基于 XLNet 的法律核心要素识别应用

张 棋[1] 胡亚谦[1] 赵 耀[1] 吉艳利[1] 李建歧[1] 洪通亮[1]
ZHANG Qi HU Yaqian ZHAO Yao JI Yanli LI Jianqi HONG Tongliang

摘 要 法律核心要素的精准识别，有助于提升案件判决的准确度及效率。然而，现有深度学习方法的准确率受限于案件信息的复杂度，通常无法有效提取事实描述中的上下文关系。对此，提出了将极长网络（extra-long network，XLNet）应用于法律核心要素识别。利用 CAIL2019 提供的要素识别任务数据集进行分案由训练和预测，分案由 divorce、labor 和 loan 下性能评估指标 F_1 值分别达到 71.79%、57.31%、72.79%，均为最佳，比第二名分别高 4.8、20.4、10.0 个百分点。实验结果表明，XLNet 模型在法律核心要素的多标签二分类任务中具有良好表现。

关键词 要素识别；多标签分类；极长网络

doi：10.3969/j.issn.1672-9528.2024.09.011

0 引言

法律作为社会规范的重要组成部分，对于维护社会秩序、保障公民权益、促进公平正义具有至关重要的作用。在法律实践中，准确识别法律核心要素是实现法律正确适用、确保法律效果的关键环节。

法律核心要素是指在以裁判文书为主的法律文书共有的案件基本维度之外的、描述重要法律事实的内容，需要结合法律知识背景进行理解的、可归属于业务预设的事实描述要素类别[1]。法律核心要素是厘清法律争议问题的重要考察因素。高质量的法律核心要素识别，有助于实现对类案的快速、精准检索和推荐，具有指导司法审判、统一法律适用、细化裁判标准、推动理论研究等重要作用。在司法实践中，深度学习技术尚不能游刃有余地发挥作用，以至于无法在应用中被足够信任。这主要是由于现有的法律要素识别在关注

1. 中国司法大数据研究院有限公司 北京 100041
[基金项目] 国家重点研发计划项目“司法知识推理与服务引擎构建技术”（2021YFC3340103）

然有一些方面可以进一步改进和探索，如引入更多的环境因素、优化算法参数等。未来将继续深入研究，进一步完善算法，以满足实际柑橘采摘需求。

参考文献：

[1] 郑文赢 . 移动机器人技术现状与展望 [J]. 信息记录材料 , 2020, 21(10):24-25.
[2] 郑娟毅 , 付姣姣 , 程秀琦 . 面向物流车辆路径规划的自适应蚁群算法 [J]. 计算机仿真 ,2021,38(4):477-482.
[3] 王猛 , 邢关生 . 基于改进蚁群算法的机器人路径规划 [J]. 电子测量技术 ,2020,43(24):52-56.
[4] 康玉祥 , 姜春英 , 秦运海 , 等 . 基于改进 PSO 算法的机器人路径规划及实验 [J]. 机器人 ,2020,42(1):71-78.
[5] 宋宇 , 张浩 , 程超 . 基于改进蚁群算法的物流机器人路径规划 [J]. 现代制造工程 ,2022(11):35-40+47.
[6] 凌海峰 , 谷俊辉 . 带软时间窗的多车场开放式车辆调度 [J]. 计算机工程与应用 ,2017,53(14):232-239.
[7] 戚远航 , 蔡延光 , 黄戈文 , 等 . 带固定半径近邻搜索 3-opt 的离散烟花算法求解旅行商问题 [J]. 计算机应用研究 , 2021, 38(6):1642-1647.
[8] 孙东艳 . 基于模拟退火 - 蚁群算法的输油管道碳排放量优化研究 [J]. 石油石化节能与计量 ,2024,14(3):53-57+69.

【作者简介】

陈淑玲（2003—），女，江西九江人，本科，研究方向：计算机算法、人工智能。

王士信（1979—），男，江西南昌人，硕士，高级工程师，研究方向：软件工程、计算机算法、人工智能。

（收稿日期：2024-06-04）

核心要素的识别之余，对包含法律领域特有的基本案件要素关注不足，导致识别结论存在一定偏差。同时，现有的识别法律关键要素的方法考察的是要素所在位置本身的含义，缺少对上下文语义的考察，影响了要素语义在全局语义下的理解[2]。

近年来司法领域的人工智能研究，核心要素识别作为新兴热点，被放入文本分类进行问题研究，现已经历了以下三个阶段：（1）基于规则方法；（2）基于统计的机器学习；（3）基于深度学习。在基于规则的方法研究中，刘炜等人[3]根据规则进行要素的推理、填充和修正，但是受限于推理规则的制定以及事件的结构。Zhang 等人[4]介绍了基于 RoBERTa-GlobalPointer 的法律文书命名实体识别，但是对特定领域数据存在依赖。在基于统计的机器学习中，大多是进行神经网络学习，Yang 等人[5]将长短时记忆网络应用在多标签文本分类上。在基于深度学习中，学者们采用了更为复杂的神经网络结构用于提升模型的准确率，E Ahmadzadeh 等人[6]将门控循环单元和长短时记忆网络进行融合，将其应用到文本分类任务中。然而，基于法律案件的独有特点，现有的核心要素识别方法提取到的内容无法涉及信息的上下文关联性，使得在法律领域的实际应用中受限。Lyu 等人[7]利用 BERT 模型和强化学习进行针对刑事案件进行要素识别，但是在处理复杂逻辑时的泛化能力受限。Li 等人[8]利用深度学习技术应用于事实描述中提取决策要素，受限于文本语境理解。

为此，本文将极长网络（extra-long network，XLNet）应用于核心要素识别中。采用 CAIL2019 要素识别任务公布的数据，开展基于法律的核心要素识别研究。通过实验结果表明，本文所提方法相比已有方法，获得了更高的识别性能，具有良好的应用前景。

1 法律核心要素

法律中的核心要素是指在法律适用过程中，对构成法律问题的关键要素进行准确界定和深入剖析的过程。这些要素包括但不限于法律事实、法律关系、法律规则等，它们是构成法律问题的基本单位，也是法律推理和法律解释的基础。

法律核心要素识别的过程实际上是对法律问题进行精细化、系统化的处理过程。它要求我们在处理法律问题时，能够深入剖析问题的本质和关键所在，准确把握问题的核心要素，从而为问题的解决提供有力的法律支撑。

（1）提高法律适用的准确性

法律适用的准确性是法律实践的基本要求。法律核心要素识别作为法律适用的关键环节，对于确保法律适用的准确性具有重要意义。通过对法律核心要素的准确识别，人们能够更加清晰地理解法律问题的本质和关键所在，避免对法律问题的误解和误判。同时，核心要素的识别也有助于人们准确把握法律规则的适用范围和条件，确保法律规则的正确适用。

（2）促进司法公正与效率

司法公正与效率是法律实践的重要目标。法律核心要素识别对于实现这一目标具有积极的推动作用。通过对法律核心要素的深入剖析和准确界定，能够更加客观地分析案件事实，避免主观臆断和偏见的影响，从而确保司法判决的公正性。同时，核心要素的识别也有助于快速把握案件的关键点，避免在次要问题上纠缠不清，进而提高司法效率。

（3）增强法律解释的说服力

法律解释是法律适用过程中的重要环节。法律核心要素识别对于增强法律解释的说服力具有重要作用。通过对法律核心要素的深入剖析和准确界定，能够更加清晰地阐述法律规则的含义和适用条件，使法律解释更加具有说服力和可信度。这有助于增强公众对法律的理解和认同，提高法律的权威性和公信力。

（4）推动法律体系的完善与发展

法律核心要素识别对于推动法律体系的完善与发展具有重要意义。通过对法律核心要素的深入研究和识别，能够发现现有法律体系中存在的不足和缺陷，为法律的修订和完善提供有益的参考。同时，核心要素的识别也有助于探索新的法律理念和解决方案，推动法律体系的不断创新和发展。

在现代社会，随着科技的发展和社会的变迁，新的法律问题不断涌现。通过对这些新问题中核心要素的识别和分析，能够发现现有法律体系的滞后和不足，应推动法律体系与时俱进。同时，核心要素的识别也有助于借鉴其他国家和地区的法律经验，吸收先进的法律理念和技术手段，为法律体系的完善和发展提供新的思路和方向。

核心要素的识别是通过深入分析问题的背景、原因及影响，识别出影响问题、情况或决策的重要要素。此类要素通常直接或间接地决定问题的发展方向和最终结果。核心要素识别逻辑较为复杂，需要运用系统思维、逻辑思维和创造性思维等多种思维方式，进而准确地识别并把握问题的核心。

2 基于 XLNet 的法律核心要素识别模型

本文使用 XLNet[9] 模型来进行法律核心要素识别，模型将案件事实描述作为输入，通过模型训练来有效识别指定的核心要素标签。XLNet 结合了 Transformer-XL 和 BERT 的优势，通过自回归方式建模双向上下文信息，同时克服了 BERT 中存在的预训练和微调不一致问题，更灵活地建模上下文关系，从而提供更高的分类准确性。

具体来说，XLNet 与 BERT[10] 模型不同，XLNet 采用自

回归方法对序列进行建模。最大化所有可能的词序列联合概率，在捕捉上下文信息的同时又可以避免 BERT 的独立掩码问题。给定一个序列 $\boldsymbol{X}=(x_1, x_2, \cdots, x_T)$，XLNet 通过最大化所有可能排列的目标词的对数似然来训练：

$$\max_{\theta} \mathbb{E}_{\mathbf{z}\sim P(Z)}\left[\sum_{t=1}^{T}\log P\left(x_{Zt} \mid X_{\mathbf{Z}<t};\theta\right)\right] \tag{1}$$

式中：Z 表示序列的所有可能排列，Z_t 是排列后的第 t 个位置。

XLNet 集成了 Transformer-XL 架构，通过引入相对位置编码和记忆机制，捕捉词语间的相对位置关系和处理长序列时保持的记忆。给定两个位置 i 和 j，其相对位置编码为：

$$a_{ij}=W_{\text{pos}}(i-j) \tag{2}$$

式中：$\boldsymbol{W}_{\text{pos}}$ 是位置嵌入矩阵。

XLNet 中的注意力机制与 Transformer 类似。给定查询 Q、键 K 和值 V，计算注意力权重：

$$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{\sqrt{d_k}}+a_{ij})V \tag{3}$$

式中：d_k 是键的维度，a_{ij} 是相对位置编码。

最后，XLNet 在段落级别进行预训练，通过记忆机制捕捉长距离依赖。给定一个段落，使用记忆单元来存储之前时间步的信息：

$$M^l=\left[H^l_{\text{prev}};H^l\right] \tag{4}$$

式中：H_{prev} 是前一时间步的隐藏状态，H_l 是当前时间步的隐藏状态。

因此，该模型在处理长距离依赖的关系上具有明显的优势，可以学习到复杂的语义信息，实现对案件事实描述中上下文信息的有效捕捉识别，从而达到核心要素识别准确率提升的目的。

3 实验结果和分析

3.1 数据集

本文采用CAIL2019要素识别任务提供的数据进行训练，其中包含三个案由：劳动争议纠纷、离婚纠纷、民间借贷纠纷。各案由的数据量均达到 1.5 万条以上，每条数据均由案件事实描述、标签组成，各案由下的标签尽管内容不同，但均为 20 个。

每条记录结构如下所示，均包含标签的命中信息，以及案件事实描述信息：{ “labels": ["DV10", "DV3"], "sentence": " 原告龙某甲诉称，原告于 2008 年 6 月经他人介绍与被告认识，在媒人撮合下，在衡阳县体育花苑购房一套，房主系被告，当时购房款、装修费用全部是借款，原告婚后省吃俭用，将借款还清。"}。

3.2 评价指标

本文中采用准确率、召回率以及 F_1 分数对模型进行评价判断。这三个指标在机器学习领域中被广泛应用，它们各自从不同的角度反映了模型的性能，综合使用可以更加全面、准确地评估模型的优劣。

（1）准确率（Accuracy）

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{5}$$

Accuracy 是最直观且易于理解的性能评估指标之一，指模型正确分类的样本数占总样本数的比例。它反映了模型分类的精确程度。它通过简单的计算方法，可以告诉人们模型在所有样本上的整体表现，使得结果直观易懂，但在样本类别分布不均衡的情况下，准确率可能会产生误导。其不仅适用于二分类问题，也可以更广泛地应用于多类别分类和回归问题。

（2）召回率（Recall）

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{6}$$

Recall 是指模型正确识别出的正样本数占实际正样本数的比例。它衡量了模型找出所有正样本的能力。在一些对漏检率要求较高的应用场景中，如疾病检测、安全监控等，召回率是一个非常重要的指标。通过关注召回率，可以提示漏报风险的高低，更多地发现潜在的正例。

（3）F_1 分数

$$F_1=\frac{2\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{7}$$

F_1 是结合 Precision 和 Recall 结果的综合指标，它综合考虑了模型在精确度和召回率两个方面的表现，在不均衡的数据集上全面评估模型性能。它具有较高的平衡性，通过平衡 Precision 和 Recall 的数值，在模型有较高精确率的同时，仍然保留较高的召回率。F_1 分数越高，说明模型在这两个指标上表现得越好，整体性能越优。

3.3 实验设置

3.3.1 数据预处理

针对三种案由分别进行数据的预处理工作，处理流程均为如下步骤：首先，针对数据集的所有数据进行乱序排序；之后，获取 80% 的数据作为训练数据集，剩余 20% 数据作为验证数据集。训练数据为中文形式，采用结巴分词。

3.3.2 参数设置

本实验基于 huggingface 的预训练模型 chinese-xlnet-base 进行训练，并在此基础上进行以下五项内容的设置。

（1）基于不同案由限制生成编码序列的最大长度（max_length）：离婚纠纷的长度为 110，劳动争议纠纷的长度为 135，民间借贷纠纷的长度为 187。

（2）将控制编码后序列的填充基于不同案由设置为 max_length 的长度。

（3）输出的数据类型设定为返回 PyTorch 张量。

（4）输出的内容中需要包括返回的注意力掩码。

（5）控制编码前根据指定的长度 max_length 进行截断。

3.4 实验结果

本文采用基于 Transformer 的 XLNet 模型与机器学习中的随机森林（random forest，RF）以及 Transformer 的 BERT 模型进行对比实验。结果如表 1 所示，在不同案由下的三项评价指标中，XLNet 均表现优异。

表 1 各案由模型结果表

案由	模型	Accuracy	Recall	F_1
divorce	RF	80.81%	59.55%	67.04%
	BERT	79.54%	50.49%	52.21%
	XLNet	**81.22%**	**70.91%**	**71.79%**
labor	RF	75.71%	24.37%	36.88%
	BERT	**84.92%**	33.92%	36.16%
	XLNet	78.64%	**50.72%**	**57.31%**
loan	RF	79.78%	49.84%	62.83%
	BERT	75.85%	36.18%	37.25%
	XLNet	**87.25%**	**67.19%**	**72.79%**

4 总结与展望

在法律实践中，核心要素的识别是确保法律正确适用、实现法律效果的关键环节，核心要素识别的准确性直接关系到法律适用的正确性和公正性。本文将 XLNet 模型应用于要素识别的研究中，从三项评估指标的结果可以看出，XLNet 模型可以较好地理解上下文复杂关系，有效地提高识别的准确率，从而达到通过案件事实描述识别指定标签的目的。

受数据量和质量的影响，不同案由下的评估结果具有较大的差异，未来可以在不同案由下准确率的均衡性方面进行提升。通过明确法律的核心要素，司法机构能够更准确地理解和适用法律，确保案件得到公正、公平的审理，进而提高司法效率，减少不必要的争议和误解。

参考文献：

[1] 王得贤 . 法律文书中的要素识别方法研究 [D]. 太原 : 山西大学 , 2020.

[2] 张和伟 . 基于深度学习的法律文本关键要素识别 [D]. 太原 : 太原理工大学 , 2022.

[3] 刘炜 , 刘菲京 , 王东 , 等 . 一种基于事件本体的文本事件要素提取方法 [J]. 中文信息学报 ,2016,30(4):167-175.

[4]ZHANG X, LUO X, WU J.A RoBERTa-GlobalPointer-Based method for named entity recognition of legal documents[C]//2023 International Joint Conference on Neural Networks (IJCNN).Piscataway:IEEE, 2023:1-8.

[5]YANG P, SUN X, LI W, et al.SGM:sequence generation model for multi-label classification[C]//27th international conference on computational linguistics,vol. 6.Stroudsburg, PA:Association for Computational Linguistics, 2018:3915-3926.

[6]AHMADZADEH E, KIM H, JEONG O, et al.A Deep Bidirectional LSTM-GRU network model for automated ciphertext classification[J].IEEE access, 2022,10:3228-3237.

[7]LYU Y, WANG Z, REN Z, et al.Improving legal judgment prediction through reinforced criminal element extraction[J]. Information processing & management, 2022(1):102780.

[8]LI X, LI Q.Deep learning for decision element extraction in fact description of legal documents[J].Advances in multimedia, 2022,2022:1-11.

[9]YANG Z, DAI Z, YANG Y, et al.XLNet: generalized autoregressive pretraining for language understanding[C]//Advances in Neural Information Processing Systems 32,Volume 8 of 20.Red Hook:Curran Associates, Inc., 2020:5730-5740.

[10]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL].(2017-06-12)[2024-04-22].https://arxiv.org/abs/1706.03762.

【作者简介】

张棋（1995—），女，北京人，硕士，工程师，研究方向：计算机与法律融合。

胡亚谦（1989—），男，安徽马鞍山人，博士，工程师，研究方向：司法大数据分析与应用。

赵耀（1992—），女，北京人，硕士，工程师，研究方向：法律信息化与数字化。

吉艳利（1995—），女，河北承德人，本科，工程师，研究方向：法律信息化与数字化。

李建歧（1988—），男，黑龙江绥化人，本科，工程师，研究方向：计算机在法律领域的应用。

洪通亮（1984—），男，陕西咸阳人，本科，高级工程师，研究方向：计算机在法律领域的应用。

（收稿日期：2024-06-04）