【电子与信息科学／Electronics and Information Science】

面向法律领域的实体和关系抽取

刘美玲，梁龙昌

东北林业大学计算机与控制工程学院，黑龙江哈尔滨 150040

摘　要：中文司法领域的实体和关系抽取技术在提高办案效率方面具有重要作用，但现有的关系抽取模型缺乏领域知识且难以处理重叠实体，造成难以准确区分和提取实体与关系等问题．通过引入领域知识，提出一种法律信息增强模块，增强了用所提法律潜在关系与全局对应(legal potential relationship and global correspondence，LPRGC)模型理解法律文本中术语、规则和上下文信息的能力，从而提高了实体和关系的识别准确性，进而提升了实体和关系抽取算法的性能．为解决重叠实体问题，设计了一种基于潜在关系和实体对齐的关系抽取方法．通过精确标注实体位置，筛选潜在关系，并利用全局矩阵对齐实体，解决重叠实体的关系抽取问题，能够更准确地捕捉到重叠实体之间的关系，并有效地将其映射到正确的实体对上，从而提高抽取结果的准确性．在中国法律智能技术评测数据集上进行实体和关系抽取实验，结果表明，LPRGC模型的准确率、召回率和F_1值分别为85.21%、81.19%和83.15%，均优于对比模型，特别是在处理实体重叠问题时，LPRGC模型在单实体重叠类型的抽取中，F_1值达到了81.45%；在多实体重叠类型的抽取中，F_1值达80.67%．LPRGC模型在实体和关系抽取的准确性上较现有方法有明显改进，在处理复杂法律文本中的实体重叠问题上取得了显著效果．

关键词：人工智能；自然语言处理；司法领域关系抽取；深度学习；信息增强；重叠实体

中图分类号：TP391.1；TP183　　**文献标志码：**A　　**DOI：**10.3724/SP.J.1249.2025.01077

Entity and relation extraction in the legal domain

LIU Meiling and LIANG Longchang

College of Computer and Control Engineering, Northeast Forestry University, Harbin 150040, Heilongjiang Province, P. R. China

Abstract: Entity and relation extraction technology in the Chinese judicial field plays an important role in improving case-handling efficiency. However, existing models lack domain knowledge and encounter challenges in handling overlapping entities, leading to difficulties in accurately distinguishing and extracting relationships. By introducing domain knowledge, we propose a legal information enhancement module that enhances the ability of the legal potential relationship and global correspondence (LPRGC) model to understand legal terms, rules, and contextual information, thereby improving the performance of entity and relation extraction algorithms. To address the issue of overlapping entities, we design a relationship extraction method based on latent relationships and entity alignment. By precisely annotating entity positions, filtering potential relationships, and aligning entities using a global matrix, the method accurately captures the relationships between overlapping entities and effectively maps them to the correct entity pairs, improving the accuracy of extraction results. Experiments conducted on the model using the China AI and Law Challenge (CAIL) dataset demonstrate that the model outperforms other compared models in terms of accuracy (85.21%), recall (81.19%), and F_1 score (83.15%). In particular, the proposed model achieves an F_1

Received: 2024-07-06; **Accepted:** 2024-08-20; **Online** (CNKI)**:** 2024-11-20
Foundation: Natural Science Foundation of Heilongjiang Province (LH2022F002)
Corresponding author: Associate professor LIU Meiling (mlliu@nefu.edu.cn)
Citation: LIU Meiling, LIANG Longchang. Entity and relation extraction in the legal domain [J]. Journal of Shenzhen University Science and Engineering, 2025, 42(1): 77-84. (in Chinese)

http://journal.szu.edu.cn

score of 81.45% for single overlapping entities, and an F_1 score of 80.67% for multiple overlapping entities. The experimental results show that the proposed LPRGC model significantly improves the accuracy of entity and relation extraction compared to existing methods, proving its effectiveness in enhancing model performance and addressing the issue of overlapping entities in complex legal texts.

Key words: artificial intelligence; natural language processing; judicial field relationship extraction; deep learning; information enhancement; overlapping entities

2022-12-09中华人民共和国最高人民法院发布《关于规范和加强人工智能司法应用的意见》[1]，提出加快智慧法治和智慧法院建设，推动人工智能与司法的深度结合，法律领域人工智能算法的开发现已成为研究热点．随着时间推进，司法机关积累了大量法律文书，传统的文书处理方式效率低，针对法律文本的自然语言处理，旨在将非结构化的法律文本转化为结构化信息，从而提高办案效率．信息抽取在法律案件处理、知识图谱构建及法律问答系统中发挥了重要作用，但现有法律信息平台深度挖掘功能缺乏，中文法律信息抽取任务发展缓慢，模型缺乏领域知识，尤其在处理重叠实体时表现不足．现有的信息抽取方法包括流水线方法和联合学习方法，前者易受错误传播影响，后者通过同时建模实体和关系来提高精度[2]．

命名实体识别(named entity recognition，NER)是自然语言处理中的关键任务，用于识别文本中的特定实体．早期的NER方法依赖于规则和模板，如RAU[3]提出的模板方法能从金融新闻中提取公司名．后来，机器学习方法如条件随机场(conditional random field，CRF)[4]、决策树(decision tree，CT)[5]、隐马尔可夫模型(hidden Markov models，HMM)[6]、最大熵模型(maximum entropy models，ME)[7]和支持向量机(support vector machine，SVM)[8]等被应用于NER．近年来，深度学习特别是递归神经网络(recursive neural network，RNN)[9]和长短期记忆网络(long short-term memory，LSTM)[10-11]在NER任务中得到了广泛应用．混合模型如双向长短期记忆(bi-directional LSTM，Bi-LSTM)和卷积神经网络(convolutional neural network，CNN)[12]进一步提升了NER整体性能[13]．诸如双向编码器表示的转换器(bidirectional encoder representations from Transformers，BERT)等预训练模型的应用，也显著提高了NER的识别性能[14]．

关系抽取(relation extraction，RE)是自然语言处理中的另一重要任务，旨在从文本中提取实体之间的关系．传统的RE方法基于规则，需预定义描述实体和关系的规则，但可移植性差[15]．基于统计的RE方法包括无监督、半监督和监督学习[16]．远程监督技术也被广泛应用，但限制在固定知识库的关系集合中[17-18]．随着深度学习的发展，神经网络特别是RNN[9]和CNN[12]在关系抽取尤其是在处理复杂的长距离依赖关系时表现出色．基于转换器(Transformer)的模型，如RoBERTa和生成式预训练转换器(generative pre-trained Transformer，GPT)，在关系抽取中表现出了处理复杂的长距离依赖关系、特征自动学习能力和上下文理解等显著优势[19-20]．

实体和关系联合抽取旨在同时识别实体及其关系．传统的流水线方法可独立处理实体和关系，容易忽视它们之间的依赖性[21]．近年来，端到端的联合抽取模型成为主流，特别是基于Transformer的模型能够同时处理实体识别和关系抽取任务，提高了关系抽取的性能[22-23]．此外，级联提取框架[24]和基于标记的模型[25]等方法也有助于解决多实体重叠问题．

1 面向法律领域的实体关系抽取方法

基于司法领域的特殊性，实体和关系抽取模型在特定领域会出现性能差异，为解决因缺乏领域知识导致模型性能不佳的问题，本研究构建了法律信息增强模块．同时，为了解决实体重叠问题，提出一种法律潜在关系与全局对应(legal potential relationship and global correspondence，LPRGC)模型进行司法领域的关系抽取，模型整体框架示例如图1，主要包含中文法律BERT(LegBERT-Chinese)、法律信息增强模块和解码器．

1.1 LegBERT-Chinese

BERT是一种深度神经网络架构，使用多层双向的Transformer编码器来从大型文本语料库中学习词语的上下文表示．传统的单向语言模型只能看到前面的词，因此在预测后面的词时受限于前面的上下文信息，而BERT采用双向的Transformer模型，

http://journal.szu.edu.cn

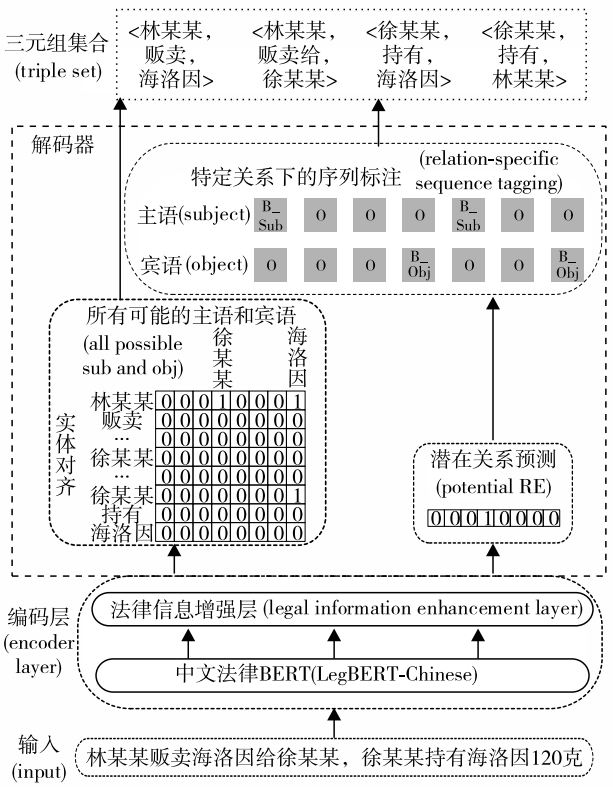

图 1　LPRGC 模型整体架构示例

Fig. 1　LPRGC model overall architecture.

可以同时考虑上下文的信息，使其能够更好地进行语言理解．本研究使用的 LegBERT-Chinese 是对 BERT 模型的一种特定改进版本，专门针对法律领域的中文文本进行训练或优化，将其作为模型的编码器，并得到模型的编码序列．

1.2　法律信息增强层

为加强模型的领域知识，本研究设计了法律信息增强层以更加适配该任务．由于识别实体的准确性对三元组提取的性能至关重要，本研究构建了法律实体词典 $\text{Legal}_{\text{Word}}$ 作为法律特征．首先，构建一个包含法律术语和实体的特征字典，这些术语和实体用于增强模型的领域知识．对于词序列 $\boldsymbol{S}=(w_1, w_2, \cdots, w_t, \cdots, w_N)$，将 $\boldsymbol{S}$ 与 $\text{Legal}_{\text{word}}$ 进行匹配，找到可能形成词典中名词表达式的所有子序列，其中，w_t 为输入序列中的第 t 个词；N 为序列长度；定义 $\boldsymbol{S}_{ij}=(w_i, w_{i+1}, w_{i+2}, \cdots, w_j)$．本研究利用一个大小为 $N\times N$ 的掩码矩阵 $\boldsymbol{M}_D=(m_{ab})$ 表示名词的法律特征，其中，D 为矩阵的维度；m_{ab} 为子序列 $\boldsymbol{S}_{ij}$ 是否是法律名词的表达式，

$$m_{ab}=\begin{cases}0, & \boldsymbol{S}_{ij}\in \text{Legal}_{\text{word}}\\ 1, & \text{其他}\end{cases} \tag{1}$$

使用 Transformer 编码层(Transformer encoder layer)计算输入句子的法律领域特定表示．该 Transformer 编码层有多头自注意力机制和位置前馈网络两个子层．每个子层之后是 1 个残差连接和 1 个层归一化操作．首先，将输入句子 $\boldsymbol{H}_L$ 通过线性变换得到矩阵 $\boldsymbol{Q}_{Dh}$、$\boldsymbol{K}_{Dh}$ 和 $\boldsymbol{V}_{Dh}$．其中，$\boldsymbol{L}$ 为输入序列的嵌入向量或表示；h 为多头注意力的第 h 个头；$\boldsymbol{Q}_{Dh}$、$\boldsymbol{K}_{Dh}$ 和 $\boldsymbol{V}_{Dh}$ 的大小分别为 $N\times d_{\text{q}}$、$N\times d_{\text{k}}$ 和 $N\times d_{\text{v}}$，其中，d_{q}、d_{k} 和 d_{v} 分别为查询向量、键向量和值向量的维度．将法律特征掩码矩阵与自注意力函数结合，得到融合法律特征的表示．自注意力函数的可用合法特征掩码矩阵 $\boldsymbol{M}_D$ 计得，即

$$\boldsymbol{A}_{Dh}=\text{softmax}\left(\boldsymbol{M}_D\times\frac{\boldsymbol{Q}_{Dh}\boldsymbol{K}_{Dh}}{\sqrt{d_k}}\right)\boldsymbol{V}_{Dh} \tag{2}$$

其中，softmax 为归一化指数函数．

将每个注意力头计算的结果 $\boldsymbol{Q}_{Dh}$、$\boldsymbol{K}_{Dh}$ 和 $\boldsymbol{V}_{Dh}$ 连接起来，并将结果通过前馈子层传递，最终输出与词典集成的特征，记为 $\boldsymbol{H}_D$．最后，对编码层得到的表示 $\boldsymbol{H}_L$ 和特征融合表示 $\boldsymbol{H}_D$ 进行加权平均，得到法律特征增强表示为

$$\boldsymbol{H}_{\text{E}}=\gamma\boldsymbol{H}_L+(1-\gamma)\boldsymbol{H}_D \tag{3}$$

其中，γ 为加权参数．

1.3　解码器

解码器包括潜在关系预测组件、特定关系序列标注组件及实体对齐组件．输入由 n 个词组成的句子 $\boldsymbol{S}$，期望的输出是关系三元组 $T(s)=\{(s,r,o)\,|\,s,o\in E, r\in R\}$，其中，$s$、$r$、$o$ 分别是三元组中的主语实体、关系和宾语实体，E 是实体集，R 是关系集．对于给定句子 $\boldsymbol{S}$，本节将对解码器各个组件进行介绍．对于潜在关系预测组件，给定一个由 n 个词组成的句子 $\boldsymbol{S}=(x_1, x_2, \cdots, x_l, \cdots, x_n)$，$x_l$ 为句子中的每个词或 token(标签)，该子任务负责预测句子 $\boldsymbol{S}$ 中可能存在的潜在关系是 $Y_{\text{rel}}(\boldsymbol{S})=\{r_1, r_2, \cdots, r_m, r_z\in R\}$，其中，$m$ 为潜在关系子集的大小．对于实体提取组件，针对给定的句子 $\boldsymbol{S}$ 和预测的潜在关系 r_i，该子任务使用 BIO(即 begin、inside 和 outside)标签方案来识别每个 token 的标签．在实体对齐组件中，对于给定的句子 $\boldsymbol{S}=(x_1, x_2, \cdots, x_n)$，该子任务预测主体和客体的起始标记之间的对应得分，只有真正三元组的会得高分，假的三元组将得低分．设 $\boldsymbol{M}$ 为全局对应矩阵，则此任务的输出为 $\boldsymbol{Y}_s(\boldsymbol{S})=\boldsymbol{M}\in R^{n\times n}$．

http://journal.szu.edu.cn

1.3.1 潜在关系预测组件

在本研究中，潜在关系预测组件与以往的逐一关系提取的关系抽取方法有所不同．以往的工作大多数都是对每个关系进行提取，这种方法往往会导致关系的冗余．相比之下，对于给定的句子，本研究提出的潜在关系预测组件会先预测可能存在的关系子集，然后只提取与这些潜在关系对应的实体．这种方法不仅可有效减少提取关系的冗余，还可更准确地捕捉到句子中的实体关系．

对于给定的经过编码层编码的由n个词组成的句子，其序列可表示为$\boldsymbol{h} \in R^{n \times d}$．使用式(4)和式(5)表示计算句子中可能存在的关系集合P_{rel}，过滤无关的关系，以减少计算量．

$$\boldsymbol{h}_{\text{avg}} = p_{\text{avg}}(h) \in R^{d \times 1} \tag{4}$$

$$\boldsymbol{P}_{\text{rel}} = \sigma\left(\boldsymbol{W}_{\text{rel}}\boldsymbol{h}_{\text{avg}} + \boldsymbol{b}_{\text{rel}}\right) \tag{5}$$

其中，p_{avg}为平均池化操作；$\boldsymbol{W}_{\text{rel}} \in R^{1 \times d}$为可训练权值；$\sigma$为sigmoid函数；$\boldsymbol{b}_{\text{rel}}$为对应的偏置项．

本研究提出的潜在关系预测组件将句子中的关系预测视为多标签的二元分类任务．在二元分类任务中，需要为每个可能的关系分配1个标签，若某个关系的概率超过了预设阈值，表示存在该关系，该关系的标签值将被分配为1；否则，该关系的标签值将被分配为0(如图1所示)．这种方法可以有效地将关系分类问题转化为多标签的二元分类问题．需要注意的是，这里只对预测的关系而非所有关系应用特定关系的序列标记．这是因为，对于一个给定的文本，可能只存在部分关系，而非所有可能的关系．因此，仅对预测为存在关系的潜在关系分配标签，可减少标签数量，从而提高分类效率．

1.3.2 特定关系下序列标注组件

在判断句子中潜在的关系后，接下来分别执行两个序列标注操作来提取主体(主语)和客体(宾语)．分开提取主体和客体是为了处理三元组中实体重叠的问题．为保证模型的简单性和计算速度，选择全连接神经网络进行序列标注，针对每个关系进行2次独立的序列标注操作，分别提取主语(subjects，sub)实体和宾语(objects，obj)实体，解决了三元组中实体重叠的问题，标注采用的是BIO标注方式，同时在标注过程中对每个token向量加入了关系向量，识别在特定关系下的实体，具体计算方式为

$$\boldsymbol{P}_{i,j}^{\text{sub}} = \text{softmax}\left(\boldsymbol{W}_{\text{sub}}\left(h_i + u_j\right) + \boldsymbol{b}_{\text{sub}}\right) \tag{6}$$

$$\boldsymbol{P}_{i,j}^{\text{obj}} = \text{softmax}\left(\boldsymbol{W}_{\text{obj}}\left(h_i + u_j\right) + \boldsymbol{b}_{\text{obj}}\right) \tag{7}$$

其中，$\boldsymbol{P}_{i,j}^{\text{sub}}$为第$i$个token第$j$个关系被预测为主语的概率分布，$\boldsymbol{P}_{i,j}^{\text{obj}}$为第$i$个token第$j$个关系被预测为宾语的概率分布，$h_i \in R^{1 \times d}$为第$i$个token的编码表示；$u_j \in R^{1 \times d}$为可训练嵌入矩阵$\boldsymbol{U} \in R^{d \times n_{\text{rel}}}$中的第$j$个关系表示，$n_{\text{rel}}$为全部关系集合的大小；$\boldsymbol{W}_{\text{sub}}$和$\boldsymbol{W}_{\text{obj}} \in R^{d \times 3}$为可训练的权重矩阵，其中3对应标记集合{B，I，O}的3种标签；$\boldsymbol{b}_{\text{sub}}$和$\boldsymbol{b}_{\text{obj}}$为偏置向量．

1.3.3 实体对齐组件

序列标注后，获得了关于句子关系的所有可能的主语和宾语，然后使用一个全局对应矩阵来确定正确的主语和宾语对．需要注意的是，由于全局对应矩阵独立于关系，因此可以与潜在关系预测同时学习．具体过程如下：首先枚举所有可能的主体-客体对；然后在全局矩阵中检查每对主体-客体对的分数，若超过阈值，则保留该主体-客体对，否则将其过滤掉．

如图1中的矩阵所示，给定一个有n个token的句子，其全局对应矩阵$\boldsymbol{M}$的形状为$R^{n \times n}$．其中，n为句子中token的数量；M_{ij}为第i个token作为主语和第j个token作为宾语的得分．该矩阵的每个元素都是关于一个配对的主体和客体的起始位置，代表了一个主体-客体对的置信度，该值越高，则表示该主体-客体对属于三元组的置信度越高．例如，位于第1行、第4列的“林某某”和“徐某某”的M_{14}值很高，它们是正确的三元组组合“<林某某，贩卖(给人)，徐某某>”的置信度就很高．针对某一类关系提取除了句子中所有可能的subjects和objects后，使用全局关联矩阵来确定正确的subject-object对，计算公式为

$$\boldsymbol{P}_{i\_\text{sub},j\_\text{obj}} = \sigma\left(\boldsymbol{W}_{\text{g}}\left[\boldsymbol{h}_i^{\text{sub}};\ \boldsymbol{h}_j^{\text{obj}}\right] + \boldsymbol{b}_g\right) \tag{8}$$

其中，$\boldsymbol{h}_i^{\text{sub}}$，$\boldsymbol{h}_j^{\text{obj}} \in R^{2d \times 1}$为输入句子中第$i$个和第$j$个token经过编码层的向量表示，二者形成一个潜在的主语和宾语对；$\boldsymbol{W}_g \in R^{2d \times 1}$为一个可训练的权重；$\boldsymbol{b}_g$为一个可训练的偏置向量，$g$为全局对齐矩阵的参数．

1.4 损失函数

交叉熵损失亦称为对数损失，是机器学习中常用于分类任务的损失函数，用于衡量预测概率分布与实际概率分布之间的差异．本研究的任务是一个多分类问题，将采用交叉熵损失函数．具体而言，使用联合训练该模型，在训练期间优化组合目标函

http://journal.szu.edu.cn

数，并共享编码器的参数．模型总的损失(L_{total})由关系判断的损失L_{rel}、序列标注的损失L_{seq}和实体对齐的损失L_{global}3部分组成，均为交叉熵损失函数，表达式为

$$L_{\text{total}} = L_{\text{rel}} + L_{\text{seq}} + L_{\text{global}} \tag{9}$$

$$L_{\text{rel}} = -\frac{1}{n_{\text{rel}}}\sum_{i=1}^{n_{\text{rel}}}\left[y_i \ln P_{\text{rel}} + \left(1 - y_i\right)\ln\left(1 - P_{\text{rel}}\right)\right] \tag{10}$$

$$L_{\text{seq}} = -\frac{1}{2 \times n \times n_{\text{r}}^{\text{pot}}}\sum_{t \in \text{sub, obj}}\sum_{j=1}^{n_{\text{r}}^{\text{pot}}}\sum_{i=1}^{n} y_{i,j}^{t} \ln P_{i,j}^{t} \tag{11}$$

$$L_{\text{global}} = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_{i,j}\ln P_{i_{\text{sub}},j_{\text{obj}}} + \left(1 - y_{i,j}\right)\ln\left(1 - P_{i\_\text{sub},j\_\text{obj}}\right)\right) \tag{12}$$

其中，n_{rel}为全部关系集合的大小；y_i为第i类关系的实际概率；P_{rel}为第i类关系的预测概率；$n_{\text{r}}^{\text{pot}}$为句子的潜在关系子集的大小；$y_{i,j}^{t}$为从$i$到$j$的词为实体$t$的实际概率；$P_{i,j}^{t}$为$n$从$i$到$j$的词为实体$t$的预测概率；$n$为全局矩阵的大小；$y_{i,j}$表示$i$为主体，$j$为客体的实际概率；$P_{i_\text{sub},j_\text{obj}}$表示$i$为主体，$j$为客体的预测概率．

2　实验结果与分析

2.1　数据集

本研究使用中国法律智能技术评测(China AI and Law Challenge 2022，CAIL2022，http://cail.cipsc.org.cn)信息抽取任务的数据集，该数据集数据信息源自公开的若干涉毒类罪名法律文书，选择涉毒类刑事案件中最具代表的贩卖毒品罪、非法持有毒品罪和容留他人吸毒罪3个罪名，总计1 750条．参考《中华人民共和国刑法》中3个罪名的审判依据，针对涉毒类案件的法律实体关系抽取任务，预定义5种实体和4种关系类型．实体分别为人名实体(Nh)、地名实体(Ns)、时间实体(NT)、毒品类型实体(NDR)和毒品重量实体(NW)．关系分别为贩卖(给人)(sell_drug_to)、贩卖(毒品)(traffic_in)、持有(possess)和非法容留(provide_shelter_for)．4种关系涵盖了3类涉毒类案件中的各种犯罪行为，定义为：

1）贩卖(给人)关系．关系三元组的头尾实体均为涉案人，在案件事实中出现头实体将某毒品贩卖给尾实体的情节；

2）贩卖(毒品)关系．关系三元组的头实体为涉案人，尾实体为涉案物品，在涉毒类案件中特指毒品类型实体，在案件事实中出现头实体进行了尾实体相关毒品交易的情节．

3）持有关系．关系三元组的头实体为涉案人，尾实体特指毒品类型实体，在案件事实中出现头实体持有或藏匿一定量尾实体的情节．

4）非法容留关系．关系三元组的头尾实体均为涉案人，在案件事实中出现了头实体提供场所容留尾实体摄入或注射某毒品的情节．

2.2　模型对比

针对CAIL2022数据集，分别采用LPRGC模型、NovelTagging模型[22]、CopyRE模型[25]、GraphRel模型[26]、ETL-span模型[27]、TPLinker模型[28]、TPLinker模型[28]和CasRel模型[24]进行实验，比较各模型的准确率、召回率和F_1值等关键指标，重点测试各模型在处理重叠实体和复杂法律文本中的关系抽取能力，从而评估这些模型在实体关系抽取任务中的表现．NovelTagging模型[22]将标注方法与端到端方法结合，解决联合抽取任务，但在处理复杂法律文本时效果不佳，准确率和召回率较低．CopyRE模型[25]是一种基于复制机制的端到端模型，解码过程中采用联合解码器和多个分离解码器，但因法律文本中实体关系复杂，该模型性能受限．GraphRel模型[26]是一种基于图卷积网络(graph convolutional network，GCN)的联合抽取模型，利用GCN学习命名实体和关系，但在大规模数据下的计算量较大，性能提升有限．ETL-span模型[27]通过将关系抽取任务分解为多个序列标记任务来解决实体重叠问题，但在解码过程中信息丢失和冗余问题依然存在．TPLinker模型[28]通过预测词头和词尾的对应关系，解决实体重叠和暴露偏差问题，但在关系复杂度高的文本中表现稍显不足．CasRel模型[24]先抽取主语实体，再同时抽取关系及其对应的客体实体，针对复杂关系的法律文本仍可能识别错误．

2.3　评价指标

本研究使用准确率(η_{pre})、召回率(η_{recall})和F_1值作为模型性能评价指标．只有当两个实体的开始和结束以及三元组之间的关系都正确时，才认为三元组的提取是正确的．评价指标的计算式为

$$\eta_{\text{pre}} = \frac{N_{\text{corr}}}{N_{\text{pred}}} \times 100\% \tag{13}$$

http://journal.szu.edu.cn

$$\eta_{\text{recall}} = \frac{N_{\text{corr}}}{N_{\text{T}}} \times 100\% \tag{14}$$

$$F_1 = \frac{2 \times \eta_{\text{pre}} \times \eta_{\text{recall}}}{\eta_{\text{pre}} + \eta_{\text{recall}}} \times 100\% \tag{15}$$

其中，N_{corr}、N_{pred}和N_{T}分别为正确提取的三元组个数、提取的三元组的总数和真实的三元组的个数.

2.4 结果与分析

表1给出了采用不同模型对CAIL2022数据集进行关系抽取的结果. 由表1可见，NovelTagging模型的准确率和召回率较低，原因是该模型仅使用简单的标注策略，未能有效处理复杂的实体关系和重叠实体问题，所以在处理具有多个关系的复杂法律文本时表现不佳；CopyRE模型在法律领域的文本中，实体关系的复杂性导致模型难以准确复制相关信息，并且在联合解码过程中存在的多个解码器增加了模型的复杂性和计算量，影响了整体性能；GraphRel模型在处理大规模数据时，计算开销较大，影响了性能提升；ETL-span在多任务解码过程中，信息丢失和冗余问题仍然存在；TPLinker在关系复杂度高的文本中表现略显不足；CasRel在关系多样且复杂的法律文本中，仍存在识别错误的可能；LPRGC模型的准确率、召回率和F_1值都显著高于其他模型. 这说明通过引入法律领域特征，LPRGC模型的法律信息增强模块能更好地理解法律文本，进而提高实体和关系的识别准确性.

表1 关系抽取对比实验结果

Table 1 Relational extraction comparative experimental results %

模型	η_{pre}	η_{recall}	F_1
NovelTagging	67.13	62.48	64.72
CopyRE	63.09	60.36	61.69
GraphRel	74.54	71.59	73.04
ETL-span	75.37	72.04	73.67
TPLinker	78.82	76.31	77.53
CasRel	82.65	80.03	81.32
LPRGC	85.21	81.19	83.15

为验证LPRGC模型在重叠实体上抽取的效果，与目前解决实体重叠比较有效的CasRel模型分别在正常、单实体重叠(single-entity overlap，SEO)和多实体重叠(entity-pair overlap，EPO)3种情况下进行了对比实验，结果如表2.

表2 三元组实体重叠对比实验结果

Table 2 Triplet entity overlapping comparison experiment results %

重叠实体	CasRel			LPRGC		
	η_{pre}	η_{recall}	F_1	η_{pre}	η_{recall}	F_1
正常	85.36	82.19	83.75	86.07	83.24	84.63
SEO	80.57	78.92	79.74	83.35	79.63	81.45
EPO	79.46	77.58	78.51	81.74	79.62	80.67

从表2可见，LPRGC模型在正常类型的抽取结果的准确率为86.07%，F_1值为84.63%，而CasRel模型的准确率为85.36%，F_1值为83.75%，两者相差不大，LPRGC模型的准确率和F_1值都仅仅较CasRel模型提高了不到1%，属于略有提升，说明两模型对于正常类型的关系抽取性能差别甚微. 但是，从在单实体重叠类型的抽取结果来看，LPRGC模型在各个指标上的结果都要明显优于CasRel的抽取结果，说明在对单实体重叠类型的抽取上，LPRGC模型较现有模型具有很大的改进，抽取效果有显著提升. 而针对多实体重叠类型的抽取，LPRGC模型的准确率和F_1值都比CasRel模型的高，说明在多实体重叠的类型抽取上，LPRGC模型的抽取效果较现有模型具有很大的改进. 综合来看，LPRGC模型在解决三元组关系抽取任务中的单实体重叠以及多实体重叠问题都优于现有模型，证明了这种先判断潜在关系再在特定关系下抽取实体，然后进行实体对齐的关系抽取方法，能有效解决三元组实体重叠问题.

结 语

针对目前中文司法案件关系抽取技术目前存在两个问题，一方面提出了法律信息增强模块用于增强模型的领域信息，另一方面提出基于潜在关系和实体对齐的联合抽取方法以解决法律文本中实体关系重叠的问题. 实验证明，所提LPRGC模型能够提高实体和关系抽取的准确性，并能有效解决三元组提取中的实体重叠问题. LPRGC模型能在法律技术平台上自动提取法律文书中的关键信息，构建法律知识图谱，支持智能问答和自动化文书分析，提高办案效率. 它还可用于智能法律检索、推荐系统和司法大数据分析，显著提升法律技术平台的智能化水平，支持智慧法治发展. 不足的是，本研究

http://journal.szu.edu.cn

使用的是源自于中国法律智能技术评测(CAIL2022)信息抽取任务的法律领域数据集，该数据集仅预定义了5种实体和4种关系类型，导致识别的实体数目和关系数目有限，后续可以继续完善数据集并重现训练模型，增强系统能力.

基金项目：黑龙江省自然科学基金资助项目(LH2022F002)
作者简介：刘美玲(mlliu@nefu.edu.cn)，东北林业大学副教授、博士．研究方向：大模型开发应用、人工智能生成技术AIGC、大数据挖掘和分析、自然语言处理、社交媒体、智能交通和智能城市等.
引　　文：刘美玲，梁龙昌，朱天胜，等. 面向法律领域的实体和关系抽取[J]. 深圳大学学报理工版，2025，42(1)：77-84.

参考文献 / References：

[1] 佚名．最高人民法院关于规范和加强人工智能司法应用的意见：法发〔2022〕33号[N]．人民法院报．2022-12-10(004)．
Anon. Opinions of the supreme people′s court on regulating and strengthening the judicial application of artificial intelligence: Fa Fa [2022] No. 33 [N]. People′s Court Daily. 2022-12-10(004). (in Chinese)

[2] WANG Peng. A survey of research on deep learning entity relationship extraction [J]. Natural Language Processing and Speech Recognition, 2019, 1(1): 1-5.

[3] RAU L F. Extracting company names from text [C]// Proceedings the 7th IEEE Conference on Artificial Intelligence Application. Piscataway, USA: IEEE, 1991: 29-32.

[4] SUTTON C, MCCALLUM A. An introduction to conditional random fields [J]. Foundations and Trends® in Machine Learning, 2012, 4(4): 267-373.

[5] SONG Yanyan, LU Ying. Decision tree methods: applications for classification and prediction [J]. Shanghai Archives of Psychiatry, 2015, 27(2): 130-135.

[6] MOR B, GARHWAL S, KUMAR A. A systematic review of hidden Markov models and their applications [J]. Archives of Computational Methods in Engineering, 2021, 28(3): 1429-1448.

[7] RATNAPARKHI A. A simple introduction to maximum entropy models for natural language processing [EB/OL]. (1997-05-13) [2022-07-11]. http://faculty. washington. edu/fxia/courses/LING572/maxent_adwait97. pdf.

[8] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[9] ZHANG Dongxu, WANG Dong. Relation classification via recurrent neural network [EB/OL]. (2015-12-25) [2022-07-11]. https://arxiv.org/abs/1508.01006

[10] SOCHER R, LIN C C Y, NG A Y, et al. Parsing natural scenes and natural language with recursive neural networks [C]// Proceedings of the 28th International Conference on Machine Learning. Madison, USA: Omnipress, 2011: 129-136.

[11] VAN HOUDT G, MOSQUERA C, NÁPOLES G. A review on the long short-term memory model [J]. Artificial Intelligence Review, 2020, 53(8): 5929-5955.

[12] ZENG Daojian, LIU Kang, LAI Siwei, et al. Relation classification via convolutional deep neural network [C]// Proceedings of the 25th International Conference on Computational Linguistics. Stroudsburg, USA: ACL, 2014: 2335-2344.

[13] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs [C]// Transactions of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2016: 357-370.

[14] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: ACL, 2016: 260-270.

[15] CUI Meiji, LI Li, WANG Zhihong, et al. A survey on relation extraction [C]// Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence. Singapore: Springer Singapore, 2017: 50-58.

[16] 谢德鹏，常青．关系抽取综述[J]．计算机应用研究，2020，37(7)：1921-1924，1930.
XIE Depeng, CHANG Qing. Review of relation extraction [J]. Application Research of Computers, 2020, 37(7): 1921-1924, 1930. (in Chines)

[17] AUGENSTEIN I, MAYNARD D, CIRAVEGNA F. Distantly supervised web relation extraction for knowledge base population [J]. Semantic Web, 2016, 7(4): 335-349.

[18] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data [C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg, USA: ACL, 2009: 1003-1011.

http://journal.szu.edu.cn

[19] GUPTA V, SINGH P, SHARAN A. A table-filling approach to joint entity and relation extraction [C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S. l. : s. n.], 2021: 1234-1245.

[20] ZHENG Suncong, HAO Yuexing, LU Dongyuan, et al. Joint entity and relation extraction based on a hybrid neural network [J]. Neurocomputing, 2017, 257: 59-66.

[21] PAWAR S S, BHATTACHARYYA P, PALSHIKAR G K. Investigations in entity relationship extraction [M]. Singapore: Springer Nature Singapore, 2023.

[22] ZHENG Suncong, WANG Feng, BAO Hongyun, et al. Joint extraction of entities and relations based on a novel tagging scheme [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2017: 1227-1236.

[23] ZENG Daojian, ZHANG Haoran, LIU Qianying. CopyMTL: copy mechanism for joint extraction of entities and relations with multi-task learning [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2020: 9507-9514.

[24] WEI Zhepei, SU Jianlin, WANG Yue, et al. A novel cascade binary tagging framework for relational triple extraction [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2020: 1476-1488.

[25] ZENG Xiangrong, ZENG Daojian, HE Shizhu, et al. Extracting relational facts by an end-to-end neural model with copy mechanism [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2018: 506-514.

[26] BEKOULIS G, DELEU J, DEMEESTER T, et al. Joint entity recognition and relation extraction as a multi-head selection problem [J]. Expert Systems with Applications, 2018, 114: 34-45.

[27] FU T J, LI P H, MA Weiyun. GraphRel: modeling text as relational graphs for joint entity and relation extraction [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2019: 1409-1418.

[28] YU Bowen, ZHANG Zhenyu, SHU Xiaobo, et al. Joint extraction of entities and relations based on a novel decomposition strategy [C]// ECAI 2020. Amsterdam, Netherlands: IOS Press, 2020: 2282-2289.

【中文责编：英子；英文责编：木柯】

http://journal.szu.edu.cn