◆计算科学◆

基于多头指针的司法事件检测方法*

张小丽[1,2,3],黄　辉[1,2,3],黄瑞章[1,2,3],秦永彬[1,2,3],陈艳平[1,2,3**]

(1.贵州大学,文本计算与认知智能教育部工程研究中心,贵州贵阳　550025;2.贵州大学,公共大数据国家重点实验室,贵州贵阳　550025;3.贵州大学,计算机科学与技术学院,贵州贵阳　550025)

摘要:针对如何解决中文司法事件检测中触发词与上下文关系不足以判定事件实例、案件触发词表述相似以及同一个案件中多个触发词识别和分类模糊的问题,本研究提出一种基于多头指针的司法事件检测方法。首先,该方法将上下文信息和罪名特征融合作为输入,使用双向长短期记忆(Bi-directional Long Short-Term Memory,BiLSTM)网络捕获数据依赖关系,深入提取特征;然后,使用多头指针网络对字符间的依赖关系进行建模,有效捕捉句子中的触发词;最后,利用指针标注技术抽取触发词,实现司法事件的有效检测。在公开司法数据集 LEVEN 上实验验证该方法的有效性,其中微平均和宏平均的 $F1$ 指标达到了 87.53%和 78.05%,优于现有模型。该方法不仅显著提高了事件触发词的识别精度,而且也增强了对复杂司法文本中事件上下文关系的把握能力。

关键词:司法事件检测;触发词;上下文关系;罪名特征;多头指针

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2024)02-0335-11

DOI:

随着信息时代的发展,新闻、教育、医疗、金融、司法等领域每天都会发生各种事件。然而,这些领域中发生的事件都被记录在无结构的文本文档中,使得快速掌握事件的信息变得困难[1]。事件抽取任务致力于从这些非结构化的文本中提取信息,并将其转换为结构化的格式,以捕获事件中的关键元素,如“谁(姓名)、何时(时间)、何地(地点)、做了什么(事件)、为什么(原因)”以及“如何(方式)”等[2]。而事件检测是事件抽取任务的重要组成部分,其主要任务是从非结构化文本中识别事件触发词并对其进行分类。触发词是指一段文本中最能表明事件发生的词[3],例如:“周三上午张三在某某停车场偷了一辆价值 5 000 元的自行车”和“张三偷了一辆自行车”,第二句缺失了时间、地点等信息,如果确定了触发词为“偷”,就可以通

收稿日期:2022-11-18　　修回日期:2023-03-30

*国家自然科学基金项目(62066008),贵州省科学技术基金重点资助项目(黔科合基础〔2020〕1Z055)和贵州省教育厅高等学校科学研究项目(青年项目)(黔教技〔2022〕149 号)资助。

【第一作者简介】

张小丽(1996—),女,在读硕士研究生,主要从事自然语言处理事件抽取研究。

【**通信作者简介】

陈艳平(1980—),男,教授,主要从事自然语言处理、人工智能等研究,E-mail:ypench@gmail.com。

【引用本文】

张小丽,黄辉,黄瑞章,等.基于多头指针的司法事件检测方法[J].广西科学,2024,31(2):335-345.

ZHANG X L,HUANG H,HUANG R Z,et al.Judicial Event Detection Method Based on Multi-head Pointer [J].Guangxi Sciences,2024,31(2):335-345.

过触发词判断为一个盗窃事件。

在司法领域,一个案件由多个相互关联的基本事件组成。司法办案的一个重要任务就是对复杂案件进行事件分析,识别出基本事件,从而为案件的理解、分析、裁定提供支撑[4]。一个完整的法律案件通常使用包含了多个事件的较长文本进行描述。法官通过阅读这些案情文本,依据相关法律条文来确定最终罪名。例如图 1 所展示的案例,根据触发词所触发的相应事件类型,可以判断 A 引发了交通事故,随后的“抛弃(Desertion)”和“逃逸(Escaping)” 共同导致了死亡事件(Died),这将 A 的指控从交通肇事罪变成了故意杀人罪,并增加了相应的处罚。然而,人工阅读这些文本内容并从中准确识别各个案件是一个耗时的工作。

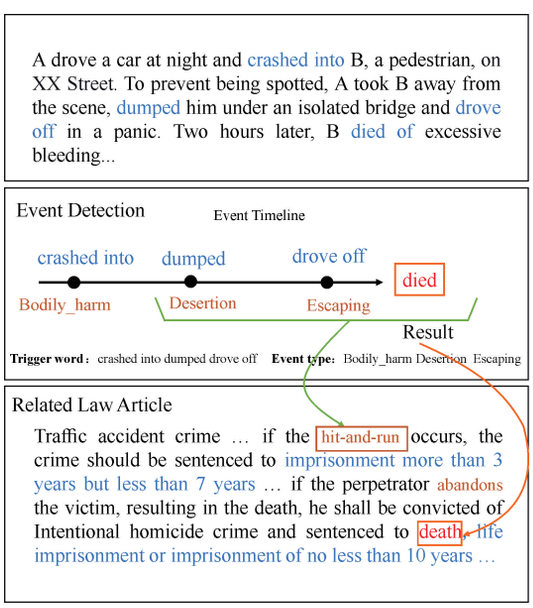


图 1　法律文档示例

Fig. 1　Example of a legal document

此外,一个案件中可能存在多个事件的描述[5],如何高效地对司法事件进行检测使其更加贴合案件的原本行为描述面临着较大困难。

目前,事件检测方法主要有以下 3 个模型:①字/词分类。Grishman 等[6]和 Ahn 等[7]对单词的词性和语法特征等进行提取,并使用最大熵方法进行事件检测。Ji 等[8]利用相关主题设计出模式匹配的方法,以改善事件检测和论元抽取的能力。Li 等[9]提出一种基于结构感知器的联合学习方式来实现对事件的检测和论文的识别。②动态最大池化。Chen 等[10]提出动态多路卷积神经网络(Deep Multi-Scale Convolutional Neural Network,DMCNN),该网络能自动提取词语级和句子级特征。Chen 等[11]提出一种使用双向长短期记忆加条件随机场(Bidirectional Long Short-Term Memory-Conditional Random Field,BiLSTM-CRF)和卷积神经网络(Convolutional Neural Network,CNN)进行句子分类的方法,通过将语义和句法依赖特征整合至词向量,有效提高句子分类的准确性并提升识别效率。③基于序列的方法。嵌入式语言模型(Embeddings from Language Models,ELMo)[12]、生成式预训练转换模型(Generative Pre-trained Transforme,GPT)[13]以及双向编码器表征法(Bidirectional Encoder Representations from Transformers,BERT)[14]等在自然语言处理领域取得了显著的成果。Wadden 等[15]使用一种名为“Dynamic Span Graph for Interaction Extraction++(DYGIE++)”的方法来实现包括事件检测在内的多项任务。Lin 等[16]基于 BERT 模型,使用条件随机场(Conditional Random Fields,CRF)对整个序列进行建模来捕捉不同事件之间的相关性。

基于字/词分类的模型,每个维度的词向量主要用于捕捉字或词的潜在语义信息,通过词向量之间的相似度来确定语义之间的关联性[17]。然而,这种方法忽略了在司法领域中触发词的识别和分类不仅依赖于单个字/词的语义信息,而且还极度依赖于上下文信息这个问题[18]。DMCNN 在一定程度上提取到句子级特征,但是存在触发词词块之间不匹配等问题。基于序列的方法主要通过标记序列中每个元素的起始位置来进行识别。然而,层与层之间可能会发生错误的传递,而且这类模型通常无法并行执行,限制了模型训练的效率。因此,在司法领域,对案件触发词的准确识别和分类,仍然存在以下 3 个方面的问题。

①缺乏上下文感知预测。许多触发器要求模型整合来自参数实体或其他句子的复杂上下文信息,以预测相应的事件类型。例如:“在 A 匆忙致电 B 告知情况”的句子中,如果 B 是警察或 110,则触发呼叫的事件类型为“向警察报告”;而如果是其他人,则事件类型为“报告”。

②触发词表述相似导致的事件类型分类模糊。例如:“挪用公款罪”和“私分国家资产罪”中会包含“转移公有资产”和“转移国家资产”事件类型,虽然这两种事件都涉及触发词“转移资产”,但是它们本质上属于不同的事件类型。

③忽略了罪名名称跟案件类型的相关性。在司

法案件描述中,罪名名称对案件描述具有一定的概括性,可在一定程度上辅助事件类型的分类。

为了解决上述问题,本研究提出基于多头指针的司法事件检测方法。该方法使用双向长短期记忆(Bi-directional Long Short-Term Memory, BiLSTM)网络[19]捕获上下文信息特征,并将每个文本输入转化为一个矩阵。此外,通过引入多头指针机制同时关注文本中不同部分的信息,从而更加精准地识别和分类文本中的触发词,以实现对司法事件信息的深度挖掘,提高法律文档处理的效率和精确度,为法律专业人士提供强大的支持,使其能够更快、更准确地理解案件细节,从而做出更加明智的决策。

1 多头指针的司法事件检测模型

本研究提出的司法事件检测模型主要分为3部分:输入层、编码层和多头指针解码层。依据司法数据集抽取任务要求高准确度的特点,采用多头指针方法,利用全局归一化的思路对案件触发词进行精准识别和分类。模型的整体框架如图2所示。

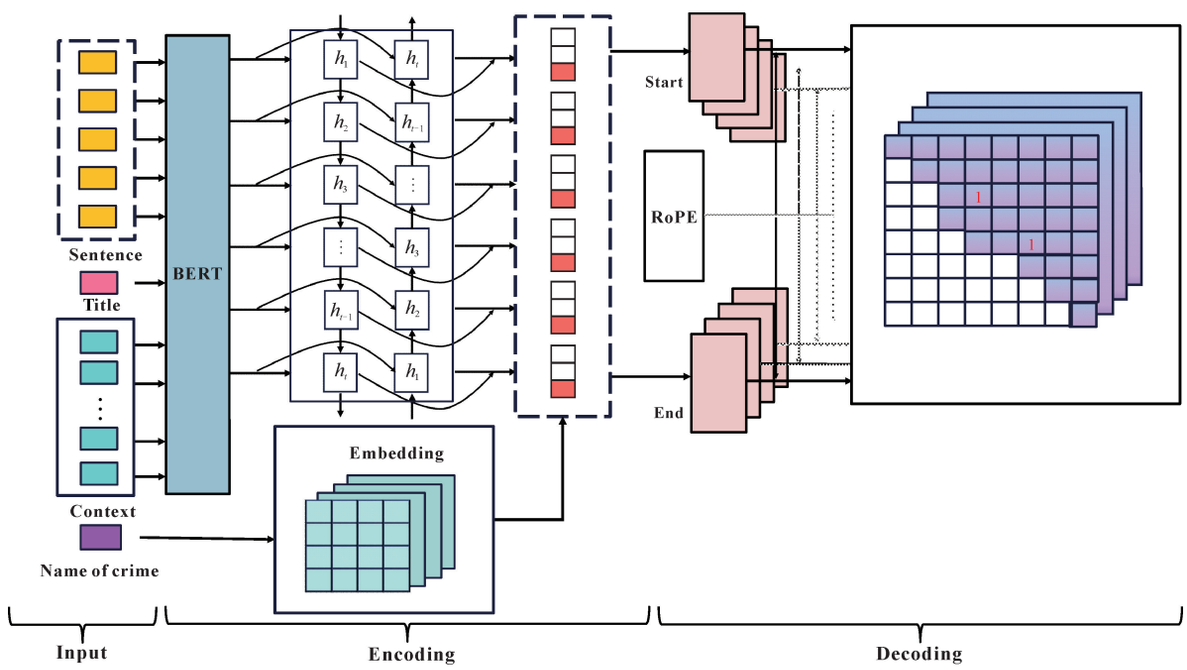


图2 多头指针模型框架

Fig. 2 Multi-head pointer model framework

1.1 模型输入

在司法案件案情描述中,案件标题、罪名名称与整个案件触发的事件类型具有一定的相关性,本研究事件检测的输入包括词汇级特征和句子级特征。其中,词汇级特征是由每个案件的标题名称和罪名名称词向量特征组成;句子级特征是通过选取案件主句和上下文相关的特征信息拼接而成。根据数据集的特点,人工选取每个案件描述的上下文信息,结合该案件的标题名称和罪名名称作为模型的输入,然后经BERT进行编码以实现字符序列到分布式序列的转换。具体输入操作如公式(1)所示:

$$S=[S_s;S_t;S_c], \tag{1}$$

式中,S表示输入的总文本,S_s表示输入的目标句子,S_t表示描述该案件文本的标题,S_c表示目标句子的上下文信息。

1.2 编码层

1.2.1 BERT模型

2018年,Devlin等[14]提出BERT模型,该模型是基于多层Transformer[20]的双向编码表征模型。在此之前,主要是通过基于Word2vec[21]、Glove[22]的方法训练静态的词向量,而BERT模型可以通过动态地训练词向量来充分地学习深层表征信息。因此,本研究采用预先训练的BERT模型作为词向量嵌入层来编码上下文信息。在编码层(Encoding)和解码层(Decoding)部分都使用了Transformer,使得一个文本中的每个字无论方向前后或距离远近,都能直接和句子中的任何一个字进行编码,每个字都能融合字左右两边的信息。文本输入BERT后,在嵌入层会得到3种嵌入向量表示:词嵌入、位置嵌入、句子类型嵌入,然后输入多层Tansformer提取特征。

1.2.2 罪名名称特征

在司法领域的数据集中,案件描述通常采用长文本形式,而罪名名称则是对案件类型的精准概括。罪名名称所蕴含的信息量较为集中和关键,与触发词的提取和分类密切相关。为了增强文本中的语义结构,本研究将罪名名称转化为稠密的分布式抽象表示。通过随机初始化一个矩阵来表示罪名名称,该矩阵的行数代表罪名的数量,每行对应一个罪名的稠密向量,从而构建了一个罪名嵌入矩阵。如公式(2)所示:

$$\boldsymbol{S}_n=\mathrm{Random}(C), \tag{2}$$

式中,C 表示初始化罪名嵌入矩阵所需的参数集,$\boldsymbol{S}_n$ 表示根据这些参数随机生成的罪名嵌入矩阵,Random 函数表示随机初始化一个矩阵。整个过程不仅简化了特征的提取和转换操作,而且通过与 BiLSTM 层的结合,进一步强化了对上下文信息的捕捉。

1.2.3 BiLSTM 模型

由于原始输入的噪声信息太多,需要通过特征提取转换层过滤噪声信息,并提取与任务目标相关度较高的高度抽象特征。BiLSTM 是由两个并行的长短期记忆网络(LSTM)层组成,其核心结构包括遗忘门、输入门、输出门和细胞状态[23]。如果在 t 时刻用 f_t、i_t、o_t 分别表示遗忘门、输入门、输出门,C_t 表示细胞状态,则其主要结构可表示为式(3)至式(8):

$$\text{遗忘门:}f_t=\sigma(\boldsymbol{W}_f[h_{t-1},x_t]+\boldsymbol{b}_f), \tag{3}$$

$$\text{输入门:}i_t=\sigma(\boldsymbol{W}_i[h_{t-1},x_t]+\boldsymbol{b}_i), \tag{4}$$

$$\widetilde{C}_t=\tanh(\boldsymbol{W}_c[h_{t-1},x_t]+\boldsymbol{b}_c), \tag{5}$$

$$\text{更新细胞状态:}C_t=f_t*C_{t-1}+i_t\widetilde{C}_t, \tag{6}$$

$$\text{输出门:}o_t=\sigma(\boldsymbol{W}_o[h_{t-1},x_t]+\boldsymbol{b}_o), \tag{7}$$

$$h_t=o_t*\tanh(C_t), \tag{8}$$

式中,σ 是 sigmoid 激活函数,tanh 是双曲正切激活函数,$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_c$、$\boldsymbol{W}_o$ 分别是遗忘门、输入门、细胞状态和输出门的权重矩阵,$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_c$、$\boldsymbol{b}_o$ 分别是遗忘门、输入门、细胞状态和输出门的偏执向量,$[h_{t-1},x_t]$ 表示时刻 $t-1$ 的隐藏状态 h_{t-1} 和输入 x_t 的组合,$\widetilde{C}_t$ 表示时刻 t 的候选细胞状态,C_t 表示时刻 t 的细胞状态,C_{t-1} 是时刻 t 的上一个细胞状态,h_t 表示时刻 t 的隐藏状态,$*$ 是元素乘法操作。

在 BiLSTM 中,分别正向和反向执行 LSTM 操作,对于正向 LSTM 层,得到正向隐藏状态序列 $\overrightarrow{h_t}=(\overrightarrow{h_1},\overrightarrow{h_2},\cdots,\overrightarrow{h_n})$;对于反向 LSTM 层,得到反向隐藏状态序列 $\overleftarrow{h_t}=(\overleftarrow{h_1},\overleftarrow{h_2},\cdots,\overleftarrow{h_n})$。最终,将两个隐藏状态序列进行拼接,形成 BiLSTM 在每个时间步的完整输出。

1.3 解码层

1.3.1 多头指针

在本研究中引入多头指针机制[24],为每个事件类别分配一个独立指针。“多头指针”实质上是指每个触发词类别与一个特定的“头”相对应,而多个不同的触发词则各自对应不同的独立指针。当面对多个事件类别时,这种机制能够确保每个事件类别都拥有其专用的指针进行标识。首先构造一个上三角矩阵,以指针形式遍历输入文本中的所有触发词,并使用“开始”和“结束”两个模块分别识别触发词的首尾;然后,将首尾视作一个文本中的整体进行判断,以获得更全局的信息。图 3 展示了多头指针的触发词识别过程。

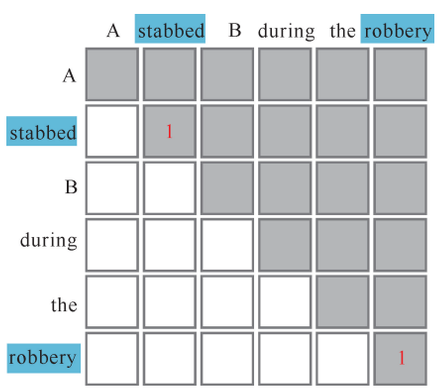


图 3 多头指针的触发词识别

Fig. 3 Trigger word recognition of multi-head pointer

若输入的文本序列长度为 n,并且每个待识别的触发词均视为该序列中的一个连续子序列,长度不限,则得出候选触发词有 $n(n+1)/2$ 个,即长度为 n 的序列有 $n(n+1)/2$ 个不同的连续子序列,再从 $n(n+1)/2$ 个候选触发词里识别出真正的触发词。在本研究模型中,将长度为 n 的输入序列经过编码层转化为对应的向量序列。每个文本序列中基本单元(Token)的编码向量被输入到两个线性变换层,即“开始”(Start)层和“结束”(End)层,以生成每个事件类别的起始和终止向量。在特定的位置 i 和 j,模型使用权重 $\boldsymbol{W}_{q,\alpha}$ 和 $\boldsymbol{W}_{k,\alpha}$ 转换当前的隐藏状态 h_i 和 h_j,通过全连接层 $l_{q,\alpha}$ 和 $l_{k,\alpha}$ 的变换,生成查询向量 $\boldsymbol{q}_{i,\alpha}$ 和键向量 $\boldsymbol{k}_{j,\alpha}$。这两个向量随后用于计算位置 i 和 j 之间的注意力分数,以评估文本各部分之间的相关性。其过程如式(9)和(10)所示:

$$\boldsymbol{q}_{i,\alpha}=\boldsymbol{W}_{q,\alpha}h_i+l_{q,\alpha}, \tag{9}$$

$$\boldsymbol{k}_{j,\alpha}=\boldsymbol{W}_{k,\alpha}h_j+l_{k,\alpha}, \tag{10}$$

式中,α 表示事件类别,其中 $\boldsymbol{W}_{q,\alpha}$ 是查询向量的权重

矩阵，h_i 是位置 i 的隐藏状态。$\boldsymbol{W}_{k,\alpha}$ 是键向量的权重矩阵，h_j 是位置 j 的隐藏状态。利用这些向量，通过内积和 Softmax 运算得到每个候选触发词的得分，如式(11)所示：

$$S_\alpha(i,j)=\mathrm{Softmax}(\boldsymbol{q}_{i,\alpha}^{\mathrm{T}},\boldsymbol{k}_{j,\alpha}), \tag{11}$$

式中，$S_\alpha(i,j)$ 表示从位置 i 到 j 的子序列中事件类别 α 的概率。

1.3.2 旋转位置编码

在 BERT 架构中，为了将位置信息融入到注意力机制中，传统方法通过向输入词向量中添加绝对位置编码。然而，这种方法不能有效地反映词与词之间的相对位置关系。不同于传统方法，旋转位置编码(Rotary Positional Encoding，RoPE)[25]提供了一种动态的位置编码方式，通过将位置编码与词向量进行旋转变换结合，使得位置信息通过改变词向量在高维空间中的方向来进行编码，从而提高模型对序列内词汇间关系的敏感度。在 RoPE 中，变换矩阵 $\boldsymbol{R}_i$ 和 $\boldsymbol{R}_j$ 被定义用于编码每个位置的信息，并通过矩阵乘法形式来表达位置间的相对关系，如式(12)所示：

$$\boldsymbol{R}(i,j)=\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j^{-1}, \tag{12}$$

然后，使用相对位置矩阵 $\boldsymbol{R}(i,j)$，将查询向量 $\boldsymbol{q}_{i,\alpha}$ 和键向量 $\boldsymbol{k}_{j,\alpha}$ 进行相互作用，计算每个位置对的得分 $S(i,j)$，如式(13)所示：

$$S(i,j)=\mathrm{Softmax}(\boldsymbol{q}_{i,\alpha}^{\mathrm{T}}\boldsymbol{R}(i,j)\boldsymbol{k}_{j,\alpha}), \tag{13}$$

式中，$\boldsymbol{q}_{i,\alpha}^{\mathrm{T}}$ 和 $\boldsymbol{k}_{j,\alpha}$ 分别是位置 i 和位置 j 的查询向量和键向量，且 α 表示事件类别。RoPE 的旋转位置示意如图 4 所示。其中，X 是没有位置嵌入的输入序列。通过 RoPE，模型将原始输入向量 (x_1,x_2) 进行转换，从而有效地引入了位置敏感性。图中向量通过旋转矩阵 $\boldsymbol{m\theta}_1$ 被映射到复数平面，旋转矩阵根据预设的角度 $\theta_\lambda(\lambda\in[1,d])$ 调整向量方向。这一映射过程将向量的实部(Re)和虚部(Im)展示在复数平面上，从而编码位置信息。转换后的向量 (x_1',x_2') 包含了经过位置编码后的特征。X_p 表示经过 RoPE 处理后的向量序列。RoPE 的加入确保模型不仅考虑到词汇的嵌入信息，而且还能通过旋转矩阵有效捕获和利用词汇的相对位置信息，提高了模型对文本结构的理解能力。

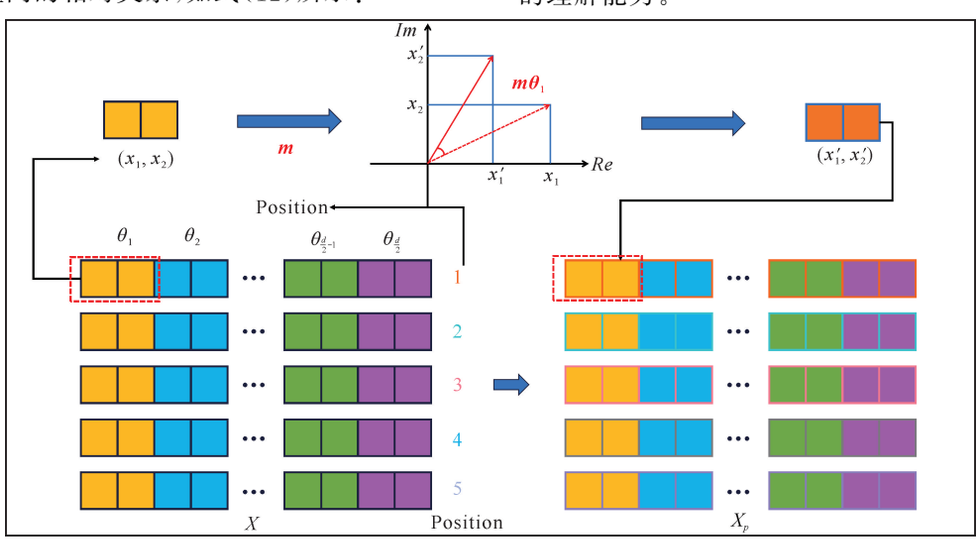


图 4 旋转位置示意图

Fig. 4 Rotation position diagram

1.4 损失函数

本研究采用了一种改进的 Softmax 损失函数[26]，通过考虑目标事件类别的预测得分和其他非目标事件类别得分的相对大小，有效增强了模型对触发词的敏感性和事件分类的准确度。具体公式如下：

$$L=-\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)+\log\left(1+\sum\nolimits_{j\neq y_i} e^{s_j-s_{y_i}}\right), \tag{14}$$

式中，$e^{s_{y_i}}$ 表示目标事件类别 y_i 得分，而 e^{s_j} 表示非目标事件类别得分；$\sum_j e^{s_j}$ 是所有非目标事件类别得分的指数和，用于对非目标事件类别得分进行归一化；$\sum_{j\neq y_i} e^{s_j-s_{y_i}}$ 是所有非目标事件类别得分与目标事件类别得分之差的指数和。

通过应用 Softmax 函数，将各类别的得分转换成一个概率分布，其中每个类别的概率都是非负的且总和为 1。这种转换不仅提升了模型对主要目标事

件类别的敏感性,而且也增强了其对非目标事件类别的区分能力,尤其适用于样本分布不均匀的数据集。通过调节事件类别间得分的相对大小,提高模型在复杂环境中的准确度和鲁棒性。

2 结果与分析

本研究使用清华大学于 2022 年发布的大规模司法数据集 LEVEN[27] 进行实验。LEVEN 数据集涵盖多种事件类型模式,提供了丰富的事件实例,专为法律事件检测设计。在此数据集中,司法事件定义为涉及参与方、具有特定属性的事情,通常被描述为状态的变化。为了评估本研究模型在司法事件检测上的效果,实验采用精确率(Precision,P)、召回率(Recall,R)和 $F1$ 值作为性能评价指标,通过对比实验结果验证模型的效能。

2.1 实验设置

2.1.1 数据集分析

在事件检测中,有些事件的基本信息来自其他句子,这就需要模型能够捕获跨句子的依赖性。本研究模型采取拼接的方式融入上下文信息作为输入,再通过模型集成来自参数实体或其他句子的复杂上下文信息,以预测相应的事件类型。此项工作需要考虑句子长度问题,而 LEVEN 数据集的长度和数量的分布情况如图 5 所示。数据集按照官方发布时的划分比例 0.65∶0.15∶0.20 划分为训练集、验证集和测试集,其中,训练集和验证集上的句子平均长度分别约为 35.38 和 34.87,训练集和验证集上的句子长度在 100 以下的占了大量分布,这就证明本研究模型所采取的输入最大长度为 512 是合理的。

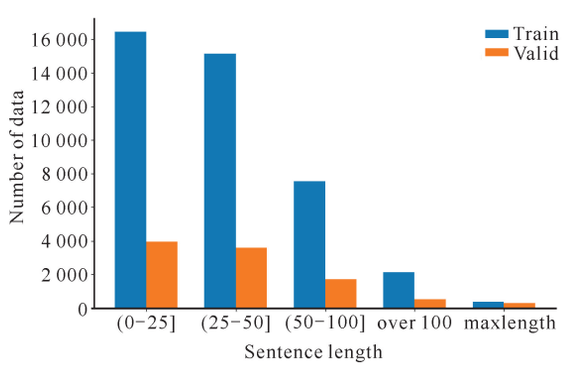


图 5 句子长度分布

Fig. 5 Sentence length distribution

2.1.2 实验环境与参数设置

本研究使用 Python3.6、Pytorch 1.2.0 深度学习框架,在 NVIDIA Tesla A100 GPU 平台上进行实验,其中 RoBERTa-wwm-ext-base[28] 用作预训练模型。实验参数设置如表 1 所示。

表 1 实验参数设置

Table 1 Experimental parameter settings

参数名 Parameter name	值 Value
Batch size	8
Epoch	16
Learning rate	5e−05
Head size	200
Dropout	0.5

2.1.3 评价指标

本研究采用 $F1$ 值来衡量模型的性能。$F1$ 值是一种基于精确率和召回率的调和平均数,用于评价模型的准确性和覆盖率。精确率是指模型正确识别事件触发词的能力,召回率则用来衡量模型识别所有相关事件触发词的能力,而 $F1$ 值的计算公式结合了这两个指标,旨在平衡模型在精确度和召回率之间的性能,确保模型不会错过重要的信息。$F1$ 值计算公式如式(15)−(17)所示,其值是一个介于 0 到 1 之间的分数,分数越高表示模型性能越好。在实际应用中,$F1$ 值意味着寻找精确率和召回率之间的最佳平衡点,以达到最优的事件触发词抽取效果。

$$P=\frac{\text{识别正确的触发词个数}}{\text{识别出的触发词个数}}\times 100\%, \tag{15}$$

$$R=\frac{\text{识别正确的触发词个数}}{\text{样本中的所有触发词个数}}\times 100\%, \tag{16}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\%。 \tag{17}$$

2.2 仿真实验

2.2.1 实验结果分析

为了验证基于多头指针模型对事件检测的有效性,在 LEVEN 数据集上将多头指针模型与 BERT[14]、BERT + CRF[16]、DMBERT[29]、BiLSTM[19]、BiLSTM+CRF[30]、DMCNN[10] 6 种基线模型与当前的 Span-Regression[31]、OneEE[32] 两种模型进行对比(表 2)。另外,分别采用双向门控循环单元(BiGRU)[33] 和 BiLSTM 来提取文本的上下文特征,并与多头指针网络相结合得到 BiGRU-Multi-Head 和 BiLSTM-Multi-Head 两个模型,用以探讨不同的特征提取方式对实验的影响,其结果如表 2 所示:①本研究模型与其他模型对比,微平均的 $F1$ 值得到一定提升,性能为 87.53%,而宏平均的 $F1$ 值略低于

DMBERT,这是因为本研究模型在进行触发词识别时特地将候选触发词信息排除在外,而基线模型均利用了候选触发词信息,但在实际应用中知道句子中的候选触发词信息是不现实的。②与基于 CRF 的模型相比,采用基于多头指针机制(Multi-Head)的模型显示出更优越的性能表现。因为 BiLSTM-Multi-Head 模型通过在单句中有效识别并关联多个事件,极大地增强了其对复杂事件的处理能力。这一策略不仅提高了模型的应用范围,而且还增强了其在实际环境中的适用性和准确性。③Span-Regression 是一种基于跨度回归对触发词进行抽取的模型,通过回归调整候选跨度的边界来准确定位触发词,而本研究直接使用指针定位识别,识别结果更加精确。④OneEE 模型使用词的直接映射进行触发词识别,主要解决重叠和嵌套事件,然而本研究所使用的数据集中并没有显著的嵌套事件,因此其性能低于本研究模型。

表 2　事件检测模型性能对比

Table 2　Performance comparison of event detection models

Unit:%

模型 Model	微平均 Micro-average			宏平均 Macro-average		
	P	*R*	*F*1	*P*	*R*	*F*1
BERT	84.35	83.80	84.07	80.21	76.08	77.38
BERT+CRF	83.72	84.13	83.93	78.38	75.39	76.01
DMBERT	83.40	86.76	85.05	79.18	79.28	78.42
BiLSTM	83.01	84.30	83.65	78.45	73.39	74.27
BiLSTM+CRF	84.63	83.10	83.86	80.99	73.39	75.73
DMCNN	86.15	79.27	82.52	79.42	69.77	73.00
Span-Regression	83.87	80.99	82.41	79.51	71.84	74.28
OneEE	85.64	83.19	84.39	82.39	74.68	77.34
BiGRU-Multi-Head	87.91	85.57	86.72	81.88	76.17	77.93
BiLSTM-Multi-Head	**88.14**	**86.92**	**87.53**	**83.55**	**74.79**	**78.05**

Note:bold numbers indicate the experimental results of the proposed model.

BiGRU 是基于 BiLSTM 的一种变体,两者都可捕获到双向语义,所以本研究分别对基于 BiGRU 和 BiLSTM 的模型进行了实验对比。基于 BiLSTM 实验的微平均和宏平均的性能均高于基于 BiGRU 的模型,该差异跟两者的结构类型、本研究数据集规模有一定关系。GRU 参数更少,收敛更快,但是由于本研究数据量很大,LSTM 效果相对更好一些,同时 LSTM 参数也比 GRU 参数多一些,在其后面接入多头指针更加适宜。

在基线模型中,DMBERT 使用预训练语言模型 BERT 来提取序列特征,并使用动态池化层来获得每个候选触发词特定表示,实验性能是基线模型中最好的。但是,该模型在训练过程中耗费了大量的时间。使用 DMBERT 模型和本研究的多头指针模型在 Epoch 都为 4、Batch size 都为 8 的条件下进行时间性能对比,结果如表 3 所示。在相同的实验参数设置下,本研究模型训练的总时间约为 40.55 min,平均每个轮数(Epoch)训练时间约为 10.14 min,而 DMBERT 训练的总时间约为 1 681.20 min,平均每个 Epoch 训练的时间为 420.30 min,该时间约是本研究模型的 41 倍。因此,本研究模型在保证实验性能的基础上大大缩短了模型的训练时间。

表 3　实验时间性能对比

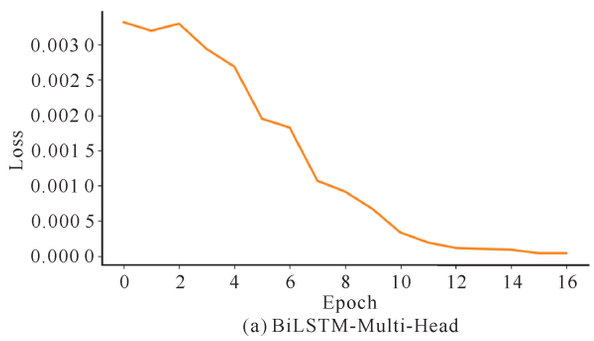
Table 3　Performance comparison of experimental time

模型 Model	轮数 Epoch	批次 Batch size	总时间/分 Total time/min	平均每轮时间/分 Average time per epoch/min
DMBERT	4	8	1 681.20	420.30
BiLSTM-Multi-Head	4	8	40.55	10.14

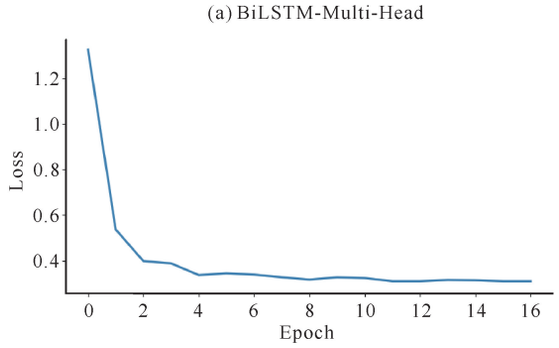
此外,从图 6 可以看出,DMBERT 模型和本研究模型的损失值在初始 Epoch 中都出现了较大的波动,但后来逐渐稳定,开始收敛。然而,本研究模型的平均损失值约为 0.001 036,方差为 0.000 82;而 DMBERT 的平均损失值约为 0.486 5,方差为 0.180 1。DMBERT 模型的损失值明显高于本研究模型,说明本研究模型变化趋势更平稳,变化速度更快,模型更容易收敛,训练时间更短,模型更稳定,并且能更准确地预测目标值。

在 BERT 模型被提出之前,研究者一般都是通过 Glove、Word2vec 训练词向量。其中,Glove 是一个基于全局词频统计的词表征工具,通过高维空间的向量捕捉词汇间的语义关系,如相似性和类比性。相对地,Word2vec 基于局部文本窗口构建共现矩阵,支持在线更新词向量,适合动态语料学习。然而,这两种方法生成的都是静态词向量,无法有效处理多义词问题。在中文数据集上,基于 Word2vec 的词向量表现优于基于 Glove 的词向量[34],这种差异可能源于 Word2vec 在处理中文文本时能更有效地利用局部上下文信息,适合于中文语言的特定语义和结构特征,而 Glove 侧重于全局统计数据。因此,本研究对比了基于 Word2Vec 和 BERT 的文本特征表示方法,以探索不同词向量技术在中文数据集上的表现,其结果

详见表 4。



(a) BiLSTM-Multi-Head



(b) DMBERT

图 6　不同模型的损失变化曲线

Fig. 6　Loss curve of different models

表 4　基于 Word2vec 和 BERT 的实验结果对比

Table 4　Comparison of experimental results based on Word2vec and BERT　**Unit:%**

模型 Model	微平均 Micro-average		
	P	*R*	*F*1
Base Word2vec	83.69	86.32	84.98
Base BERT	88.14	86.92	87.53

从表 4 可知,对比基于 Word2vec 和基于 BERT 的文本特征表示模型的性能,基于 Word2vec 的模型表现较差,精确率下降了 4.45 百分点,召回率下降了 0.60 百分点,*F*1 值下降了 2.55 百分点。Word2vec 通过预训练加微调的先进思想实现,但是,它仅仅从低维的角度进行词到索引的映射,无法解决一词多义的问题。进行词向量化的过程实际上如同对数据进行空间坐标化,Word2vec 仅仅从平面的角度去解决问题,其关注点主要在于如何获得更好的词向量,将任务同词向量结合起来,忽视了词向量在不同维度上的潜在差异,以及这些差异对数据表征精度的具体影响。在 BERT 模型中,通过引入高维度的位置编码,有效捕捉到词汇在不同语境中的语义变化,从而显著提高模型在各种复杂文本处理任务中的表现。在复杂司法案件描述中,由于触发词的表述存在一定的相似性,所以本研究选择基于 BERT 来做词向量的深层表征。

2.2.2　消融实验

在真实场景中,由于案情事件构成的复杂性和语言表述的多样性,大量的案件是以多个句子表达的,且一个事件往往涉及多个事件元素,而事件对应的元素往往分散在多个语句中,都出现在同一个语句中的理想情况并不常见,即案件元素分散。要准确地捕获散落的案件信息,需要模型捕获跨句子依赖性。为了验证上下文信息特征对司法事件检测的有效性,本研究设计了消融实验,如表 5 所示,其中 w/o 表示去掉某模块。

表 5　微平均实验结果对比

Table 5　Comparison of micro-average experimental results　**Unit:%**

模型安装 Model setup	微平均 Micro-average		
	P	*R*	*F*1
BiLSTM-Multi-Head	88.14	86.92	87.53
w/o contextual information	86.59	84.04	85.30
w/o above information	87.12	84.80	85.94
w/o following information	86.88	85.36	86.11
w/o BiLSTM	86.91	85.56	86.23
w/o crime name feature vector	87.04	84.19	85.59

由表 5 可知,上下文信息对于增强模型预测的敏感性和对事件触发词及其类型的准确分类至关重要。当模型去掉上下文信息特征之时,*F*1 值下降了 2.23 百分点,精确率下降了 1.55 百分点,召回率下降了 2.88 百分点。为进一步探索上下文语义完整性对性能的影响,本实验尝试分别去除上文和下文信息,结果显示 *F*1 值分别下降了 1.59 百分点和 1.42 百分点。此外,BiLSTM 通过过滤输入的过量噪声信息,聚焦于抽取与司法案件时序相关的高度抽象特征,证明其在处理时序信息中的重要性。去除 BiLSTM 组件后,模型的 *F*1 值、精确率和召回率分别下降了 1.30、1.23 和 1.36 百分点。同时,考虑到罪名名称与案件主题及触发词类型的强关联性,移除罪名名称特征矩阵后对复杂司法案件中的触发词识别和分类能力产生负面影响,导致 *F*1 值、精确率和召回率分别下降了 1.94、1.10 和 2.73 百分点。这些结果表明,整合上下文信息及罪名名称特征向量对提升触发词的识别和分类准确度具有显著的效果。

3 结论

本研究设计一种基于多头指针的司法事件检测方法,首先融合待预测文本的上下文信息、罪名特征作为输入,然后结合 BiLSTM 进行特征提取,最后通过指针标注识别触发词,将触发词的首尾视作一个整体去做预测。该方法在丰富预训练语言模型的语义表征能力的同时,有效地强化了触发词的识别和分类能力,提升了司法事件检测的性能。未来可进一步优化模型结构,对其他领域的事件检测展开研究。

参考文献

[1] 贾阵,丁泽华,陈艳平,等.面向司法数据的事件抽取方法研究[J].计算机工程与应用,2023,59(6):277-282.

[2] LI Q,PENG H,LI J X,et al. A comprehensive survey on schema-based event extraction with deep learning [Z/OL]. (2021-08-23)[2023-03-10]. https://arxiv.org/pdf/2107.02126v4.

[3] AFYOUNI I,AL AGHBARI Z,RAZACK R A. Multi-feature,multi-modal,and multi-source social event detection:a comprehensive survey [J]. Information Fusion,2022,79:279-308.

[4] 谢伟.复杂司法案件事件抽取方法[D].上海:上海大学,2021.

[5] BISCANI F,IZZO D. Reliable event detection for Taylor methods in astrodynamics [J]. Monthly Notices of the Royal Astronomical Society,2022,513(4):4833-4844.

[6] GRISHMAN R,WESTBROOK D,MEYERS A. NYU's English ACE 2005 system description [C]. [S. l.]:ACE 2005 Evaluation Workshop,2005.

[7] AHN D. The stages of event extraction [C]//Proceedings of the Workshop on Annotating and Reasoning about Time and Events-ARTE '06. Stroudsburg,PA:Association for Computational Linguistics,2006:1-8.

[8] JI H,GRISHMAN R. Refining event extraction through cross-document inference [C]//MOORE J D,TEUFEL S,ALLAN J,et al. Proceedings of ACL-08:Hlt. Stroudsburg,PA:Association for Computational Linguistics,2008:254-262.

[9] LI Q,JI H,HUANG L. Joint event extraction via structured prediction with global features [C]//SCHUETZE H,FUNG P,POESIO M. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers). Stroudsburg,PA:Association for Computational Linguistics,2013:73-82.

[10] CHEN Y B,XU L H,LIU K,et al. Event extraction via dynamic multi-pooling convolutional neural networks [C]//ZONG C Q,MICHAEL S. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1:Long Papers). Stroudsburg,PA:Association for Computational Linguistics,2015:167-1776.

[11] CHEN T,XU R F,HE Y L,et al. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN [J]. Expert Systems with Applications:An International Journal,2017,72(c):221-230.

[12] DING R X,LI Z J. Event extraction with deep contextualized word representation and multi-attention layer [C]//GAN G G,LI B H,LI X,et al. Advanced Data Mining and Applications. Berlin:Springer,2018:189-201.

[13] DONG L,YANG N,WANG W H,et al. Unified language model pre-training for natural language understanding and generation [C]//WALLACH H M,LAROCHELLE H,ALINA B,et al. NIPS'19:Proceedings of the 33rd International Conference on Neural Information Processing Systems. NY,United States:Curran Associates Inc.,2019:13063-13075.

[14] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of deep bidirectional transformers for language understanding [C]//BURSTEIN J,DORAN C,SOLORIO T. 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). Stroudsburg,PA:Association for Computational Linguistics,2019:4171-4186.

[15] WADDEN D,WENNBERG U,LUAN Y,et al. Entity,relation,and event extraction with contextualized span representations [C]//INUI K,JIANG J,NG V,et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg,PA:Association for Computational Linguistics,2019:5784-5789.

[16] LIN Y,JI H,HUANG F,et al. A joint neural model for information extraction with global features [C]//JURAFSKY D,CHAI J,SCHLUTER N,et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg,PA:Association for Computational Linguistics,2020:7999-8009.

[17] LI S,LIU L Y,XIE Y Q,et al. P4E:few-shot event detection as prompt-guided identification and localization [Z/OL]. (2022-02-15)[2023-03-10]. https://arxiv.org/pdf/2202.07615.

[18] LI Q,ZHANG Q,YAO J,et al. Event extraction for criminal legal text [C]//2020 IEEE International Conference on Knowledge Graph (ICKG). Piscataway,NJ:IEEE,2020:573-580.

[19] ZHAO Y,YANG Y. Sports news relationship extraction based on BERT′s BiLSTM and attention [C]//International Conference on Internet of Things and Machine Learning (IoTML 2021). Bellingham,Washington USA:SPIE 12174,2022:282-287.

[20] YE H,ZHANG N,DENG S,et al. Contrastive triple extraction with generative transformer [C]//Proceedings of the AAAI conference on artificial intelligence. Stroudsburg,PA:Association for Computational Linguistics,2021,35(16):14257-14265.

[21] STYAWATI S,NURKHOLIS A,ALDINO A A,et al. Sentiment analysis on online transportation reviews using word2vec text embedding model feature extraction and support vector machine (SVM) algorithm [C]//2021 International Seminar on Machine Learning,Optimization,and Data Science (ISMODE). Piscataway,NJ:IEEE,2022:63-167.

[22] PENNINGTON J,SOCHER R,MANNING C D. Glove:global vectors for word representation [C]//MOSCHITTI A,PANG B,DAELEMANS W. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg,PA:Association for Computational Linguistics,2014:1532-1543.

[23] ZHOU Z Q,HUANG K J,QIU Y,et al. Morphology extraction of fetal electrocardiogram by slow-fast LSTM network [J]. Biomedical Signal Processing and Control,2021,68:102664.

[24] VINYALS O,FORTUNATO M,JAITLY N. Pointer networks [C]//CORTES C,LEE D D,SUGIYAMA M,et al. NIPS′15:Proceedings of the 28th International Conference on Neural Information Processing Systems:Volume 2. Cambridge,MA,United States:NIPS,2015:2692-700.

[25] SU J L,LU Y,PAN S F,et al. Roformer:enhanced transformer with rotary position embedding [Z/OL]. (2021-04-20)[2023-03-14]. https://arxiv.org/pdf/2104.09864.

[26] SU J L,MURTADHA A,PAN S F,et al. Global pointer:novel efficient span-based approach for named entity recognition [Z/OL]. (2022-08-05)[2023-03-14]. https://arxiv.org/pdf/2208.03054.

[27] YAO F,XIAO C J,WANG X Z,et al. LEVEN:a large-scale Chinese legal event detection dataset [C]//MURESAN S,NAKOV P,VILLAVICENCIO A. Findings of the Association for Computational Linguistics:ACL 2022. Stroudsburg,PA:Association for Computational Linguistics,2022:183-201.

[28] CUI Y M,CHE W X,LIU T,et al. Pre-training with whole word masking for Chinese BERT [J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2021,29:3504-3514.

[29] WANG X Z,HAN X,LIU Z Y,et al. Adversarial training for weakly supervised event detection [C]//BURSTEIN J,DORAN C,SOLORIO T. 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). Stroudsburg,PA:Association for Computational Linguistics,2019:998-1008.

[30] MU X F,XU A P. A character-level BiLSTM-CRF model with multi-representations for Chinese event detection [J]. IEEE Access,2019,7:146524-146532.

[31] 赵宇豪,陈艳平,黄瑞章,等. 基于跨度回归的中文事件触发词抽取[J]. 应用科学学报,2023,41(1):95-106.

[32] CAO H,LI J Y,SU F F,et al. OneEE:A one-stage framework for fast overlapping and nested event extraction [J/OL]. (2022-09-06)[2023-03-15]. https://arxiv.org/pdf/2209.02693.

[33] LIU J,YANG Y H,LV S Q,et al. Attention-based BiGRU-CNN for Chinese question classification [J]. Journal of Ambient Intelligence and Humanized Computing,2019,10:1-12.

[34] CAO S S,LU W,ZHOU J,et al. cw2vec:learning Chinese word embeddings with stroke *n*-gram information [C]//MCILRAITH S A,WEINBERGER K Q. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. Menlo Park:AAAI Press,2018,32(1):5053-5061.

Judicial Event Detection Method Based on Multi-head Pointer

ZHANG Xiaoli[1,2,3], HUANG Hui[1,2,3], HUANG Ruizhang[1,2,3], QIN Yongbin[1,2,3], CHEN Yanping[1,2,3**]

(1. Engineering Research Center of Text Computing and Cognitive Intelligence, Ministry of Education, Guizhou University, Guiyang, Guizhou, 550025, China; 2. State Key Laboratory of Public Big Data, Guizhou University, Guiyang, Guizhou, 550025, China; 3. School of Computer Science and Technology, Guizhou University, Guiyang, Guizhou, 550025, China)

Abstract: This article aims to solve the problem that the relationship between trigger words and context in Chinese judicial event detection is not enough to determine the case instance, and the case trigger words are similar in expression, and the identification and classification of multiple trigger words in the same case are fuzzy. A judicial event detection method based on multi-head pointer is proposed. Firstly, the method integrates context information and crime features as input, and utilizes a Bi-directional Long Short-Term Memory (BiLSTM) network to capture data dependencies and extract features in-depth. Then, the multi-head pointer network is used to model the dependency relationship between characters, and the trigger words in the sentence are effectively captured. Finally, trigger words are extracted by pointer annotation technology to realize effective detection of judicial events. Experiments on the public judicial dataset LEVEN validate the effectiveness of this method, in which the $F1$ index of micro-average and macro-average reaches 87.53% and 78.05%, which is better than the existing model. This method not only significantly improves the recognition accuracy of event trigger words, but also enhances the ability to grasp the context relationship of events in complex judicial texts.

Key words: judicial event detection; trigger; word context; crime feature; multi-head pointer

责任编辑:米慧芝