软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(12):5724−5736 [doi: 10.13328/j.cnki.jos.006903]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

全局和局部信息融合的案情关键要素识别*

毛星亮[1], 陈晓红[1], 宁 肯[2], 李芳芳[2], 张师超[2]

[1](湖南工商大学 大数据与互联网创新研究院, 湖南 长沙 410205)

[2](中南大学 计算机学院, 湖南 长沙 410038)

通信作者: 陈晓红, E-mail: csu_cxh@163.com; 张师超, E-mail: zhangsc@csu.edu.cn

摘　要: 司法人工智能中主要挑战性问题之一是案情关键要素识别, 现有方法仅将案情要素作为一个命名实体识别任务, 导致识别出的多数信息是无关的. 另外, 也缺乏对文本的全局信息和词汇局部信息的有效利用, 导致要素边界识别的效果不佳. 针对这些问题, 提出一种融合全局和局部信息的关键案情要素识别方法. 所提方法首先利用BERT 模型作为司法文本的输入共享层以提取文本特征. 然后, 在共享层之上建立司法案情要素识别、司法文本分类 (全局信息)、司法中文分词 (局部信息) 这 3 个子任务进行联合学习模型. 最后, 在两个公开数据集上测试所提方法的效果, 结果表明: 所提方法 *F*1 值均超过了现有的先进方法, 提高了要素实体分类的准确率并减少了识别边界错误问题.

关键词: 信息融合; 多任务联合学习; 关键案情; 要素识别

中图法分类号: TP18

中文引用格式: 毛星亮, 陈晓红, 宁肯, 李芳芳, 张师超. 全局和局部信息融合的案情关键要素识别. 软件学报, 2023, 34(12): 5724–5736. http://www.jos.org.cn/1000-9825/6903.htm

英文引用格式: Mao XL, Chen XH, Ning K, Li FF, Zhang SC. Global and Local Information Integration for Recognizing Key Case Elements. Ruan Jian Xue Bao/Journal of Software, 2023, 34(12): 5724–5736 (in Chinese). http://www.jos.org.cn/1000-9825/6903.htm

Global and Local Information Integration for Recognizing Key Case Elements

MAO Xing-Liang[1], CHEN Xiao-Hong[1], NING Ken[2], LI Fang-Fang[2], ZHANG Shi-Chao[2]

[1](Institute of Big Data and Internet Innovation, Hunan University of Technology and Business, Changsha 410205, China)

[2](School of Computer Science and Engineering, Central South University, Changsha 410038, China)

Abstract: One of the main challenges in judicial artificial intelligence is the identification of key case elements. The existing methods only take the identification of case elements as an identification task of named entities, and thus, the recognized information is mostly irrelevant. In addition, due to the lack of effective use of global and local information in texts, the effect of element boundary recognition is poor. To solve these problems, this study proposes a recognition method of key case elements by integrating global and local information. Specifically, the BERT model is used as the input-sharing layer of judicial texts to extract text features. Then, three sub-task networks of judicial case element recognition, judicial text classification (global information), and judicial Chinese word segmentation (local information) are established on the sharing layer for joint learning. Finally, the effectiveness of this method is tested on two public data sets. The results show that the *F*1 value of the proposed method exceeds the existing optimal method, improves the classification accuracy of element entities, and reduces boundary recognition errors.

Key words: information fusion; multi-task joint learning; key facts of the case; element recognition

近年来, 随着司法大数据不断公开, 司法领域各类文本数据获得爆发性增长[1]. 这些文本数据主要是各级法院公开的裁判文书数据, 而裁判文书是记载人民法院审理过程及结果的承载文书[2]. 如何有效利用海量的司法文本

* 基金项目: 国家重点研发计划 (2020YFC0832700); 国家自然科学基金 (62172449, 62006251); 湖南省自然科学基金 (2022JJ30211, 2021JJ30870, 2021JJ40783); 长沙市自然科学基金 (kq2202300); 长沙市科技计划 (kq2107004)
收稿时间: 2022-09-30; 修改时间: 2022-11-17; 采用时间: 2023-01-03; jos 在线出版时间: 2023-07-12
CNKI 网络首发时间: 2023-07-13

辅助司法工作, 成为司法人工智能的一个关键性问题. 关键案情要素识别是司法人工智能中不可或缺的重要一环, 其结果一方面可以给司法工作者带来直观的结构化信息, 简化其查阅文书资料的过程; 另一方面可以给非司法工作者带来关键的案情信息, 辅助其快速了解司法要素. 此外, 其结果还是司法文书结构化的组成部分, 可用于构建司法知识图谱, 或为司法问答等其他下游任务提供数据支持, 并直接影响着下游司法任务的表现[3].

司法案情要素是指文书中所涉及的、具有司法领域特色且与案情密切相关的名词或短语, 如犯罪嫌疑人姓名、作案时间、涉案金额等. 相对于传统命名实体识别 (named entity recognition, NER) 方法提取所有的实体[4,5], 司法案情要素识别更关注与案情相关的命名实体. 例如, 司法领域对于人名的定义可能会分为嫌疑人、受害人、证人、相关人等, 其定义带有司法领域性质. 对于地名、时间的定义往往关注的是与案情密切相关的要素, 而与案情本身不相关的一些地名、时间, 如判决时间、开庭时间等, 则一般不是所关注的案情要素实体, 识别时不应将其提取出来. 总的来说, 目前司法领域的要素识别方法中, 存在以下两个主要问题: 1) 缺乏对司法文本全局信息的考虑. 司法案情要素的出现常常与上下文的表述, 乃至文本全局信息有关, 若忽视上下文和全局信息, 将导致要素识别结果的分类错误. 2) 缺乏对司法词汇局部信息的考虑. 中文文本不具有天然分词的标志, 且中文词语种类繁多, 难以收集详尽完整的词表, 故现有模型多是基于汉字字符进行处理的. 若模型缺乏词汇局部信息, 将导致要素识别结果的边界错误.

针对上述两个问题, 本文提出了一种全局和局部信息融合的关键案情要素识别. 首先, 根据字符的上下文语义, 利用预训练语言模型 BERT 动态生成语义向量, 作为文本编码的输入, 即多任务中的共享层; 然后, 在共享层之上建立司法案情要素识别、司法文本分类 (全局信息)、司法中文分词 (局部信息) 这 3 个子任务进行联合学习, 通过司法文本分类子任务学习司法文本类别, 通过司法中文分词子任务学习分词边界; 其次, 将文本全局信息以及词汇局部信息融入要素识别序列中, 达到增强司法案情要素识别子任务学习效果的目的; 最后, 利用条件随机场计算并输出最优标注序列, 得到最终案情要素识别结果.

本文在 2021 年中国法律智能技术评测 (challenge of AI in law, CAIL) 信息抽取赛道数据集 (简称 CAIL2021), 以及涉毒案件数据集 (简称 Drug) 上进行了实验, 结果表明: 本文提出的方法 $F1$ 值均超过了现有的先进方法, 提高了司法要素实体分类的准确率并减少了识别边界错误问题, 可以有效地对司法文书中的关键案情要素信息进行识别.

1 相关工作

对于案情要素识别任务, 大部分学者将其作为命名实体识别任务来进行, 也有少部分学者针对司法领域的特点提出了一些算法. 早期命名实体识别主要使用规则或词典的方式, 此种方法适用性与扩展性差, 但也没有完全被摒弃, 在不少场景仍与其他方法一起发挥作用. 此后, 机器学习、深度学习的方法逐渐被运用于 NER 当中, 深度学习较机器学习方法而言省去了人工特征, 能够自动提取有效的潜在数据特征[6]. 将深度学习网络用于 NER 时, 学者们常在网络之后加上一层条件随机场 (conditional random field, CRF) 用于规范标签输出的合理性, 如长短时记忆 (long short term memory, LSTM) 网络加上 CRF 是如今典型的 NER 任务搭配方法[7].

基于命名实体识别任务进行案情要素识别方面: 由于深度学习方法通常需要大量有监督语料来达到比较好的学习效果, 而在一些专有领域, 有监督的语料数据常常稀缺且难以标注. 预训练模型的出现可以使得模型对于有监督语料的需求大大减少. 2018 年, Devin 等人[8]提出由多层 Transformer[9]编码器组成的预训练语言模型 BERT, 其优异的特征抽取能力刷新了包括 NER 任务的各大自然语言处理任务成绩. 此后, 不少学者把目光聚焦于 BERT 的改进增强上, 如 Sun 等人[10]提出的 ERNIE 模型, 通过外部知识库在 BERT 掩码语言模型任务中掩码实体或短语, 以增强 BERT 预训练阶段的语义学习效果; Cui 等人[11]提出的 MacBERT 模型, 该模型没有使用 BERT 中的下一句进行预测, 而是针对中文词语加入了全词掩码并使用近义词来替换掩码, 以此缩小预训练和微调时的训练差异; Clark 等人[12]提出的 ELECTRA 模型, 该模型采用了产生器 (generator) 和鉴别器 (discriminator) 相结合的预训练方式, 同时将 BERT 中的掩码语言模型任务改为了检测是否是被替换字符.

此外, 由于中文文本没有天然分词标志使得其与英文文本处理不一样, 以往的 NER 任务常先将中文文本进行分词处理[13,14]. 多项研究[15−17]表明, NER 采用基于字符进行处理要优于基于词进行处理, 但仅基于字符的方式模

型容易忽略词语的语义信息. 不少学者针对这一突出问题进行了长足的研究. Zhang 等人[18]在 LSTM 模型的结构上提出栅格 LSTM (Lattice-LSTM) 结构, 该结构通过匹配外部词典在两个 LSTM 细胞间以有向无环图的方式加入词汇信息; Gui 等人[19]提出 LGN 结构, 该结构以图的方式将每一个字符作为节点, 匹配到的词汇信息构成边, 通过图结构实现局部信息的聚合和全局信息融入; Li 等人[20]提出的 FLAT 结构基于 Lattice-LSTM 的思想, 设计了一种位置向量来融合 Lattice 结构, 其中对于每一个字符和词汇都构建两个头位置向量和尾位置向量; Ma 等人[21]利用软词典的思想, 对当前中文字符根据外部词典依次获取词头、词中、词尾、单字词对应所有的词汇集合, 然后再进行编码表示.

基于司法领域特点进行案情要素识别方面: 李春楠等人[22]提出了一个基于盗窃案件的司法命名实体识别数据集, 定义了与案件相关的人名、时间、地点、物品等 10 余种实体, 并使用 BERT 加上有序 LSTM (ordered neurons LSTM, ONLSTM)[23]的结构对数据集中的司法实体进行识别, 取得了良好的效果. Leitner 等人[24]采集了大量的德文法律相关文书, 定义了 19 种细粒度法律相关的命名实体类型并进行了标注, 同时将经典的 NER 模型 CRF 及 BiLSTM-CRF 对数据集进行了测评. Luz 等人[25]提出了一个巴西法律文书命名实体识别数据集, 并使用 LSTM-CRF 作为数据集的基线模型进行了实验测试. Quaresma 等人[26]使用支持向量机用于法律文书的分类, 随后使用语言信息和机器学习相结合的方法来识别其中重要的实体, 该模型在欧盟法律文书数据集上获得了较好的结果.

综上, 这些工作要么将关键案件要素识别作为命名实体识别任务来进行, 要么重点关注了司法领域实体的特点, 通过预先定义相关专业实体来进行识别, 而没有深入考虑案件要素识别需要关联的文本全局和局部信息. 因此, 本文在前人研究的基础之上, 从关键案件要素识别任务本身出发, 综合考虑裁判文书中的文本全局信息和词汇局部信息, 来进一步提升关键案情要素识别的精度.

2 全局和局部信息融合的关键案情要素识别

2.1 问题描述

关键案情要素识别任务是司法人工智能中一个兼具重要性和挑战性的工作, 其任务目标可定义为: 给定法律裁判文书集合 $D=\{(X_i,Y_i)|1\leqslant i\leqslant N\}$, 其中 X_i 为集合中任一给定法律裁判文书序列, Y_i 为对应的标注序列, N 为集合中法律裁判文书数. 详细地, 对于任一法律裁判文书序列 X_i, 其具体表示形式定义为由 m 个字组成的序列 $X_i=\{x_1,x_2,\ldots,x_m\}$, 而其对应的标注序列 Y_i, 其具体表示形式定义为由 m 个字对应标注类别组成的序列 $Y_i=\{y_1,y_2,\ldots,y_m|y_i\in\{B,I,O\}\}$. 关键要素识别任务的目标则是学习映射函数 $f:X\rightarrow Y=\{(y_1,y_2,\ldots,y_m)|y_i\in\{B,I,O\}\}$, 其中任一关键要素 $E_k=\{x_i,\ldots,x_j|0\leqslant i<j\leqslant m\}$, 其对应标注 $Y_k=\{y_i,\ldots,y_j|0\leqslant i<j\leqslant m\}$ 中, $y_i=B,y_{i+1},\ldots,y_j=I$.

区别地, 为了更清晰地描述关键案情要素识别任务与通用领域命名实体识别之间的区别, 本文查阅对比了大量法律裁判文书后, 选取了部分实际样本来进一步展开对于问题的描述, 采用不同的下划线标识各类实体, ###代表文字, 具体如表 1 所示.

表 1 法律裁判文书样例

序号	样例
1	2017年11月28日15时许, 被告人 郑某某 潜入被害人 习某某 位于灞桥区**区**号的家中, 盗窃 两部手机 (其中一部价值1 100元)、 金戒指 价值1 208元、 银项链 、 现金6 000元, 共计8 308元
2	由于此路段的路灯需要更换, 该公司将所有的路灯灯杆放倒, 将灯杆和灯头分离后, 摆放在绿化带内待处置
3	在该超市二楼将一袋墨鱼上的防盗扣卸下后, 将价值110元的 干墨鱼 夹带回家食用
	被告人: ### 被害人: ### 被盗物品: ### 物品价值: ###

样例 1 为盗窃案文书抽取的描述文本, 其中“郑某某”为“犯罪嫌疑人”, “习某某”为“受害人”, “1 100 元”和“1 208 元”为“物品价值”, “两部手机”“金戒指”“银项链”“现金 6 000 元”为“被盗物品”. 可以看出, 裁判文书的关键案情要

素相比通用领域命名实体描述更为细致复杂, 分类难度更高. 进一步地, 样例 2 中, “路灯灯杆”虽是物品类的实体, 但不是目标任务需要的实体, 很容易将其识别并提取出来而造成分类错误. 样例 3 中, 是“干墨鱼夹”还是“干墨鱼”被带走? 从上下文看, 应是“干墨鱼”被带走, “夹带”是一个动词. 而识别算法极有可能将“干墨鱼夹”识别为一个实体, 这便是边界错误问题.

2.2 模型框架

为了增强司法命名实体识别模型的健壮性, 本文通过查阅法律文书样例, 注意到司法实体分类常受司法文本类别影响, 而司法文本类别信息与司法实体类别信息紧密关联. 同时, 为了缓解广泛存在于命名实体识别任务中的实体边界错误问题, 本文设计了基于司法命名实体识别、司法文本分类、司法中文分词的多任务联合学习网络. 本文提出的融合全局和局部信息的案情要素识别方法, 其结构如图 1 所示. 为了结合各个任务的特点, 本文采用多任务共享方式中硬参数共享的方式建立 BERT 共享层, 作为多任务架构中各个任务网络共享的部分, 以用于学习司法文本的特征表示. 同时, 在目标任务上, 本文提出了序列全局和局部信息融合的方法以增强关键要素识别的能力. 进一步地, 为了解决问题描述中样例 2、样例 3 中所出现的司法文本分类错误和实体边界错误的问题, 本文提出了司法文本分类网络和司法中文分词网络, 分别利用全局信息和局部信息两个不同粒度的信息进行学习. 其中司法文本分类网络, 通过共享层输出的文本全局信息学习司法文本的类别 (全局信息), 而司法中文分词网络主要学习司法文本的中文分词边界 (局部信息). 下面将分别介绍各个子网络以及将子网络组合成多任务学习的整体过程.

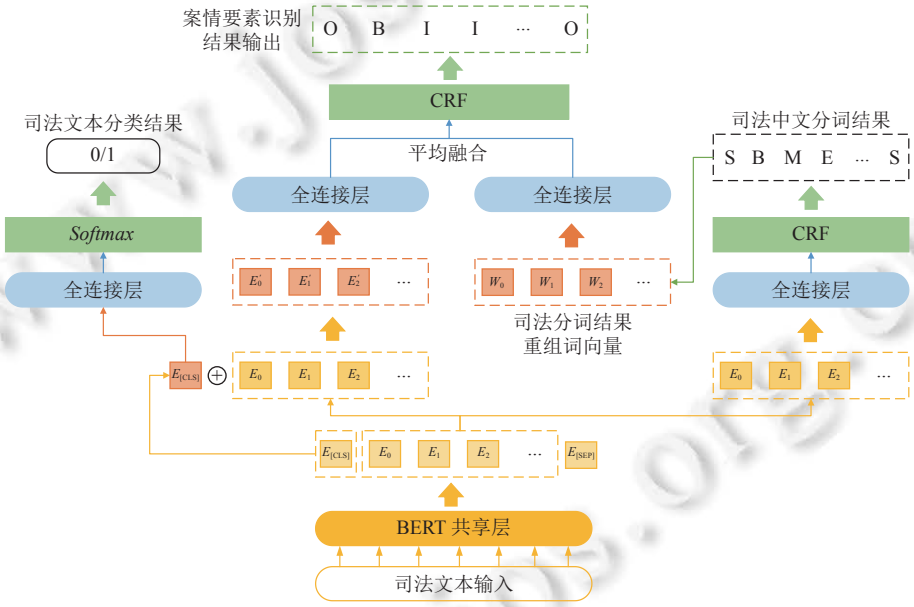


图 1 融合全局和局部信息的案情要素识别多任务学习框架

2.3 BERT 共享层

本文采用基于字符的方式将司法文本输入 BERT 模型. 在输入模型时, 根据 BERT 文本向量化的原理, 首先通过查词表将输入的字符转为一个独有的 ID, 根据 ID 将每个字符对应转化为字符向量, 再根据段标记和位置查到段向量和位置向量, 3 个向量相加后即输入模型, 如图 2 所示.

对于该任务, 每次输入一个司法文书句子 (样本), 在 BERT 的段向量中只需标注为同一段, 即在输入序列中段向量采用同一个, 同时句子的头尾分别加入一个开头标记 [CLS] 和结尾标记 [SEP]. 假设其输入序列为 X={[CLS], c_1, c_2, c_3,…, c_n, [SEP]}, 其中 C_i 表示一个字符, 则向量化后的序列为 E={$e_{[CLS]}$, e_1, e_2, e_3,…, e_n, $e_{[SEP]}$}, 其输出 H 可

表示为公式 (1), 其中 θ 为 BERT 的相关参数.

$$H = \text{BERT}(E, \theta) \tag{1}$$

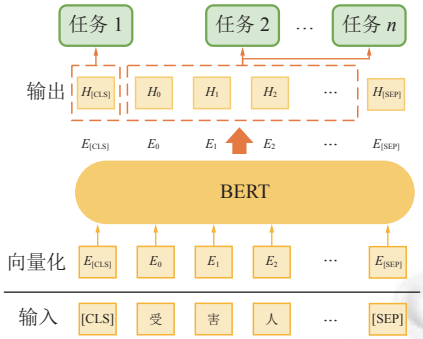


图 2 BERT 共享层

2.4 司法文本分类子网络

输入的司法文本经过底层的 BERT 共享层, 并进行司法文本特征抽取之后, 将得到一串输出序列的向量表示 H. 接下来, 司法文本分类子网络将根据不同的数据集构建不同的司法类别标签, 并通过该网络学习文本中蕴含的司法全局信息. 其中根据 BERT 的实现方式, BERT 在每条输入文本的第 1 个位置都添加了一个 [CLS] 标签, 意义为“分类”, 在原始的任务中此标签的作用是: 在下一句预测任务中判别输入的句子是否为第 1 个句子的下一句. 而在本文任务中, 输入是单个句子, 此时经共享层输出的 [CLS] 标签表示了整个司法文本句子的全局信息, 因此可以直接作为司法文本分类任务的输入, 其后再接入一个独立的全连接层转化维度, 最后经过 *Softmax* 层即得到每个分类的概率, 计算的具体过程描述如下.

对于 BERT 共享层中所得到的向量表示 $H = \{h_{[CLS]}, h_2, \ldots, h_n, h_{[CLS]}\}$, 司法文本分类任务选取 $h_{[CLS]}$ 作为该任务的输入特征, 并通过公式 (2) 和公式 (3) 计算各个类别的概率.

$$Y_{[CLS]} = w_{\text{dim}} \cdot h_{[CLS]} + b \tag{2}$$

其中, w_{dim} 为全连接层中的权重矩阵, $\text{dim} = c$, b 为偏置值, $Y_{[CLS]} = [y_1, y_2, \ldots, y_c]$, c 为分类类别数, y_i $(i = 0, 1, \ldots, c)$ 为当前输入司法文本属于 i 类别的初步概率.

$$p(y_i | i = 0, 1, \ldots, c-1, c) = Softmax(y_i) = \frac{\exp(y_i)}{\sum_{k=0}^{c} \exp(y_k)} \tag{3}$$

本文使用交叉熵损失函数 (cross entropy, CE) 计算法文本分类任务的损失值, 计算方式如公式 (4):

$$Loss_{\text{tc}} = -\sum_{i=0}^{c} (y_c \log[p(y_i)] + (1 - y_i) \log[1 - p(y_i)]) \tag{4}$$

其中, C 表示司法文本分类的类别集合, y_c 表示当前类别是否为文本实际的类别, 是则 y_i=1, 反之 y_i=0, $p(y_i)$ 表示模型预测出当前类别的概率, 由 *Softmax* 计算所得.

2.5 司法中文分词子网络

基于司法文本分类子网络学习全局信息之后, 接下来基于司法中文分词子网络学习局部信息. 对于司法中文分词任务, 本文将其转化为序列标注任务进行. 在数据预处理阶段, 使用外部分词工具对原始司法文本进行分词作为标签. 经过前述 BERT 共享层后得到的输出序列, 将直接作为司法中文分词子网络的输入. 司法中文分词子网络由独立的全连接层和 CRF 层组成: 首先, 通过全连接层将序列维度与标签数相对应; 然后, 将对应结果作为 CRF 层的输入即发射矩阵; 最后, 通过该子网络的 CRF 层预测司法分词标注序列. 例如, 长度为 n 的输入序列 X 预测为各种标签构成的序列时, 其分数是所有时间步分数的总和, 如公式 (5):

$$s(X,Y)=\sum_{i=0}^{n}E_{x_i,y_i}\sum_{i=0}^{n}T_{y_i,y_{i+1}} \tag{5}$$

其中, X 表示输入序列, Y 表示观测序列, E 表示 CRF 发射矩阵, 此处为上层网络的输出结果, T 表示 CRF 转移矩阵, E_{x_i,y_i} 表示序列中第 i 个字符预测为某类标签的概率, $T_{y_i,y_{i+1}}$ 表示从输入序列当前第 i 个节点标签预测为某类标签时, 下一个节点标签预测为某类标签的概率.

对于给定输入序列 X, 计算其标签序列概率的方式如下所示:

$$P(Y|X)=\frac{\exp(s(X,Y))}{\sum_{\tilde{Y}\in Z_X}\exp\left(s\left(X,\tilde{Y}\right)\right)} \tag{6}$$

其中, $\tilde{Y}$ 为任意可能的标签序列, Z_X 为输入序列 X 的所有可能的标签序列的集合. 本文采用维特比算法求解输入序列 X 的最优全局标签序列.

司法中文分词子网络的训练目标, 即损失函数定义为: 当前 CRF 的转移状态矩阵在正确路径的分数与所有路径分数之和的负对数似然, 如公式 (7):

$$loss_{\text{cws}}=-\log(p(Z_{\text{true}}|X))=-s(X,Z_{\text{true}})+\log\left(\sum_{z\in Z_x}e^{s(X,z)}\right) \tag{7}$$

其中, X 表示输入的司法文本序列, Z_{true} 表示真正的司法分词标签序列, $S(X, Z_{\text{true}})$ 表示真正司法分词标签序列的得分, $s(X, z)$ 为某种预测序列 Z 的得分, Z_x 表示输入序列 X 输出的所有可能的预测序列.

2.6 司法案情要素识别子网络

司法案情要素识别子网络主要包括 3 个部分: 融入 CLS 司法文本全局信息、融入司法中文分词结果、司法案情要素识别子网络.

(1) 融入 CLS 司法文本全局信息

上文已提到, 根据 BERT 处理司法文本的方式, 每条输入文本的第 1 个位置都添加了一个 [CLS] 标签. 经过 BERT 编码及文本表示学习后输出的 [CLS] 标签向量, 可以直接为司法案情要素识别提供文本全局信息. 本文将 [CLS] 标签向量分别与每个司法文本中的字符输出向量相加, 得到一个融合全局信息的序列, 即假设输出向量序列为 $H=\{h_{[\text{CLS}]}, h_1, h_2, h_3,\ldots, h_n, h_{[\text{SEP}]}\}$, 每个字符向量加入 [CLS] 向量后得到的结果为 H', 该结果将作为司法案情要素识别子网络的输入, 其中 $H'=\{h_{[\text{CLS}]}, h_1+h_{[\text{CLS}]}, h_2+h_{[\text{CLS}]}, h_3+h_{[\text{CLS}]},\ldots, h_n+h_{[\text{CLS}]}, h_{[\text{SEP}]}\}$.

(2) 融入司法中文分词结果

司法案情要素识别子网络中, 司法词语信息不仅可以提供实体边界的局部信息, 还可以为司法要素实体分类带来重要参考. 本文从减小损失词汇信息的角度出发考虑, 当司法分词任务完成后, 进一步将词汇信息利用起来, 将分词结果融入司法案情要素识别任务中. 融合方式如下: 将分词结果中属于一个词的字符根据其向量进行重组, 形成“词向量”, 如公式 (8):

$$W_p=\frac{1}{j+1-i}\sum_{c=i}^{j}h'_c \tag{8}$$

其中, p 代表当前位置, i 代表当前组合中 B 的位置, j 代表当前组合中 E 的位置.

由此将得到一个与输出序列长度相同的“伪词向量序列”, 将其设为 $W=\{w_{[\text{CLS}]}, w_1, w_2, w_3,\ldots, w_n, w_{[\text{SEP}]}\}$. 接下来进行以下操作: 首先将输出序列 H'与 W 分别输入全连接层, 然后将两个全连接层的输出序列再相加求和并求算数平均, 最后将得到的算法平均作为司法案情要素识别子网络 CRF 层的输入.

(3) 司法案情要素识别子网络

将上述 (1)、(2) 两点进行组合后, 得到最终的司法案情要素识别子网络. 训练时将计算出标签路径的得分. 在计算出所有可能的预测路径得分后, 根据得分结果计算损失值, 再进行反向传播等操作. 参考司法中文分词子网络中序列得分的计算式 (5), 以及损失函数的计算公式 (7), 得出司法案情要素识别子网络的损失函数, 如公式 (9):

$$loss_{\text{cei}} = -\log(p(t_{\text{true}}|X)) = -s(X,t_{\text{true}}) + \log\left(\sum_{t\in T_x} e^{s(X,t)}\right) \tag{9}$$

其中, X 表示输入的文本字符序列, t_{true} 表示真正的实体标签序列, $S(X, t_{\text{true}})$ 表示真正实体标签序列的得分, $s(X, t)$ 为某种预测序列 t 的得分, T_x 表示输入序列 X 输出的所有可能的预测序列.

2.7 损失函数

多任务学习网络的训练目标即最小化总体损失 $Loss_{\text{total}}$, 综合公式 (4)、公式 (7) 和公式 (9) 中各个子网络的损失函数, 可得最终总的损失函数, 如公式 (10):

$$Loss_{\text{total}} = \lambda_1 loss_{\text{cei}} + \lambda_2 loss_{\text{tc}} + \lambda_3 loss_{\text{cws}} \tag{10}$$

其中, λ_1、λ_2、λ_3 分别为上述 3 个子网络损失函数的权重系数, 取值范围在 [0.5, 2], 通过调整权重大小可调整 不同任务的权重, 此 3 个参数将作为超参数加入训练过程中, 经过实验后最终设置各任务权重为 $\lambda_1=\lambda_2=\lambda_3=1$, 权重参数设置相关实验将于消融实验部分介绍.

3 实验与结果分析

3.1 数据集

本文选取了两个开源司法文书信息抽取数据集 (如表 2), 各个数据集的情况如下.

(1) 2021 年中国法律智能技术评测 (CAIL) 信息抽取赛道数据集, 简称 CAIL2021. 此数据集由 2021 年中国法律智能技术评测赛事举办方开源 (http://cail.cipsc.org.cn/task9.html?raceID=7), 其数据主要来自于网络公开的法律裁判文书, 以“盗窃罪”刑事案件作为研究主体, 并根据相关司法解释定义了 10 个业务相关的要素实体分类. 数据集中每条数据样本为文书中的截取描述片段, 且包含任意数目的案情要素实体标注. 源数据中包含了一定量的嵌套实体数据, 本文选取了其中仅含平滑实体的数据组成平滑实体数据集.

(2) 涉毒案件数据集, 简称 Drug. 此数据集由大连理工大学信息检索研究室开源 (https://github.com/DUTIR-LegalIntelligence/JointExtraction4Legal), 其数据同样来源于网络公开的法律裁判文书, 以“吸毒贩毒罪”刑事案件即涉毒案件为研究主体, 预定义了 5 类业务相关要素实体, 其中源数据中包含了案情要素实体标注和关系抽取任务的标注, 本文仅选取其中的实体标注作为研究和测评.

表 2 实验数据集

数据集	指标	训练集	验证集	测试集
CAIL2021	样本数	4 197	525	525
	字符数	268 004	34 158	32 848
	要素实体数	20 470	2 598	2 501
Drug	样本数	1 417	176	178
	字符数	239 541	29 189	30 786
	要素实体数	15 509	1 879	1 933

3.2 数据预处理

由于每个数据集都没有官方公布的划分标准, 本文对其源数据按 8:1:1 比例划分了训练集、验证集和测试集, 之后采用 BIO (B-begin, I-inside, O-outside) 三位标注方式对数据进行标注以供模型学习, 其中, “B-X”表示此元素所在的片段属于 X 类型并且此元素在此片段的开头; “I-X”表示此元素所在的片段属于 X 类型并且此元素在此片段的中间位置; “O”表示此元素不属于任何类型. 此外, 由于司法文本分类任务、司法中文分词任务在数据集中不存在相应的有监督标签, 故本文对这两个任务的标签进行了处理: 对于司法文本分类任务, 从刑事文书的角度, 根据该条文本对犯罪事实等的陈述, 标注其司法文本类别, 同时制定了一套标注规则用于自动标注; 对于司法中文分

词任务, 本文使用了开源的 LTP 分词工具 (https://github.com/HIT-SCIR/ltp), 加上外部司法领域词典对原始司法文书数据进行了分词操作. 本文在数据预处理中的分类标注规则具体如表 3 所示.

表 3 司法文本分类标注规则

分类标签	说明	标注规则
0	非犯罪事实	① 某条数据不含有任何实体 ② 对于审判结果、执法行为等的文本描述, 一般不是犯罪事实描述, 本文通过设定关键词和相关描述匹配类似数据(如“宣判结果如下”“被捕获”“缉拿归案”) ③ 除①、②外, 对于本文所使用的数据, 如对于CAIL2021数据集, 当某条数据中仅出现有嫌疑人、受害人、组织机构实体时, 如对于Drug数据集, 当某条数据中仅出现有涉毒人员实体时, 亦标记为0
1	犯罪事实	① 有关犯罪的各种情况的描述, 包括过程、结果、带来的危害等, 同样, 本文通过设定关键词和相关描述匹配类似数据(如“盗窃得xxx”“使用xxx工具作案”) ② 不属于0中的情况时, 则标记为1

3.3 实验设置及评价指标

本文 BERT 模型选用谷歌开源的中文模型 BERT-base-Chinese, 其基本结构为 12 层 Transformer 编码器, 隐藏层维度为 768, 自注意力头数量为 12, 总共 11 亿参数量. 根据 BERT 原文对于微调阶段调整超参的建议, 并根据本文进行的调参试验, 当训练轮数 (epoch) 设置为 10 并训练选取最优的验证集结果, 批量 (batch size) 设置为 16, 最大输入序列长度 (max length) 设置为 512, 学习率 (learning rate) 设置为 0.000 05, CRF 学习率 (CRF learning rate) 和全连接层的学习率 (linear learning rate) 设置为 0.001, 优化器 (optimizer) 选用 Adam, 权重衰减 (weight decay) 为 0.01, 多任务权重比均为 1 时, 效果最好.

对于实验结果的评价指标, 本文采用机器学习中常用的精确率、召回率和 $F1$ 值进行评测, 且采用严格指标的方式, 即对于一个要素实体, 当且仅当其开始位置、结束位置以及类别均正确时才判为正确. 评测指标的具体定义如下:

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{13}$$

其中, TP、FP、FN 分别表示预测正确的正例数量、预测错误的正例和预测错误的负例, 在 NER 任务中, 正例即为实体, 负例为其他非实体文本内容.

3.4 基线模型

为了验证本文方法的可行性和有效性, 将本文方法与现有先进的 9 个命名实体识别方法进行了对比, 这 9 个基线模型分别如下.

(1) Lattice-LSTM[18]: 通过在 LSTM 结构上加入有向无环图, 引入外部词汇信息辅助中文 NER, 在当时中文通用领域 NER 上取得当时最优的效果, 可称为中文 NER 词汇增强的“开山之作”.

(2) LGN[19]: LGN 是基于 Lattice-LSTM 改进的, 与 Lattice-LSTM 不同的是, LGN 采用图神经网络结构, 以两个字符为点外部词汇信息为边的方式构建图网络, 同时加入一个全局节点整合全局信息.

(3) FLAT[20]: 基于 Lattice-LSTM 的思想设计了一种位置向量来融合 Lattice 结构, 其中对于每一个字符和词汇都构建两个头位置向量和尾位置向量, 更有利于定位实体, 同时引入了外部词向量有利于增强模型对词汇的理解. FLAT+BERT 即上述方法基于 BERT 输出向量作为输入向量的情况.

(4) SoftLexicon[21]: 此方法利用软词典 (soft-lexicon) 的思想, 对当前中文字符匹配外部词典, 依次获取该字符词头、词中、词尾、单字词对应所有词汇的集合, 然后再将这些词向量进行编码表示, 增强了模型的表达能力. SoftLexicon(LSTM)+BERT 即基于 LSTM 结构使用 BERT 输出向量作为其输入向量的情况.

(5) BOCNER[22]: 针对中文司法领域命名实体识别提出的模型, 其在 BERT 的基础上加入 ONLSTM 模块, 对于司法领域的上下文特征提取有很大的帮助, 有利于挖掘更深层次的信息, 提升了部分长实体的准确度.

(6) BERT-xs, BERT-ms (https://github.com/lichunnan/OpenCLaP): 分别为清华大学人工智能研究院利用 663 万刑事文书和 2 654 万篇民事文书, 在 BERT 结构下进行预训练得到的预训练模型.

(7) ERNIE[10]: 其通过外部知识库在 BERT 掩码语言模型任务中, 掩码匹配到实体或短语, 以增强预训练阶段的语义学习的效果.

(8) MacBERT[11]: 去除 BERT 的下一句预测, 对中文加入全词掩码, 同时提出使用近义词替换的方式代替掩码, 以缩小预训练和微调时的训练差异.

(9) ELECTRA[12]: 改变了 BERT 预训练的方式, 使用了产生器 (generator) 和鉴别器 (discriminator) 相结合, 将 BERT 中掩码的方式, 改为由鉴别器检测当前是否是被产生器替换掉的字符. legal-ELECTRA 是哈工大讯飞实验室 (https://github.com/ymcui/Chinese-ELECTRA) 将 ELECTRA 在司法语料上进行进一步预训练得到的.

3.5 实验结果

3.5.1 对比实验及结果分析

为了对比目前中文命名实体识别方法在司法案情要素识别任务上的效果, 本文方法与基线模型在 CAIL2021 数据集和 Drug 数据集上进行了比较, 实验结果如表 4 所示, 其中 Ours 表示本文模型.

表 4 两个数据集上不同方法的对比实验结果

序号	方法	CAIL2021数据集			Drug数据集		
		精确率	召回率	*F*1值	精确率	召回率	*F*1值
1	Lattice-LSTM	0.8819	0.8836	0.8828	0.8720	0.8702	0.8711
2	LGN	0.8543	0.8629	0.8586	0.8737	0.8836	0.8786
3	SoftLexicon(LSTM)	0.9010	0.9024	0.9017	0.8772	0.8831	0.8801
4	SoftLexicon(LSTM)+BERT	0.9150	0.9216	0.9183	0.8874	0.8888	0.8881
5	FLAT	0.8860	0.9008	0.8933	0.8671	0.8743	0.8706
6	FLAT+BERT	0.8828	0.9036	0.8931	0.8901	0.8883	0.8892
7	BOCNER	0.9142	0.9248	0.9195	0.8916	0.8934	0.8925
8	BERT	0.9011	0.9108	0.9059	0.8730	0.8929	0.8829
9	BERT-xs	0.9041	0.9120	0.9080	0.8837	0.8888	0.8863
10	BERT-ms	0.8856	0.8980	0.8918	0.8760	0.8883	0.8821
11	ERNIE	0.9114	0.9176	0.9145	0.8836	0.8996	0.8916
12	MacBERT	0.9098	0.9236	0.9167	0.8868	0.8960	0.8914
13	ELECTRA	0.9098	0.9192	0.9145	0.8677	0.8888	0.8781
14	legal-ELECTRA	0.9139	0.9296	0.9217	0.8880	0.8898	0.8889
15	Ours	**0.9204**	**0.9340**	**0.9272**	**0.8929**	**0.9053**	**0.8990**

根据表 4 中的实验结果可以得到以下结论.

(1) Lattice-LSTM 是目前中文 NER 领域, 使用外部词典增强来增强 LSTM 的经典基线模型.

(2) LGN 是基于 Lattice-LSTM 的思想改进的, 其采用图神经网络结果加入词汇信息, 然而在 CAIL2021 的效果却不如 Lattice-LSTM, 可能的原因是图神经网络对于司法文书的编码能力不及基于序列处理的 LSTM, 也侧面说明 Lattice-LSTM 是一个强劲的模型.

(3) SoftLexicon(LSTM) 使用软词典的思想加入外部词汇信息, 表现优于 Lattice-LSTM 和 LGN, 但不及预训练模型 BERT. SoftLexicon(LSTM)+BERT 使用了 BERT 的编码向量作为其模型输入, 结果优于 BERT.

(4) FLAT 同样是在 Lattice-LSTM 基础上, 加入外部词向量以及头尾向量形成展平的结构, 加入 BERT 向量后也较 BERT 有了提升. 以上两点说明 SoftLexicon 和 FLAT 进行词汇辅助增强的方式优于 LGN 和 Lattice-LSTM, 而加上 BERT 取得了更好的结果, 说明其词汇增强方式在司法领域同样适用于 BERT.

(5) 不同于上述方法, BOCNER 是针对司法领域提出的模型, 在两个数据集上较上述模型中均取得了最好的效果, 说明其在 BERT 的基础上加入 ONLSTM 模块, 对于司法领域实体识别的层级上下文特征提取, 有很大的帮助.

为了对比不同预训练模型在司法领域数据产生的效果, 本文将目前不同的 BERT 模型用于上述两个数据集中, 实验结果显示:

(1) BERT-xs 略优于利用通用语料预训练的 BERT, 因使用的两个数据集几乎都来源于刑事文书, 而基于民事文书的 BERT-ms 表现不佳, 其原因可能是语料相关性较刑事文书低的缘故. 同时 BERT-xs 和 BERT-ms 皆未在通用语料上进行训练, 这可能是其表现并无特别突出的一个原因.

(2) ERNIE 性能表现优于 BERT 模型, 说明对预训练模型的预训练阶段做一定合理的优化, 可以进一步提升模型效果, 对下游任务提升也有帮助.

(3) MacBERT 采用了近义词替换的掩码方式重新做预训练, 取得了更优的效果, 再次说明了对预训练任务的改进可以提升模型效果.

(4) ELECTRA 在预训练阶段采用不同的预训练方式, 自身结果优于 BERT. 而将 ELECTRA 在司法语料上进行进一步预训练, 得到的 legal-ELECTRA 较 ELECTRA 取得了很大的提升, 说明对预训练模型进行合理有针对性的领域预训练, 可以提升领域下游任务的结果.

(5) 本文方法虽然没有进行进一步的司法领域预训练, 但本文方法是端到端的实现方式, 更为简捷. 本文方法在表 4 中的各项指标, 均超过了对比方法中的最好指标. 说明针对领域问题考虑领域特征是十分重要的, 司法领域的文本全局信息给司法案情要素实体分类提供了很大的辅助, 而司法分词信息可以帮助司法实体边界的判定以减少边界错误, 同时融入词汇信息可以增强对上下文的理解, 提升实体分类的准确度. 本文针对司法领域特征的考虑, 带来了边界错误减少、分类准确提高的效果, 从而带来了最终结果的提升.

3.5.2 消融实验及结果分析

本文提出的方法, 对于模型要素识别效果的提升主要通过以下两种方式: 1) 在司法案情要素识别任务的基础上, 新增司法文本分类和司法中文分词模块辅助学习; 2) 在要素识别的输入序列融入全局信息和词汇局部信息. 为验证所各个模块、方法的有效性, 本文在 CAIL2021 数据集上通过消融实验进一步分析. 消融实验以要素识别的效果为基准. 首先, 验证加入司法文本分类、司法中文分词两个任务对于模型整体性能的提升; 其次, 在司法文本分类、司法中文分词两个任务均加入的基础上, 验证加入 CLS 全局信息、分词词汇信息对于模型要素识别效果的提升. 该消融实验的结果报告如表 5 所示, 其中“√”表示实验加入相应的模块.

表 5 在 CAIL2021 数据集上加入不同模块对多任务网络结果的影响

序号	消融模块				指标		
	司法文本分类任务	司法中文分词任务	CLS加入识别序列	分词结果加入识别序列	精确率	召回率	*F*1值
1	—	—	—	—	0.9011	0.9108	0.9059
2	√	—	—	—	0.9051	0.9224	0.9137
3	—	√	—	—	0.9137	0.9184	0.9161
4	√	√	—	—	0.9106	0.9248	0.9177
5	√	√	√	—	0.9183	0.9216	0.9200
6	√	√	—	√	0.9165	0.9300	0.9232
7	√	√	√	√	0.9204	0.9340	0.9272

根据表 5 中的实验结果可以得到以下结论.

(1) 当单独加入司法文本分类任务时, 司法案情要素识别的 *F*1 指标从 0.9059 提升至 0.9137, 结果有所提升, 说明此任务对要素识别产生了辅助效果, 减少了结果的分类错误从而带来提升; 当单独加入司法中文分词任务时, 司法案情要素识别的 *F*1 指标从 0.9059 提升至 0.9161, 结果也有所提升, 且相较于司法文本分类任务提升稍微大些 (*F*1 指标增加了 0.0024), 说明此任务减少了要素识别中的边界错误且带来的效果更好; 当同时加入司法文本分类、司法中文分词两个任务时, 相较于加入单个任务的情况, 结果没有下降反而有了略微的提升, *F*1 值分别从 0.9137

和 0.9161 提升至 0.9177, 这一结果虽然没有较大幅度的提升, 但说明了两个任务并没有产生互相抵触的效果. 该观测现象表明, 在要素识别单任务中引入司法文本分类任务和司法中文分词任务时的网络均好于要素识别单任务网络, 由此带来了模型效果的提升.

(2) 在同时加入司法文本分类任务和司法中文分词任务的情况下: 当向要素识别的输入序列加入 [CLS] 全局信息时, 有了新的提升 (*F*1 值进一步提升至 0.9200), 说明在多任务情况下加入全局信息, 对要素识别产生了有益的效果, 可以帮助网络更好地进行要素分类; 当向要素识别的输入序列中融入分词结果时, 又有了新的提升 (*F*1 值进一步提升至 0.9232), 且提升效果更大, 说明本文融入分词结果的方法给要素识别引入了词级信息, 在减少边界错误的同时还可进一步提升分类效果; 同时向要素识别的输入序列加入 [CLS] 全局信息和分词结果时, 达到了最佳的效果 (*F*1 值进一步提升至 0.9272), 说明所加模块共同叠加在一起时, 促进了最终要素识别任务的提升, 也印证了本文针对司法领域数据集所进行改进的有效性.

此外, 在表 5 第 7 号实验的基础上, 即将所有消融模块全部加入多任务框架时, 本文继续进行了多任务中各个子任务不同权重赋值时的实验, 以检测并验证 3 个任务的重要性. 当各个子任务被赋予不同权重时对最终实验结果的影响如表 6 所示.

表 6　在 CAIL2021 数据集上不同子任务权重对结果的影响

序号	超参数			指标		
	λ_1	λ_2	λ_3	精确率	召回率	*F*1值
基线	1.0	1.0	1.0	0.9204	0.9340	0.9272
1	1.5	1.0	1.0	0.9134	0.9240	0.9187
2	1.0	1.5	1.0	0.9103	0.9212	0.9157
3	1.0	1.0	1.5	0.9121	0.9256	0.9188
4	0.5	1.0	1.0	0.9117	0.9252	0.9184
5	1.0	0.5	1.0	0.9176	0.9216	0.9196
6	1.0	1.0	0.5	0.9139	0.9172	0.9156
7	2.0	1.0	1.0	0.9036	0.9252	0.9143

表 6 展示了调整不同子任务的损失函数权重带来的最终评价指标的变化, 可见: 当分别增大各个任务的权重时, 相比于 1:1:1 基线的情况最终结果都有所下降, 当降低司法案情要素识别任务权重时, 结果也有所下降, 由此分析, 当偏重某个任务时, 最终结果将有所下降, 降低司法案情要素识别任务的权重也会带来结果的下降, 可以说, 各个子任务对于司法案情要素识别最终结果起到了近乎相当的作用.

4 总　结

针对当前中文司法领域案情要素识别任务中, 存在缺乏文本全局信息导致产生要素分类错误, 以及缺乏词汇局部信息导致产生识别边界错误的问题, 本文提出了一种融合全局和局部信息的多任务学习方法来进行解决. 该方法利用 BERT 作为司法文本输入的共享层, 在共享层之上建立司法案情要素识别、司法文本分类、司法中文分词 3 个子任务网络, 并融入全局信息和词汇局部信息辅助案情要素识别的学习. 实验结果表明, 本文方法在两个公开的司法领域数据集 CAIL2021 和 Drug 上取得了当前最好效果, 有效减少了要素识别任务中分类和边界的错误, 说明在司法领域进一步考虑司法文本全局信息和局部信息, 可以减少类似错误, 带来结果的提升. 展望未来, 司法领域仍存在许多值得更进一步深入研究的空间, 在后续的研究工作中, 将重点对本文提出的多任务学习方法中关于司法文本分类任务中类别如何定义, 如何更好地平衡多个任务的进行, 以及是否还存在其他可以有效辅助命名实体识别进行的下游任务等多个方面展开研究.

致谢　感谢中南大学高性能计算中心提供的计算资源.

References:

[1] Liu YH. Practical application and prospects of AI technology in the construction of smart court. Journal of Comparative Law, 2022(1): 1–11 (in Chinese with English abstract).

[2] Li GD. Value orientation of judgment documents online system in China and its jurisprudential reflection. Studies in Law and Business, 2022, 39(2): 22–35 (in Chinese with English abstract). [doi: 10.16390/j.cnki.issn1672-0393.2022.02.015]

[3] Li CN. Research on named entity identification of legal documents [MS. Thesis]. Dalian: Dalian University of Technology, 2021 (in Chinese with English abstract). [doi: 10.26991/d.cnki.gdllu.2021.000961]

[4] Chinchor N, Robinson P. Appendix E: MUC-7 named entity task definition (version 3.5). In: Proc. of the 7th Conf. on Message Understanding. Fairfax: Association for Computational Linguistics, 1998. 1–21.

[5] Sang EFTK, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proc. of the 7th Conf. on Natural Language Learning at HLT-NAACL 2003. Edmonton: Association for Computational Linguistics, 2003. 142–147.

[6] Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 2018, 13(3): 55–75. [doi: 10.1109/MCI.2018.2840738]

[7] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 260–270. [doi: 10.18653/v1/N16-1030]

[8] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[10] Sun Y, Wang SH, Li YK, Feng SK, Chen XY, Zhang H, Tian X, Zhu DX, Tian H, Wu H. ERNIE: Enhanced representation through knowledge integration. arXiv:1904.09223, 2019.

[11] Cui YM, Che WX, Liu T, Qiu B, Wang SJ, Hu GP. Revisiting pre-trained models for Chinese natural language processing. In: Proc. of the Findings of the 2020 Association for Computational Linguistics. Association for Computational Linguistics, 2020. 657–668. [doi: 10.18653/v1/2020.findings-emnlp.58]

[12] Clark K, Luong MT, Le QV, Manning CD. ELECTRA: Pre-training text encoders as discriminators rather than generators. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[13] Peng NY, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015. 548–554. [doi: 10.18653/v1/D15-1064]

[14] He HF, Sun X. F-score driven max margin neural network for named entity recognition in Chinese social media. In: Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics: Vol. 2, Short Papers. Valencia: Association for Computational Linguistics, 2017. 713–718.

[15] Li HB, Hagiwara M, Li Q, Ji H. Comparison of the impact of word segmentation on name tagging for Chinese and Japanese. In: Proc. of the 9th Int'l Conf. on Language Resources and Evaluation. Reykjavik: European Language Resources Association, 2014. 2532–2536.

[16] Liu ZX, Zhu CH, Zhao TJ. Chinese named entity recognition with a sequence labeling approach: Based on characters, or based on words? In: Proc. of the 6th Int'l Conf. on Intelligent Computing. Changsha: Springer, 2010. 634–640. [doi: 10.1007/978-3-642-14932-0_78]

[17] He JZ, Wang HF. Chinese named entity recognition and word segmentation based on character. In: Proc. of the 6th SIGHAN Workshop on Chinese Language Processing. Hyderabad: Association for Computational Linguistics, 2008. 128–132.

[18] Zhang Y, Yang J. Chinese NER using lattice LSTM. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1554–1564. [doi: 10.18653/v1/P18-1144]

[19] Gui T, Zou YC, Zhang Q, Peng ML, Fu JL, Wei ZY, Huang XJ. A lexicon-based graph neural network for Chinese NER. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 1040–1050. [doi: 10.18653/v1/D19-1096]

[20] Li XN, Yan H, Qiu XP, Huang XJ. FLAT: Chinese NER using flat-lattice transformer. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 6836–6842. [doi: 10.18653/v1/2020.acl-main.611]

[21] Ma RT, Peng ML, Zhang Q, Wei ZY, Huang XJ. Simplify the usage of lexicon in Chinese NER. In: Proc. of the 58th Annual Meeting of

the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 5951–5960. [doi: 10.18653/v1/2020.acl-main.528]

[22] Li CN, Wang L, Sun YY, Lin HF. BERT based named entity recognition for legal texts on theft cases. Journal of Chinese Information Processing, 2021, 35(8): 73–81 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0077.2021.08.010]

[23] Shen YK, Tan S, Sordoni A, Courville AC. Ordered neurons: Integrating tree structures into recurrent neural networks. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[24] Leitner E, Rehm G, Moreno-Schneider J. Fine-grained named entity recognition in legal documents. In: Proc. of the 15th Int'l Conf. on Semantic Systems. Karlsruhe: Springer, 2019. 272–287. [doi: 10.1007/978-3-030-33220-4_20]

[25] Luz de Araujo PH, de Campos TE, de Oliveira RRR, Stauffer M, Couto S, Bermejo P. LeNER-Br: A dataset for named entity recognition in brazilian legal text. In: Proc. of the 13th Int'l Conf. on Computational Processing of the Portuguese Language. Canela: Springer, 2018. 313–323. [doi: 10.1007/978-3-319-99722-3_32]

[26] Quaresma P, Gonçalves T. Using linguistic information and machine learning techniques to identify entities from juridical documents In: Francesconi E, Montemagni S, Peters W, Tiscornia D, eds. Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language. Berlin, Heidelberg: Springer, 2010. 44–59. [doi: 10.1007/978-3-642-12837-0_3]

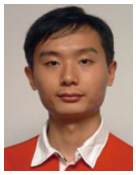
附中文参考文献:

[1] 刘艳红. 人工智能技术在智慧法院建设中实践运用与前景展望. 比较法研究, 2022(1): 1–11.

[2] 李广德. 裁判文书上网制度的价值取向及其法理反思. 法商研究, 2022, 39(2): 22–35. [doi: 10.16390/j.cnki.issn1672-0393.2022.02.015]

[3] 李春楠. 法律文书命名实体识别研究 [硕士学位论文]. 大连: 大连理工大学, 2021. [doi: 10.26991/d.cnki.gdllu.2021.000961]

[22] 李春楠, 王雷, 孙媛媛, 林鸿飞. 基于BERT的盗窃罪法律文书命名实体识别方法. 中文信息学报, 2021, 35(8): 73–81. [doi: 10.3969/j.issn.1003-0077.2021.08.010]



毛星亮(1979－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理, 文本挖掘.

李芳芳(1983－), 女, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为机器学习, 自然语言处理, 文本挖掘.

陈晓红(1963－), 女, 教授, 博士生导师, 主要研究领域为管理科学与工程, 工程管理, 数据智能.

张师超(1962－), 男, 博士, 教授, 博士生导师, 主要研究领域为数据挖掘, 知识发现.

宁肯(1995－), 男, 硕士生, 主要研究领域为自然语言处理, 信息抽取.