陈昌佳, 代丽, 金致楷, 等. 基于实体识别及要素匹配的事件抽取方法研究[J]. 智能计算机与应用,2025,15(9):90-95.
DOI:10.20169/j.issn.2095-2163.250914

基于实体识别及要素匹配的事件抽取方法研究

陈昌佳, 代 丽, 金致楷, 王欣悦
(浙江理工大学 经济管理学院, 杭州 310018)

摘 要: 环境和动作是事件的重要特征元素,是抽取特定事件的重要依据。为实现事件抽取,论文提出一种基于 BERT、双向门循环控制单元(Bidirectional Gated Recurrent Unit,BiGRU)、条件随机场(CRF)并添加注意力机制的命名实体识别模型用以精确识别文本中的环境和动作要素,经验证,模型对时间要素、地点要素及事件触发词的识别 $F1$ 值分别达到了 0.852 7、0.886 2 和 0.820 1,具有较好的识别效果。创新性提出事件环境要素匹配算法,用以完成事件-时间-地点的对应划分,从而实现基于特定环境的单一事件抽取。此研究方法可为事件类语料库的构建及应用提供技术参考。

关键词: 事件抽取; 要素匹配; 注意力机制; 命名实体识别
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0090-06

Research on event extraction based on entity recognition and elements matching

CHEN Changjia, DAI Li, JIN Zhikai, WANG Xinyue
(School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Specific events can be clearly distinguished based on environment and action elements. To achieve event extraction, the named entity recognition model is proposed. The named entity recognition model with BERT, BiGRU, CRF and added attention mechanism can accurately identify the environment and action elements in the text. It is verified that the $F1$ values of the model for time elements, place elements and event trigger words reach 0.852 7, 0.886 2 and 0.820 1, respectively, which has the good recognition effect. An innovative event environment element matching algorithm is proposed to complete the corresponding division of event-time-place, so as to realize single event extraction based on specific environment. This research method can provide technical reference for the construction and application of event corpus.

Key words: event extraction; element matching; attention mechanism; named entity recognition

0 引 言

事件抽取是自然语言处理领域的关键任务,传统的事件抽取基于文本特点构建匹配模式,即通过人为设定好的规则从非结构化文本中抽取事件内容,并将其进行结构化表示[1]。当前,人工智能及深度学习技术飞速发展,基于神经网络的事件抽取方法逐渐成为主流[2-4]。另外,随着信息传播媒介的不断变化和技术的推陈出新,灾害相关新闻报道、年鉴记录等文本数据在网络上海量积累[5],利用事件抽取技术规范化处理事件文本数据,将其转化为事件语料库,对于提升灾害文本的应用价值以及知识转化率具有重大意义,并可为后续事件关系挖掘及事理知识图谱的构建提供基础数据支持及技术参考。

命名实体识别[6]作为提取语料实体、挖掘文本信息的重要工具,在事件抽取领域具有广泛应用。传统的命名实体识别方法基于符合文本特征的匹配规则,具有较高的召回率,但构建特定词典和规则费时费力且后续维护困难[7]。现在,深度学习技术的应用也愈加普及,研究者们将长短时记忆网络(LSTM)、循环神经网络(CNN)、人工神经网络(RNN)等模型接入 CRF 层 ,使模型获取了能够预测句子级别标签的能力,神经网络技术则凭借其优

作者简介: 陈昌佳(2000—),男,硕士研究生,主要研究方向:信息管理与信息系统研发; 金致楷(2001—),男,硕士研究生,主要研究方向: 信息管理与信息系统研发; 王欣悦(2004—),女,本科生,主要研究方向:信息管理与信息系统研发。
通信作者: 代 丽 (1977—),女,博士,副教授,主要研究方向:机械机构分析优化与综合,管理信息系统,数据库技术。Email:dlzist@163.com。

收稿日期: 2023-12-24

异的性能逐渐占据命名实体识别主导地位。彭相等学者[8]构建了 LSTM 和 CRF 的联合识别模型帮助完成命名实体识别及分词任务，大幅提高了实体识别效能。此外，将卷积神经网络及混合神经网络应用至命名实体识别任务，也都取得了良好的实体识别效果[9-10]。

在一段事件文本报道中，为保证整体文段的可读性及叙事连贯性，往往把多个类似或具有因果关系的事件放在同一文段中进行描述，如何划分此类事件，将其从易读型的报道逆向转化为客观型的语料，是提高文本数据知识转换能力与应用设计的重要途径。刘宗田等学者[11]提出了事件六元组的概念，即认为事件是由动作、对象、事件、环境、断言、语言这 6 个元素组成，其中动作及环境要素是大多数事件所共有的特征可用以识别单一完整事件，采用事件触发词[12]和时间地点实体分别表示事件内的动作和环境要素。为实现此类事件的抽取，本文在利用命名实体识别技术获取事件触发词及时间、地点要素的基础上，设计了事件文句的时间地点要素匹配算法，实现了以环境要素为划分指标的事件抽取。同时又基于自然灾害事件数据进行了实验验证，证明了方法的可行性，为事件领域语料库的构建及应用提供了参考。

1　环境要素识别算法

添加注意力机制的 BERT-BiGRU-CRF 中文事件触发词及环境要素识别模型由 BERT 层、BiGRU 层、注意力层及 CRF 层四部分组成，分别用来实现中文灾害文本的向量转化嵌入、文本编码、注意力权重分配及解码输出部分的功能。

1.1　BERT 层

BERT 网络结构如图 1 所示。图 1 中，一段灾害文本输入示例经[CLS]句首向量和[SEP]句尾向量划定首尾后，进行处理并形成包含词向量(token embedding)、段落向量(segment embedding)、以及位置向量(position embedding)三部分的 BERT 模型最终输入序列 $\{X_1, X_2, X_3, \cdots, X_n\}$。具体地，Trm 表示 Transformer 编码处理，多层双向 Transformer encoder 构成了 BERT 的核心网络结构。$T = \{T_1, T_2, T_3, \cdots, T_n\}$ 为 BERT 模型输出词向量列表，其中 $T \in R^{n \times d}$，d 表示向量维度。

1.2　BiGRU 层

门控制单元(Gated Recurrent Unit，GRU)是对 RNN 循环神经网络的改进，相较于 LSTM 其结构更加精简，只保留了重置门以及更新门两个单元，GRU 单元结构如图 2 所示[13]。BiLSTM 及 BiGRU 均为 RNN 的一种变形结构。GRU 和 LSTM 均通过“门”结构来控制信息通过，相较于 LSTM 包括遗忘门、输入门以及输出门三个门，GRU 通过将输入门以及遗忘门相结合为更新门(update gate)，另一个门称为重置门[14]。相对来说，GRU 结构更加简单，参数更少，因此训练起来更加容易，训练效率得到了大幅提升。

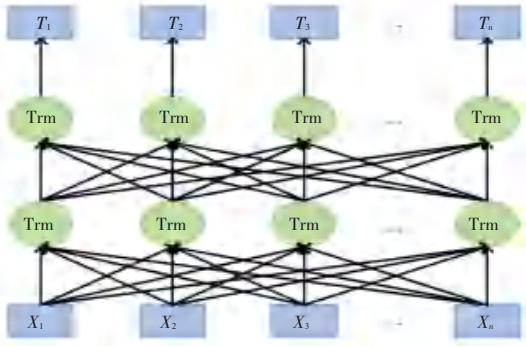

图 1　BERT 模型架构

Fig. 1　BERT model architecture

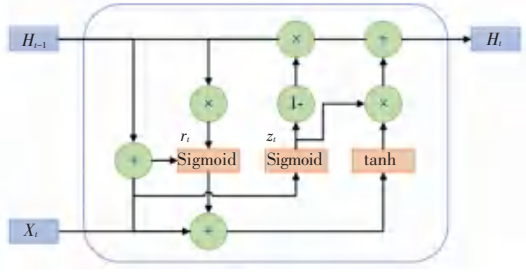

图 2　GRU 模型架构图

Fig. 2　GRU model architecture

图 2 中，r_t 表示重置门，X_t 表示更新门，h_{t-1} 和 h_t 分别表示前一时刻隐藏层状态以及当前时刻隐藏层状态，在 t 时刻，GRU 单元的更新状态可由以下公式推理得到[15]：

$$z_t = \sigma(\boldsymbol{w}_z \times [h_{t-1}, X_t] + b_z) \tag{1}$$

$$r_t = \sigma(\boldsymbol{w}_r \times [h_{t-1}, X_t] + b_r) \tag{2}$$

$$\widetilde{h_t} = \tanh(\boldsymbol{W}_h \times [r_t, h_{t-1}] + b_h) \tag{3}$$

$$h_t = (1 - z_t) h_{t-1} + z_t \widetilde{h_t} \tag{4}$$

其中，σ 表示 Sigmoid 函数，通过这个函数将数据转化为 0 ~ 1 区间的值作为门控制信号；$\boldsymbol{w}_z$、$\boldsymbol{w}_r$、$\boldsymbol{w}_h$ 均为权重矩阵；$\widetilde{h_t}$ 表示当前单元中需要更新的信息。

普通的 GRU 单元是单向的，因此只可从前往后

地使用文本信息,存在文本特征信息利用不全、甚至信息丢失的情况。而在中文文本中,上下文描述之间存在密切的联系,即同时联系上下文语义信息可以更精准地识别事件触发词及环境实体要素,因此,本文选择双向 GRU 进行实体识别研究。利用双向 GRU 网络弥补普通 GRU 网络单元的单向传递缺陷,实现文本信息的双向识别,从而更加充分地捕获语句文本特征,增强长文本实体的识别效果。

1.3 注意力层

一段完整事件描述因包括事件发生背景、环境、事件类型、事件内容及后果等多个要素而常常需要较长的文本描述,因此可以通过引入注意力机制弥补 BiGRU 模型无法提取长距离文本特征的缺点。注意力机制在使用时表现为序列到序列的处理方式[16],可对识别实体相关的特征分配较大的特征权重,其他无关特征分配较小的特征权重,从而提升模型对重要局部信息的捕捉能力,且进一步结合注意力权重不受词间距离的影响的特点,有效解决了 BiGRU 模型长距离依赖问题[17]。研究推得的系列公式如下:

$$\boldsymbol{m}_t = \tanh(\boldsymbol{W}_v \times \boldsymbol{h}_t + \boldsymbol{b}_v) \tag{5}$$

$$\boldsymbol{\alpha}_t = \frac{\exp(\boldsymbol{m}_t^{\mathrm{T}} \times \boldsymbol{k}_v)}{\sum_t \exp(\boldsymbol{m}_t^{\mathrm{T}} \times \boldsymbol{k}_v)} \tag{6}$$

$$\boldsymbol{C} = \sum_t \boldsymbol{\alpha}_t \times \boldsymbol{h}_t \tag{7}$$

其中,$\boldsymbol{h}_t$ 表示 BiGRU 网络层的输出;$\boldsymbol{m}_t$ 表示权重向量,由 $\boldsymbol{h}_t$ 经全连接层后得到,显示了当前信息与上下文信息的相关性;$\boldsymbol{W}_v$ 表示注意力模型可调整的权重矩阵;$\boldsymbol{b}_v$ 表示偏置项;$\boldsymbol{\alpha}_t$ 表示注意力权重矩阵;$\boldsymbol{C}$ 表示经过注意力模型权重分配输出加权后的特征向量。

1.4 CRF 层

条件随机场可以在 BiGRU 编码层得到各个标签具体分值的基础上考虑标签之间的依赖关系,从而输出一个全局最优的合理标签序列[18]。将句子序列表示为 $\{X_1, X_2, X_3, \cdots, X_n\}$,其对应预测标签 $Y = \{Y_1, Y_2, Y_3, \cdots, Y_n\}$,预测列标签分数可由下式计算求得:

$$s(X, Y) = \sum_{i=1}^{n} (W_{y_i, y_{i+1}} + p_{i+1, y_{i+1}}) \tag{8}$$

其中,$\boldsymbol{W}$ 表示转换矩阵;$W_{i,j}$ 表示标签转移分数;$p_{i+1, y_{i+1}}$ 表示当前字符第 y_{i+1} 个标签的分数。此后,通过全局归一化处理,标签序列 Y 的产生概率可由下式计算求得:

$$P(X \mid Y) = \frac{e^{s(X,Y)}}{\sum_{Y \in Y_X} e^{s(X,Y)}} \tag{9}$$

2 基于要素匹配的事件抽取方法

在事件相关的记录和报道中,一个段落可能包含多个事件描述。这些事件往往具有一定的相关性,一个事件要素可能是多个事件的共同要素,这就需要运用一定的规则方法划分出具有相同时间地点要素的农灾事件,实现事件-时间-地点的一一对应。

文章提出一种基于环境要素及事件触发词的事件抽取方法,在命名实体识别模型识别出环境实体及触发词事件的基础上,具体包括事件要素匹配以及共现事件融合两个步骤。

2.1 要素匹配算法设计

以地点要素为例,匹配算法流程描述如下。

算法 语句地点要素匹配算法

输入 Sen = $\{S_1, S_2, \cdots, S_i\}$ #Sen 为事件文本按句号划分后的语句列表;$L_i = \{L_{i,1}, L_{i,2}, \cdots, L_{i,j}\}$ #L_i 是语句 s 中识别出的地点要素列表,$L_{i,j}$ 表示 S_i 根据逗号分隔后第j句中识别出的地点要素,若该句不包含地点要素,则$L_{i,j}$ = None;$V_i = \{V_{i,1}, V_{i,2}, \cdots, V_{i,j}\}$ #V_i 表示语句 s 中识别出的事件触发词列表,$V_{i,j}$ 表示S_i 根据逗号分隔后第j句中识别出的事件触发词,若该句不包含地点要素,则 $V_{i,j}$ = None。

输出 匹配更新后的 $L_i = \{L_{i,1}, L_{i,2}, \cdots, L_{i,j}\}$

执行:

1. for i in range(len(Sen)):

#设置该句的初始默认地点为该句识别出的第一个事件触发词所在句之前识别出的地点要素

2. set default_L

#按逗号划分当前语句并遍历划分列表

3. seq = Sen[i]. split(',')

4. for j in range(len(seq)):

#若当前语句不包含地点要素,则设置该句地点要素为当前默认地点

5. if L_{ij} is None:

6. L_{ij} = default_L

#若当前语句包含地点要素(识别出新地点),且前一句包含事件触发词(前一语句同时具有环境要素和事件触发词,构成完整事件),则用当前语句地点要素替换当前默认地点(前一地点不再修饰后续语句,进行替换)

7. if $L_{i,j}$ is not None and $V_{i,j-1}$ is not None:

8.　set default_L= $L_{i,j}$

#若当前语句包含地点要素(识别出新地点)，前一句不包含事件触发词，则在默认地点上添加当前语句地点要素(前一语句未与事件触发词构成完整事件，继续修饰后续语句，无需替换)

9. elif ：

10.　　default_L += $L_{i,j}$

11. return L_i

2.2　实体融合

中文表达方式具有多样性，为使得同一内容在不同的语境下得到更加贴切的表达，文本作者在描述时往往采用更适合当前描述的表达方式，造成了在中文语境下多词一意现象，例如合肥市、合肥、安徽省会等地点均指代同一地点要素。基于此，就需利用实体融合技术实现实体的共指消歧[19]。时间地点是体现灾害事件共现关系的重要指标，基于时间地点要素实现单一事件抽取，对后续共现灾害事件的挖掘，提升事件语料库的应用性能具有重要意义。

基于环境要素，本文结合中国行政区划数据库对识别出的环境地点进行实体消歧以及数据补充。该数据库记录了中国各省及直辖市的行政区划地点的名称，每条数据的最低行政区划等级为居委会，该条数据还记录居委会所处的省(直辖市)、市、县(区、旗)以及街道，共计 5 490 亿条数据，每条数据有唯一 ID 标识。地点实体融合流程如图 3 所示。应用该方法，对识别出的地点实体进行中文分词、数据清洗、数据库检索、正则表达式匹配，将匹配的地点实体融合为统一地点描述，并实现其行政区划等级的完善。

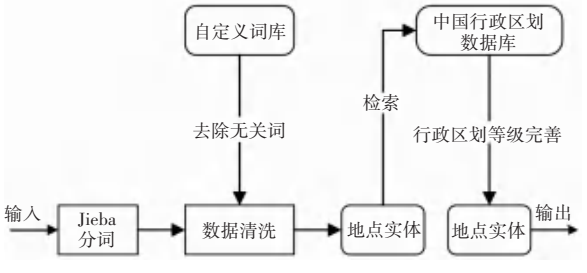

图 3　地点实体融合流程图

Fig. 3　Flow chart of location entity integration

3　实验与分析

3.1　实验数据

本文以灾害事件为例，验证上述方法对于相关事件的抽取效果，数据来源于《中国气象灾害大典》。该书记录了有文字记载以来的古代、近代以及现代各种文字途径中有据可查的气象灾害，时间跨度截止至 2000 年。文章从《大典》中选取了国内旱灾多发的 20 个省份 1949～2000 年的旱灾文本描述数据，对选取数据进行文字识别以及人工校验，总共获取旱灾文本 4 902 条，共计 736 064 字。

采用经典的 BIO(B-begin，I-inside，O-outside)三位序列标注法通过精灵标注助手平台对文本中的时间、地点要素进行标注，'B'，'I'，'O'分别表示实体词的开始、中间、非实体词位置，则针对时间、地点及事件触发词要素对应的标签集合为{O，B-Loc，I-Loc，B-Time，I-Time，B-V，I-V}。随机选取旱灾文本中的 1 052 条数据进行环境要素人工标注，标注地点实体 5 148 个，时间实体 2 138 个，选取 411 条旱灾文本数据进行事件触发词人工标注，标注触发词实体 2 695 个，均按照 8∶1∶1 的比例划分训练集、验证集以及测试集。

3.2　评价体系

对于命名实体识别任务评估的精确匹配，是指要求同时识别出实体的类型以及精确的边界，才视为正确识别。文中采用精确匹配作为要素识别任务的评估方法，包括识别准确率(Precision)，识别召回率(Recall)以及识别 $F1$($F1$ - Score) 值三个指标，公式如下：

$$\text{Precision} = \frac{T_1}{T_1 + T_2} \tag{10}$$

$$\text{Recall} = \frac{T_1}{T_1 + T_3} \tag{11}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

其中，T_1 表示识别正确的实体数；T_2 表示识别错误的实体数；T_3 表示未识别出的实体数。

3.3　实验结果分析

实验基于 Window10 操作系统，NVIDIA GeForce TRX 3050 Laptop GPU 显卡配置，Python3. 9. 12 编程语言及 Pytorch 0. 15. 2 深度学习框架。

进行命名实体识别实验的相关参数配置见表 1。

文中构建 BERT-BiGRU-CRF-Att 模型完成命名实体识别任务，为验证添加注意力机制对于事件触发词及环境实体识别效果的提升，选择 BERT-BiGRU-CRF 模型作为对比参照。模型实体识别结果见表 2。

基于表 2 中 BERT-BiGRU-CRF 模型实体识别结果显示，该模型识别灾害事件中时间及地点要素

$F1$ 值均达到了 0.8 以上,具有良好的识别效能。通过添加注意力机制,该模型表现出了更好的实体识别效果,3 类要素识别 $F1$ 值均得到了提升,验证了注意力机制对于长文本实体识别的优化效果。

表 1 实验参数设置

Table 1 Experimental parameter settings

参数名	参数值
batch_size	50
epoch	100
学习率(lr)	2×10^{-4}
优化器	Adam(Adaptive Moment Estimation)
Max_sentence_len	252
Dropout	0.5
Hidden_size	128

表 2 模型实体识别结果

Table 2 Model entity recognition results

模型	实体	Precision	Recall	$F1$
BERT-BiGRU-CRF	地点实体(Loc)	0.846 2	0.873 0	0.859 4
	时间实体(Time)	0.832 8	0.863 9	0.848 1
	事件触发词(V)	0.758 1	0.755 4	0.756 8
BERT-BiGRU-Att-CRF	地点实体(Loc)	0.893 4	0.879 0	0.886 2
	时间实体(Time)	0.833 3	0.873 0	0.852 7
	事件触发词(V)	0.795 3	0.846 4	0.820 1

经命名实体识别、实体融合处理,从 4 902 条旱灾文本数据中获取时间要素 2 138 个,地点要素 5 148 个。通过要素匹配算法处理,匹配每句描述文本所对应的时间及地点要素,随后将具有相同时间及地点要素的语句合并为同一事件,并按照原语序排序,得到时间地点单一对应的灾害事件语料共计 17 598 条。

4 结束语

针对如何从新闻报道、文献记录等语料实现单一事件抽取,构建规范的事件语料库的问题,本文提出了 BERT-BiGRU-Att -CRF 事件环境要素识别模型,通过引入注意力机制,解决了 BiGRU 模型针对长文段的事件文本特征提取不足的问题,提升了识别效果。基于文本特点,设计灾害语句环境要素匹配算法,通过实体融合,按顺序结合灾害描述语句,得到特定环境下发生的具体完整的事件,并通过实验,验证了研究方法对于事件抽取的有效性。

语料库在自然语言识别领域具有广泛的应用[20],事件领域语料库为事例研究、事件关系挖掘的相关问题提供了新的解决方案。后续将基于事件语料库,构建事理知识图谱和领域问答系统,进一步结合领域专家系统构建辅助决策系统。研究成果对于提高相应领域智能化应用水平具有重要意义。

参考文献

[1] 李熠,耿朝阳,杨丹. 基于 Fin-BERT 的中文金融领域事件抽取方法 [J]. 计算机工程与应用,2024,60(14): 123-132.

[2] DU Xinya, CARDIE C. Document - level event role filler extraction using multi - granularity contextualized encoding [C]// Proceedings of Association for Computational Linguistics (ACL). ACL, 2020:634-644.

[3] YANG Hang, CHEN Yubo, LIU Kang, et al. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data [C]//Proceedings of Association for Computational Linguistics (ACL). ACL, 2018: 50-55.

[4] ZHENG Shun,CAO Wei, XU Wei, et al. Doc2EDAG:An end-to-end document - level framework for Chinese financial event extraction[C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). ACL,2019:337-346.

[5] 杜志强,李钰,张叶廷,等. 自然灾害应急知识图谱构建方法研究[J]. 武汉大学学报(信息科学版),2020,45(9):1344-1355.

[6] 陈曙东,欧阳小叶. 命名实体识别技术综述[J]. 无线电通信技术,2020,46(3):251-260.

[7] 宋卫强, 李焰. 基于 BERT-IDCNN-CRF 的医疗命名实体识别[J]. 电脑与信息技术, 2023, 31 (6): 53-57.

[8] PENG Xiang, DRESER M. Learning based on word segmentation representation improves named entity recognition in Chinese social media[C]// Annual Meeting of the Association for Computational Linguistics. ACL,2016:149-155.

[9] DONG Xu, QIAN Ling, GUAN Ying, et al. A multi-class classification method for electronic medical record named entity recognition based on deep learning[C]//Proceedings of 2016 New York Scientific Data Summit (NYSDS). Piscataway, NJ: IEEE, 2016:1-10.

[10]廖 涛,勾艳杰,张顺香. 融合注意力 机制的 BERT-BiLSTM-CRF 中文命名实体识别[J]. 阜阳师范大学学报(自然科学版),2021,38(3):86-91.

[11]刘宗田,黄美丽,周文,等. 面向事件的本体研究[J]. 计算机科学,2009,36(11): 189-192.

[12]仲伟峰, 杨航, 陈玉博,等. 基于联合标注和全局推理的篇章级事件抽取[J]. 中文信息学报, 2019, 33 (9): 88-95.

[13]杨国田,何雨晨,李鑫,等. 基于梯度提升决策树改进双向门限循环单元的锅炉变负荷燃烧系统建模[J]. 热力发电,2021,50(12):6-12.

[14]刘家利. 基于门控循环单元神经网络的结构智能控制算法研究[D]. 武汉:武汉理工大学,2022.

[15]黄忠祥,李明. BiGRU 结合注意力机制的文本分类研究[J]. 北京联合大学学报,2021,35(3):47-52.

[16]GUO Xuchao, ZHOU Han, SU Jie, et al. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism[J] . Computers and Electronics in Agriculture, 2020,179: 105830.

[17]陈娜,孙艳秋,燕燕. 结合注意力机制的 BERT-BiGRU-CRF 中文电子病历命名实体识别[J]. 小型微型计算机系统,2023,44(8):1680-1685.

[18] YANG Hang, CHEN Yubo, LIU Kang, et al. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data[C]//Annual Meeting of the Association for Computational Linguistics. ACL,2018:50-55.

[19]张吉祥,张祥森,武长旭,等. 知识图谱构建技术综述[J]. 计算机工程,2022,48(3):23-37.

[20]凌天, 焦阳, 狄碧云,等. 面向机器学习的智慧诊疗语料库构建研究[J]. 医学信息, 2023, 36 (10): 6-10.