基于区域顶点标注的司法文本实体关系联合抽取

乐滢滢, 孙媛媛†, 林鸿飞
大连理工大学，计算机科学与技术学院，大连，116024
29354772@mail.dlut.edu.cn,{syuan,hflin}@dlut.edu.cn,

摘要

司法领域中的实体关系联合抽取在许多下游任务中（如量刑预测、知识库构建等）具有重要意义。然而，由于垂直领域中的数据资源稀缺，而且司法文本中存在复杂的长句以及关系重叠现象，这使得信息抽取工作颇具挑战性。为应对这一挑战，我们首先标注了一个包含多个罪名的司法领域的专有数据集，然后提出了一种基于三元组区域顶点的联合抽取填表法。我们采用多标签分类对三元组的边界进行标注，以此提取三元组，从而充分利用实体的边界信息。此外，为融入实体对之间的距离信息，我们引入了距离嵌入，并采用扩张卷积来捕捉多尺度上下文信息。我们在司法数据集上对模型进行了评估。实验结果表明，我们的模型在这个数据集上均取得了最先进的性能。

关键词： 实体关系联合抽取 ； 扩张卷积

Joint Extraction of Judicial Text Entities and Relations Based on Regional Vertices

Yingying Le, Yuanyuan Sun†, Hongfei Lin
School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
29354772@mail.dlut.edu.cn,{syuan,hflin}@dlut.edu.cn,

Abstract

Joint extraction of entity relations in the judicial field holds significant importance in many downstream tasks (such as sentencing prediction, knowledge base construction, etc.). However, due to the scarcity of data resources in vertical domains and the presence of complex long sentences and overlapping relations in judicial texts, information extraction becomes quite challenging. To address this challenge, we first annotated a proprietary dataset in the judicial field covering multiple crime types, and then proposed a table-filling method for joint extraction based on the vertices of the triple region. We employed multi-label classification to annotate the boundaries of triples for extracting them, thereby fully utilizing the boundary information of entities. Additionally, to incorporate distance information between entity pairs, we introduced distance embedding and utilized dilated convolutions to capture multi-scale contextual information. We evaluated our model on a judicial dataset. The experimental results demonstrate that our model achieves state-of-the-art performance on this dataset.

Keywords: Joint extraction , Dilated convolution

† 通讯作者

1 引言

法律文本中的实体和关系抽取是一项重要的研究任务，旨在从非结构化文本中抽取由实体对及其关系组成（头实体，关系，尾实体）的三元组，这是信息提取的一项重要任务。这项任务面临着两个主要挑战，即如何解决重叠问题以及如何有效利用实体和关系之间的信息。

在早期的研究中，传统的流水线方法 (Zelenko et al., , 2003; Zhou et al., , 2005; Chan and Roth, , 2011)通常将三元组抽取任务分解为命名实体识别和关系抽取两个独立的步骤。流水线方法虽然很灵活，但它忽略了两个任务之间的相互依存关系，会受到错误传播的影响导致性能下降 (Ren et al., , 2022)。为了解决这个问题，一些研究人员试图使用端到端的实体关系联合抽取模型来解决三元组抽取任务。基于标签的方法 (Zheng et al., , 2017) 将联合提取任务转化为序列标记问题，然而，他们的方法只考虑每个实体只属于一个三元组的情况；一种新的级联二进制标记框架也被提出 (Wei et al., , 2020)，但面临着错误传播等局限性；填表法例如TPLinker (Wang et al., , 2020)和OneRel (Shang et al., , 2022)将任务转化为矩阵填充问题，它们能够有效地捕捉文本中的实体和关系信息，显著提升了抽取的准确性和效率。但它们在处理相互依赖的信息方面仍显不足，当文本中存在多个相互关联的实体和关系时，这些方法可能难以准确捕捉它们之间的复杂依赖关系，从而导致抽取结果的完整性和准确性受到影响。尽管现有方法在通用领域的三元组抽取任务中已展现出显著性能，但在处理司法文书等复杂场景时，其对于复杂多变的实体的建模能力仍显不足。司法文书特有的领域特性对现有方法在关系建模的鲁棒性及跨领域泛化能力方面提出了更高要求。

相对于通用领域，司法领域文本的实体关系复杂多样，而且常常出现实体重叠（Single Entity Overlap，SEO）三元组和实体对重叠（Entity Pair Overlap，EPO）三元组，该文使用了一种新的实体关系联合抽取方法，有效地利用实体之间及其关系的信息，着重解决重叠三元组的问题。受OD-RTE (Ning et al., , 2023)的启发，该文为每个实体关系构建了一个表格，三元组中的两个实体在关系的表格中形成一个矩形区域，通过识别该区域的四个顶点（即左上角、右上角、左下角、右下角）来抽取三元组。同时，在该文的方法中，引入了距离嵌入来捕获与token之间的距离有关的信息。受Li等人 (Li et al., 2022)的启发，该文应用扩展卷积来捕获多尺度上下文信息，提高了实体关系联合抽取的性能。

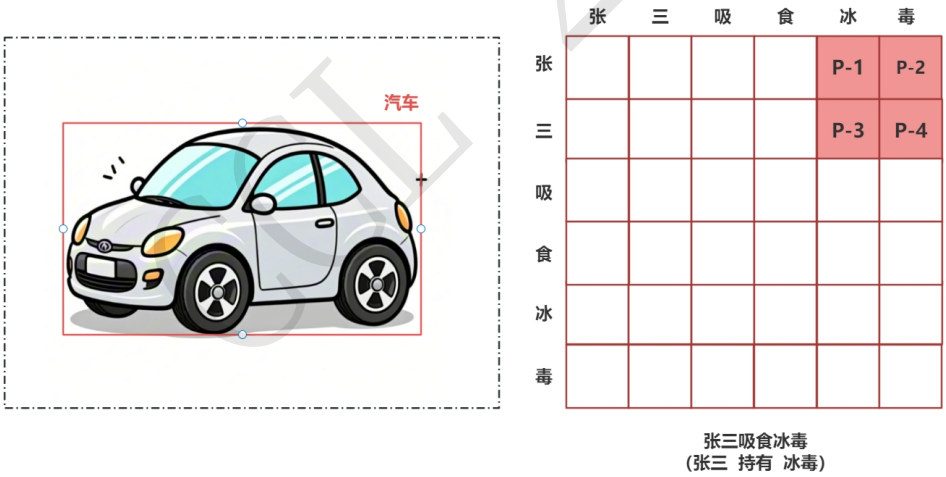

Figure 1: 目标检测和基于填表法的三元组抽取的比较，P代表关系持有

在法律文本处理领域，应用实体关系联合抽取技术能够显著提升信息处理效率，并为下游司法任务（诸如罪名精准分类、司法辅助性文件自动生成、司法知识图谱体系化构建等）提供关键性支撑。相对于通用领域，司法领域因其文本形态的多样性、专业术语的复杂性及案件类型的广泛性，导致相关领域数据集的稀缺性成为制约技术发展的核心瓶颈。针对这一现状，该文研究团队系统性地构建了司法数据集，该数据集聚焦于盗窃罪、诈骗罪及涉毒罪三大高发、高关注度罪名，通过多源数据整合与专业标注，实现了罪名类别的均衡覆盖与语义特征的深度挖掘，为法律文本智能化分析提供了高质量的标注语料基础。

本文工作的主要贡献概括如下：

- 构建了一套涵盖盗窃罪、诈骗罪及涉毒犯罪三类典型司法领域的数据集。该数据集包含5,209条盗窃罪相关数据样本、1,520条诈骗罪相关数据样本以及1,548条涉毒犯罪相关数据样本。在数据关系维度上，盗窃罪与涉毒犯罪数据集中均包含四类核心关系，而诈骗罪数据集则包含六类数据核心关系，充分彰显了该司法数据集在犯罪类型与关系特征层面的多元性与复杂性；
- 创新性地提出了一种基于多标签分类机制的三元组边界标注策略，为复杂文本场景下的三元组抽取任务提供了高效解决方案；在此基础上，进一步融合距离感知嵌入技术与多尺度扩张卷积模块，构建了具备多维信息融合能力的深度学习模型，该模型能够有效捕获文本序列中的长距离依赖特征及空间结构化信息，显著增强了模型对上下文语义关联和实体空间关系的建模精度；
- 在构建的司法多罪名数据集和一个广泛使用的通用数据集上评估了该文提出的模型，结果表明，该模型不仅在司法领域专有数据集上达到了最先进的性能，相较于最好的基线模型提升了1.0%，在通用领域数据上也达到了最高的的F_1分数，相较于最好的基线模型提升了0.1%。

2 实体关系联合抽取相关工作

实体关系联合抽取方法可根据其抽取思路大致分为以下四类。第一类是基于标注（tagging-based）的方法，该方法利用多个相互关联的序列标注模块来标注头实体、尾实体甚至关系。例如，Zheng等人 (Zheng et al., , 2017)提出的标注方案将关系三元组抽取任务转化为标注问题，随后，Wei等人 (Wei et al., , 2020)提出的CASREL方法将关系建模为将主语映射到宾语的函数，从而自然地处理了重叠问题。此外，Zheng等人 (Zheng et al., , 2021)提出了一种基于潜在关系和全局对应的抽取器，以缓解关系预测中的冗余问题。BiRTE (Ren et al., , 2022)则提出了一种基于双向抽取框架的方法，该方法根据从两个互补方向提取的实体对来抽取三元组。第二类方法是表格填充（table-filling）方法，该方法通过分类词对之间的关系来确定头实体和尾实体。表格填充方法的典型代表是TPLinker (Wang et al., , 2020)，该方法引入了一种新颖的握手标记方案，该方案能够针对每一种关系类型，对齐实体对的边界词。此外，OneRel (Shang et al., , 2022)也是一种有效的方法，它将联合抽取视为一个细粒度的三元组分类问题。第三类方法是文本生成方法(Zeng et al., , 2018; Zeng et al., , 2020; Ye et al., , 2021)，该方法采用序列到序列（seq2seq）的结构来生成三元组。第四类方法是图链接方法(Shang et al., , 2022b)。

近年来，随着大规模预训练语言模型的快速发展及其在自然语言处理任务中展现出的强大能力，研究者们开始积极探索如何有效利用这些模型的语义理解和生成能力来提升实体关系联合抽取任务的性能。GPT-RE (Wan et al., , 2023)提出了一种基于上下文学习的训练无关范式，直接利用预训练语言模型（如GPT-3）进行关系分类。该方法通过动态示例选择和结构化提示模板，在少样本设定下显著优于传统监督学习模型（如BERT），并在跨领域任务中表现出良好的泛化能力。LAL-JER (He and Bai, , 2023)提出了一种创新的标签感知学习框架，通过结合大语言模型数据增强技术，通过精心设计的提示模板生成多样化的合成数据，并采用对抗性过滤机制确保数据质量，解决了传统联合抽取任务中的标签分布不平衡和标注数据稀缺问题。AutoRE (Xue et al., , 2024) 提出了一种基于大语言模型的文档级关系抽取框架，通过层次化提示机制解决长文档处理难题。该方法创新性地采用两阶段流程：先进行实体感知的文档摘要，再利用结构化提示引导关系推理。研究提出了动态实体链指机制维护全局实体状态，有效提升推理能力。尽管大语言模型在实体关系抽取任务中展现出了卓越的性能，但其实际应用仍面临若干关键性局限。首先，领域适应性不足显著制约了模型在专业场景中的表现，特别是对领域专有术语和低频关系的识别准确率明显下降。其次，受限于上下文窗口长度和注意力机制的计算复杂度，模型在处理长文本时的实体关系推理能力有限。再者，模型对提示工程的高度敏感性导致性能波动显著，给实际部署带来了额外调试成本。资源消耗方面，大参数量带来的高计算成本使得模型训练和推理效率成为瓶颈，此外，模型生成的幻觉问题会导致预测结果存在实体指代错误或虚构关系。

3 方法

本节中将详细介绍该文提出的联合抽取方法，算法框架图如图二所示。给定一个司法文

本中的句子$S = \{w_1, w_2, \ldots, w_N\}$，其中$N$ 为句子长度。实体关系联合抽取任务旨在从S 中提取所有潜在关系三元组$T = \{T_i = (h_i, r_i, t_i),\ \ i = 1, \ldots, M\}$，其中$M$ 为句子S 中三元组的个数，$h_i, t_i \in E$，分别代表三元组的头实体和尾实体，E 是S 中所有实体的集合，$r_i \in R$，$R = \{r_1, \ldots, r_K\}$，代表预定的K 个关系。

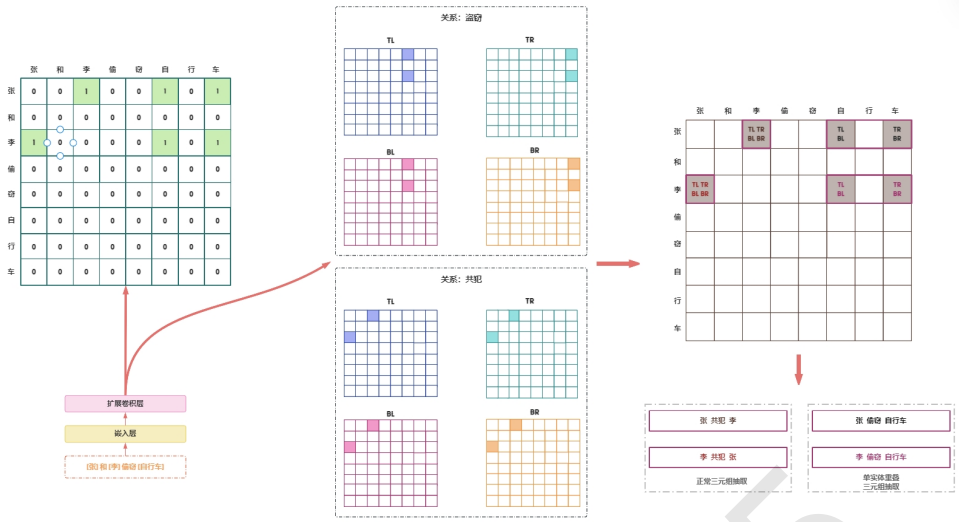

Figure 2: 算法框架图

图二演示了该文提出的模型是如何解决两个单实体重叠（Single Entity Overlap，SEO）三元组的过程，在司法领域数据集中还存在实体对重叠（Entity Pair Overlap，EPO）三元组，图三演示了该文的模型解决EPO的过程。

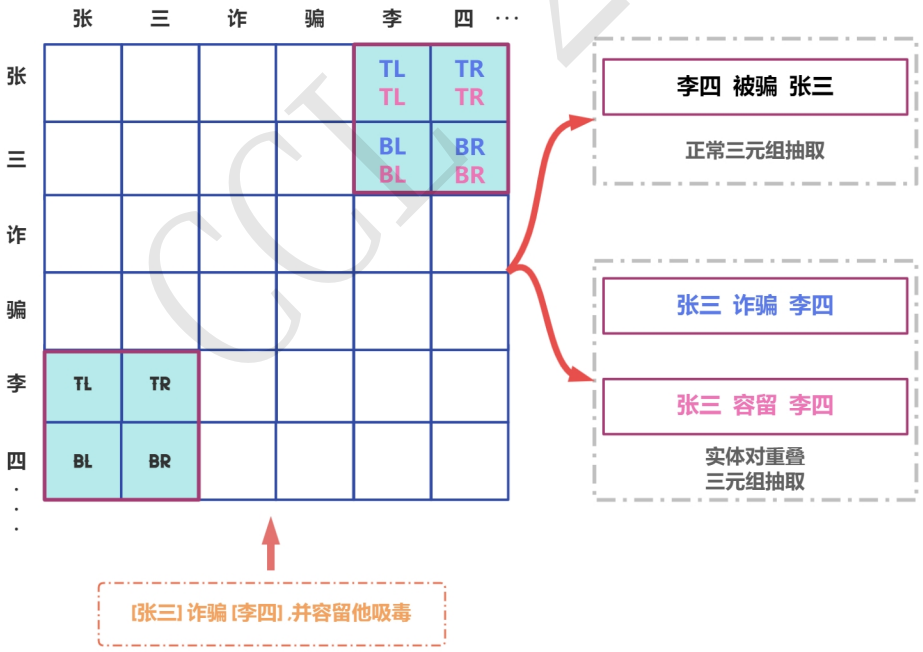

Figure 3: 实体对重叠三元组抽取

3.1 嵌入层

嵌入层的作用是将输入文本转换为嵌入表示。对于一个给定的司法文本S =

$\{w_1, w_2, \ldots, w_N\}$，首先使用预训练的语言模型BERT 获得该句子的token 表示：

$$H = \{h_1, h_2, \ldots, h_N\} = BERT(\{w_1, w_2, \ldots, w_N\}) \tag{1}$$

通过对数据集的统计分析，发现三元组中头实体和尾实体之间的距离分布是不均匀的。因此，该文使用距离嵌入来捕捉这一信息。具体来说，首先将所有可能的距离分为十个组别，然后用十个距离嵌入来表示它们。对于词对(w_i, w_j)，词对的嵌入表示$h(w_i, w_j)$ 的计算方法如下，其中W_{tp} 和b_{tp} 是可学习的参数，d_{ij} 表示词对(w_i, w_j) 对应的距离嵌入，

$$h(w_i, w_j) = W_{tp}[h_i; h_j] + d_{ij} + b_{tp} \tag{2}$$

3.2 扩展卷积层

通过嵌入层可以得到带有距离信息的词对的表示。由于这些表示形成一个二维表H_{tp}，包含$N \times N$ 令牌对表示，该文利用CNNs 进一步增强词对表示。受Li等人(Li et al., 2022) 的启发，采用不同扩张速率的多个扩张卷积来捕获多尺度上下文信息。一个扩张卷积的输出计算如下：

$$C^l = DCNN_t(H_{tp}) \tag{3}$$

其中C^l表示扩张速率为l的扩张卷积输出。将所有不同扩张速率的卷积输出进行拼接，以获得多重扩张卷积的输出C。最终的token 对的表示Q 计算如下：

$$Q = ReLU(H_{tp}; C) \tag{4}$$

3.3 标签预测

设计了一种方法来提取实体和关系，并处理三元组重叠的问题。算法为每个实体关系构建了一个表格。三元组中的两个实体在关系的表格中形成一个矩形区域，通过识别该区域的四个顶点来识别三元组。(1) 标签“TL”是区域的左上顶点，代表头部实体和尾部实体的起始位置；(2) 标签“TR”是区域的右上顶点，代表头部实体的起始位置和尾部实体的结束位置；(3) 标签“BL”是区域的左下顶点，代表头部实体的结束位置和尾部实体的起始位置；(4) 标签“BR”是区域的右下顶点，代表头部实体和尾部实体的结束位置。该文提出的方法可以解决三元组的重叠问题，包括SEO（单一实体重叠）和EPO（实体对重叠），当一个实体仅由一个标记组成时，相应的标记对具有多个标签。因此，该文采用多标签分类来处理这种情况。对于关系r_m 下所有顶点对应的每个标记对的概率分数计算如下：

$$\text{Score}_{ijm}^{TL} = \sigma\left(W_{r_m} W_{TL} q(w_i, w_j) + b_{r_m}^{TL}\right) \tag{5}$$

$$\text{Score}_{ijm}^{TR} = \sigma\left(W_{r_m} W_{TR} q(w_i, w_j) + b_{r_m}^{TR}\right) \tag{6}$$

$$\text{Score}_{ijm}^{BL} = \sigma\left(W_{r_m} W_{BL} q(w_i, w_j) + b_{r_m}^{BL}\right) \tag{7}$$

$$\text{Score}_{ijm}^{BR} = \sigma\left(W_{r_m} W_{BR} q(w_i, w_j) + b_{r_m}^{BR}\right) \tag{8}$$

其中：r_m 表示实体关系的类别，σ 表示sigmoid 函数，用于将输出压缩到0 和1 之间，用于二分类问题中将输出转换为概率。$W_{r_m}, W_{TL}, W_{TR}, W_{BL}, W_{BR}$ 是与特定关系类别r_m 相关的权重矩阵，$b_{r_m}^{TL}, b_{r_m}^{TR}, b_{r_m}^{BL}, b_{r_m}^{BR}$ 是与特定关系类别r_m 相关的偏置项，这些权重和偏置项都是可学习的参数。$q(w_i, w_j)$ 表示token 对(w_i, w_j) 的最终表示，Score 是(w_i, w_j) 对应于不同顶点标签(TL, TR, BL, BR) 的概率分数。当某个标签的概率分数高于设定的阈值γ 时，(w_i, w_j) 将被分配该标签，如果所有标签的概率分数都低于阈值γ，则(w_i, w_j) 不会被分配任何标签。

3.4 损失函数

该文使用BCE损失作为令牌对标签预测任务的损失函数。该文模型的损失函数定义如下：

$$\mathcal{L} = \frac{1}{4 \times L \times L \times N} \sum_{D \in \mathcal{R}} \sum_{i=1}^{L} \sum_{j=1}^{L} \sum_{m=1}^{N} BCE_{ijm}^{D} \tag{9}$$

$$BCE_{ijm}^{D} = y_{ijm}^{D} \log(Score_{ijm}^{D}) + (1 - y_{ijm}^{D}) \log(1 - Score_{ijm}^{D}) \tag{10}$$

其中$R = \{TL, TR, BL, BR\}$，N 表示实体关系的数量。如果token对(w_i, w_j) 在类别r_m 下被分配了标签’TL’ 和’BL’，那么$y_{ijm}^{TL} = 1, y_{ijm}^{TR} = 0, y_{ijm}^{BL} = 1$ 和$y_{ijm}^{BR} = 0$。

3.5 双向解码

算法预测三元组的顶点，解码目标是基于预测结果提取三元组。为了解决实体嵌套问题，该方法同时从两个方向进行解码，即（TL, TR, BR）和（BR, BL, TL），然后将提取的三元组合并为最终结果。以方向（TL, TR, BR）为例，从标记为“TL”的标记对开始，然后按照这个方向找到相应的最近标记为“TR”和“BR”的标记对。通过识别这些顶点，确认三元组的区域并提取三元组。

具体步骤如下:

1.**起始点（TL）：**首先识别并选择所有标记为“TL”（左上角）的标记对作为起始点;

2.**寻找TR（右上角）：**从每个“TL”标记对出发，按照（TL, TR, BR）的方向，寻找最近的标记为“TR”（右上角）的标记对;

3.**寻找BR（右下角）：**接着，从每个“TR”标记对出发，寻找最近的标记为“BR”（右下角）的标记对。

4.**确认区域并提取三元组：**通过识别“TL”, “TR”, 和“BR”这三个顶点，能够确定三元组的区域，并据此提取出三元组。

5.**双向解码：**为了更全面地提取三元组，该方法不仅从（TL, TR, BR）方向进行解码，还从（BR, BL, TL）方向进行解码，即从右下角开始，寻找左下角（BL）和左上角（TL）。

6.**合并结果：**最后，将两个方向解码得到的三元组结果合并，形成并集，作为最终的提取结果。

这种双向解码的方法有助于更准确地处理实体嵌套的情况，确保三元组的完整性和准确性。

4 实验

4.1 数据集

本次实验所采用的数据集来自聚法网上公开的盗窃、诈骗和涉毒类案件文本，在公开司法领域数据集CAIL 2022和CAIL 2023上进行进一步扩充和标注，下面将依次介绍这三类数据集的关系定义。

盗窃类数据集的训练集，验证集和测试集分别有4168，520和521条实例，在该数据集中一共定义了四种关系：“盗窃”是指嫌疑人盗窃物品的行为，头尾实体分别是嫌疑人和被盗窃的物品；“所有”表示受害人与被盗窃物品之间的所有权关系，头尾实体分别是受害人和被盗窃物品；“交易”是指嫌疑人处置被盗物品的去向，头尾实体分别是嫌疑人和销赃去向；“共犯”是犯罪嫌疑人之间合作关系的集中体现，犯罪主体之间存在共同犯罪关系，盗窃数据集的具体信息如表1所示。

Table 1: 盗窃数据集关系数量统计表

关系	盗窃	所有	交易	共犯
Train	2941	2143	77	730
Valid	337	230	10	98
Test	365	260	7	70
All	3643	2633	94	898

诈骗类数据集的训练集，验证集和测试集分别有1216，152和152条实例，在该数据集一共定义了六种关系：“获利”是指嫌疑人通过诈骗获得的钱财或者物品，头尾实体分别是嫌疑人和钱财或者物品；“损失”是指被害人损失的钱财或者物品，头尾实体分别是被害人和钱财或者物品；“利用”是指嫌疑人利用某种物品实施诈骗的行为，头尾实体分别是嫌疑人和物品；“购买”是指被害人购买某种物品而被诈骗的行为，头尾实体分别是班被害人和物品；“诈骗”是指嫌疑人对被害人实施诈骗的行为，头尾实体分别是嫌疑人和被害人，“被骗”是指被害人被诈骗的过程，头尾实体分别是被害人和嫌疑人，诈骗数据集的具体信息如表2所示。

涉毒类数据集的训练集，验证集和测试集分别有1242，154和152条实例，在该数据集一共定义了四种关系：“贩毒”是指嫌疑人贩卖毒品的行为，头尾实体分别是嫌疑人和毒品；“贩卖”是指嫌疑人之间交易毒品的行为，头尾实体分别是贩卖和购买的嫌疑人；“持有”是指嫌疑人

Table 2: 诈骗数据集关系数量统计表

关系	获利	损失	利用	购买	诈骗	被骗
Train	1398	1270	92	79	1155	1157
Valid	174	165	10	12	165	165
Test	196	170	11	10	163	161
All	1768	1605	113	101	1483	1483

吸食持有毒品的行为，头尾实体分别是嫌疑人和毒品；“容留”描述了一种为他人提供吸毒场所的犯罪行为，头尾实体分别为是提供场所的嫌疑人和在该场所吸毒的嫌疑人，涉毒数据集的具体信息如表3所示。

Table 3: 涉毒数据集关系数量统计表

关系	贩毒	贩卖	持有	容留
Train	719	769	409	764
Valid	96	98	55	107
Test	90	94	41	98
All	905	961	505	969

4.2 评价指标

在评估模型时，一般将得到的结果分成四种实验样本，分别是：真实正例(True Positive，TP)、真实负例(True Negative，TN)、错误正例(False Positive，FP)和错误负例(False Negative，FN)。其中，TP 表示模型预测为正例且真实值为正例的样本数，TN表示模型预测为负例且真实值为负例的样本数，FP 表示模型预测为正例但真实值是负例的样本数，FN 表示模型预测为负例但真实值是正例的样本数。

本次实验对模型的评估指标采用了一种常用的方法，即用准确率、召回率和F_1值对模型进行性能评估的方法。精确率（Precision）用于计算预测值是正例的样本中预测正确的样本数量占比，代表了模型的查准率；召回率（Recall）用于计算预测值是正例的样本中预测正确的正例样本数量占比，代表了模型的查全率；F_1值是精确率和召回率的调和平均值，用F_1值来代表模型的最终效果好坏。

$$Precision = TP/(TP + FP) \tag{11}$$

$$Recall = TP/(TP + FN) \tag{12}$$

$$Precision = 2 * Precision * Recall/(Precision + Recall) \tag{13}$$

4.3 主实验结果

Table 4: 单罪名模型性能对比

模型	盗窃			诈骗			涉毒		
	P	R	F_1	P	R	F_1	P	R	F_1
BiLSTM	63.0	53.5	57.7	82.2	67.9	74.4	80.1	66.5	72.7
Tplinker	82.1	79.4	80.7	91.5	83.4	87.3	84.7	82.9	83.8
OD-RTE	85.7	84.4	85.0	**94.7**	88.9	91.8	86.4	**88.1**	87.2
OURS	**86.4**	**84.5**	**85.4**	**94.7**	**89.8**	**92.2**	**89.6**	86.1	**87.8**

为了评估该文的模型，将该模型与一些基线方法进行了比较。在实验中使用的预训练模型是BERT-base-chinese，该模型适合处理实体名称和关系中的中文信息，并且该模型使用字符级别和词级别混合的输入表示，避免了传统中文分词（如Jieba）可能引入的边界错误。

对于流水线方法，该文选择使用BiLSTM先抽取实体后抽取关系的方法，同时也选择了两种先进的适用于中文数据集的填表法：TPLinker (Wang et al., , 2020)和OD-RTE (Ning et al., , 2023)。

单个罪名的实验结果如表4所示，从中可以看出，该文提出的方法优于所有基线，在所有数据集的F_1分数方面都达到了最先进的性能。

根据实验结果分析：与流水线方法相比，该文提出的方法在三元组抽取上的性能有了显著的提高。这种改进归因于该方法能够充分利用来自两个子任务的信息并减轻错误传播。与TPLinker相比，该方法仍然具有优越的性能。这验证了基于三元组区域四个顶点的联合抽取方法可以更好地利用三元组的区域信息。与OD-RTE相比证明该方法可以利用距离嵌入和扩展卷积来捕捉上下文和空间关系，从而更好地抽取三元组。

Table 5: 多罪名模型性能对比

模型	精确率(P)	召回率(R)	$\mathbf{F}_1$
BiLSTM	75.1	57.4	65.1
Tplinker	82.9	80.3	81.6
OD-RTE	87.6	85.2	86.4
OURS	**88.0**	**86.8**	**87.4**

多罪名的实验结果如表5所示，从中可以看出，该文提出的方法优于所有基线，在整体观察数据后，发现将三个数据集集中实验后，性能相较于涉毒罪和诈骗罪有所下降，相较于盗窃罪有所上升，导致这种现象的原因可能是涉毒最和诈骗罪数据集比较少，模型偏向于盗窃类案件学习。

4.4 通用数据集实验结果

为了显示该文提出的方法的鲁棒性和可迁移性，该文在一个广泛使用的通用数据集NYT (Riedel et al., , 2010)上评估了该文提出的方法。这个数据集包含两个不同的版本：版本1标注整个实体跨度，版本2仅标注实体的最后一个单词。在本实验中，该文使用版本2进行实验，基线模型选择了三种填表法：TPLinker (Wang et al., , 2020)、OneRel (Shang et al., , 2022)和OD-RTE (Ning et al., , 2023)，实验结果如表六所示。

Table 6: 通用数据集模型性能对比

模型	精确率(P)	召回率(R)	$\mathbf{F}_1$
Tplinker	91.3	92.5	91.9
OneRel	92.8	92.9	92.8
OD-RTE	93.5	**93.9**	93.7
OURS	**93.7**	**93.9**	**93.8**

与三种填表法相比，该文提出的模型仍然具有优越的性能。这证明基于区域四个顶点的三元组抽取方式可以更好地利用三元组的区域信息，利用距离嵌入和扩张卷积来捕获额外信息，也可以增强模型理解上下文和空间关系的能力。

4.5 大模型实验结果

该文选取Qwen-7B-Chat 和Qwen-plus 大语言模型作为实验对象，Qwen-7B-Chat 采用0-Shot 学习方法对盗窃罪、诈骗罪和涉毒罪三个单一罪名数据集进行三元组抽取实验，实验设计中，0-Shot 提示模板仅包含任务指令、关系定义和输出格式要求，未提供任何标注示例；Qwen-plus采用5-Shot 学习方法对盗窃罪、诈骗罪和涉毒罪三个单一罪名数据集进行三元组抽取实验，实验设计中，5-Shot 提示模板不仅包含上述任务指令、关系定义和输出格式要求，还为每种关系定义提供了5个标注示例。实验结果表7所示，实验结果表明，大语言模型虽然在通用领域的NLP任务中表现出卓越的性能，但本方法在司法领域结构化知识抽取任务中展现出显著的性能优势。

Table 7: 大模型在单罪名数据集上的性能对比

模型	盗窃			诈骗			涉毒		
	P	R	F_1	P	R	F_1	P	R	F_1
Qwen-7B-Chat	9.8	7.1	8.4	25.3	20.7	22.8	21.6	17.7	19.5
Qwen-plus	58.6	50.8	54.5	76.5	72.5	74.4	77.8	71.2	73.4
OURS	**86.4**	**84.5**	**85.4**	**94.7**	**89.8**	**92.2**	**89.6**	**86.1**	**87.8**

4.6 案例分析

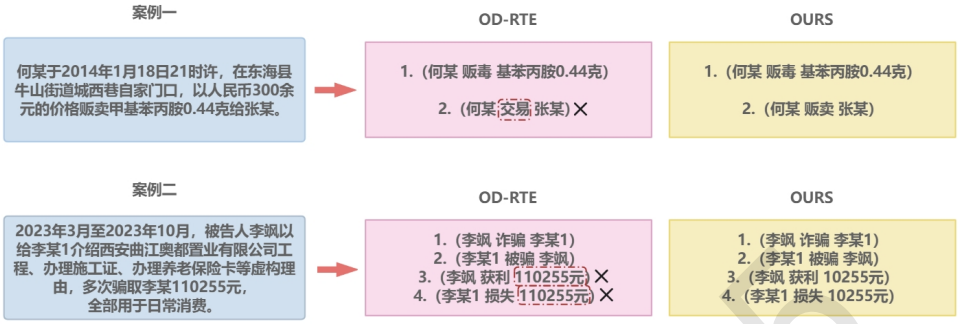

Figure 4: 错误案例

对基线方法中性能最好的OD-RTE方法 (Ning et al., , 2023)的错误案例进行了分析，并与该文的模型的识别结果进行了对比，如图四所示。对于实例1，OD-RTE方法错误地将何某和张某的关系计算为盗窃类数据中的“交易”关系，该文的模型正确地识别出两者为“贩卖”关系，说明不同罪名的案件会对彼此产生了干扰，相比之下，该文的模型能够较好地减轻不同数据集给彼此带来的噪声，正确地识别出文本中所包含的三元组。对于实例2，OD-RTE方法错误地将尾实体10255元识别为110255元，该文的模型正确地识别出尾实体，说明OD-RTE方法没有精确地识别出准确的实体边界，该文的模型融合距离感知嵌入技术与多尺度扩张卷积模块，能够有效捕获文本序列中的长距离依赖特征及空间结构化信息，显著增强了模型对上下文语义关联。

该文发现错误多发于实体之间距离过长，导致模型没有识别出正确的关系。因为算力有限，本文提出的模型对于输入的句子限制了token的数量，少数样例因为token过多被提前截断，从而无法识别出正确的关系，对于这个问题，在未来的工作中进一步改进算法。

5 总结

本文的核心工作聚焦于从司法文本中精准抽取三元组，以有效应对重叠三元组抽取的挑战，并为数据资源相对匮乏的司法领域研究提供有力支持。具体而言，研究团队首先通过人工标注的方式，构建了一个涵盖多种罪名的司法领域数据集。在此基础上，通过确定每个三元组所对应的四个关键顶点，实现了实体关系三元组的提取。进一步地，研究引入了距离嵌入技术，以充分利用实体对之间距离所蕴含的信息；同时，采用扩展卷积策略，增强了模型对多尺度上下文信息的捕捉能力。实验结果表明，所提出的方法在司法数据集上取得了SOTA的性能表现。

致谢

本研究得到国家自然科学基金项目（62276043, 62302076）资助。

参考文献

J. Li, H. Fei, J. Liu, S. Wu, M. Zhang, C. Teng, D. Ji, and F. Li. 2022. *Unified Named Entity Recognition as Word-Word Relation Classification*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10965–10973.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. *Kernel methods for relation extraction. Journal of machine learning research*, 3(Feb):1083–1106.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. *Exploring various knowledge in relation extraction.* In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 427–434.

Yee Seng Chan and Dan Roth. 2011. *Exploiting syntactico-semantic structures for relation extraction.* In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560.

Feiliang Ren, Longhui Zhang, Xiaofeng Zhao, Shujuan Yin, Shilei Liu, and Bochao Li. 2022. *A simple but effective bidirectional framework for relational triple extraction.* In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining*, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022, pages 824–832.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. *Joint extraction of entities and relations based on a novel tagging scheme.* In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1227–1236.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. *A novel cascade binary tagging framework for relational triple extraction.* In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488.

Y. Wang, B. Yu, Y. Zhang, T. Liu, H. Zhu, and L. Sun. TPLinker: Singlestage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1572–1582.

Y.-M. Shang, H. Huang, and X. Mao. OneRel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11285–11293.

J. Ning, Z. Yang, Y. Sun, Z. Wang, and H. Lin. OD-RTE: A one-stage object detection framework for relational triple extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 11120–11135.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. Prgc: Potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6225–6235.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 506–514.

Daojian Zeng, Haoran Zhang, and Qianying Liu. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2020, pp. 9507–9514.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Mosha Chen, Chuanqi Tan, Fei Huang, and Huajun Chen. Contrastive triple extraction with generative transformer. In *Proceedings of the AAAI conference on artificial intelligence*, 2021, pp. 14257–14265.

Yu-Ming Shang, Heyan Huang, Xin Sun, Wei Wei, and Xian-Ling Mao. Relational triple extraction: One step is enough. *arXiv preprint arXiv:2205.05270*, 2022.

Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., & Kurohashi, S. GPT-RE: in-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*, 2023.

He, M., & Bai, Y. (2023, October). *LAL-JER: Label-Aware Learning for Adaptive Joint Entity and Relation Extraction with LLM data augmentation.* In *Proceedings of the 2023 International Conference on Communication Network and Machine Learning*, 2023, pp. 414–419.

Xue, L., Zhang, D., Dong, Y., & Tang, J. AutoRE: Document-level relation extraction with large language models. *arXiv preprint arXiv:2403.14888*, 2024.