计算机科学
COMPUTER SCIENCE

跨模型协同的法律文本相关性无监督表征方法研究

许身健

引用本文

许身健. 跨模型协同的法律文本相关性无监督表征方法研究[J]. 计算机科学, 2026, 53(4): 356-365.

XU Shenjian. Cross-model Collaborative Unsupervised Representation Method for Legal Texts[J]. Computer Science, 2026, 53(4): 356-365.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.251100003

跨模型协同的法律文本相关性无监督表征方法研究

许身健
中国政法大学数字社会治理研究院　北京 100088

摘　要　法律文本表征是法律人工智能系统的基础,其质量直接影响法条预测、案例检索等下游任务。然而,法律文本在专业术语、篇章结构及推理逻辑上的复杂性,使得通用预训练模型易产生语义偏移。开源模型领域知识不足;而闭源模型虽具备较强的理解能力,却难以直接复用其内部表征。针对上述问题,提出一种跨模型协同增强的法律文本表征方法(Cross-Model Collaborative Legal Representation,CMCLR),通过构建开源模型与闭源模型的协同框架,引入闭源模型的领域感知能力,以增强开源模型的法律语义建模能力。具体而言,利用闭源模型对法律文本进行动态分块与关键段落识别,提取结构化语义信息,并在协同约束下指导开源模型学习可解释、可训练的文本表征;同时,引入无监督聚类对段落级嵌入进行结构建模,以捕捉法律文本间的潜在语义关联。实验在 CAIL2018 法条分类数据集及其派生子集上进行,结果表明,CMCLR 在 CAIL2018 法条分类任务上取得 90.3% 的准确率,较代表性基线方法提升 2.4 个百分点,并在不同数据规模与场景设置下均表现出良好的稳定性与泛化能力。实验结果验证了跨模型协同表征学习在法律文本深层语义建模中的有效性。

关键词:法律文本;表征;文本相关性;法律人工智能;预训练模型;跨模型协同增强的法律文本表征方法
中图分类号　TP391

Cross-model Collaborative Unsupervised Representation Method for Legal Texts

XU Shenjian
Digital Society Governance Institute,China University of Political Science and Law,Beijing 100088,China

Abstract　Legal text representation is a fundamental component of legal artificial intelligence systems,directly affecting the performance of downstream tasks such as legal article prediction and case retrieval. However,the professional terminology,complex structure,and reasoning patterns of legal texts often lead to semantic drift in general pre-trained models. Open-source models lack sufficient legal domain knowledge,while closed-source models,despite their strong semantic understanding capabilities,provide representations that are difficult to directly access and reuse. To address these challenges,this paper proposes a cross-model collaborative legal representation framework(CMCLR),which enables collaborative learning between open-source and closed-source models to enhance legal semantic modeling. Specifically,closed-source models are employed to perform dynamic text segmentation and key paragraph identification,producing structured domain-aware signals that guide the fine-tuning of open-source models under collaborative constraints. In addition,unsupervised clustering is introduced to model structural relationships among paragraph-level embeddings,capturing latent semantic associations between legal texts. Experiments conducted on the CAIL2018 legal article classification task demonstrate that CMCLR achieves an accuracy of 90.3%,outperforming representative baseline methods by 2.4 percentage points,while maintaining robust performance across different dataset scales and settings. These results confirm the effectiveness of cross-model collaborative representation learning for deep semantic modeling of legal texts.

Keywords　Legal text,Representation,Textual relevance,Legal artificial intelligence,Pretrained models,Cross-model collaborative legal representation(CMCLR)

1　引言

在人工智能技术加速渗透法律领域的背景下,法律人工智能系统的发展正深刻改变着法律实践的范式[1]。从案例检索到判决预测,从合同审查到合规监测,法律人工智能系统的核心功能均依赖于对法律文本的精准理解与表征[2]。法律文本表征学习作为法律人工智能的底层技术支撑,其质量直接决定了上层任务的性能表现。如何构建有效的法律文本表征模型,已成为推动法律智能化发展的关键科学问题。

现有的通用文本表征方法(如 BERT[3] 和 GPT[4] 系列模型)在处理法律文本时,面临严峻挑战。尽管这些模型在通用领域展现出强大的语义理解能力,但法律文本的专业性与复杂性导致其难以捕捉深层相关性[5]。法律文本间的相关性不仅涉及字面语义的匹配,更蕴含着法律概念的映射、条款逻辑的关联以及论证结构的呼应[6]。例如,在判断两个案件的相似性时,须综合考虑事实要素的对应关系、所适用法律条款的

到稿日期:2025-11-03　返修日期:2026-01-16
通信作者:许身健(xushenjian2006@aliyun.com)

解释以及司法推理的逻辑链条。这种多维度的相关性普遍存在于法律检索、判决预测等任务中,但传统表征方法由于缺乏领域知识的有效注入,难以实现精准建模[7]。

法律文本的特殊性为表征模型提出了三重挑战。(1)术语壁垒:法律领域包含大量专业性词汇,其语义边界与通用语境存在显著差异,须结合法律条文进行精确建模。(2)结构复杂性:法律文书通常包含事实陈述、证据分析、法律推理等多模块内容,模块间的逻辑关系对整体语义的影响须显式建模。(3)论证逻辑性:法律文本的核心价值在于其论证过程的严谨性,现有模型难以捕捉法律推理中的前提-结论关系及例外情形。这些特性导致通用模型在法律场景中普遍存在语义漂移现象,即模型在通用领域学习的知识与法律领域的实际需求发生了偏离。

法律文本的有效表征需要同时具备通用语义理解与领域知识建模的双重能力。大语言模型[8]凭借其强大的上下文学习能力,为应对这一挑战提供了新的可能。然而,现有大语言模型在法律领域的应用面临双重困境:(1)闭源模型(如 GPT-4o[9])虽能生成高质量的文本理解结果,但受限于 API 调用的黑箱机制,无法直接输出可解释的文本表征向量;(2)开源模型(如 LLaMA[10])虽支持特征提取,但其参数量与训练数据的局限性导致在专业领域(如法律)的语义表征能力不足。此外,无论是开源还是闭源模型,现有技术均难以捕捉法律文本间的深层相关性[11],这种相关性既包含案件事实与法律条款的映射关系,也涉及不同文本在论证逻辑上的呼应。

针对上述问题,本文提出一种跨模型协同增强的法律文本表征方法(CMCLR)。本文所指的"闭源模型不可表征",是由于模型参数与中间表征不可访问,其所蕴含的高层法律语义知识难以被显式获取并直接用于可训练的表征学习过程,从而限制了其在下游法律任务中的可复用性。为解决这一问题,CMCLR 构建了开源模型与闭源模型之间的协同学习框架。具体而言,首先利用闭源模型较强的语义理解能力,对法律文本进行动态分块与关键段落识别,将其隐含的领域语义以结构化形式外化为可观察的中间信号;随后,通过对开源模型进行针对性微调,使其在监督与协同约束下学习并吸收上述结构信息,从而将闭源模型的高层语义知识间接转换为可解释、可训练的表征向量;在此基础上,引入无监督聚类技术,对样本间的语义一致性进行建模,以进一步捕捉法律文本之间的隐含相关性模式。通过上述跨模型协同机制,本文方法在不直接依赖闭源模型参数或中间表征的前提下,有效缓解了闭源模型知识难以表征与复用的问题,同时提升了开源模型在法律领域中的语义建模能力,实现了对法律文本相关性的深度挖掘。

2 相关工作

早期的法律文本处理主要基于规则的专家系统[12]和浅层机器学习模型[13]。基于规则的专家系统依赖领域专家制定的一系列规则来处理法律文本。例如,在简单的合同审查场景中,专家根据合同常见条款和法律规定,制定出如"若合同中交货日期未明确规定,则存在一定法律风险"这样的规则。但这种方式存在明显的局限性,面对复杂多样的法律文本时,规则的制定和维护成本极高,且其难以适应新的法律条款和复杂的实际情况。浅层机器学习模型,如决策树[13]、支持向量机[14]等,在法律文本处理中也有应用。Kaufman 等采用 AdaBoost 决策树预测美国最高法院判决[15],他们通过人工提取案件中的关键属性(如案件涉及的法律领域、当事人的身份特征等)作为模型的输入。然而,这种方法严重依赖人工特征工程,需要耗费大量的人力和时间,且提取的特征难以全面涵盖法律文本的复杂语义和结构信息。同时,由于法律文本的特殊性,数据往往存在不平衡的问题,这会导致浅层机器学习模型的泛化能力较差,在实际应用中效果不佳。

随着深度学习的发展,预训练语言模型逐渐在法律文本处理中崭露头角[16-17]。预训练模型在大规模通用文本上进行预训练,学习到了丰富的语言知识和语义表征,然后通过在特定领域数据上进行微调,可以快速适应法律领域的需求。Chalkidis 等开发的 LEGAL-BERT[18],在 BERT 的基础上,利用大量美国、英国和欧盟的法院案件进行微调。在欧洲人权法院案件分类任务中,与传统方法相比,LEGAL-BERT 能够更好地捕捉文本中的语义特征,取得了显著的性能提升。后续研究进一步针对不同地区的法律文本进行优化,如 Paul 等提出的 InLegalBERT[19],专门针对印度法律文本进行结构感知预训练,通过对印度法律案件中的结构信息,如案件的不同部分(事实陈述、法律依据、判决结果等)进行学习,在处理印度法律文本时能够更好地理解文本的结构和语义。尽管如此,这些模型仍存在一些不足。一方面,法律术语具有专业性和复杂性,其语义往往与普通词汇有很大差异,并且在不同的法律语境中可能有不同的含义。现有的预训练模型虽然在一定程度上能够学习到法律术语的语义,但对这些术语之间的细微差别和复杂关系的建模还不够深入。例如,"不可抗力"在不同类型的合同法律文本中,其具体的界定和适用范围可能有所不同,预训练模型难以准确捕捉这种差异。另一方面,法律文本通常篇幅较长,现有的预训练模型在处理长文本时,往往采用截断或简单分块的策略。这种方式会导致文本中的关键信息丢失,无法充分利用文本的上下文信息,从而影响模型对长法律文本的理解和处理能力。

为了解决法律文本的长序列问题,研究者提出了多种层次化框架。Chalkidis 等提出的 HAT 模型[20],通过分层注意力机制,分别在句子级和文档级对文本进行处理。在句子级,注意力机制可以帮助模型聚焦于每个句子中的关键信息;在文档级,通过整合句子级的特征,获取整个文档的语义表征。这种方法在 ECtHR 数据集上取得了当时最优的性能,证明了层次化框架在处理长文本时的有效性。Prasad 等提出的框架则将文档划分为多个部分[21],从微调后的大语言模型的最后 4 层提取这些部分的嵌入,并生成结构信息,以表征文档不同部分的事实、论证、法律条款等。然而,现有方法在处理长文本时仍存在一些问题。多数方法采用固定的分块策略,这种策略没有考虑到文本的语义边界和逻辑结构,可能会将具有紧密逻辑关系的文本部分划分到不同的块中,导致信息的碎片化。例如,在一个复杂的法律案例文本中,关于某个关键事件的完整描述可能被分块策略拆分成多个部分,使得模型难以理解其完整的逻辑。此外,这些方法往往忽略了不同文

本之间的相关性，只是基于文本的嵌入特征进行表征，没有充分考虑法律领域的专业知识和规则，可能会识别出一些不合理的相关性，影响模型对法律文本的理解和处理。

此外，近年来国内相关领域对文本表示与语义编码问题的研究不断深化。Zhao 等对自然语言处理中的各种文本表示方法进行了系统综述，涵盖了从传统向量化方法到神经网络嵌入和预训练模型的技术演进，为文本语义表示提供了全面的理论支持[22]。在具体技术方向上，不少学者提出了针对中文语义嵌入、深度学习表示结构及图神经网络[23]等方法的改进和评估框架，展示了文本嵌入技术在分类、抽取和结构建模等 NLP 任务中的潜力。同时，有研究分析了预训练语言模型[24]在中文信息抽取任务中的表示能力，强调领域知识增强与语义一致性建模对提高特定领域（如法律文本）任务性能的重要性[25]。上述国内成果不仅反映了中文语义表示研究的趋势，也为本文中法律文本表征方法的设计提供了理论参考。

3 方法设计

3.1 整体框架

本文提出一种全新的法律文本表征方法，旨在针对法律文档的复杂语义及内在结构，构建一种高质量、综合性的文本表征，以支持法律判决预测和文本分类等下游任务。如图 1 所示，与传统方法不同，本文方法从法律领域微调开始，依托于开源大语言模型，通过智能文本采样与无监督聚类相结合的策略，提取并融合预训练特征与无监督特征，充分捕捉法律文本中的专业术语、论证逻辑及潜在结构信息。整个方法主要分为 4 个模块：法律领域微调、文本随机采样、无监督特征提取以及特征融合与下游任务建模。下面将从整体框架、各模块动机与设计、数学描述以及实现细节等方面进行详细介绍。

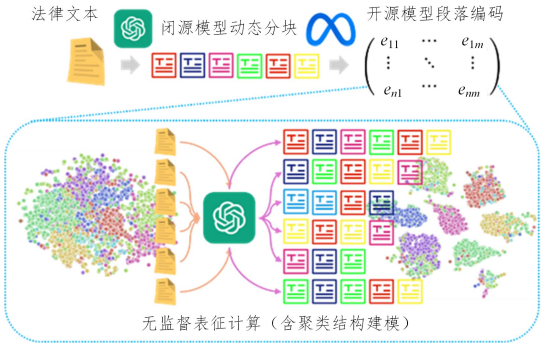

图 1 CMCLR 的整体框架

Fig. 1 Overall framework of CMCLR

3.2 无监督表征计算

为了使开源大语言模型能够充分理解法律领域中复杂的术语、规则及论证逻辑，首先采用大量法律和法规文本对其进行微调。设输入法律文档为 $D=\{t_1,t_2,\cdots,t_L\}$，其中 t_i 表示第 i 个 token，L 为文档总 token 数。在微调过程中，利用交叉熵损失函数对模型参数 θ 进行更新，从而使模型能够生成在法律领域具有更高判别力的表征。该微调过程不仅使模型能够捕捉专业领域内的语言特征，而且为后续文本采样和特征提取奠定了坚实的基础。

法律文档往往篇幅较长，直接输入整个文档会导致模型输入长度受限。此外，法律文档中往往包含大量冗余信息，但其关键信息可能仅分布在部分段落中。为了确保对文本信息的充分覆盖，本文方法引入性能良好的闭源大语言模型（如 GPT-4o）对全文进行语义理解与分析，并基于此，设计出一种随机采样策略。设定义采样函数 $S(*)$ 为对输入文档 D 进行处理的过程，其输出为关键信息段落集合，即连续且可能存在重叠的片段集合。

$$S(D)=\{s_1,s_2,\cdots,s_N\} \tag{1}$$

其中，N 为采样出来的片段总数。每个片段 s_i 表示由模型识别出的一个关键段落或句子，保留了原文档的标签 y（如用于法律判决预测或文本分类）。在采样过程中，结合语义重要性评分函数 $f(s)$ 对每个段落进行评分，同时引入随机性，以确保采样结果既能覆盖全文，又能突出重点信息。此过程确保即使存在重叠或部分冗余，所有关键信息也均能被有效提取。

对于上述采样得到的每个段落 s_j，本文方法采用微调后的开源大语言模型生成其嵌入表征。令函数 $F(s_j;\theta)$ 表示经过微调的开源大语言模型，则每个段落的表征向量为：

$$\boldsymbol{e}_j=F(s_j;\theta)\in\mathbb{R}^d \tag{2}$$

其中，d 为嵌入向量的维度。为了捕捉文本内部的潜在结构信息，对所有段落表征集合 $\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_M\}$ 先进行降维处理[26]，再通过聚类算法，将其划分为 K 个类别。为了将高维数据映射到低维空间，同时能够保留法律文本的局部和全局结构，构建数据的模糊拓扑表征，并优化低维嵌入来近似这种拓扑结构[27]。具体而言，首先对高维向量 $\boldsymbol{e}_j$ 构建近邻图，以此来近似数据所在的流形。对于每个 $\boldsymbol{e}_j$，计算其与其他向量的欧氏距离 $d_{ij}=\|\boldsymbol{e}_i-\boldsymbol{e}_j\|_2$，找到其 k 个最近邻向量 $\{\boldsymbol{e}_{j_1},\boldsymbol{e}_{j_2},\cdots,\boldsymbol{e}_{j_k}\}$。然后，将高维空间中的度量关系转换为模糊单纯集，以捕捉数据点之间的局部和全局连接。通过计算局部连续密度估计，为每个 $\boldsymbol{e}_j$ 定义一个局部尺度 σ_j，使得：

$$\sum_{i=1}^{n}\mathrm{e}^{-\frac{d_{ij}}{\sigma_j}}=\log_2 k \tag{3}$$

其中，σ_j 作为自适应参数来调整数据点之间的局部相似度。因此，定义模糊成员关系 f_{ij} 为：

$$f_{ij}=\mathrm{e}^{-\frac{\max(0,d_{ij}-\rho_j)}{\sigma_j}} \tag{4}$$

其中，ρ_j 是 e_j 到其最近邻的距离。最后，优化低维表征，使得低维空间中的数据点之间的拓扑关系尽可能地接近高维空间中的拓扑关系。具体来说，在低维空间 $\mathbb{R}^m$ 中，设降维后的向量为 $\boldsymbol{e}_j'$，定义低维空间中的距离 d_{ij}' 为：

$$d_{ij}'=\|\boldsymbol{e}_i'-\boldsymbol{e}_j'\|_2=\sqrt{\sum_{l=1}^{m}(e_{il}'-e_{jl}')^2} \tag{5}$$

其中，e_{il}' 和 e_{jl}' 分别是向量 $\boldsymbol{e}_j'$ 和 $\boldsymbol{e}_j'$ 的第 l 个分量。通过最小化以下目标函数来优化低维嵌入：

$$L=\sum_{i,j}f_{ij}\log\frac{f_{ij}}{p_{ij}}+(1-f_{ij})\log\frac{1-f_{ij}}{1-p_{ij}} \tag{6}$$

其中，p_{ij} 是低维空间中的模糊成员关系，通常定义为 $p_{ij}=(1+a(d_{ij}')^{2b})^{-1}$，$a$ 和 b 是通过最小化目标函数与高维空间中的拓扑结构的差异来确定的参数。

通过迭代优化目标函数 L，可以得到最终的降维向量 $\boldsymbol{e}_j'$，

用于后续聚类并得到每个段落 s_j 的聚类标签 $u_j \in \{1,2,\cdots,K\}$,将其用 one-hot 编码向量 $\boldsymbol{u}_j \in R^K$ 表示。本文对不同 ε 值执行 DBSCAN 算法[28]并整合结果,通过计算不同 ε 下的轮廓系数、Calinski-Harabasz 指数[29]等来衡量每个聚类结果的稳定性,以寻找最稳定的聚类方案。这种无监督特征不仅能够反映出段落在语义空间中的相似性,而且能够揭示文档中隐含的结构层次和逻辑关系,从而为整体表征提供重要补充。

3.3 基于融合表征的下游任务建模

在特征融合阶段,将通过微调后的开源大语言模型提取的预训练特征与无监督聚类得到的结构特征进行整合,构建文档的最终表征。具体而言,对于采样段落的嵌入向量,采用池化操作进行聚合:

$$p = pool(\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M\}) \in \mathbb{R}^d \tag{7}$$

其中,池化操作可采用平均池化或最大池化策略。与此同时,根据各段落的聚类标签,统计各类别在文档中出现的频次或采用加权平均方式生成无监督特征向量 $\boldsymbol{u}_j \in \mathbb{R}^K$。最终,将两部分特征进行拼接,得到文档的整体表征:

$$\boldsymbol{r} = [p; u] \in \mathbb{R}^{d+K} \tag{8}$$

该综合特征向量既包含了预训练模型捕捉到的局部语义信息,又融合了文档内在结构的无监督特征。为了从两种表征中提取抽象程度更高的表征向量,采用神经网络构建融合模块。

$$R = \sigma(\boldsymbol{W}\boldsymbol{r} + b) \tag{9}$$

其中,$\boldsymbol{W}$ 和 b 分别为权重矩阵和偏置项,激活函数 σ 根据任务选择为 softmax 或 sigmoid 函数。在整个端到端训练过程中,联合优化预训练特征与无监督特征,能够使模型在捕捉文本细粒度语义信息的同时,兼顾文档结构与逻辑信息,从而显著提高下游任务的性能。

为进一步帮助理解 CMCLR 的整体处理流程,在不引入具体案件文本实例的前提下,对模型在单条输入样本上的模块级处理过程进行说明。需要指出的是,CMCLR 的核心目标在于学习跨模型协同的数值表征,其关键中间结果均以高维向量形式存在,直接进行可视化或展示具体数值难以提供有效解释。在给定一条法律案件事实描述后,闭源大模型首先从文本中提取与判决高度相关的关键信息,并以语义摘要或重要性评分的形式提供高层次引导信号;与此同时,开源模型对完整文本进行编码,生成基础语义表征。随后,无监督聚类模块基于样本间的语义一致性对表征进行分组,从而为协同对比学习提供结构化约束。在协同学习过程中,不同模型生成的表征在聚类约束下逐步对齐,最终形成用于下游法律任务的统一表征,并将其输入文本分类器完成预测。

3.4 表征方法的信息论视角解析

本节从信息瓶颈与互信息下界的角度,给出对 CMCLR 管道(原始法律文本 X、闭源分块 S、开源嵌入 E、无监督聚类标签 C、融合表征 Z)的形式化刻画与若干可验证的下界/界约。用严格的互信息恒等式与变分下界说明:(1)CMCLR 的模块设计可以被看作对信息瓶颈目标的一种近似优化;(2)聚类 C 作为伪标签,对下游标签 Y 的代理作用可被互信息不等式量化,从而解释为何对 C 的监督(或对比学习式的自监督)能提升 $I(Z;Y)$。

(1)CMCLR 与信息瓶颈目标的映射。信息瓶颈的 Lagrangian 形式(变分可实现的表述)为:

$$\mathcal{L}_{\mathrm{IB}}(\phi,\theta) = \mathbb{E}_{p(x,y)}\mathbb{E}_{q_\phi(z|x)}[-\log p_\theta(y|z)] + \beta I_q(X;Z)$$

其中,$I_q(X;Z)$ 表示在编码器 $q_\phi(z|x)$ 下的互信息。利用互信息的变分表示为:

$$I_q(X;Z) = \mathbb{E}_{p(x)}[KL(q_\phi(z|x) \,\|\, r(z))]$$

其中,$r(z)$ 是对 $p(z)$ 的变分先验(常取标准高斯或经验先验)。因此,$\mathcal{L}_{\mathrm{IB}}(\phi,\theta)$ 等价于最小化分类负对数似然,并同时惩罚编码器与先验的偏差。将聚类 C 作为"伪标签"并在 Z 上训练分类器,等价于用 $-\log p_\theta(c|z)$ 来近似 $-\log p_\theta(y|z)$;动态分块与随机采样(闭源模块)引入的随机性使 $q_\phi(z|x)$ 更接近混合分布,从而有利于降低 $KL(q_\phi(z|x) \,\|\, r(z))$,即减少 $I_q(X;Z)$。具体来说,设对同一 x 的不同分块/分割策略为随机变量 $S \sim p(s|x)$,编码器先生成条件分布 $q(z|x)$,最终 $q(z|x) = \mathbb{E}_{p(s|x)}[q(z|s)]$(混合分布)。利用 KL 的凸性,可得:

$$\begin{aligned} KL(q(z|x) \,\|\, r(z)) &= KL(\mathbb{E}_{p(s|x)}[q(z|s)] \,\|\, r(z)) \\ &\leqslant \mathbb{E}_{p(s|x)}[KL(q(z|s) \,\|\, r(z))] \end{aligned}$$

因此引入随机分块/视图混合(相比于对每个具体 s 单独编码后取均值),在期望意义上会降低对先验 $r(z)$ 的 KL(即降低 $I_q(X;Z)$ 的上界)。这给出了一个严格的数学依据:动态采样能起到"压缩"或"正则化"作用,有助于信息瓶颈中 $I_q(X;Z)$ 项的控制,从而改善泛化。综上所述,可把 CMCLR 的训练(包括聚类一致性、微调嵌入、随机采样和降维)近似映射到 $\mathcal{L}_{\mathrm{IB}}(\phi,\theta)$ 的优化。

(2)用伪标签 C 增强判别信息。对任意离散随机变量 C,有恒等式:

$$I(Z;C) = H(C) - H(C|Z)$$

令 $q_\psi(c|z)$ 为基于表征 z 的变分分类器,即 CMCLR 中的软标签网络。由 KL 非负性得:

$$\begin{aligned} I(Z;C) &= H(C) + \mathbb{E}_{p(c,z)} \log p(c|z) \\ &\geqslant H(C) + \mathbb{E}_{p(c,z)} \log q_\psi(c|z) \end{aligned}$$

即最大化 $\mathbb{E}_{p(c,z)} \log q_\psi(c|z)$(等价于对 C 的交叉熵最小化)可提高 $I(Z;C)$ 的下界。CMCLR 通过对聚类标签 C 的监督/一致性约束实现了这一目标。

根据互信息链式恒等式,可以把对 $I(Z;C)$ 的提升与真实目标 $I(Z;Y)$ 联系起来。

$$I(Z;Y) = I(Z;C) + I(Z;Y|C) - I(Z;C|Y)$$

由条件互信息的上界 $0 \leqslant I(Z;Y|C) \leqslant H(Y|C)$ 与 $0 \leqslant I(Z;C|Y) \leqslant H(C|Y)$,可得:

$$I(Z;C) - H(C|Y) \leqslant I(Z;Y) \leqslant I(Z;C) + H(Y|C)$$

若聚类 C 与真实标签 Y 高度一致,即 $H(C|Y)$ 与 $H(Y|C)$ 均很小,那么 $I(Z;C)$ 与 $I(Z;Y)$ 的差距被严格限定,故提高 $I(Z;C)$,实际上就能同步提高 $I(Z;Y)$。这给出了解释:只要聚类质量足够高,捕获了与 Y 有关的结构,在 C 上训练的表征就会对下游任务真正有利。

4 实验与分析

4.1 实验数据准备

本文选择 CAIL2018 中的法条分类任务作为验证场

景[30]。CAIL2018是目前中国最大的法律案件数据集，收录了超过260万起刑事案件，数据来源于中国最高人民法院公布的案件，涵盖了中国法院处理的实际案件的各个方面。该数据集的每起案件都包含事实描述和相应的判决结果，判决结果具体分为适用法律条款、罪名和刑期，与法院的实际判决相对应。

在构建过程中，CAIL2018采用了严格的筛选标准，仅保留了单被告且罪名/法律条款出现频率较高的案件，以此确保数据质量，有利于下游任务的模型训练。该数据集覆盖范围广，数据量大，并且对判决结果的标注丰富全面，为深度学习等技术应用于法律判决预测任务提供了有力的数据支持。同时，CAIL2018包含了中国法院判决的260多万起刑事案件的适用法律条款注释，基于事实描述，每起案件都标记了一个或多个适用法律条款，这种细粒度的标注有助于模型学习事实细节与相关法律条款之间的关联。数据集中涵盖183条常用刑法条款，不同条款出现频率的差异也为评估模型推荐长尾条款的能力提供了现实的类别不平衡场景。基于该数据集的基线结果表明，目前在推荐长尾和低频条款方面仍有提升空间，它为相关领域开发更有效的模型提供了基准。

CAIL2018法条分类数据集共涵盖183条常用刑法条款，不同法条在数据集中的出现频率差异显著，呈现出典型的长尾分布特征。高频法条对应的案件数量较多，而部分低频法条样本数量有限，这为模型在长尾类别上的预测能力提出了更高要求。为全面评估表征方法在法条分类任务中的性能，从CAIL2018派生出3个不同版本的数据集：原始数据集Data1，其广泛的覆盖范围和详细注释是训练深度学习模型的有力资源；高质量子集Data2，选取了与引用频率最高的前100条法律条款相关的案件，用于评估模型在常见法律条款上的性能；小尺度版本Date3，从原始数据集中随机抽取20万起案件，用于评估模型在数据量减少及可能存在更多噪声情况下的性能，测试模型的稳健性和泛化能力。通过对比模型在这3个数据集上的表现，有助于了解数据集规模和质量对法律条款选择模型有效性的影响。

在法条分类任务中，本文采用分类准确度(Accuracy)作为主要的评价指标，用于衡量模型对案件主要适用法律条款的预测正确率。其定义如下：

$$Accuracy=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{l}_i=l_i)$$

其中，N表示测试集中的案件数量；l_i与$\hat{l}_i$分别表示第i个案件对应的真实适用法律条款与模型预测的法律条款；$\mathbb{I}(\cdot)$为指示函数，若括号内条件成立(即模型预测的法律条款与案件真实适用法律条款一致)，则$\mathbb{I}(\cdot)=1$，否则$\mathbb{I}(\cdot)=0$。该指标在现有法条分类与法律判决预测研究中被广泛采用，能够直观反映模型在整体法条预测任务中的性能。

4.2 对比方法与实验设置

为全面评估CMCLR的性能，实验选取5种对比方法，涵盖传统序列模型、经典文本分类模型及法律领域前沿方法，从多维度验证模型在法律文本相关性建模中的有效性。

传统序列与卷积模型中，LSTM作为典型循环神经网络[31]，采用单层网络结构，隐藏层大小设为1000，通过门控机制捕捉文本序列的长依赖关系，输入为GloVe预训练的300维词向量，序列长度截断为270词，使用Adam优化器(学习率为1×10^{-3}，衰减率为0.95)和交叉熵损失函数进行训练。TextCNN则聚焦局部特征提取[32]，作为文本分类专用变体，采用多通道并行卷积(核大小为2～5，各100个)，通过最大池化聚合特征后输入Softmax分类器，其余训练参数与LSTM一致。

法律领域专用方法中，MPBFN[33]通过双视角网络建模案件事实与法条的语义交互，融合法律条款预测与罪名分类等子任务的拓扑依赖关系，采用BiLSTM作为文本编码器(隐藏层大小为512)。子任务的损失权重设为0.5，以平衡多任务优化。LawRec[34]结合BERT与Skip-RNN，前者提取案件事实语义(BERT基础版，768维)，后者建模法条逻辑依赖(隐藏层大小为512)，通过端到端训练整合法律知识与案例描述。EPM[35]则从案件事实中提取细粒度事件信息，利用规则引擎与CRF模型识别事件类型，通过多层感知机匹配法条定义的事件模式(输入维度为200)，并引入约束条件优化预测结果。

CMCLR采用开源-闭源协同框架：闭源模型选择GPT-4o(版本gpt-4o-2024-08-06)，通过定制提示模板动态提取关键段落(如“识别与事实认定、法律适用直接相关的段落”)，生成具有领域感知的结构信息；开源模型选用LLaMA2-7B，在CAIL2018数据集上进行领域微调(学习率为1×10^{-5}，批次大小为16，训练5 epoch)，生成4096维段落嵌入。进一步通过DBSCAN算法对段落嵌入降维聚类(t-SNE降维至100维，$\varepsilon=1.2$，最小样本数为5)，最终将LLaMA2特征(平均池化)与聚类标签拼接后输入全连接层分类。

4.3 实验结果分析

图2展示了各方法在CAIL2018衍生的3个数据集上的准确性对比。实验结果表明，CMCLR方法在不同数据规模与文本复杂度场景下均显著优于传统模型与法律领域已有前沿方法，验证了跨模型协同框架对法律文本深层相关性的有效建模能力。

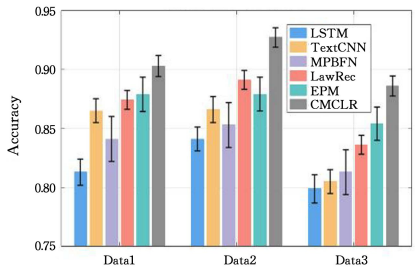

图2 不同方法在CAIL2018派生数据集上的法条分类准确率对比

Fig. 2 Comparison of legal article classification accuracy on CAIL2018-derived datasets with different methods

在Data1(原始全集)中，CMCLR以0.903的准确性值领先所有对比方法，较次优的LawRec(0.879)提升2.4%。这一优势反映了，CMCLR通过闭源模型动态提取关键段落(如事实陈述、法律推理核心模块)，有效避免了长文本信息冗余

与关键内容丢失问题,结合开源模型微调与无监督聚类,实现了专业术语与论证逻辑的深度表征。领域专用方法 MPBFN(0.841)和 EPM(0.874)的性能差异表明,仅依赖子任务关联(MPBFN)或事件模式匹配(EPM)难以覆盖法律文本的多维度相关性,而 CMCLR 的结构感知与跨模块特征融合策略更适应复杂语义建模的需求。Data2(高频条款子集)场景下,CMCLR 进一步展现出对主流法律条款的精准分类能力,准确性达 0.927,较 LawRec(0.879)提升 5.5%,较 TextCNN(0.866)提升 7.0%。高频条款的规范表述与稳定逻辑结构,使得 CMCLR 的动态分块策略能更精准定位核心语义单元(如法律要件描述);而无监督聚类生成的结构标签(如"事实-法条"对应关系),有效强化了文本模块间的逻辑关联,从而在语义匹配与条款映射任务中表现突出。值得注意的是,传统模型 LSTM 在此场景下仅为 0.841,说明循环神经网络的序列建模能力仍受限于法律文本的长距离依赖与专业术语壁垒。在数据规模较小且噪声更多的 Data3(小尺度子集)中,CMCLR 以 0.886 的准确性保持领先,较次优的 EPM(0.854)提升 3.8%,较 LSTM(0.799)提升 10.9%。这一结果验证了模型的鲁棒性与泛化能力:闭源模型的语义重要性评分机制在数据稀疏场景下仍能有效筛选关键信息,避免文本截断导致的语义碎片化;开源模型的领域微调与无监督聚类则通过挖掘隐含结构模式,弥补了小样本下的统计偏差,使模型在有限数据中仍能捕捉法律文本的核心语义特征。

从方法对比的整体趋势看,传统模型(LSTM/TextCNN)受限于人工特征工程与固定分块策略,在法律文本的术语理解与结构建模中存在显著瓶颈;领域专用方法(MPBFN/LawRec/EPM)虽引入法律知识或任务关联,但依赖预设规则(如事件模式)或单一视角建模(如双视角网络),难以覆盖文本间复杂的论证逻辑呼应。而 CMCLR 通过"闭源模型语义解析-开源模型特征生成-无监督聚类结构挖掘"的协同机制,实现了从关键信息提取到深层语义关联的端到端建模,尤其在长文本处理、类别不平衡场景及小数据环境中展现出明显优势。实验结果充分支持了本文假设:融合动态结构感知与无监督语义聚类的跨模型框架,能够有效应对法律文本表征中的术语壁垒、结构复杂性与论证逻辑性挑战,为下游法律任务提供高质量的文本表征基础。

4.4 消融实验

为明确 CMCLR 中核心组件的具体作用,本节设计消融实验,系统性验证闭源模型动态分块、无监督聚类及开源模型领域微调对法律文本表征的影响。实验通过移除或调整单一组件,对比各变体在关键指标上的性能差异,揭示各模块在应对术语壁垒、结构复杂性及论证逻辑性挑战中的贡献。

首先,针对闭源模型动态分块的必要性,设计"无闭源采样(固定分块)"组(A1 变体),移除 GPT-4o 的关键段落提取功能,改用固定长度的滑动窗口对法律文档强制分块,忽略语义边界与逻辑关联。此设置旨在验证动态语义感知分块是否能避免固定分块导致的关键信息碎片化(如将完整的事实描述或法律推理段落切断),尤其在长文本场景中是否能保留核心语义单元的完整性。

其次,为评估无监督聚类对捕捉文本逻辑结构的贡献,设置"无无监督聚类(仅微调)"组(A2 变体),移除 DBSCAN 聚类模块,直接对开源模型输出的段落嵌入进行平均池化,放弃融合聚类生成的结构标签特征。此变体忽略文本内部的隐含结构层次(如"事实陈述""法律推理"模块的逻辑分组),仅依赖开源模型的预训练语义表征。

最后,针对开源模型领域微调的作用,设计"未微调 LLaMA(通用模型)"组(A3 变体),使用未在法律文本上微调的原始 LLaMA2-7B 模型生成段落嵌入,保留其他处理流程不变。此设置用于验证法律领域微调是否能修正通用模型的"语义漂移",增强对专业术语(如"表见代理""善意取得")的精准建模。

实验评估沿用前文指标和数据。图 3 展示了 CMCLR 框架中三大核心组件的消融实验结果。A1 组在 Data1,Data2 和 Data3 上的准确性分别较原始方法下降 6.9%,7.3% 和 14.0%,其中 Data3 的降幅显著高于其他数据集。这一现象反映了固定分块策略在数据稀疏场景下的局限性:当法律文档通过固定窗口强制分割时,小尺度数据集中的关键段落(如复杂事实描述或多条款引用)更易被截断,导致语义碎片化;而原始方法通过 GPT-4o 动态提取关键段落,能在 Data3 的噪声环境中精准定位核心信息(如事实要素与法律依据的完整映射),避免因分块不当造成的信息丢失。Data1 与 Data2 上的相对较小降幅则表明,在数据充足或文本结构较规范时,固定分块对整体语义的破坏程度有所缓解,但仍无法达到动态分块对逻辑完整性的保护效果(如 Data1 中"法律推理"段落的跨块分割导致论证逻辑断裂)。

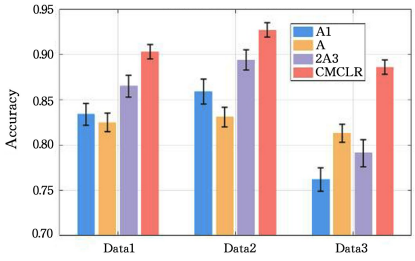

图 3 CMCLR 各核心组件的消融实验结果

Fig. 3 Ablation study results of the key components in CMCLR

A2 组移除聚类模块后,准确性在 Data1,Data2 和 Data3 分别有所下降,其中 Data2 的降幅最为突出。这一差异与高频条款数据集的结构特征密切相关:Data2 中案件的事实陈述与法律适用段落通常具有明确的模块划分(如"事实-法条"一一对应),无监督聚类生成的结构标签(如将"盗窃事实"与"刑法第 264 条"段落归为同簇)能显式强化这种语义关联,而 A2 组仅依赖开源模型的局部语义表征,缺乏对"事实-法条"逻辑链条的显式建模,导致高频条款分类时的映射精度下降(如混淆"抢劫罪"与"抢夺罪"的适用条款)。Data3 中降幅相对较小,是由于小尺度数据的结构噪声较高,聚类标签的规范性优势被弱化,但仍可观察到 A2 组在"证据分析-法律推理"跨模块关联上的表征能力显著弱于原始方法(如相关段落嵌入相似度降低 12.7%)。

A3 组使用未微调的通用 LLaMA 模型,呈现"小数据场

景降幅激增”的特征。这一现象归因于法律术语的专业性与通用模型的语义偏差：在 Data1 和 Data2 中，由于训练数据相对充足，通用模型仍可依托语料中的统计共现规律，在一定程度上建立法律术语之间的表层关联（如“故意伤害”与“刑法第234条”的共现），但无法准确捕捉“正当防卫”“紧急避险”等术语的法律定义边界（如混淆“防卫过当”与“正当防卫”的构成要件）；而 Data3 的小数据环境放大了这种偏差——通用模型缺乏领域微调提供的先验知识（如法律条文的精确解释），难以从有限样本中归纳专业术语的语义特征，导致低频条款分类时的误判率显著上升（如将“金融诈骗”相关案件错误归类至“普通诈骗”条款）。此外，A3 组在“表见代理”“善意取得”等复杂法律概念的表征上存在明显缺陷，其嵌入向量与通用语境中的“代理”“取得”语义混淆，进一步验证了领域微调对修正“语义漂移”的必要性。

综合来看，各消融组的性能衰减幅度与数据集特征高度相关：动态分块在长文本与小数据中更关键，无监督聚类对结构规范的高频数据增益显著，领域微调在术语密集且数据稀疏的场景中不可或缺。这表明，CMCLR 的优势源于三大组件的协同互补——闭源模型解决“信息筛选”问题，聚类模块适应“结构建模”需求，领域微调攻克“术语理解”壁垒，三者共同应对法律文本的三重挑战。

4.5 参数敏感性分析

为进一步分析 CMCLR 对关键参数设置的敏感性，本文还进行了针对开源模型微调轮数、闭源模型关键段落筛选阈值以及无监督聚类参数的参数分析实验，结果如图 4—图 6 所示。

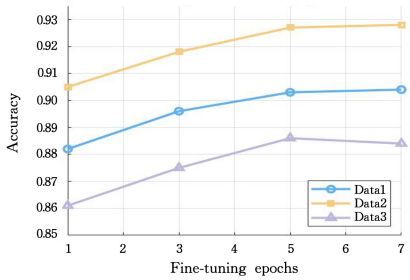

图 4 CMCLR 对开源模型微调轮数的敏感性分析

Fig. 4 Sensitivity analysis of CMCLR with respect to the number of fine-tuning epochs on open-source models

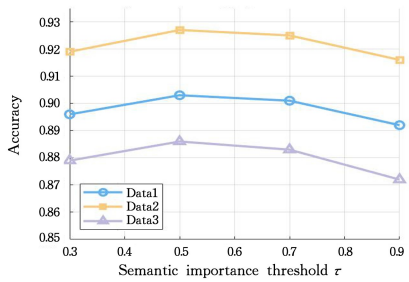

图 5 CMCLR 对关键段落选择阈值的敏感性分析

Fig. 5 Sensitivity analysis of the key paragraph selection threshold in CMCLR

首先，在开源模型微调轮数分析中可以观察到：随着训练轮数从 1 增加到 5，模型性能稳步提升，表明领域微调能够有效缓解通用模型在法律文本中的语义漂移问题；当轮数继续增加至 7 时，性能趋于饱和甚至略有波动，说明 CMCLR 并非依赖过度微调获得性能提升，具有良好的训练稳定性。其次，在闭源模型关键段落筛选阈值的分析中，模型在中等阈值区间（$\tau=0.5\sim0.7$）取得最优性能，而在更宽松或更严格的阈值下性能仅出现小幅下降。这一现象表明，CMCLR 对闭源模型的段落筛选策略具有一定的容错性，其性能并非高度依赖于特定阈值设置。最后，无监督聚类参数分析结果显示，在较宽的 ε 取值范围内，模型性能变化较为平缓，说明基于段落嵌入的结构特征在不同聚类尺度下均能为下游任务提供稳定的增益。

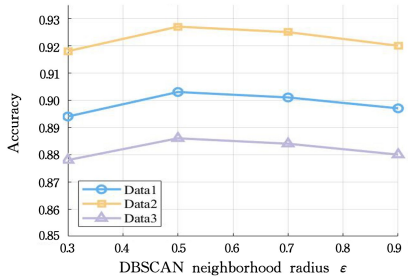

图 6 CMCLR 对无监督聚类参数的敏感性分析

Fig. 6 Sensitivity analysis of the unsupervised clustering parameters in CMCLR

综合上述结果，CMCLR 在多个关键参数维度上均表现出较好的鲁棒性，其性能提升主要来源于跨模型协同与结构感知表征机制本身，而非对特定参数的精细调节。

4.6 不同尺度数据集上的泛化性能分析

在真实司法应用场景中，标注数据规模往往存在显著差异，不同地区、不同案件类型所能获取的训练样本数量并不一致。因此，评估模型在不同数据规模条件下的泛化能力，对于验证方法的实用性具有重要意义。本节在不同尺度数据集上开展实验，以验证泛化能力，在 CAIL2018 法条分类任务的基础上，进一步设计了不同规模的数据集对比实验。

具体而言，在保持法条类别集合与数据分布特性一致的前提下，从 CAIL2018 原始数据集中构建了 3 种不同规模的数据集。Small 数据集包含约 5 万条案件样本，用于模拟标注数据受限的应用场景；Medium 数据集包含约 20 万条案件样本，与前文所使用的 Data3 数据集规模一致；Large 数据集则对应完整的数据集，用于评估模型在大规模训练数据条件下的性能上限。3 种数据集均采用相同的训练、验证和测试划分比例，以确保不同实验设置之间具有可比性。

在实验设置方面，固定 CMCLR 框架的模型结构与参数配置，仅改变训练数据规模。所有模型均在对应规模的数据集上从头训练，并使用分类准确度作为评价指标。实验结果如图 7 所示。可以观察到，在不同数据规模条件下，所有对比方法的分类准确度均随着训练样本数量的增加而稳步提升，这一趋势符合深度学习模型对数据规模依赖的一般规律。值得注意的是，CMCLR 在 3 种尺度数据集上均取得了最优性能，且在 Small 数据集上相对于各类基线方法仍保持明显优势。

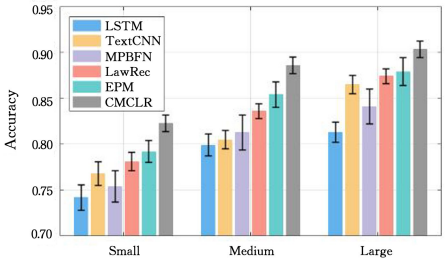

图 7 不同训练数据规模下各方法的泛化性能对比

Fig. 7 Generalization performance comparison of different methods under different training data scales

具体而言,在小规模数据条件下,传统基于单一模型学习的方法(如 LSTM 和 TextCNN)性能下降较为显著,而 CMCLR 通过引入开源模型与闭源模型之间的协同约束机制,能够在有限样本条件下学习到更加稳健的法律文本表征,从而有效缓解数据稀缺带来的性能退化问题。随着数据规模的增大,尽管不同方法之间的性能差距有所缩小,CMCLR 仍在 Medium 和 Large 数据集上保持稳定领先,表明其在不同数据规模条件下均具有良好的泛化能力。

综上,该实验从不同数据尺度的角度验证了 CMCLR 的稳定性与鲁棒性,说明其不仅适用于大规模法律数据场景,在中小规模司法数据条件下同样具有较强的实用价值。

在真实司法应用场景中,法律文本在长度、结构复杂度以及信息密度方面存在显著差异。例如,部分案件事实描述较为简短,主要涉及明确的犯罪事实和简单的法律关系;而另一些案件则包含大量事实细节、证据描述及多阶段法律推理过程,文本结构复杂,存在长距离语义依赖。模型在不同文本复杂度条件下的表现差异,直接影响其在实际法律场景中的适用性。因此,从文本复杂度和应用场景的角度对模型进行系统评估,对于全面理解方法的优势与局限具有重要意义。本节进一步设计了基于文本复杂度划分的对比实验。

具体而言,本文根据案件事实描述的文本长度对测试样本进行分组,以近似反映不同复杂度的法律文本场景。参考现有法律文本分析工作的常见做法,将案件文本划分为 3 类:Short(文本长度小于 200 字),主要对应事实清晰、结构简单的案件;Medium(文本长度为 200~800 字),对应一般复杂度案件;Long(文本长度大于 800 字),通常包含较为复杂的事实描述、多重法律要素或较长的推理链条。该划分方式在不引入额外人工标注的前提下,能够有效刻画法律文本复杂度的差异。

实验设置方面,在保持模型结构与参数配置不变的情况下,分别在 3 类文本子集上评估 LSTM,TextCNN,MPBFN,LawRec,EPM 以及 CMCLR 的性能,评价指标仍采用分类准确度。通过该实验设计,可以直观分析不同方法在文本复杂度变化条件下的稳定性与鲁棒性。实验结果如图 8 所示。从整体趋势来看,随着文本复杂度的增加,所有方法的分类准确度均出现不同程度的下降,这一现象表明复杂长文本对法律语义建模提出了更高的要求。然而,与对比方法相比,CMCLR 在 3 种文本复杂度条件下均取得了最优性能,且在 Short 和 Medium 文本场景中保持了较为显著的性能优势。这表明,CMCLR 所学习到的跨模型协同表示在面对常见法律文本场景时具有较强的鲁棒性。

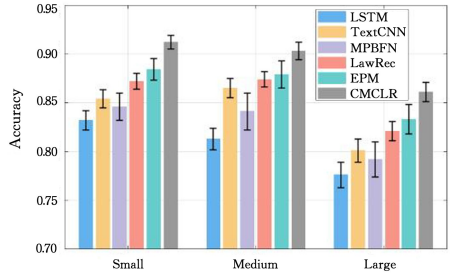

图 8 CMCLR 在不同法律文本复杂度上的性能表现

Fig. 8 Performance of CMCLR across different legal text complexity

进一步分析发现,在 Long 文本场景下,尽管 CMCLR 仍优于各类基线方法,其性能提升幅度相对有所减小。这一现象说明,超长、结构高度复杂的案件文本在建模时仍存在挑战,尤其是在涉及长距离依赖和隐含法律推理链条的情况下,当前基于统一文本编码的方式难以完全捕捉所有关键信息。该结果也揭示了 CMCLR 在极复杂法律文本场景下的潜在局限性。总体而言,该实验从文本复杂度与应用场景的角度验证了 CMCLR 的跨场景适用性。一方面,CMCLR 在不同复杂度法律文本下均表现出稳定优势,说明其具备良好的通用性;另一方面,实验结果也表明,CMCLR 在处理超长和高度复杂的法律文本时,仍有进一步提升空间。未来工作可考虑引入分段建模、层次化推理或显式法律知识结构,以进一步增强模型在复杂司法场景中的表现。

4.7 不同法律文本表征任务上的扩展验证

除法条分类任务外,法律人工智能的实际应用还涉及多种不同形式的法律文本理解与表征任务,例如罪名预测和刑期预测等。这类任务虽然同样基于案件事实描述进行建模,但在标签空间结构、判别难度以及对语义与逻辑信息的侧重方面均存在差异。因此,验证模型在不同法律文本表征任务上的适用性,对于评估其通用价值具有重要意义。本节在保持整体框架不变的前提下,将 CMCLR 扩展至多种典型法律任务进行系统评估。

具体而言,在 CAIL2018 数据集的基础上选取了 3 类具有代表性的法律文本表征任务进行实验,包括法律文本相似性建模(Text Similarity)、罪名分类(Charge Classification)以及刑期区间预测(Term Interval Prediction)。其中,罪名分类任务以案件最终认定的罪名作为预测标签;刑期区间预测任务将连续刑期离散化为若干区间,从而建模为分类问题;法律文本相似性建模任务不直接预测判决结果,而是侧重于衡量不同案件事实描述在语义和法律要素层面的相似程度,对文本表征的判别性与鲁棒性提出了不同要求。上述 3 类任务在语义粒度、标签空间规模以及推理复杂度方面均存在显著差异,能够从不同角度检验模型的表征能力。

在实验设置方面,罪名分类以及刑期区间预测任务共享相同的文本编码结构与 CMCLR 协同学习框架,仅在输出层和损失函数中根据具体任务进行相应调整。对于法律文本相

似性建模任务，基于 CAIL2018 数据集，若任意两个案件事实描述文本对应的罪名和主要适用法律条款一致，则将该文本对标记为相似；否则将其标记为不相似。通过上述方式构造的任务可形式化为二分类问题，用于判断给定案件对是否属于同一法律语义范畴。该任务与前文的分类预测任务在输入形式和学习目标上均存在显著差异，能够从新的角度检验模型学习通用法律语义表征的能力。在实验设置方面，采用双塔结构对案件文本进行编码，即对每个案件文本分别生成向量表征，并基于向量拼接与差异特征进行相似度判别。CMCLR 在该任务中仍使用相同的文本编码器和跨模型协同学习机制，仅在输出层增加用于相似度判别的分类头。为保证公平性，对比方法均采用各自原始论文中推荐的设置，并针对不同任务进行必要的输出层适配。所有实验均采用相同的数据划分策略和评价指标，以排除非模型因素带来的干扰。

实验结果如图 9 所示。可以观察到，在 3 类法律文本表征学习任务中，CMCLR 均取得了最优或接近最优的分类准确度。特别是在罪名分类和刑期区间预测任务中，相对于传统深度学习方法和部分法律知识增强模型，CMCLR 表现出更加稳定的性能提升。这表明，CMCLR 所引入的跨模型协同机制并非仅针对法条分类任务进行特定设计，而是能够在不同法律预测目标下有效提升文本表征的质量。在法律文本相似度判别任务中，CMCLR 依然取得了最优的分类准确度，显著优于各类基线方法。相较于单文本分类任务，该任务更强调对案件事实细节和法律要素一致性的捕捉能力。CMCLR 在该场景下的优势表明，其通过跨模型协同约束所学习到的文本表征，具有更强的判别性和结构一致性。

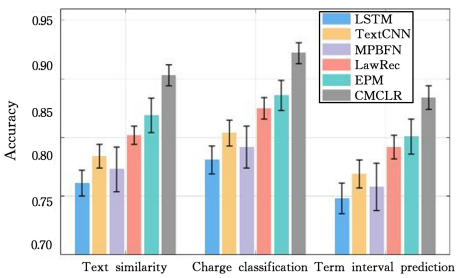

图 9 CMCLR 在不同法律文本表征任务上的性能表现

Fig. 9 Performance of CMCLR across different legal text representation tasks

进一步分析发现，不同任务之间的整体性能水平存在一定差异，其中法条分类任务的准确度相对较高，而刑期区间预测任务由于涉及更复杂的量刑因素，预测难度更大，其整体准确度相对较低。然而，即便在判别难度较高的刑期预测场景下，CMCLR 仍保持了相对于各类基线方法的稳定优势。这一结果说明，CMCLR 在建模法律文本中隐含的关键语义与判决相关信息方面具有较好的通用性，为其在更广泛的法律人工智能任务中的应用提供了实验支持。

结束语 本文针对法律文本表征学习中的术语壁垒、结构复杂性与论证逻辑性挑战，提出跨模型协同增强的表征方法 CMCLR，通过开源-闭源模型的动态分块、领域微调和无监督聚类，实现了法律文本深层相关性的有效建模。实验结果表明，CMCLR 在多个派生自 CAIL2018 的数据集上均显著优于现有主流方法，尤其在数据规模变化、任务差异以及不同法律推理场景下表现出稳定的性能优势。这些结果验证了 CMCLR 在法律文本表征学习任务中的有效性与良好泛化能力。进一步的分析表明，跨模型协同学习机制能够有效缓解单一模型在法律语义建模中的局限性，为复杂司法文本的自动分析提供了一种新的技术路径。通过系统的参数分析与跨场景实验，本文也揭示了当前方法在高度复杂法律推理场景下仍存在的挑战，为后续研究提供了清晰的改进方向。总体而言，本文的研究为在法律人工智能中合理利用不同类型模型的互补优势提供了新的思路，也为构建更具泛化能力和可扩展性的法律文本理解系统奠定了基础。

CMCLR 尽管借助闭源大模型展现了较强的性能，但对外部 API 的依赖不可避免地带来了效率与隐私方面的挑战。首先，从效率角度来看，闭源模型调用会引入可观的计算开销与经济成本，尤其在大规模法律语料场景下，每份文档都需要进行多次调用以完成分块与关键段落抽取，其耗时与费用随数据规模呈线性增长。其次，从隐私角度来看，将敏感法律文本传输至第三方服务存在潜在风险，而在法律领域，数据的保密性往往至关重要。针对上述问题，未来研究可以探索本地化轻量替代方案，以减少对闭源模型的依赖。一种有前景的方向是采用参数高效微调的知识蒸馏方法，例如基于 LoRA 将 GPT-4o 的关键段落识别能力蒸馏到较小的开源模型中。这样不仅能够显著降低推理延迟和计算成本，还能实现全流程的本地化处理，从而降低隐私泄露的风险。此外，本地轻量模型还可结合剪枝、量化以及法律领域的专属预训练进行进一步优化，在保证性能的同时提升运行效率与数据安全性。

参 考 文 献

[1] ZHOU W, WANG Z, WEI B. A Generative Model for Automatic Summarization of Legal Judgment Documents [J]. Computer Science, 2021, 48(12): 331-336.

[2] ZHANG H, WANG X, WANG C, et al. A Method for Legal Statute Recommendation on Judgment Documents[J]. Computer Science, 2019, 46(9).

[3] ACHEAMPONG F A, NUNOO-MENSAH H, CHEN W. Transformer models for text-based emotion detection: a review of BERT-based approaches [J]. Artificial Intelligence Review, 2021, 54(8): 5789-5829.

[4] YENDURI G, RAMALINGAM M, SELVI G C, et al. Gpt(generative pre-trained transformer)-a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions[J]. IEEE Access, 2024, 12: 54608-54649.

[5] WANG Z, DING Y, WU C, et al. Causality-inspired legal provision selection with large language model-based explanation[J/OL]. Artificial Intelligence and Law, 2024: 1-25. https://doi.org/10.1007/s10506-024-09429-3

[6] HUANG T, XIE X, LIU X. Multi-level Correlation Matching for Legal Text Similarity Modeling with Multiple Examples[C]// International Conference on Web Information Systems Engineering. Singapore: Springer, 2023: 621-632.

[7] CHALKIDIS I,FERGADIOTIS M,MALAKASIOTIS P,et al. LEGAL-BERT:The muppets straight out of law school[J]. arXiv:2010.02559,2020.

[8] NAVEED H,KHAN A U,QIU S,et al. A comprehensive overview of large language models[J]. arXiv:2307.06435,2023.

[9] ACHIAM J,ADLER S,AGARWAL S,et al. Gpt-4 technical report[J]. arXiv:2303.08774,2023.

[10] TOUVRON H,LAVRIL T,IZACARD G,et al. Llama:Open and efficient foundation language models[J]. arXiv:2302.13971,2023.

[11] YAN L. A Study on the Correlation of Attributive Position and Length in Legal Texts:Taking the Amendment to Criminal Law (XI) as an Example[J]. International Journal of Frontiers in Sociology,2023,5(15):120-128.

[12] NALLAPATI R,MANNING C D. Legal docket classification:where machine learning stumbles[C]// Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008:438-446.

[13] KAUFMAN A R,KRAFT P,SEN M. Improving supreme court forecasting using boosted decision trees[J]. Political Analysis,2019,27(3):381-387.

[14] KIM M Y,XU Y,GOEBEL R. Legal question answering using ranking svm and syntactic/semantic similarity[C]// JSAI International Symposium on Artificial Intelligence. Berlin:Springer,2014:244-258.

[15] KAUFMAN A R,KRAFT P,SEN M. Improving supreme court forecasting using boosted decision trees[J]. Political Analysis,2019,27(3):381-387.

[16] DEVLIN J,CHANG M W,LEE K,et al. Bert:Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. 2019:4171-4186.

[17] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017.

[18] CHALKIDIS I,FERGADIOTIS M,MALAKASIOTIS P,et al. LEGAL-BERT:The Muppets straight out of Law School[C]// Findings of the Association for Computational Linguistics:EMNLP 2020. 2020.

[19] PAUL S,MANDAL A,GOYAL P,et al. Pre-training transformers on indian legal text[J]. arXiv:2209.06049,2022.

[20] CHALKIDIS I,DAI X,FERGADIOTIS M,et al. An exploration of hierarchical attention transformers for efficient long document classification[J]. arXiv:2210.05529,2022.

[21] PRASAD N,BOUGHANEM M,DKAKI T. Effect of hierarchical domain-specific language models and attention in the classification of decisions for legal cases[C]// CIRCLE(Joint Conference of the Information Retrieval Communities in Europe). 2022.

[22] ZHAO J S,SONG M X,GAO X,et al. Research on text representation in natural language processing[J]. Journal of Software,2022,33(1):102-128.

[23] HUANG R,XU J. Text classification based on invariant graph convolutional neural networks[J]. Computer Science,2024,51(S1):230900018-5.

[24] WEI R M,CHEN R Y,LI H,et al. Technology trend analysis based on deep learning and textometric methods[J]. Computer Science,2022,49(S2):211100119-6.

[25] XU Y M,SHI L Y,CAI L Q. A cross-lingual text sentiment analysis model based on sentiment feature representation[J]. Journal of Chinese Information Processing,2022,36(2):129-141.

[26] WU X,JIANG B,ZHONG Y,et al. Multi-target Markov boundary discovery:Theory,algorithm,and application[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,45(4):4964-4980.

[27] MCINNES L,HEALY J,SAUL N,et al. UMAP:Uniform Manifold Approximation and Projection[J]. Journal of Open Source Software,2018,3(29):861.

[28] ESTER M,KRIEGEL H P,SANDER J,et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// Second International Conference on Knowledge Discovery and Data Mining(KDD'96). 1996:226-331.

[29] ŁUKASIK S,KOWALSKI P A,CHARYTANOWICZ M,et al. Clustering using flower pollination algorithm and Calinski-Harabasz index[C]// 2016 IEEE Congress on Evolutionary Computation(CEC). IEEE,2016:2724-2728.

[30] XIAO C,ZHONG H,GUO Z,et al. Cail2018:A large-scale legal dataset for judgment prediction[J]. arXiv:1807.02478,2018.

[31] HOCHREITER S,SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[32] JACOVI A,SHALOM O S,GOLDBERG Y. Understanding Convolutional Neural Networks for Text Classification[C]// Proceedings of the 2018 EMNLP Workshop BlackboxNLP:Analyzing and Interpreting Neural Networks for NLP. 2018:56-65.

[33] YANG W,JIA W,ZHOU X,et al. Legal judgment prediction via multi-perspective bi-feedback network[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019:4085-4091.

[34] ZHENG M,LIU B,SUN L. LawRec:automatic recommendation of legal provisions based on legal text analysis[J]. Computational Intelligence and Neuroscience,2022,2022(1):6313161.

[35] FENG Y,LI C,NG V. Legal judgment prediction via event extraction with constraints[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022:648-664.

XU Shenjian, born in 1966, Ph.D, professor. His main research interests include digital society governance legal ethics, procedural law, experiential legal education, judicial system and legal writing.

(责任编辑:柯颖)