijing and Tsinghua University. Roberts et al. (2013) used a combination of supervised, unsupervised and rule-based method, and the task ranked third. First, it uses the CRF classifier to identify event boundaries. Then use an independent SVM classifier for type detection. Kovacevic et al. (2013) combined rules and machine learning and achieved F1 measure of 79.85%, it proposed the event CRF models were trained on relevant (type-specific) subsets of the training data and they all shared some feature groups. Although the rule-based method has high classification accuracy, it does not have the ability to learn from experience and is difficult to maintain. The rule making requires professional participation, time-consuming and labor-intensive, poor scalability, and it is difficult to promote and use in multiple fields. Traditional machine learning does not need to manually write rule templates, it can effectively solve the problems in rule-based methods. But, it is time consuming to extract features, and its consumption tends to grow as the size of the data set becomes larger, which is prone to dimensional disasters. (2) Methods based on deep learning: The emergence of deep learning greatly reduces the difficulty of obtaining text features. Zhu et al. (2019) proposed a bidirectional LSTM-CRF model is trained for clinical concept extraction using the contextual word embedding model, it achieved the best performance among reported baseline models on the i2b2 2010 challenge dataset and the result is higher than this article, this is due to the dataset of this article has added three new event types three: evidential, occurrence and clinical department, in particular, the evidential and occurrence event types seem more difficult to detect than other event types (2013). Recently, research on the 2012 i2b2 data set has decreased, but the NER task has been widely studied. The LSTM and CRF models greatly improve the performance of the NER task (Akhundov et al., 2018). Transformer is widely used in NER tasks due to its parallelism and advantages in modeling long-range context (Yan et al., 2019). Chen et al. (2019) proposed a simple but effective CNN-based network for NER, Gated Relation Network (GRN), which is more capable than common CNNs in capturing long-term context. Lin et al. (2020) also used Self Attention when solving NER tasks. Graph Neural Networks (GNNs) are also widely used in NER tasks (Liu et al., 2019; Luo and Zhao, 2020).

2.2 BioBERT

In this section, we briefly introduce the pre-trained language model (BioBERT) used in this article. Direct application of NLP advancements to clinical text mining often yields unsatisfactory results due to a word distribution shift from general domain corpora to clinical corpora. Lee et al. (2020) investigate how the recently introduced pre-trained language model BERT can be adapted for biomedical corpora, and proposed a domain-specific language representation model pre-trained on large-scale biomedical corpora (BioBERT) to solve this problem. BioBERT initializes weights from BERT, which is pre-trained on the English Wikipedia and Books Corpus general domain corpus. Then, BioBERT is pre-trained on PubMed abstract and PMC full-text article biomedical corpus.

Resolve name	**content**
Sentence	She had a CT scan
Part-of-Speech	[('She', 'PRP'), ('had', 'VBD'), ('a', 'DT'), ('CT', 'NN'), ('scan', 'VB')]
Constituency Parsing	(ROOT (S (NP (PRP She)) (VP (VBD had) (S (NP (DT a) (NN CT)) (VP (VB scan))))))
Dependency Parsing	[('ROOT', 0, 2), ('nsubj', 2, 1), ('det', 4, 3), ('nsubj', 5, 4), ('ccomp', 2, 5)]

Table 1: Stanford CoreNLP semantic analysis example.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, August 13 - 15, 2021.

2.3 Stanford CoreNLP

In this section, we will briefly introduce the semantic parsing tool Stanford CoreNLP used in this article. It is a natural language processing toolkit. CoreNLP enables users to derive linguistic annotations for text, including token and sentence boundaries, Part-of-Speech, named entities, numeric and time values, dependency and constituency parses, coreference, sentiment, quote attributions, and relations. CoreNLP currently supports 6 languages: Arabic, Chinese, English, French, German, and Spanish. This article mainly uses its Part-of-Speech, dependency and constituency parses functions. Since this tool is suitable for sentences in the general domain, that is, the model used by the tool is trained on a large number of general domain corpora, so this article calls it the introduction of external knowledge. Taking the sentence "She had a CT scan" in the medical field as an example, The Part-of-Speech, dependency and constituency parses of the sentences parsed by the Stanford CoreNLP tool are shown in table 1. When we train the model, we need to encode the information in Table 1 into a matrix form, which will be described in detail in a later part.

3 Model

The model of this article mainly includes four parts: character-word embedding module, Part-of-Speech attention module, semantic GCN module, CRF decoding module. Figure 1 shows an overview of our model, where CNN is a character-level encoder for words, BioBERT is a word-level encoder for words, and the semantic analysis tool uses Stanford CoreNLP. First, input a sentence to generate character-level embedding of words through CNN, and generate word-level embedding of words through biomedical pre-training language model BioBERT, the two are spliced together. And then perform Part-of-Speech attention coding for character-level embedding, perform semantic graph convolutional network coding for the spliced character-word embedding. Finally, the information of these three parts is fused as CRF input to generate the sequence label of the word. We will introduce each part in detail in the following chapters.

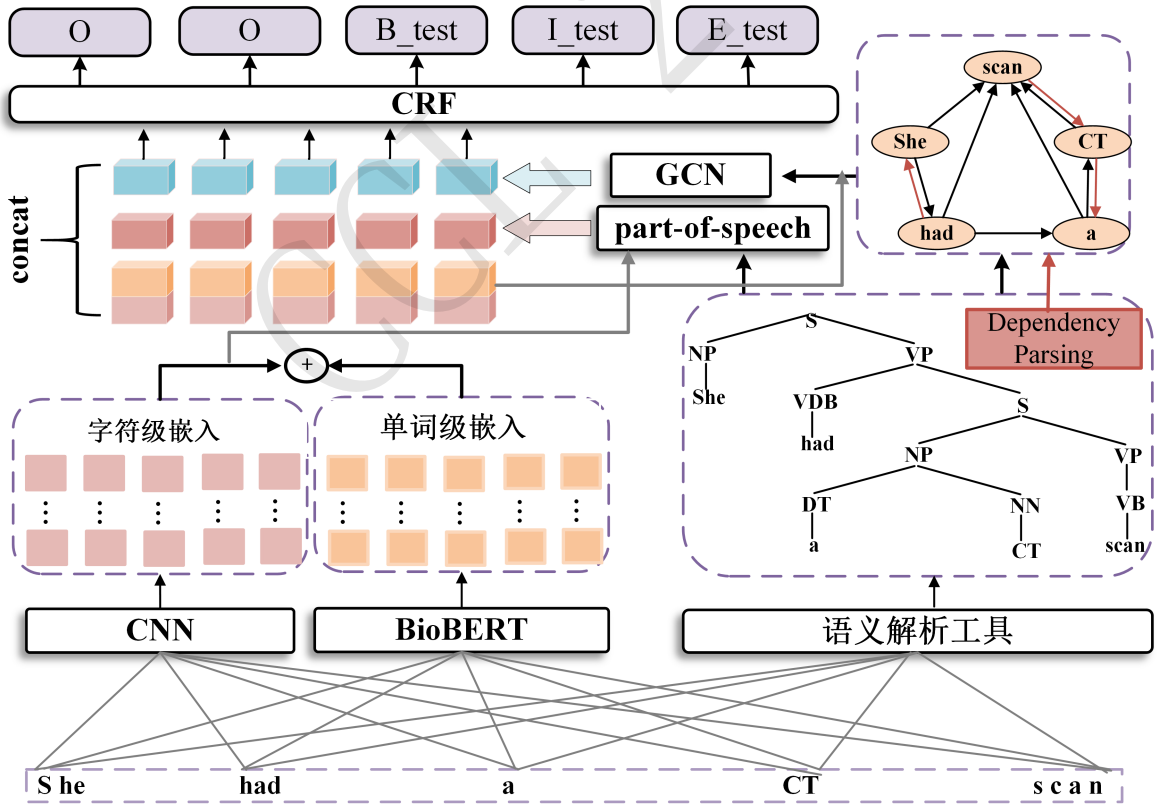

Figure 1: The framework of our model.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, August 13 - 15, 2021.

3.1 Character-word Embedding

Given a sequence of N tokens $H = [H_1, H_2, ..., H_N]$, For each token H_i , We first splicing word-level and character-level embedding $H_i = [C_i; W_i]^T$, where C_i is character-level embedding, W_i is word-level embedding. And provide it to subsequent modules. So a sentence can be expressed as $H = [[C_1; W_1]^T; [C_2; W_2]^T; ...; [C_N; W_N]^T]^T$, $H \in R^{N\times(cd+wd)}$, where cd is the character-level encoding dimension, wd is the word-level encoding dimension. The following describes the details of word-level embedding and character-level embedding of tokens in detail.

Character-level embedding: For a token character-level embedding C_i, As shown in Figure 2, for the character sequence of the token i in a sentence $x_i = x_i^1, x_i^2, ..., x_i^{cd}$, ($cd$ represents the number of characters in the longest word, namely, the maximum length of the word). We vectorize it and use the character embedding method to get the vector representation of each character $c^i = e^c(x^i)$, Then the character-level embedding of the word is expressed as a matrix $c = [c^1; c^2; ...; c^{wl}]^T$, $c \in R^{wl\times cd}$, Let's perform a convolution operation. Assuming that there are cd convolution kernels, the formula for the n-th convolution operation of the m-th convolution kernel is as follows: $h^{mn} = w \cdot c^{n:n+k-1} + b$, The size of the sliding window contains k characters, which is represented by the symbol $c^{n:n+k-1}$, w represents the convolution kernel, each time the feature is obtained by sliding k characters h^{mn}, that is, the red box and the yellow box in the figure 2. The m-th convolution kernel generates feature vectors for all characters sliding $h^m = [h^{m1}; h^{m2}; ...; h^{m(wl-k+1)}]^T$, So, the character-level features of the words generated by a set of convolution kernels are $h = [h^1; h^2; ...; h^{cd}]$, Then perform maximum pooling to get the character-level representation of the token $C_i = [max(h^1), max(h^2), ..., max(h^{cd})]$, $C_i \in R^{N\times cd}$.

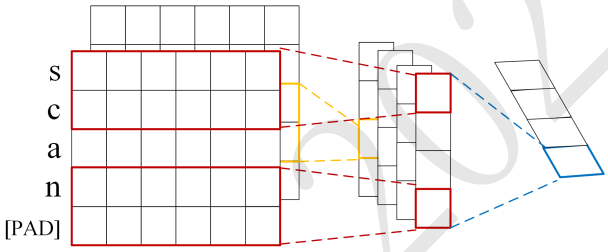

Figure 2: The architecture of character-level embedding

Word-level embedding: In order to solve the problem of the particularity of clinical data, this article uses BioBERT (a domain-specific language representation model pre-trained on large-scale biomedical corpora). BioBERT initializes weights from BERT, which is pre-trained on the English Wikipedia and Books Corpus general domain corpus. Then, BioBERT is pre-trained on PubMed abstract and PMC full-text article biomedical corpus. Specifically, it uses pre-trained BioBERT on PubMed for 1M steps model, this version as BioBERT v1.1 (+PubMed). Fine-tuned based on this model, the BioBERT model takes in word sequence $s = w_1, w_2...w_N$ (N represents the maximum sentence length), Calculate the output of the BioBERT layer using the equations below:

$$H^w = BioBERT(s) \tag{1}$$

where $H^w \in R^{l\times wd}$, wd is the word-level embedding dimension, the model limits it to a multiple of 768. This article is the last BioBERT layer, so the size is 768.

3.2 Part-of-Speech Attention

We found that most event words are basically verbs or nouns, so Part-of-Speech features are helpful to event recognition. In order to learn sentence representations based on Part-of-Speech attention, we follow the self-attention method introduced by Lin et al. (2017), this method has also been used in named entity recognition tasks (Lin et al., 2020). it uses attention to convert the sentence into multiple vectors to extract different parts of the sentence, and uses a matrix to represent the sentence embedding. This article replaces the self-attention of the sentence with Part-of-Speech attention. The specific calculation

formula is as follows:

$$Pos = SoftMax(W_2 tanh(W_1 P^T)) \tag{2}$$

$$H^{'} = Pos \cdot C \tag{3}$$

Where P represents the Part-of-Speech matrix corresponding to the sentence, we use one hot to represent, $P \in R^{N \times 37}$, 37 is the number of all types of Part-of-Speech of the Stanford CoreNLP tool, and C represents the character-level embedding of the word. W_1 and W_2 are two trainable parameters for calculating the Part-of-Speech attention score vector Pos. The W_2 first dimension is fixed to N for information fusion later. We get a matrix of weighted sums of token vectors for a Part-of-Speech attention $H^{'}$, $H^{'} \in R^{N \times cd}$.

3.3 Semantic GCN

Since the original input sentences are plain texts without inherent graphical structure, we first construct graphs based on the sequential information of texts and the semantic information in sentences parsed using Stanford CoreNLP tool. Then, we apply GCN (Luo and Zhao, 2020; Kipf and Welling, 2017; Qian et al., 2019) which propagates information between neighboring nodes in the graphs, to extract extract events.

Graph construction: We create three kinds of information fusion people graphs for each sentence, each graph is defined as $G = (V, E)$, where V is the node set (word) and E is the edge set. Figure 3 shows the structure of the sentence "She had a CT scan." The process of constructing a graph is divided into three steps.

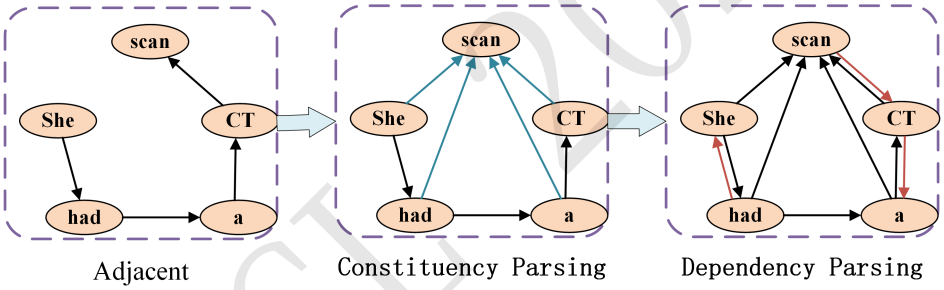

Figure 3: Semantic analysis graph.

- Adjacent graph: For each pair of adjacent words in the sentence, we add one directed edge from the left word to the right one, allowing local contextual information to be utilized.
- Constituency parsing graph: According to the constituency parsing obtained by the Stanford CoreNLP tool, we connect the leftmost node of all subtrees to the last edge node according to the established tree structure.
- Dependency parsing graph: The dependency parsing obtained by the Stanford CoreNLP tool, we build a dependency parsing graph based on the dependencies between words.

Bi-GCN: In order to consider both incoming and outgoing features for each node, we use Bi-GCN to extract graph features (Marcheggiani and Titov, 2017; Fu et al., 2019),. Given a graph $G = (V, E)$, and the word representation $H = [H_1, H_1, ..., H_N]^T$, the graph feature $H^{''} \in R^{N \times 2d_f}$ learned from Bi-GCN is expressed as follows, we use them the same as Luo et al. (2020) :

$$\overrightarrow{f_i} = ReLU(\sum_{e_{ij} \in E} \overrightarrow{W_f} H_j + \overrightarrow{b_f}) \tag{4}$$

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, August 13 - 15, 2021.

$$\overleftarrow{f_i} = ReLU(\sum_{e_{ji} \in E} \overleftarrow{W_f} H_j + \overleftarrow{b_f}) \tag{5}$$

$$H^{''} = [\overrightarrow{f_i}; \overleftarrow{f_i}] \tag{6}$$

where $W_f \in R^{2d_f \times (cd+wd)}$ and $b_f \in R^{d_f}$, d_f are trainable parameters, it is the hidden size of GCN, $ReLU$ is the non-linear activation function. e_{ij} represents the edge outgoing from token H_i, and e_{ji} represents the edge incoming to token H_i.

3.4 CRF Decoding

The character-word embedding representation, Part-of-Speech attention coding representation and semantic graph convolutional coding representation of the word are spliced together to obtain a matrix $\hat{H} = [H; H^{'}; H^{''}], \hat{H} \in R^{N \times (2cd+wd+2d_f)}$. it is input to the CRF layer to predict the corresponding tag sequence. The probability of a label sequence $Y = y_1, y_2 \ldots y_l$ is

$$P(y|H) = \frac{\exp(\sum_i (W_{CRF}^{y_i} h_i + b_{CRF}^{(y_{i-1}, y_i)}))}{\sum_{y'} \exp(\sum_i (W_{CRF}^{y_i} h_i + b_{CRF}^{(y_{i-1}, y_i)}))} \tag{7}$$

Where y' represents an arbitary label sequence, $W_{CRF}^{y_i}$ is a model parameter specific to y_i, and $b_{CRF}^{(y_{i-1}, y_i)}$ is a bias specific to $y_{i-1}$and y_i. Finally, the Viterbi Algorithm is used to find the path achieves the maximum probability.

4 Experiment

4.1 Dataset

To evaluate our proposed model, we experiment on 2012 i2b2 challenge dataset, the training corpus consists of 190 electronic medical records, which contains 2250 sentences (The number after adjusting the sentence length), and the test corpus of 120 electronic medical records, which contains 1741 sentences (The number after adjusting the sentence length), event types include clinical department, evidential, "occurrence", problem, test, treatment (Grouin et al.,2013). The 2012 i2b2 challenge dataset does not have development set, this article divides the test set into a test set and a development set at a ratio close to 1:1. Among them, there are 821 sentences in the development set, 920 sentences in the test set.

4.2 Evaluation Metrics

This paper CED is a sequence labeling task, standard precision (P), recall (R) and F1-score (F) are used as evaluation metrics. In order to prove the effectiveness of the model, this article also uses the same evaluation metrics as 2012 i2b2 challenge (Span F1-score and Type accuracy) (Grouin et al., 2013). We used the Span F1-score, the harmonic mean of precision and recall of the predict output span against the gold standard span to evaluate event span detection performance. The calculation of Span F1-score is the same as the calculation of the F1-score evaluation metrics. It is worth noting that the Span F-score is lenient matching (predict event span overlap with the gold standard span). For event types, we calculated classification accuracy, that is, the percentage of correctly identified event types for the events whose spans are detected correctly, the specific calculation process is as follows:

$$P = \frac{|pred.type \bigcap glod.type|}{|pred.span \bigcap glod.span|} \tag{8}$$

where, "pred.type" means predict type output, "gold.type" means gold standard type, " pred.span" means predict span output, "gold.span" means gold standard span.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, Augest 13 - 15, 2021.

4.3 Settings

In the process of data preprocessing, in order to solve the large gap in sentence length, we split the sentence according to several punctuation marks. such as “,” and “;” etc, these punctuation marks can be used to break the sentence, the sentence is still complete. Then join several short sentences that are adjacent but whose total length does not exceed the maximum length. Because some special characters have their own characteristics, this article will deal with them to increase accuracy, such as “’s”, splice it directly to the previous word, and we replace all digits with “0”. In the experiment based on the CED task. we use the BIOES tag schema. For character-level embedding, we set randomly initialized character embedding size to 30. For word-level embedding we only take the last layer, the dimension is 768. Since the Part-of-Speech attention coding representation and the semantic map convolution representation are obtained from the introduced external knowledge, in order to avoid negative effects and reduce the problem of error propagation. We still have to focus on coding in the medical field, supplemented by coding with external knowledge. Therefore, the dimension of the trainable parameters represented by the semantic graph convolutional coding needs to be as small as possible than the word-level coding dimension of words, we set it to 120.The batch size for training is 16, epochs is 100, We use SGD and 0.9 momentum to optimize the model. During the optimization, we use the triangle learning rate where the learning rate rises to the pre-set learning rate (0.0008) at the first 1% steps and to 0 in the left 99% steps (Smith, 2017).

4.4 Evaluation on CED

We compare proposed model with the latest model on the 2012 i2b2 challenge dataset. In addition, we will also apply the latest model for NER to the data set of this article for comparative experiments. The 2012 i2b2 challenge test results are shown in Table 2 and Table 3.

Model	**Precision**	**Recall**	**F1-score**
rule-based and machine learning (Kovacevic et al., 2013)	0.8147	0.7805	0.7985
BiLSTM_CRF	0.7484	0.5272	0.6186
ELMo_TENER (Yan et al., 2019)	0.7454	0.7805	0.7626
BioBERT_TENER_Data Augmentation	0.8101	0.7953	0.8026
Ours	**0.8152**	**0.7961**	**0.8055**

Table 2: Results of Precision, Recall and F1-score metrics.

The overall results of the model using P, R and F evaluation metrics are shown in Table 2. In the first block, the model combines rule-based and machine learning approaches that rely on morphological, lexical, syntactic, semantic, and domain specific features. In the second block, we give the model performance based on BiLSTM, the model uses TENER for character level embedding and the Glove 100d pre-trained embedding for word level embedding. In the third block of, we give the model performance based on Transformer, the model uses TENER for character level embedding and the ELMo for word level embedding. In the fourth part, we give the performance of another model experiment we did. The model uses TENER for character-level embedding, BioBERT for word-level embedding, and TENER for final encoding. In addition, the model also has data enhancements. In the last block, we give the experimental result of our proposed model. We can observe that our proposed model outperforms other models. it improves the F1-score from 79.85% to 80.55% on overall performance. Compared with the the last experiment, our model improves the F1-score from 80.26% to 80.55%.

For the results use span F1-score and type accuracy evaluation metrics which is depicted in Table 3, The first three block are the results of the top three participating in the 2012 i2b2 challenge, the forth block used a combination of supervised, unsupervised and rule-based method, and the task ranked third. First, it identifies event boundaries with a CRF classifier. Then it detects type using separate SVM classifiers. the fifth block, we give the model performance based on Transformer, the Glove 100d pre-trained embedding for word level embedding and model uses TENER for character level embedding, the TENER for word-character level embedding. In the sixth part, we give the performance of another

Model	Span F1-score	Type accuracy
Beihang University et al. (CRF) (Grouin et al., 2013)	0.9166	0.8600
Vanderbilt University (CRF _SVM) (Grouin et al., 2013)	0.9000	0.84.00
The University of Texas (CRF_SVM) (Grouin et al., 2013)	0.8900	0.8000
supervised, unsupervised and rule-based (Roberts et al., 2013)	0.8933	0.8045
TENER(the Glove 100d) (Yan et al., 2019)	0.7424	0.7505
BioBERT_TENER_Data Augmentation	0.9033	**0.9300**
Ours	**0.9168**	0.9276

Table 3: Results of Span F1-score and Type accuracy metrics.

model experiment we did. The model uses TENER for character-level embedding, BioBERT for word-level embedding, and TENER for final encoding. In addition, the model also has data enhancements. In the last block, we give the experimental result of our proposed model. We can observe that our proposed model improves Span F1-score from 91.66% to 91.68%, but our method improves the Type accuracy score from 86% to 92.76%. Compared with the the last experiment, our model improves the Span F1-score from 90.33% to 91.68%.

4.5 Ablation Study

We examine the contributions of four main components, namely, BioBERT word-level embedding, CNN character-level embedding, Part-of-Speech attention coding, and semantic GCN coding. The experimental results are shown in Table 4, Where "-" means remove component, "→" means replace component.

Model	Precision	Recall	F1-score
Ours	**0.8152**	**0.7961**	**0.8055**
BioBERT→Glove 100d	0.7308	0.5849	0.6498
-CNN	0.8025	0.7892	0.7958
-Part-of-Speech	0.7553	0.7806	0.7677
-GCN	0.7547	0.6120	0.6759

Table 4: Results of ablation study.

We can observe that BioBERT word-level embedding, CNN character-level embedding, Part-of-Speech attention coding and semantic GCN coding improved the performance of the model to varying degrees. Especially, BioBERT and semantic GCN coding representation. BioBERT has been pre-trained on a large amount of biomedical data and has lots of biomedical information. For semantic GCN coding, because the external knowledge it uses is a tool that is trained on a large amount of general predictive data. The understanding of sentence structure has played a very good help, and has a certain enlightening effect on determining the boundary of the event. Therefore, it is helpful to improve the performance of the model. The character-level embedding has less impact on the result, so analyze the reason is that the word-level encoding dimension is much smaller than the 768 of the BioBERT word-level encoding, but if the character-level encoding is also adjusted to a large value, the training efficiency of the model will be very low. For the Part-of-Speech attention module, it is helpful to the model in judging whether it is an event or not, but because its dimension is smaller than that of other information, Has little effect. Experimental results show, these four components can help the model learn medical text information better.

4.6 Type Analysis

The evaluation task challenged by i2b2 shows that event detection in 2012 seems more challenging than i2b2 in 2010. This is due to the addition of three new event types: "evidence", "occurrence" and "clinical

department". In particular, the types of "evidence" and "occurrence" seem to be more difficult to detect than other types. The results of our last experiment also found that the "evidential" and "occurrence" event types seem more difficult to detect than other event types. Especially occurrence EVENT type,

Event Type	F1-score(Last)	F1-score(Ours)	Train(Dataset)	Test(Dataset)
Clinical department	0.8164	0.8153	0.0605	0.0539
Evidential	0.7519	0.7532	0.0449	0.0438
Occurrence	0.6632	**0.6910**	0.1995	0.1838
Problem	0.8375	0.8367	0.3050	0.3170
Test	0.8409	0.8393	0.1576	0.1599
Treatment	0.8420	0.8345	0.2325	0.2417

Table 5: Results of different event types.

there is enough data volume, but the result is much lower than other types.The experimental results of our previous model and the model in this article on different types are shown in the Table 5, where the "Last" represents the result of our last experiment, where the "Ours" represents the experimental results of this article. We found that due to the "occurrence" type, it is more biased towards the general field. In the model in this article, we have introduced external knowledge, so its F value has been greatly improved compared to the previous model and the model in our last experiment, while the performance of other types has not significantly decreased.

5 Conclusion

This paper proposes a multi-granularity information fusion encoder-decoder framework.This framework uses the pre-trained language model (BioBERT) to generate word-level features, and solves the problem of poor model recognition performance caused by obscure professional terms in electronic medical records. The Graph Convolutional Network that introduces external knowledge improves the performance of the model in identifying "occurrence" type. Further improve the overall performance of the model. Experiments on the 2012 i2b2 challenge dataset show that our model achieves superior performance than other existing models.

References

Akhundov A., Trautmann D., Groh G. 2018. Sequence labeling: A practical approach. *arXiv preprint arXiv:1808.03926.*

Cortes C., Vapnik V. 1995. Support-vector networks. *Proceedings of the Twelfth International Conference on Machine Learning,* 20(3): 273-297.

Chen H., Lin Z., Ding G. et al. 2019. GRN: Gated relation network to enhance convolutional neural network for named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence,* 33(1): 6236-6243.

Chiu J., Nichols E. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics,* 357-370.

Cheng F., Miyao Y. 2017. Classifying temporal relations by bidirectional lstm over dependency paths. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 1-6.

Chikka, V. 2016. Cde-iiith at semeval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016),* 1237-1240.

Dligach D., Miller T., Lin C. et al. 2017. Neural temporal relation extraction. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers,* 746-751.

Fries J. 2016. Brundlefly at SemEval-2016 Task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction. *arXiv preprint arXiv:1606.01433.*

Fu T., Li P., Filannino M., Ma W. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1409C1418.

Grouin C., Grabar N., Hamon T. et al. 2013. Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association,* 20(5): 820-827.

Huang Z., Xu W., Yu K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991.*

Kovacevic A., Dehghan A., Filannino M. et al. 2013. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association,* 20(5): 859-866.

Kipf T., Welling M. 2017. Semi-supervised classification with graph convolutional networks. *In International Conference on Learning Representations (ICLR).*

Luo Y., Zhao H. 2020. Bipartite flat-graph network for nested named entity recognition. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6408C6418.

Lin Y., Lee D., Shen M. et al. 2020. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 8503C8511.

Lin B., Lee D., Shen M. et al. 2020. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 8503C8511

Lee J., Yoon W., Kim S. et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics,* 36(4): 1234-1240.

Lafferty J., McCallum A., Pereira F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning,* 282-289.

Liu P., Chang S., Huang X. et al. 2019. Contextualized non-local neural networks for sequence learning. *Proceedings of the AAAI Conference on Artificial Intelligence,* 33(1): 6762-6769.

Li P., Huang H. 2016. UTA DLNLP at SemEval-2016 Task 12: deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016),* 1268-1273.

Lin C., Miller T., Dligach D. et al. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis,* 165-176.

Lin Z., Feng M., Santos C. et al. 2017. A structured self-attentive sentence embedding. *In International Conference on Learning Representations (ICLR).*

Lample G., Ballesteros M., Subramanian S. et al. 2016. Neural architectures for named entity recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 260C270.

Liu L., Shang J., Ren X. et al. 2018. Empower sequence labeling with task-aware neural language model. *Proceedings of the AAAI Conference on Artificial Intelligence,* 32(1): 5253-5260.

Mikolov T., Chen K., Corrado G. et al. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Marcheggiani D., Titov I. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 1506C1515.

Ma X., Hovy E. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 1064C1074.

Pennington J., Socher R., Manning C. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing,* 1532-1543.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, Augest 13 - 15, 2021.

Qian Y., Santus E., Jin Z. et al. 2019. Graphie: A graph-based framework for information extraction. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 751C761.

Roberts K., Rink B., Harabagiu S. 2013. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association,* 20(5): 867C875.

Smith L. 2017. Cyclical learning rates for training neural networks. *In 2017 IEEE winter conference on applications of computer vision (WACV),* 464-472.

Tourille J., Ferret O., Neveol A. et al. 2017. Neural architecture for temporal relation extraction: A bi-lstm approach for detecting narrative containers. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 224-230.

Tourille J., Ferret O., Neveol A. et al. 2017. Neural architecture for temporal relation extraction: A bi-lstm approach for detecting narrative containers. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 224-230.

Xu Y., Jia R., Mou L. et al. 2016. Improved relation classification by deep recurrent neural networks with data augmentation. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics,* 1461-1470.

Yan H., Deng B., Li X. et al. 2019. Tener: Adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474.*

Zhu H., Paschalidis I., Tahmasebi A. 2019. Clinical concept extraction with contextual word embedding. *Journal of the American Medical Informatics Association,* 26(11): 1297-1304.

Zhao S., Li L., Lu H. et al. 2019. Associative attention networks for temporal relation extraction from electronic health records. *Journal of biomedical informatics,* 99: 103309.

Proceedings of the 20th China National Conference on Computational Linguistics, pages 1190-1201, Hohhot, China, Augest 13 - 15, 2021.