doi:10.6043/j.issn.0438-0479.202405016

基于 BERT 并融合法律事件信息的罪名预测方法

邱一卉*,喻瑶瑶
(厦门理工学院经济与管理学院,福建 厦门 361005)

摘要:[**目的**] 罪名预测是 AI&Law 领域的一个关键研究内容,对于提升司法领域的判决效率具有重要意义.由于法律文本的专业性和复杂性,传统罪名预测模型在提取法律文本特征时面临挑战,因此本文提出了一个基于预训练语言模型(BERT)并融合法律事件信息的罪名预测模型,通过利用法律事件信息为模型提供更多的法律案件特征,提升模型对案件的理解,从而提升罪名预测的表现.[**方法**] 首先训练了一个全局上层事件类型信息增强的法律事件检测模型,利用其对案情描述中的法律事件类型进行检测,并在此基础上构建法律事件类型序列.其次,利用双向长短期记忆模型(BiLSTM)对法律事件类型序列进行编码获取法律事件信息,并将法律事件信息与经过 BERT 编码后的案情描述的语义表示拼接融合,最后利用一层全连接层对罪名进行预测.[**结果**] 在公开的刑事案件数据集 CAIL2018-small 上的实验结果表明,相比于领域内的其他基线模型,本文提出的模型在各个关键指标上具备更好的性能,即在 Mac.F_1 上平均提升 3.12 个百分点,在 Mac.P 上平均提升 1.94 个百分点,在 Mac.R 上平均提升 3.53 个百分点.[**结论**] 验证了法律事件信息在增强模型对案件理解方面的有效性,从而提高罪名预测的准确性.

关键词:AI&Law;BERT 模型;罪名预测;法律事件信息
中图分类号:TP 391.1;TP 18 **文献标志码**:A **文章编号**:0438-0479(2025)04-0642-11

Charge prediction method based on BERT and integrated with legal event information

QIU Yihui*, YU Yaoyao
(School of Economics and Management, Xiamen University of Technology, Xiamen 361005, China)

Abstract: [**Objective**] As a key research topic in the field of AI&Law, the charge prediction plays a significant role in improving the efficiency of judicial decisions. Due to the professionalism and complexity of legal texts, traditional charge-prediction models face challenges in extracting features from legal texts. Therefore, in this paper, we propose a charge-prediction model based on BERT and integrating legal event information. By utilizing legal event information, the model is provided with numerous features of legal cases, thus enhancing the model's understanding of cases and thereby improving the performance of charge prediction. [**Methods**] First, a legal event detection model enhanced with global high-level event type information was trained to detect types of legal events in case descriptions and to construct a sequence of legal event types. Next, BiLSTM was used to encode the legal event type sequence to obtain legal event information. This information was then concatenated and fused with the semantic representation of the case description encoded by BERT. Finally, a fully connected layer was used to predict those charges. [**Results**] Experimental results based on the publicly available criminal case dataset CAIL2018-small demonstrate that, compared to other baseline models in the field, the proposed model achieves better performance on key metrics, with an average improvement of 3.12 percentage point in Mac.F_1, 1.94 percentage point in Mac.P, and 3.53 percentage point in Mac.R. [**Conclusion**] In this study, we demonstrate the effectiveness of legal

收稿日期:2024-05-17 **录用日期**:2024-08-13
基金项目:福建省社会科学基金(FJ2024B116);厦门市科技计划项目(3502Z20226036)
* **通信作者**:qiuyihui@xmut.edu.cn
引文格式:邱一卉,喻瑶瑶.基于 BERT 并融合法律事件信息的罪名预测方法[J].厦门大学学报(自然科学版),2025,64(4):642-652.
Citation: QIU Y H, YU Y Y. Charge prediction method based on BERT and integrated with legal event information[J]. J Xiamen Univ Nat Sci, 2025, 64(4): 642-652. (in Chinese)

event information in enhancing the model's understanding of cases, thereby improving the accuracy of charge prediction.

Keywords: AI&Law; BERT model; charge prediction; legal event information

1　预备知识

法律判决预测(legal judgment prediction, LJP)是 AI&Law 领域的关键研究方向之一[1],旨在利用机器学习或深度学习模型,根据案件的案情描述自动预测案件判决结果,主要研究内容包括罪名预测、法条预测和刑期预测. 罪名预测作为 LJP 中的一个关键子任务,对于提升司法系统的决策效率和优化其资源分配具有重要意义. Zhong 等[2]发布的大规模刑事案件法律数据集 CAIL2018(legal judgment prediction competition at Chinese AI and law challenge),为 LJP 的研究提供了宝贵资源,进一步推动了该领域的发展. 然而由于法律领域文本的专业性和复杂性,模型在提取其语义特征时面临挑战,导致罪名预测的准确性仍有提升空间.

近年来,深度神经网络的发展有效促进了 AI&Law 领域的发展,比如预训练语言模型(bidirectional encoder representations from transformers, BERT)[3],双向长短期记忆网络(bidirectional long short-erm memory network, BiLSTM)[4],层次注意力网络(hierarchical attention networks, HAN)[5]. 大部分研究者开始利用深度神经网络模型的自动特征提取能力,将其应用于罪名预测. 其中,Hu 等[6]根据法律先验知识预定义了 10 个罪名属性,提出了一个基于罪名属性注意力机制的罪名预测模型,该模型不仅能准确预测罪名,还能推断出与罪名紧密相关的属性. Luo 等[7]设计了一个基于注意力机制的神经网络模型,该模型同时对罪名预测和法条预测进行联合建模. Xu 等[8]为了区分易混淆罪名,利用图神经网络和注意力机制自动学习易混淆法条之间的差异. 还有学者[9-10]利用 LJP 中 3 个子任务之间的依赖关系来提升罪名预测的性能. 然而,尽管将法律先验知识融入到模型中可以显著提高罪名预测的表现,但是这一方法通常伴随着一定的标注成本[11].

近年来,法律事件信息受到了罪名预测领域内研究者的关注,他们将其融入到模型中以提升罪名预测的准确性. 其中 Feng 等[12]利用法律事件信息对案件进行判决,有效提高了法律判决预测的性能. 法律事件检测的目标是先从文本中识别事件触发词,并对这些触发词触发的事件类型进行分类,是提取法律事件信息中最关键的一步[13],但由于受限于法律事件标注的高成本,大部分研究者基于人工标注的法律数据集对其展开研究[14-15]. Li 等[14]针对盗窃案件构建了特定的事件抽取数据集,并提出了一个基于 BERT+CRF(conditional ramdom fields)的两阶段的法律事件抽取模型. 他们根据法律事件在案件中出现的顺序创建案件时间线,从而帮助法律从业者深入分析案件. 其中 CRF 作为一种统计建模方法[16],特别适用于序列数据标签的预测问题,它能考虑词汇标签之间的关系,从而更好地对序列标签做出预测. Yao 等[17]提出了一个大规模、法律事件类型全面的法律事件检测数据集 LEVEN(a large-scale Chinese legal eVENt detection dataset),有效促进了法律事件检测领域的发展. 基于 LEVEN 数据集,他们训练了一个基于 BERT+CRF 的法律事件检测模型,从而实现对法律事件类型的自动标注. 此外,他们通过在 BERT 模型的嵌入层中添加一层法律事件类型词嵌入为模型融入法律事件信息,并验证了法律事件信息能有效提升 LJP、相似案件匹配任务的表现. 基于 Yao 等[17]的研究,Yu 等[18]提出了一个基于法律事件类型注意力机制的 LJP 模型,他们利用一个法律事件类型注意力机制从案情描述中提取和法律事件类型相关的案情信息,并将其与案情描述语义表示融合后对案件做出判决. 因此,上述研究表明,案情描述中的法律事件这一细粒度的法律特征,能够增强模型对案件的理解,从而提升罪名预测任务的表现.

综上,罪名预测模型在深度神经网络的推动下取得了显著进步,但仍然存在法律文本特征提取不充分的问题. 因此,本文提出了一种基于 BERT 并融合法律事件信息的罪名预测方法. 首先,基于训练好的全局上层事件类型信息增强的法律事件检测模型,对案情描述中的法律事件进行检测,从而构建法律事件类型序列. 其次,利用 Bi-LSTM 编码法律事件类型序列获取法律事件信息,将其与经过 BERT 编码后的案情描述语义表示融合,并利用一层全连接层对案件涉及的罪名进行预测. 本文提出的模型通过利用 BERT 的深层次特征提取能力并结合法律事件信息,能够增强模型理解法律案件的能力,从而提升罪名预测的准确性.

2　问题定义

法律事件检测模型的问题定义为,给定一个法律

文本序列 $W=\{w_i\}$ 和该文本对应的标签序列为 $Y=\{y_i\}$，其中 $i\in\{1,2,\cdots,l\}$，w_i 代表 W 中的第 i 个词汇，y_i 表示 w_i 对应的标签，l 表示 W 的长度，且 $y_i\in N_t$，$N_t=\{B_1,B_2,\cdots,B_{2N_e+1}\}$，$N_e$ 表示事件类型的总数. 法律事件检测模型的目标是从已标注的数据中学习一个函数 φ，使得对于任意给定的法律文本 W，模型能够自动预测出其对应的标签序列 Y，即 $Y=\varphi(W,\theta)$，其中 $\theta=\{\theta_1,\theta_2,\cdots,\theta_{n_\beta}\}$ 表示模型中的参数集合，n_β 表示模型的参数总数.

罪名预测任务的问题定义为，给定一个案件的案情描述 $X=\{x_1,\cdots,x_i,\cdots,x_{fl}\}$，其中 x_i 代表案情描述中的第 i 个词汇，fl 表示案情描述的长度，其对应的罪名标签为 $y\in N_c$，$N_c=\{1,2,\cdots,n_c\}$ 表示罪名类别集合，n_c 表示所有罪名标签的个数. 罪名预测模型的目标是从已标注的数据中学习一个目标函数 F，使得 $y=F(X,\beta)$，其中 $\beta=\{\beta_1,\beta_2,\cdots,\beta_{n_\alpha}\}$ 代表模型中的参数集合，n_α 表示模型的所有参数量.

3 基于 BERT 并融合法律事件信息的罪名预测

图 1 展示了提出的罪名预测模型的整体框架，分为两个模块，法律事件检测模块和罪名预测模块. 法律事件检测模块基于 LEVEN 数据集训练了一个全局上层事件类型信息增强的法律事件检测模型，用于检测案情描述中的法律事件，并在此基础上构建法律事件类型序列. 罪名预测模块则利用获取到的法律事件类型序列和案情描述对案件的罪名进行预测，从而利用法律事件信息这一细粒度的案情特征提升模型对案件信息的理解，以准确对案件的罪名进行预测.

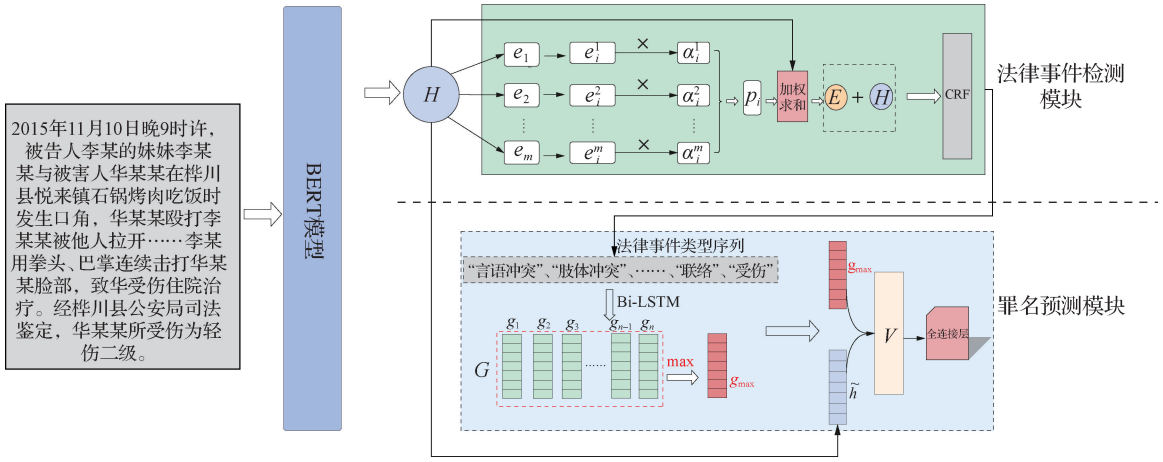

图 1 基于 BERT 并融合法律事件信息的罪名预测框架

Fig. 1 Charge prediction framework based on BERT and integrated with legal event information

3.1 法律事件检测模块

法律事件检测模块基于 LEVEN 数据集训练了一个全局上层事件类型信息增强的法律事件检测模型，该模型在序列标注的范式下对法律文本中的法律事件进行检测，采用 BIO(begin，inside，outside)标记机制对法律文本中的法律事件类型进行标注. 在 BIO 标注机制下，每个法律事件类型被拆分为“B-事件”和“I-事件”两个标签，因此，当 N 表示事件类型总数时，所有可能的分类标签总数为 $2N+1$，其中包括一个非事件标签，标记为“O”.

如图 2 所示，全局上层事件类型信息增强的法律事件检测模型分为 4 个组成部分：BERT 编码层、事件类型层次注意力网络(hierarchical attention neural network for event types，HANN-ET)[19]、全局上层事件类型信息增强和 CRF 输出层. 该模型以 BERT 和 CRF 作为基础架构，结合法律事件的层级结构[15]，利用 HANN-ET 从法律文本表示中生成全局上层事件类型信息，并利用其对法律文本表示进行增强，从而能够有效利用法律文本中不同事件的关系对法律事件进行检测，提升法律事件检测的准确性.

其中，法律事件的层级结构通常由上层事件类型及其对应的子事件类型构成，并形成一个清晰的事件分类体系. 图 3 是 LEVEN 数据集中部分事件类型的层级结构，其中“事故”、“危害结果”、“禁止性行为”、“司法相关”、“一般行为”和“自然灾害”属于上层事件类型，这些上层事件类型下都包含多个具体的子事件类型. 例如，“危害结果”这一上层事件类型下细分有“死亡”、“受伤”和“被困”等具体的子事件类型.

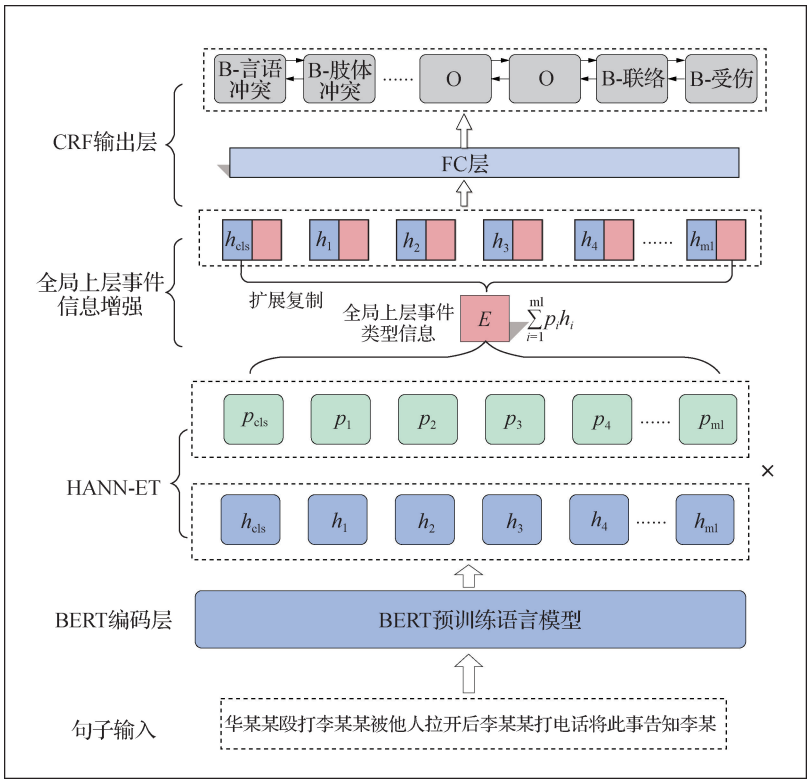

图 2　全局上层事件类型信息增强的法律事件检测模型

Fig. 2　Legal event detection model enhanced with global high-level event type information

图 3　LEVEN 中部分事件类型的层级结构

Fig. 3　Hierarchical structure of some event types in LEVEN

3.1.1 BERT 编码层

BERT 编码层基于预训练语言模型 BERT 对法律文本进行编码,从而获取法律文本的上下文语义表示.

对于预处理后的法律文本序列 $W'=\{w_i\}$ 及其对应的标签序列 $Y'=\{y_i\}$,其中 $i\in\{1,2,\cdots,n\}$,n 表示 W 的最大文本长度,w_i 代表 W 中的第 i 个词汇,y_i 表示 w_i 对应的标签. 在 BERT 编码层中,BERT 将利用内置的嵌入层和 Transformer 编码层对序列 W' 进行处理.

在 BERT 的嵌入层中,序列 W'_{word} 的嵌入表示将由 3 部分构成,分别是词嵌入 W'_{tok}、句子嵌入 W'_{seg} 以及位置嵌入 W'_{pos},BERT 的嵌入层通过对这 3 种嵌入表示进行求和,从而生成 W 序列的嵌入表示 W'_{word},如式(1)所示.

$$W'_{\text{word}}=W'_{\text{tok}}+W'_{\text{seg}}+W'_{\text{pos}}. \tag{1}$$

此后,BERT 中的多层 Transformer 编码层进一步对 W'_{word} 进行处理,从而得到序列 W'的上下文语义表示 H,如式(2)所示.

$$H=[h_1,\cdots,h_i,\cdots,h_n]=\text{Transformer}_{\text{mn}}(W'_{\text{word}}), \tag{2}$$

其中:$H\in\mathbb{R}^{n\times h}$,mn 是 BERT 中的多层 Transformer 编码器的数量,h_i 表示经过 BERT 编码后 w_i 的上下文语义向量,也称为隐藏嵌入,h 表示经过 BERT 编码后每个隐藏嵌入 h_i 的特征维度.

3.1.2 事件类型层次注意力网络

HANN-ET 通过利用法律事件的层级结构和两层注意力机制来生成全局上层事件类型信息,从而有效捕获法律文本中不同法律事件之间的关系.

首先,HANN-ET 根据 LEVEN 数据集中法律事件层级结构中的上层事件类型数定义了 m 个上层事件类型模块,且每个上层事件类型模块都有其对应的嵌入向量 $e_k\in\mathbb{R}^d$,其中 $k\in\{1,2,\cdots,m\}$,m 表示上层事件类型的数量,d 表示 e_k 的维度. e_k 在模型训练开始时随机初始化,并在训练过程中不断优化,从而能够捕捉到不同上层事件类型模块的语义表示.

其次,HANN-ET 采用两层注意力机制来捕获 W 中不同事件间的关系. 第一层注意力机制计算每个 h_i 与所有上层事件类型模块 e_k 之间的注意力分数,即 h_i 与上层事件类型模块之间的相关性. 此外,h_i 可能与多个上层事件类型相关,但它们对 h_i 的重要性可能不同,因此第二层注意力机制则被用来聚合 h_i 来自所有上层事件类型模块的注意力分数,从而为每个 h_i 赋予一个综合的权重. 因此,HANN-ET 通过两层注意力机制来捕获每个 h_i 与不同上层事件类型之间的关联,从而有效捕捉多个法律事件之间的关系. HANN-ET 的基本结构如图 4 所示.

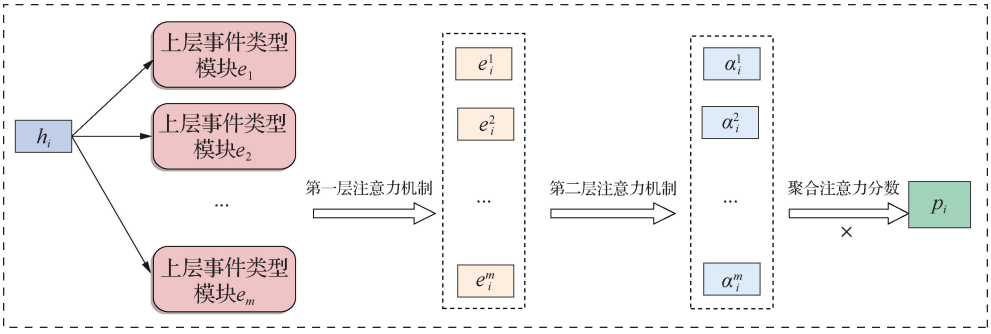

图 4 HANN-ET 的基本结构

Fig. 4 Basic structure of HANN-ET

在第一层注意力机制中,对于某个上层事件类型模块 e_k,HANN-ET 使用注意力机制来为每个 h_i 分配其对应的注意力分数 e_i^k,该注意力分数反映了每个 h_i 与上层事件类型模块 e_k 之间的相关性,其计算过程如式(3)和(4)所示.

$$h_i^k=\tanh(W_a[h_i;e_k]), \tag{3}$$

$$e_i^k=\frac{\exp(w_b^k h_i^k)}{\sum_{j=1}^{\text{ml}}\exp(w_b^k h_j^k)}, \tag{4}$$

其中:$[h_i;e_k]$ 是 h_i 与 e_k 拼接后的表示,e_i^k 是上层事件类型模块 k 对 h_i 的注意力分数,表示两者之间的相关性.

在第二层注意力机制中,为每个 h_i 对应的所有上层事件类型模块注意力分数 e_i^k 分配一个权重 α_i^k,以聚合 h_i 来自所有上层事件类型模块的注意力分数. 计算过程如式(5)~(7)所示.

$$q_i^k=\tanh(W_e e_i^k+b_e), \tag{5}$$

$$\alpha_i^k=\frac{\exp(u_e q_i^k)}{\sum_{j=1}^{m}\exp(u_e q_i^j)}, \tag{6}$$

$$p_i = \sum_{k=1}^{m} \alpha_i^k e_i^k. \tag{7}$$

因此,p_i 表示 h_i 从所有上层事件类型模块中聚合得到的全局注意力分数,表示每一个 h_i 和所有上层事件类型模块的相关性. 最后,将每个 $h_i \in \mathbb{R}^h$ 与其对应的全局注意力分数 p_i 结合,形成全局上层事件类型信息 $E \in \mathbb{R}^{1\times h}$,如式(8)所示.

$$E = \sum_{i=1}^{n} p_i h_i. \tag{8}$$

3.1.3　全局上层事件类型增强

为了利用序列中多个法律事件间的关系对事件进行检测,本部分将获取到的全局上层事件类型信息 E 对序列表示 H 进行增强,即采取两者沿特征维度拼接的方式,将 E 融入到 H 中的每一个 h_i 中,形成一个综合的序列表示 $Z \in \mathbb{R}^{n\times 2h}$,

在将两者进行拼接时,通过对 E 复制 n 次将全局上层事件类型信息扩展为 $E_{\text{expanded}} \in \mathbb{R}^{n\times h}$,从而确保每个 h_i 都能与全局上层事件类型信息 E 相融合. 其拼接如式(9)所示.

$$Z = [H; E_{\text{expanded}}]. \tag{9}$$

在此基础上,序列表示 $Z = [z_1, \cdots, z_i, \cdots, z_n]$ 被输入到全连接层中,以获取每个 z_i 属于所有分类标签 N_t 的概率分数矩阵 $\boldsymbol{P}$,称为发射矩阵,$\boldsymbol{P}$ 随后被输入到 CRF 层中,并与 CRF 中的转移矩阵 $\boldsymbol{T}$ 结合,用以计算并选择最佳的标签序列.

其中,发射矩阵 $\boldsymbol{P}$ 中每一个行向量的维度与分类标签的数量 N_t 相同,表示每一个词属于每个标签的概率分数,如式(10)所示.

$$\boldsymbol{P} = \begin{pmatrix} P_{11} & \cdots & P_{1N_t} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nN_t} \end{pmatrix}, \tag{10}$$

其中 P_{ij} 表示第 i 个词属于第 j 个标签的概率分数,$\boldsymbol{P}$ 的计算过程如式(11)所示.

$$\boldsymbol{P} = H\boldsymbol{W}_e + b_e, \tag{11}$$

其中,$\boldsymbol{P} \in \mathbb{R}^{n\times N_t}$,权重矩阵 $\boldsymbol{W}_e \in \mathbb{R}^{2h\times N_t}$ 和偏置项 b_e 是该线性连接层的参数.

3.1.4　CRF 输出层

在 CRF 输出层中,通过结合发射矩阵 $\boldsymbol{P}$ 和转移矩阵 $\boldsymbol{T}$,模型对序列 W' 所有可能标签序列的概率分数进行计算,并确定最佳的标签序列. 转移矩阵 $\boldsymbol{T}$ 如式(12)所示.

$$\boldsymbol{T} = \begin{pmatrix} T_{11} & \cdots & T_{1(N_t+2)} \\ \vdots & \ddots & \vdots \\ T_{(N_t+2)1} & \cdots & T_{(N_t+2)(N_t+2)} \end{pmatrix}, \tag{12}$$

其中 $\boldsymbol{T} \in \mathbb{R}^{(N_t+2,N_t+2)}$,$T_{ij}$ 表示在标签序列中标签 i 到标签 j 的转移分数.

在计算标签序列的概率分数时,CRF 首先对序列中每一个标签对应的发射分数和转移分数求和,以获得每一个标签对应的概率分数,并将序列中所有标签的分数求和,从而获得整个标签序列对应的概率分数. 真实标签序列 Y' 的概率分数 $S(\theta, Y')$ 的计算如式(13)所示.

$$S(\theta, Y') = T_{\text{START},y_1} + \sum_{i=2}^{n-1} (P_{i,y_i} + T_{y_{i-1},y_i}) + T_{y_n,\text{END}}, \tag{13}$$

其中 θ 为模型中的所有参数,包括 CRF 中的参数.

在测试阶段,CRF 输出层利用动态规划算法在所有可能的标签序列 $Y(\theta')$ 中找到概率分数最高的序列,即最佳标签序列,如式(14)所示.

$$Y^* = \text{argmax}(S(\theta', Y(\theta'))), \tag{14}$$

其中,θ' 表示法律事件检测模型训练更新后的参数.

因此,对经过预处理后的案件案情描述 $\widetilde{X} = \{x_0, \cdots, x_i, \cdots, x_{\text{ml}}\}$,其中 ml 表示案情描述的最大文本长度,$x_0$ 表示 CLS 标记,利用法律事件检测模块自动获取 $\widetilde{X}$ 对应的最佳标签序列 O,如式(15)所示.

$$O = \text{argmax}(S(\theta', Y(\theta'))), \tag{15}$$

其中,$O = \{o_0, \cdots, o_i, \cdots, o_{\text{ml}}\}$,即对应的 BIO 法律事件类型标签. 在此基础上,不再区分每一个法律事件类型的开始或内部,将得到的 BIO 事件类型标签转化为对应的法律事件类型标签,并保留每一个法律事件类型出现的顺序,对其进行去重处理后得到相应的法律事件类型序列为 $E = [e_1, \cdots, e_i, \cdots, e_n]$,其中 n 表示法律事件类型的个数.

3.2　罪名预测模块

罪名预测模块分 3 个部分,分别是案情描述编码层、法律事件类型序列编码层和融合及输出层.

3.2.1　案情描述编码层

为了获取案情描述的语义表示,本文利用法律事件检测模块中微调后的 BERT 对案情描述进行编码,BERT 通过在法律事件检测任务上进行微调,可以更准确地捕捉法律文本的语义表示.

在利用 BERT 对 $\widetilde{X}$ 进行编码时,$\widetilde{X}$ 将经过 BERT 的两个部分处理,一个是 BERT 的嵌入层,另一个是 BERT 中的多层 Transformer 编码层,如式(16)和(17)所示.

$$\widetilde{X}_{\text{word}} = \widetilde{X}_{\text{tok}} + \widetilde{X}_{\text{seg}} + \widetilde{X}_{\text{pos}}, \tag{16}$$

$$H = [h_0, \cdots, h_i, \cdots, h_{\text{ml}}] =$$

$$\text{Transformer}_{\text{mn}}(\widetilde{X}_{\text{word}}), \tag{17}$$

其中,mn 是 BERT 模型中多层 Transformer 编码器的数量,$h_i \in \mathbb{R}^k$ 表示 $\widetilde{X}$ 中 x_i 的上下文向量表示,即隐藏状态,h_0 是 $\widetilde{X}$ 中[CLS]标记对应的隐藏状态,即 h_{CLS},它表示 $\widetilde{X}$ 的语义表示.

为了获取 $\widetilde{X}$ 适应当前罪名预测任务的语义表示,通过一层全连接层和激活函数 tanh 对 h_{CLS} 进行处理,如式(18)所示.

$$\widetilde{h} = \tanh(\boldsymbol{W} \cdot h_{\text{CLS}} + b), \tag{18}$$

其中:$\widetilde{h} \in \mathbb{R}^k$ 表示经过全连接层处理后得到的案情描述整体语义表示,$\boldsymbol{W}$ 和 b 分别是全连接层的权重矩阵和偏置项,其中 $\boldsymbol{W} \in \mathbb{R}^{k \times k}$.

3.2.2 法律事件类型序列编码层

法律事件类型序列编码层利用 BiLSTM 对法律事件类型序列 E 进行编码,BiLSTM 通过结合每个事件类型在序列中出现的顺序,有效捕获法律事件类型序列中每个法律事件类型的上下文表示,从而实现对法律事件类型序列的全面表示.

在将法律事件类型序列 E 输入到 BiLSTM 之前,该编码层首先初始化一个法律事件类型词嵌入矩阵 $\boldsymbol{D} \in \mathbb{R}^{d \times N_e}$,将每个法律事件类型转换为对应的词嵌入表示.因此,通过该法律事件类型词嵌入矩阵 $\boldsymbol{D}$,E 被转换为了对应的词嵌入表示 $\boldsymbol{U} = [u_1, \cdots, u_i, \cdots, u_n]$,其中 $U \in \mathbb{R}^{d \times n}$,如式(19)所示.

$$[u_1, \cdots, u_i, \cdots, u_n] = E\boldsymbol{D}^{\mathrm{T}}, \tag{19}$$

其中,$u_i \in \mathbb{R}^d$ 是第 i 个事件类型的向量表示,n 是序列中事件类型的数量.在此基础上将 $\boldsymbol{U}$ 输入到 BiLSTM 中对其进行编码,通过 BiLSTM 前向和后向的编码能力,获取每个法律事件类型向量 u_i 包含其上下文信息的表示 $g_i \in \mathbb{R}^h$,如式(20)所示.

$$g_i = \text{BiLSTM}(u_i), i \in [1, n], \tag{20}$$

其中,h 表示经过 BiLSTM 编码后 g_i 的维度,因此获取到的法律事件类型序列语义表示为 $G = [g_1, \cdots, g_i, \cdots, g_n] \in \mathbb{R}^{h \times n}$.

获取到法律事件类型序列的语义表示后,模型采用最大池化操作从 G 中进一步捕获法律事件信息.最大池化操作在 G 的 h 维度上选择每个事件类型表示 g_i 的最大值作为该维度的代表值,从而保留法律事件类型序列表示中最关键的法律事件信息.因此,通过这种方式,模型得到了一个固定维度的向量 $\boldsymbol{g}_{\text{event}} \in \mathbb{R}^h$,表示法律事件类型序列表示 G 中提取的法律事件信息,如式(21)所示.

$$\boldsymbol{g}_{\text{event}} = \begin{bmatrix} \max(G_{1,1}, G_{1,2}, \cdots, G_{1,n}) \\ \max(G_{2,1}, G_{2,2}, \cdots, G_{2,n}) \\ \vdots \\ \max(G_{h,1}, G_{h,2}, \cdots, G_{h,n}) \end{bmatrix}, \tag{21}$$

其中,每个元素 $\max(G_{i,1}, G_{i,2}, \cdots, G_{i,n})$ 是 G 第 i 行中所有元素的最大值,即第 i 个特征维度在所有 n 个事件类型向量中的最大值.

3.2.3 融合及输出

为了将获取到的法律事件信息 $\boldsymbol{g}_{\text{event}}$ 与案情描述整体语义表示 $\widetilde{h}$ 有效融合,融合层将 $\boldsymbol{g}_{\text{event}}$ 与案情描述的语义表示 $\widetilde{h}$ 在特征维度上进行拼接.因此,通过这种方式,生成的最终案件信息表示 V 不仅涵盖了关键的法律事件信息,还整合了整个案情描述的语义信息,从而确保了信息的全面性和丰富性,如式(22)所示.

$$V = [\widetilde{h}; \boldsymbol{g}_{\text{event}}], \tag{22}$$

其中 $V \in \mathbb{R}^{h+k}$,表示最终的案情描述语义表示.在输出层,使用全连接层和 softmax 激活函数对罪名标签预测,从而得到罪名的预测概率分布 $\hat{y}_c$,如式(23)所示.

$$\hat{y}_c = \text{softmax}(\boldsymbol{W}_c V + b_c), \tag{23}$$

其中,$\boldsymbol{W}_c$ 和 b_c 分别是全连接层的权重矩阵和偏置项,其中 $\boldsymbol{W}_c \in \mathbb{R}^{(k+h) \times n_c}$,$n_c$ 表示所有的罪名标签总数.

3.3 训练及损失函数

在法律事件检测模型的训练过程中,正确的标签序列 Y' 应该比其他任何可能的标签序列 $Y(\theta)$ 都有更高的概率,因此模型的目标是最大化真实标签序列 Y' 的条件概率 $P(Y' \mid W')$,法律事件检测模型的损失函数如式(24)所示.

$$\text{NLL}(\theta, Y') = -S(\theta, Y') + \log \sum_{\widetilde{Y} \in Y(\theta)} \exp(S(\theta, \widetilde{Y})). \tag{24}$$

在训练过程中,模型采用小批量梯度下降法来更新模型参数 θ,即通过最小化一个批次样本的平均损失,使得真实标签序列 Y' 相对于其他所有可能的标签序列 $Y(\theta)$ 是模型输出的概率最大,一个批次样本的平均损失计算如式(25)所示.

$$\text{Loss} = \frac{1}{M} \sum_{i=1}^{M} \text{NLL}_i(\theta, Y'), \tag{25}$$

其中 M 是每个批次中的样本数.

在罪名预测模块的训练过程中,模型通过计算预测结果和真实标签之间的交叉熵损失来评估预测的性能,并采用小批量梯度下降法更新模型参数,如式(26)所示.

$$\text{Loss} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{n_c} y_{ij} \log(\hat{y}_{ij}), \tag{26}$$

其中,M 是每个批次中样本的数量,n_c 是类别总数. y_{ij} 表示第 i 个样本属于类别 j 的真实标签. $\hat{y}_{ij}$ 是模型预测第 i 个样本属于类别 j 的概率. 因此,通过最小化该损失函数,模型能够在训练过程中逐步调整参数,准确对罪名做出预测.

4　实　验

本文实验分为两部分,第一部分先利用 LEVEN 数据集验证提出的全局上层事件类型信息增强的法律事件检测模型的有效性,第二部分在此基础上验证提出的基于 BERT 并融合法律事件信息的罪名预测模型的有效性.

4.1　法律事件检测模型实验

4.1.1　LEVEN 数据集描述

本部分利用 LEVEN 数据集对提出的法律事件检测模型进行验证. 表 1 展示了 LEVEN 数据集的分布.

表 1　LEVEN 数据集分布

Tab. 1　Distribution of LEVEN datasets

数据类型	案件数	句子数	事件标注数
训练数据	5 301	41 238	98 410
验证数据	1 230	9 788	22 885
测试数据	1 585	12 590	29 682
总计	8 116	63 616	150 977

此外,根据表 2,LEVEN 数据集中存在 6 种上层事件类型,但是有两类上层事件类型“自然灾害”、“事故”占比较少,因此本文将这两者合并为一个新的上层事件类型,用以表示那些非预期且非人为直接引起的事件,因此模型中 HANN-ET 层的上层事件类型模块的数量设置为 5.

表 2　LEVEN 数据集中的事件类型分布

Tab. 2　Overview of LEVEN dataset event types

上层事件类型	子事件类型数	标注数	占比	子类型示例
一般行为	40	68 616	45.4%	销售、雇用
禁止性行为	40	43 021	28.5%	盗窃
司法相关	13	29 709	19.7%	逮捕、自首
危害结果	7	6 832	4.5%	死亡、受伤
事故	4	2 742	1.8%	交通事故
自然灾害	4	57	0.03%	干旱、洪水
总计	108	150 977	1.0	

4.1.2　对比实验分析

在法律事件检测模型的对比实验中,本文采用了和 Yao 等[17]一致的对比模型和评价指标,并选择在验证集上表现最优的模型版本在测试集上进行测试. 此外,实验中使用不同随机种子对模型进行 5 轮独立训练和测试,并对这 5 轮实验的结果计算平均值和标准差.

通过对表 3 中的实验结果进行分析,可以发现,提出的法律事件检测模型在各个指标上都超越了对比的基线模型,从而验证了提出模型的有效性,特别与 BERT+CRF 相比,它在 Mic. F_1 上提高了 1.26 个百分点,在 Mac. F_1 上提高了 1.06 个百分点,这说明 Ours 能够有效捕获法律文本序列中多个法律事件间的关系,从而显著提升事件检测的性能.

表 3　法律事件检测模型对比实验结果

Tab. 3　Comparison results of legal event detection models　%

模型	Mic. P	Mic. R	Mic. F_1	Mac. P	Mac. R	Mac. F_1
DMCNN	85.88±0.70	79.70±0.59	82.67±0.08	80.55±0.49	73.31±3.88	75.03±0.40
BiLSTM	83.09±0.89	85.16±0.95	84.11±0.24	78.70±0.92	76.67±2.23	76.65±1.42
BiLSTM+CRF	84.74±0.55	83.33±0.49	84.03±0.05	78.5±1.31	72.60±1.11	74.49±0.77
BERT	84.19±0.39	84.31±0.34	84.25±0.18	79.61±0.91	76.76±1.79	77.33±1.30
BERT+CRF	83.82±0.48	84.56±0.52	84.19±0.09	79.77±1.10	77.65±2.20	77.84±1.58
Ours	85.38±0.31	85.53±0.38	85.45±0.08	80.73±0.53	78.48±0.65	78.90±0.56

4.2　罪名预测模型实验

4.2.1　罪名预测数据集描述

本文使用 CAIL2018 数据集中的 CAIL2018-small[2]来评估提出的罪名预测模型的有效性. 由于本文聚焦于单罪名预测任务,本文采用 Xu 等[8]的数据预处理方式,过滤掉了 CAIL2018-small 数据集中涉及多个罪名的数据. 为了获取案情描述中的法律事件

类型标注,本文使用训练好的法律事件检测模型对数据集进行事件类型的标注. 表 4 展示了数据处理后数据集的相关统计信息,包括数据集的大小、罪名的数量以及法律事件类型的数量等.

表 4 CAIL2018-small 数据集分布

Tab. 4 Distribution of CAIL 2018-small dataset

统计信息	数量
训练集案件数	101 619
测试集案件数	26 749
罪名数	119
法律事件类型数	108

4.2.2 实验设置

和 Xu 等[8]保持一致,在实验设置方面,本文使用清华大学开源的 THULAC 分词器[20]对案情描述进行分词,在训练期间,学习率设置为 10^{-5},权重衰减设置为 10^{-5},批处理大小设置为 18,使用 Adam 优化器[21],训练周期为 16,模型参数的具体设置见表 5.

表 5 罪名预测模型参数设置

Tab. 5 Hyperparameter settings for the charge prediction model

参数	设置
BERT 模型的最大文本长度	512
批处理大小	18
优化器	Adam
学习率	10^{-5}
权重衰减	10^{-5}
隐藏层维度	768
事件类型的维度	768
迭代次数	16

由于 CAIL2018-small 数据集中存在着不平衡的罪名分布,因此本文使用宏观指标来评估模型的性能,其中宏观指标先为每个类别单独计算精确率(Precision,P)、召回率(Recall,R)和 F_1 值,再取这些值的平均,因此宏观精确率(Mac. P)、宏观召回率(Mac. R)和宏观 F_1 值(Mac. F_1)的计算如式(27)~(29)所示.

$$\text{Mac.}\,P=\frac{\sum_i^C P_i}{C}, \tag{27}$$

$$\text{Mac.}\,R=\frac{\sum_i^C R_i}{C}, \tag{28}$$

$$\text{Mac.}\,F_1=\frac{\sum_i^C F_{1i}}{C}, \tag{29}$$

其中 C 表示多分类问题中的类别数.

4.2.3 对比方法

在对比方法中,本节将提出的模型与一系列领域内的基线模型进行了比较,这些模型包括:

1) BERT[3]:一个在大规模的语料上训练的预训练语言模型. 利用 BERT 对案情描述编码获取其整体语义表示后,在多任务框架下对罪名进行预测.

2) Few-Shot[6]:一种基于罪名属性注意力机制的预测模型,通过人工标注 10 个相关的罪名属性,并在此基础上设计一个罪名属性注意力机制来区分相关的罪名.

3) FLA[7]:一种基于法条的罪名预测模型,FLA 利用案情描述与相关法条来预测案件的罪名.

4) LADAN[8]:为区分易混淆罪名,利用图神经网络自动学习易混淆法条之间的差异,并设计一种新的注意力机制,充分利用法条之间的差异对罪名进行预测.

5) TopJudge[9]:利用 LJP 中 3 个子任务间的拓扑依赖关系对罪名进行预测,并且利用有向无环图来建模子任务间的拓扑依赖关系,是一种多任务联合建模的方法.

6) Bert+event[17]:该模型在 BERT 的嵌入层中添加一层法律事件类型词嵌入来融入案情描述中的法律事件信息,并在多任务框架下对罪名展开预测.

7) EventAtt[18]:该模型采用 BERT 对案情描述和法律事件类型进行编码,利用一个法律事件类型注意力机制从案情描述表示中获取法律事件信息,并将其与案情描述表示融合后预测案件罪名. 为了和本文提出的模型进行对比,在该模型中,本文没有利用 BERT 对法律事件类型进行编码,而是直接采用法律事件类型的嵌入表示.

为了节省计算资源并确保实验结果的可比性与一致性,本文使用 Xu 等[8]公开的实验结果进行分析. 鉴于上述对比的大部分基准模型均在多任务框架中预测罪名,因此本文也在相同的多任务框架中对提出的模型进行验证.

为确保实验的公平性,实验中基于 BERT 的对比模型均使用法律事件检测微调后的 BERT 模型. 为了评估模型性能,实验选择在验证集上表现最优的模型版本在测试集上进行测试,以获取对模型性能的评估.

5 实验结果分析

5.1 主实验结果分析

根据表 6,和 LADAN 相比,模型在 Mac. F_1 上提

升了 2.36 个百分点,在 Mac.R 上提升了 3.38 个百分点,在 Mac.P 提升了 1.68 个百分点,这说明了利用法律事件信息和利用易混淆法条差异的方式相比,法律事件信息对罪名预测更加有效. 和 FewShot 相比,模型在 Mac.F_1 上提升了 3.55 个百分点,这充分地说明了相关的事件信息比相关的罪名属性更加有助于提高模型的泛化能力. 和 Yao 等[17]提出的 Bert+event 模型中融入法律事件信息的方式相比,本文提出的模型在 Mac.R 上提升了 0.6 个百分点,Mac.F_1 上提升了 0.4 个百分点. 和 EventAtt 相比,Ours 模型在 Mac.R 上提升了 0.7 个百分点,在 Mac.F_1 上提升了 0.4 个百分点. 此外,和 BERT 相比,Ours 模型在各个指标上都有所提升. 因此,上述实验结果说明了法律事件信息在罪名预测中的重要作用,即能够给模型提供更多的案情特征,进而增强模型对案情的理解能力,提高罪名预测的准确性.

表 6　主实验结果对比

Tab.6　Main experiment results comparison　%

模型	Mac.P	Mac.R	Mac.F_1
FLA	79.25	77.61	76.94
FewShot	80.84	82.01	81.55
TopJudge	83.60	78.42	79.05
LADAN	83.42	82.52	82.74
Bert+event	85.40	85.30	84.70
BERT	84.80	85.30	84.40
EventAtt	**85.50**	85.20	84.70
Ours	85.10	**85.90**	**85.10**

5.2　消融实验及分析

为验证本文提出罪名预测模型各个组成部分的有效性,本节设计了 4 个消融实验.

1) w/o BiLSTM:为评估 BiLSTM 编码法律事件类型序列的有效性,该方法使用法律事件类型序列的词嵌入表示.

2) w/o max:为评估利用最大池化从编码后的法律事件类型序列表示中获取法律事件信息的有效性,该方式采用注意力机制的方式.

3) w/o fact:为了评估案情描述和法律事件信息融合的作用,该方法仅使用法律事件信息对罪名做出预测,没有融入案情描述的语义表示.

4) w/o event:为了评估法律事件信息对于罪名预测的重要作用,该方法仅仅使用案情描述对罪名做出预测.

根据表 7 中的实验结果,当没有用 BiLSTM 对法律事件类型序列进行编码时,模型在 Mac.P 上下降了 0.8 个百分点,在 Mac.R 上下降了 1.2 个百分点,在 Mac.F_1 上下降了 1.2 个百分点,这说明了 BiLSTM 能够捕获法律事件类型序列的顺序性,从而获取法律事件类型序列的整体语义表示. 当没有采用最大池化的方式提取法律事件信息时,模型的 Mac.F_1 下降了 0.6 个百分点,这证明最大池化方式提取法律事件信息的有效性. 然而,当仅仅利用法律事件信息时,模型表现大幅度下降,这说明仅利用法律事件信息不能充分的对案情进行概括,法律事件信息仅代表案情描述中的关键案情特征. 此外,当不往模型中融入法律事件信息时,模型的性能有所下降,这说明法律事件信息的融入,能够给模型提供更多的法律特征,从而帮助模型更好的理解案情,有效地提升罪名预测的表现.

表 7　消融实验对比结果

Tab.7　Ablation study results　%

模型	Mac.P	Mac.R	Mac.F_1
Ours	**85.10**	**85.90**	**85.10**
w/o BiLSTM	84.30	84.70	83.90
w/o max	84.80	85.40	84.50
w/o fact	55.20	52.20	51.40
w/o event	84.80	85.30	84.40

6　结　论

本文提出了一种基于 BERT 并融合法律事件信息的罪名预测方法. 通过结合 BERT 的文本特征提取能力与法律事件信息,该模型显著提升了罪名预测的准确性. 模型首先训练并利用了一个全局事件信息增强的法律事件检测模型从案情描述中构建法律事件类型序列,并使用 Bi-LSTM 对法律事件类型序列进行编码获取其法律事件信息. 其次,将法律事件信息与基于 BERT 编码后的案情描述语义表示融合后对罪名预测. 因此,模型通过利用案情描述中细粒度的法律事件特征,从而增强模型对案件的理解,提升罪名预测的表现. 本文在一个公开发布的刑事案件数据集 CAIL2018-small 上进行了大量的对比实验和分析实验,验证了提出模型的有效性.

参考文献:

[1] ZHONG H X, XIAO C J, TU C C, et al. How does NLP benefit legal system: a summary of legal artificial intelligence[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 5218-5230.

[2] ZHONG H X, XIAO C J, GUO Z P, et al. Overview of CAIL2018: legal judgment prediction competition[EB/OL]. [2024-05-01]. https://arxiv.org/abs/1810.05851v1.

[3] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language under-standing[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: NAACL-HLT, 2019: 4171-4186.

[4] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5/6): 602-610.

[5] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016: 1480-1489.

[6] HU Z, LI X, TU C, et al. Few-shot charge prediction with discriminative legal attributes[C]// Proceedings of the 27th International Conference on Computational Linguistics. [S. l.]: COLING, 2018: 487-498.

[7] LUO B F, FENG Y S, XU J B, et al. Learning to predict charges for criminal cases with legal basis[C]// Proceedings of the 2017 Conference on Empirical Methods in NaturalLanguage Processing. Copenhagen: ACL, 2017: 2727-2736.

[8] XU N, WANG P H, CHEN L, et al. Distinguish confusing law articles for legal judgment prediction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 3086-3095.

[9] ZHONG H X, GUO Z P, TU C C, et al. Legal judgment prediction via topological learning[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018: 3540-3549.

[10] YANG W M, JIA W J, ZHOU X J, et al. Legal judgment prediction via multi-perspective bi-feedback network[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao: IJCAIO, 2019: 4085-4091.

[11] LYU Y G, WANG Z H, REN Z C, et al. Improving legal judgment prediction through reinforced criminal element extraction[J]. Information Processing & Management, 2022, 59(1): 102780.

[12] FENG Y, LI C Y, NG V. Legal judgment prediction via event extraction with constraints[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: ACL, 2022: 648-664.

[13] WANG X Z, WANG Z Q, HAN X, et al. MAVEN: a massive general domain event detection dataset[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2020: 1652-1671.

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

[15] SHEN S R, QI G L, LI Z, et al. Hierarchical Chinese legal event extraction via pedal attention mechanism[C]// Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: ICCL, 2020: 100-113.

[16] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning. [S. l.]ICML, 2001: 282-289.

[17] YAO F, XIAO C, WANG X, et al. LEVEN: a large-scale Chinese legal event detection dataset[C]// Findings of the Association for Computational Linguistics. New York: ACL 2022, 2022: 183-201.

[18] YU Y Y, QIU Y H. Enhancing legal judgment prediction with attentional networks utilizing legal event types[C]// Neural Information Processing. Singapore: Springer Nature Singapore, 2024: 393-404.

[19] JIN Y L, YE J J, SHEN L Q, et al. Hierarchical attention neural network for event types to improve event detection[J]. Sensors, 2022, 22(11): 4202.

[20] SUN M S, CHEN X X, ZHANG K X, et al. THULAC: an efficient lexical analyzer for Chinese[EB/OL]. [2024-05-01]. https://github.com/thunlp/THULAC.

[21] Kingma D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2024-05-01]. https://arxiv.org/abs/1412.6980.

(责任编辑:汪 军)