

基于大语言模型的唐蕃古道文物知识图谱构建研究

刘华瑞, 雒伟群⁺, 张兆基

西藏民族大学 网络空间安全学院, 陕西 咸阳 712000

⁺ 通信作者 E-mail: 1034228464@qq.com

摘要: 针对文物领域多源异构数据整合困难, 以及传统知识抽取方法依赖人工标注、自动化程度低等问题, 提出一种基于大语言模型的文物知识图谱构建方法。融合领域专家知识与文物行业标准, 构建涵盖文物核心概念及其关系的知识本体, 系统定义其时空属性、文化属性和物理特征, 以及文物间的关联关系, 形成多维度结构化的知识框架。设计多阶段任务分解的联合抽取策略, 利用大语言模型的语义理解与生成能力, 将知识抽取拆解为“实体属性提取”与“三元组抽取”两个子任务, 并通过提示工程引导模型逐步完成。同时, 引入多维度校验机制, 从逻辑一致性、领域规范性和事实准确性等三方面对生成的三元组进行验证, 确保知识图谱的专业性与可靠性。实验表明, DeepSeek-R1 模型在抽取任务中 F1 值达 86.25%, 较 BERT-BiLSTM-CRF 模型提升 3.13 个百分点; 消融实验也验证了各模块的有效性。该方法为文物数字化保护与跨学科研究提供了高效、自动化的技术支撑。

关键词: 大语言模型; 知识图谱; 信息抽取; 唐蕃古道; 文物数字化

文献标志码: A **中图分类号:** TP391

Research on Construction of Knowledge Graph for Tang-Bo Ancient Road Cultural Heritage Based on Large Language Models

LIU Huarui, LUO Weiqun⁺, ZHANG Zhaoji

School of Cyberspace Security, Xizang Minzu University, Xianyang, Shaanxi 712000, China

Abstract: To address the challenges of integrating multi-source heterogeneous data in the cultural heritage domain and the issues of traditional knowledge extraction methods relying on manual annotation and having a low degree of automation, this paper proposes a method for constructing cultural heritage knowledge graphs based on large language models (LLMs). Firstly, by integrating domain expert knowledge and cultural heritage industry standards, this paper constructs a knowledge ontology encompassing core cultural heritage concepts and their relationships. This ontology systematically defines spatio-temporal attributes, cultural attributes, and physical characteristics of cultural relics, as well as the associative relationships between them, forming a multi-dimensional, structured knowledge framework. Secondly, a joint extraction strategy with multi-stage task decomposition is designed. Leveraging the semantic understanding and generation capabilities of LLMs, the knowledge extraction process is decomposed into two subtasks: “entity-attribute extraction” and “triple extraction” guided step-by-step through prompt engineering. Simultaneously, a multi-dimensional verification mechanism is introduced to validate the generated triples from three aspects: logical consistency, domain normative compliance, and factual accuracy, ensuring the professionalism and reliability of the knowledge graph. Experiments show that the DeepSeek-R1 model achieves an F1-score of 86.25% in the extraction task, which is 3.13 percentage points higher than the BERT-BiLSTM-CRF model. Ablation studies also verify the effectiveness of each module. This method provides efficient and

基金项目: 国家社会科学基金(22BMZ072); 西藏自治区文物局项目(藏文物办字[2024]243号)。

This work was supported by the National Social Science Fund of China (22BMZ072), and the Project of Xizang Autonomous Region Cultural Relics Office (Xizang Cultural Relics Office Document [2024] No.243).

收稿日期: 2025-09-09 **修回日期:** 2025-11-13

automated technical support for the digital preservation of cultural heritage and interdisciplinary research.

Key words: large language models; knowledge graph; information extraction; Tang-Bo Ancient Road; cultural heritage digitization

唐蕃古道作为古代连接唐朝与吐蕃的重要文化纽带,在政治、经济及军事交流中发挥了关键作用,同时承载了丰富的历史文化遗产和深刻的民族融合印记。沿线分布的大量文物,如文献资料、器物遗存等,具有极高的学术研究价值,成为探讨古代丝绸之路文化发展、民族交流及历史演进过程的重要依据。然而,由于这些文物呈现出分布零散、结构不均、异质性高等特点,严重制约了对其系统性研究的深入开展。

知识图谱自2012年由谷歌提出以来,在跨学科理论的推动下迅速发展,并广泛应用于搜索引擎优化、知识测评以及人工智能等多个领域。作为一种重要的知识组织与表达方式,知识图谱通过整合异构数据并挖掘其潜在的语义关系,展现出广阔的应用前景^[1]。同时,作为一种语义网络技术,知识图谱能够以结构化的方式表达领域知识,从而清晰展现文物与文化之间的内在联系,为用户提供更直观、更易理解的知识认知路径,提升公众对文化遗产的理解与接受能力^[2]。根据应用范围的不同,知识图谱可分为通用知识图谱(如DBpedia^[3]、Wikidata^[4])和领域知识图谱(如医疗、金融、教育等领域)。然而,传统的知识图谱构建方法高度依赖人工标注与结构化数据,导致构建成本高昂、周期长,且难以适应快速更新的知识需求。

在文物领域,知识图谱的研究仍处于相对滞后阶段,主要受限于文物数据本身的特性,如数据结构不统一、专业性强、标注难度大等。在文物知识图谱的相关探索中,国外自2006年起便开始探索文化遗产资源的数字化整合与知识图谱构建。目前已有多种针对文化遗产的知识图谱元数据模型被开发,其中最具代表性的是国际博物馆协会提出的概念参考模型(CIDOC conceptual reference model, CIDOC CRM),该模型主要服务于博物馆及相关机构的数据建模工作。Marchand等^[5]结合机器学习与开放信息抽取技术,利用命名实体识别与关系抽取手段实现了文化遗产知识图谱的构建;Carriero等^[6]提出基于XD(extreme design)方法开发与验证领域知识图谱;Pellegrino等^[7]则设计了一种基于虚拟助手的知识图谱访问方式,使得非专业用户也能通过自然语言查询文化遗产相关信息。

在国内,近年来基于深度学习的文物知识图谱构建也取得一定进展。李永卉等^[8]对南朝陵墓石刻进行了信息资源整合与开发;周冬艳等^[9]基于语言技术平台(lan-

guage technology platform, LTP)完成实体抽取任务,构建了青瓷知识图谱;尽管上述研究在一定程度上缓解了文物数据多源异构与结构化程度不足的问题,但仍面临高质量数据标注依赖度高、知识更新效率低等挑战。

大语言模型(large language model, LLM)因其强大的语义理解与知识表达能力,逐渐成为知识抽取研究的重要方向^[10]。特别是,Liu等^[11]系统综述了语言模型在自然语言理解任务中的演进与核心进展,为其在垂直领域的应用提供了理论基础。LLM可在少样本或零样本条件下高效完成知识抽取与推理任务^[12],显著降低了对人工标注的依赖,提高了知识更新效率^[13],为知识图谱的自动化构建提供了全新的技术路径。例如Chen等^[14]提出了一种名为AutoKG的轻量级自动化知识图谱构建方法,通过自动化构建带权关系图谱,弥补了纯向量检索在复杂关系推理上的缺陷。Chen等^[15]提出一种具有协同优化的知识感知提示调优方法进行关系抽取。Yu等^[16]通过结合领域本体和LLM的零样本能力,高效构建了大规模服务计算知识图谱BEAR(bidirectional encoder adaptive reweighting),解决了数据稀缺和知识复杂性的挑战。

然而,直接将大语言模型应用于文物长文本的三元组抽取仍存在诸多限制:(1)文物描述常包含冗余信息、停用词、语气词以及非结构化叙述(如修辞、描述性语句、引文等),易引发模型幻觉^[17];(2)长文本处理消耗大量计算资源,影响抽取效率;(3)缺乏适配文物领域的语义框架,导致抽取结果准确性不高。

针对上述问题,本文结合大语言模型的语义理解能力与知识图谱的结构化表达优势,设计并实施了一套面向唐蕃古道文物知识图谱的构建方法。首先,从概念层梳理文物领域的核心概念及其语义关系,明确知识体系的语义框架;在此基础上,设计分步式知识抽取策略,借助大语言模型的提示工程实现高质量三元组的自动提取,最终构建出唐蕃古道文物知识图谱,实现文物领域知识的高效获取与结构化表达。

综上所述,本文的主要贡献包括:

(1)构建唐蕃古道文物概念层。系统梳理文物领域的核心概念、属性及其关联关系,建立涵盖文物实体、时间、地点与文化的语义网络框架,为后续知识图谱构建提供理论基础。

(2)提出适用于文物领域的分步知识抽取策略。采

用“分步抽取+验证机制”的方法,有效应对多源异构数据带来的抽取难题,提升知识抽取准确率,同时降低模型计算开销与错误率。

(3)实现唐蕃古道知识图谱的构建与评估。基于DeepSeek-R1模型完成知识图谱构建实验,结果表明所提方法F1值达到86.25%,较传统BERT-BiLSTM-CRF(bidirectional encoder representations from transformers-bidirectional long short-term memory-conditional random field)模型提升3.13个百分点。通过消融实验进一步验证各模块有效性,为文物领域知识图谱构建提供了可复用的技术方案。

1 研究方法

知识图谱的构建通常涵盖两个核心层次:概念层与实例层。其中,概念层作为知识图谱的逻辑基础,主要负责定义领域知识的语义结构和本体框架,通常通过本体(ontology)的构建与管理来实现;而实例层则由具体的事实性知识组成,以结构化三元组的形式表达特定领域的实际数据^[18]。针对文物领域的专业特性以及唐蕃古道沿线文物的语义需求,设计了一个分层结构的知识图谱体系架构,如图1所示。

概念层构建方面,在充分挖掘考古报告、历史文献及行业标准的基础上,结合国际通用的文化遗产本体CIDOC CRM中的核心类定义,并融合文物领域专家的研究经验,对本体库中所涵盖的实体类型、关系类别及属性特征进行了细化与扩展,从而构建出面向文物领域的概念层模型,为后续实例层的知识组织提供标准化的语义框架。

实例层构建方面,在概念层所提供的本体框架指导下,实例层采用自底向上的方法,对唐蕃古道沿线文物

相关数据进行采集、清洗与整合等预处理操作;在此基础上,设计了分步知识抽取策略,并将概念层中的本体信息作为模板输入至大语言模型中,辅助完成三元组的自动抽取,最终,将结构化的文物知识以三元组形式存储于Neo4j图数据库中,完成唐蕃古道文物知识图谱的构建。

2 知识图谱构建

2.1 概念层构建

概念层的构建旨在规范实例层的数据结构和语义关系,确保知识图谱的逻辑结构清晰且符合领域实际需求,并为实例层的构建提供语义标准和理论基础支持^[19]。本文采用Protégé5.6.4作为本体构建工具,使用网络本体语言(web ontology language, OWL)作为本体描述语言,通过“斯坦福七步法”对唐蕃古道沿线文物的本体进行构建。

文物知识本体是一种对文物实体、实体关系及其属性进行语义描述的框架,能够为文物领域的知识建模与语义表达提供结构化支持^[20]。通过分析文物领域的知识体系,明确了文物领域的概念范围,并评估本体复用的必要性。借助CIDOC CRM模型,明确了唐蕃古道沿线文物知识本体的核心概念,在继承CIDOC CRM国际标准核心语义框架的基础上,结合唐蕃古道文物领域的特点进行了优化与创新。一方面保留了CIDOC CRM中的核心类,如“E12_物品”(E12_Proposed_Object)、“E53_地点”(E53_Place)和“E52_时间跨度”(E52_Time-Span),并继承其语义内涵,确保了本体在数据层面与国际标准的兼容性与互操作性。另一方面基于文物领域专家知识,对本体结构进行了细化与扩展,新增了“纹饰”“制作工艺”等具有显著领域特征的核心子类。该设计不仅增强了本体对文物领域知识的表达能力,也显著提升了后续知识抽取任务的准确性与针对性。结合领

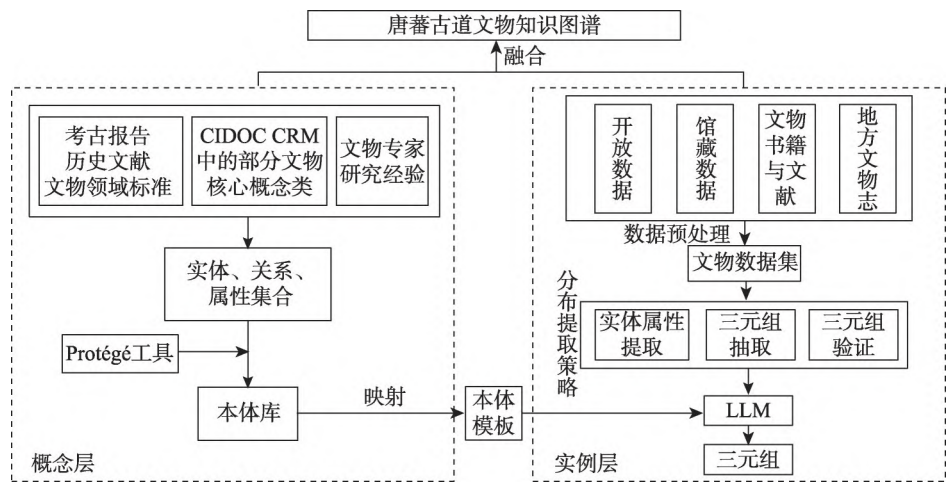


图1 唐蕃古道文物知识图谱构建架构图

Fig.1 Architecture for construction of Tang-Bo Ancient Road cultural heritage knowledge graph

域特点,列举了文物领域的相关术语,为核心类与属性的定义奠定了基础,如表1所示。

表1 文物领域术语(部分)

Table 1 Cultural relics domain terms (Partial)

分类	部分术语列举
年代(Time)	旧石器时代、新石器时代、南北朝、唐、元、清、民国、近现代……
类别(Category)	金银器、铜器、铁器、瓷器、陶器……
形状(Shape)	壶、尊、斧、杯、刀、剑、矛……
质地(Material)	金银、铜、铁、瓷、珐琅……
用途(Function)	乐器、兵器、发饰、容器、法器……
文物来源(Source)	征集购买、捐赠……
完残程度(Condition)	完好、残缺、修复过……
级别(Grading)	国家一级文物、国家二级文物……
象征意义(Symbolic meaning)	吉祥、喜庆……
收藏单位(Collecting institutions)	博物馆、考古研究所、寺庙、档案馆、民间收藏……
纹饰(Pattern)	云纹、龙纹、几何纹、动物纹、花卉纹、瑞兽纹……
制作工艺(Craftsmanship)	铸造镂空、鎏金、烧制、刺绣、篆刻……
……	……

通过系统分析文物领域的概念,本文构建了包含“文物实体”“时间”“地点”“文化”4个主要类别的唐蕃古道沿线文物知识模式体系,形成了整体的语义框架。文物实体类分为6个子类,分别是类别、用途、质地、来源、纹饰、制作工艺;时间类包括抽象时间和具体时间2个子类,其中抽象时间表示文物的历史年代,具体时间则表示出土时间或建设时间;地点类涵盖收藏机构、出土位置2个子类,收藏机构表示文物的保存与管理单位,出土位置指文物最初被发现或挖掘出来的具体地点;文化类则从抽象层面展现文物的文化内涵,涵盖历史文化类与象征意义类的内容,用以描述文物的文化特征与象征属性。类层次结构如表2所示。

在本体构建中,仅通过概念体系难以全面描述领域知识的细节,需要通过属性进一步描述概念间的关联性

表2 唐蕃古道沿线文物概念层次

Table 2 Conceptual hierarchy of cultural relics along Tang-Bo Ancient Road

核心概念类	子类
文物实体(Cultural relic entity)	类别(Category)
	用途(Function)
	质地(Material)
	来源(Source)
	纹饰(Pattern)
	制作工艺(Craftsmanship)
时间(TimeSpan)	抽象时间类(TimeAbstract)
	具体时间类(TimeSpecific)
地点(Location)	收藏机构类(Organization)
	出土位置类(PlaceCreation)
文化(Culture)	历史文化类(CultureHistorical)
	象征意义类(MeaningSymbolic)

以及文物的特征信息^[21]。在本体中通常用两种属性来表示,即数据属性和对象属性,其中数据属性用于描述概念或实体的具体特性,而对象属性用于刻画概念之间的语义关系^[22]。部分属性如表3所示。

表3 本体属性说明(部分)

Table 3 Description of ontology attributes (Partial)

类别	属性名称	属性说明
对象属性	hasCategory	所属类别
	hasTime	所属时间
	hasCulture	关联文化
	hasUsage	用途
数据属性	Title	名称
	Dimensions	尺寸大小
	ExcavationTime	出土时间

最终,本文构建的唐蕃古道文物知识本体模型共包含4个核心类、12个子类、10个对象属性和4个数据属性。唐蕃古道沿线文物的知识图谱概念层如图2所示。

本文按照知识图谱构建三要素原则,构建了“实体-关系-实体”和“实体-属性-属性值”这两种核心语义关系,并成功地搭建起文物知识图谱的骨架。为了更加清晰地展示实体间的语义关系,在模式层构建的基础上,

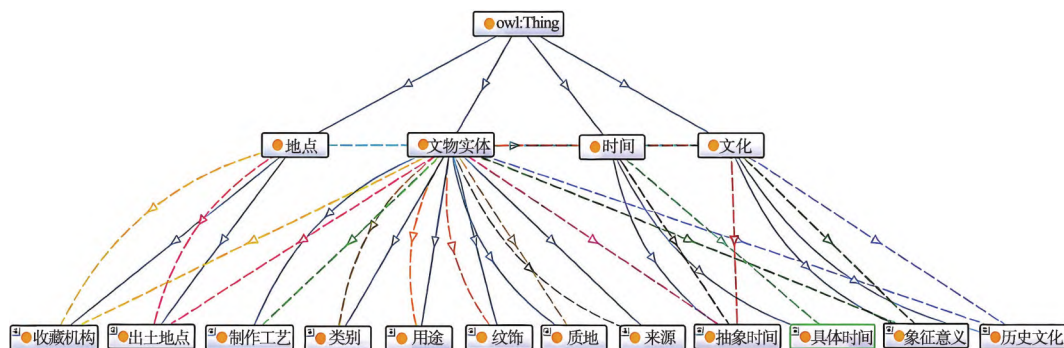


图2 唐蕃古道文物概念图

Fig.2 Conceptual illustration of cultural relics from Tang-Bo Ancient Road

进行了完善与补充,不同实体之间的语义关系列举如表4所示。

表4 “实体-关系-实体”语义关系
Table 4 Semantic relationship of entity-relation-entity

实体	关系	实体
文物实体	出土于	出土位置
文物实体	收藏于	收藏机构
文物实体	具有用途	用途
文物实体	采用材质	质地
抽象时间	包含	具体时间
出土位置	属于	历史文化
文物实体	来源方式	来源
文物实体	装饰纹样	纹饰
文物实体	所属类别	类别
文物实体	制作技术	制作工艺
制作工艺	传承自	历史文化
文物实体	所属年代	抽象时间
文化	关联文化	文物实体
用途	象征含义	象征意义

2.2 基于大语言模型的实例层构建

实例层的构建是在概念层框架基础上的实践应用,也是概念本体的进一步映射^[23],其核心目标在于将抽象的概念结构转化为具体的、结构化的实例数据,从而为知识图谱提供完整的语义支撑和数据支撑。

2.2.1 分步知识抽取

现有基于三元组抽取的自动化知识图谱构建方法在文物领域面临应用瓶颈。文物数据多呈现为包含冗余信息、非结构化叙述及描述性语言的长文本特征,大语言模型在处理此类文本时需承担多重任务负荷,既需完成文本解耦与语义辨识,又面临 Token 容量限制,易引发事实性错误生成。

为应对这一问题,研究了面向文物领域的知识抽取方法。通过大语言模型的提示工程,设计了一种能够有效融合领域特征与任务需求的高效方法:分步知识抽取方法。该方法大大降低模型运行时的计算压力与出错率,同时提升了抽取精度。分步知识抽取流程如图3所示。

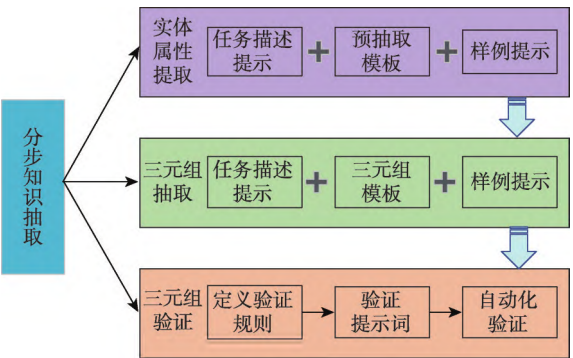


图3 分步知识抽取流程图

Fig.3 Step-by-step knowledge extraction process

第一步:实体属性提取

采用任务描述、预抽取模板、样例提示构建第一步提示词模板,引导大语言模型明确角色定位和抽取目标,确保对文物领域特征的精准理解,从而高效提取出文物的基本信息、特征及文化背景等。其中任务描述包括角色描述、任务需求、限制条件提示,第一步提示词如图4所示。

第一步提示词模板

##预抽取模板##
(["文物基本信息": {"名称": "xxx", "朝代": "xxx", "质地": "xxx", "来源": "xxx", "出土地点": "xxx", "出土时间": "xxx", "收藏单位": "xxx", "类别": "xxx", "文物特征": {"外形": "xxx", "装饰": "xxx", "工艺特点": "xxx"}, "文物文化": {"历史文化": "xxx", "象征意义": "xxx"}})

##示范案例##
输入: 圆釜宽叶倒钩铜矛, 青铜时代, 西宁市沈那遗址出土, 长61.5厘米, 宽19.5厘米, 青海省博物馆藏, 宽叶、钝锋, 厚脊, 矛阔叶状, 双面锋刃。矛中部两面有1.5厘米的脊梁, 釜与刃部结合处设一曲线倒钩。圆釜下端有一组三道圆箍, 釜内遗留残木柄痕迹。器体宽大, 铸造精美。推测该件铜矛是具有象征意义的礼器, 属于受塞伊玛-图尔宾诺文化影响的典型器物。
输出: ["文物基本信息": {"名称": "圆釜宽叶倒钩铜矛", "朝代": "青铜时代", "质地": "铜", "来源": "出土", "出土地点": "西宁市沈那遗址", "出土时间": "无", "收藏单位": "青海省博物馆藏", "类别": "礼器"}, "文物特征": {"外形": "宽叶、钝锋, 厚脊, 矛阔叶状, 双面锋刃", "装饰": "无", "工艺特点": ["厚脊", "曲线倒钩"]}, "文物文化": {"历史文化": ["青铜时代", "受塞伊玛图尔宾诺文化影响"], "象征意义": ["无"]}]

##任务描述##
1. 假如你是文物领域抽取大模型, 对文物文本数据进行抽取, 回答的所有信息必须来自文本, 同时参考示范案例, 请严格按照给出模板指定的格式输出JSON格式内容, 不能添加任何推测或额外内容。确保输出的JSON格式没有转义符号, 对格式中不存在的值填写“无”;
2. 请根据以上JSON数据, 重新提取并修复任何可能的错误或遗漏。
限制条件

图4 第一步提示词模板

Fig.4 First-step prompt template

第二步:三元组抽取

基于初步抽取的结果,采用任务描述、三元组模板构建第二步提示词模板,进一步提升抽取结果的准确性与信息覆盖度。整个流程伴随着思维链方式,通过逐层递进的设计逻辑贯穿始终,能够准确抽取出三元组,同时有效降低了对模型的高要求。第二步提示词如图5所示。

第二步提示词模板

##三元组模板##
[{"head": "", "head_label": "", "relation": "", "tail": "", "tail_label": ""}]

##示范案例##
输入: 1. ["文物基本信息": {"名称": "圆釜宽叶倒钩铜矛", "朝代": "青铜时代", "质地": "铜", "来源": "出土", "出土地点": "西宁市沈那遗址", "出土时间": "无", "收藏单位": "青海省博物馆藏", "类别": "礼器"}, "文物特征": {"外形": "宽叶、钝锋, 厚脊, 矛阔叶状, 双面锋刃", "装饰": "无", "工艺特点": ["厚脊", "曲线倒钩"]}, "文物文化": {"历史文化": ["青铜时代", "受塞伊玛图尔宾诺文化影响"], "象征意义": ["无"]}]
2. 圆釜宽叶倒钩铜矛, 青铜时代, 西宁市沈那遗址出土, 长61.5厘米, 宽19.5厘米……(文本)
输出: [{"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "所属年代", "tail": "青铜时代", "tail_label": "时代"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "出土地点", "tail": "西宁市沈那遗址", "tail_label": "出土地点"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "收藏于", "tail": "青海省博物馆藏", "tail_label": "收藏单位"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "长度", "tail": "61.5厘米", "tail_label": "尺寸"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "宽度", "tail": "19.5厘米", "tail_label": "尺寸"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "重量", "tail": "无", "tail_label": "重量"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "材质", "tail": "铜", "tail_label": "金属"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "工艺", "tail": "厚脊", "tail_label": "工艺"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "纹饰", "tail": "无", "tail_label": "纹饰"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "用途", "tail": "无", "tail_label": "用途"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "象征意义", "tail": "无", "tail_label": "象征意义"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "文化背景", "tail": "无", "tail_label": "文化背景"}, {"head": "圆釜宽叶倒钩铜矛", "head_label": "文物", "relation": "其他", "tail": "无", "tail_label": "其他"}]

##任务描述##
1. 假如你是文物领域三元组抽取大模型, 对JSON格式数据进行三元组抽取, 同时参考文本上下文和示范案例补充任何遗漏或修复可能出现的错误; 请严格按照输出模板指定的格式输出内容; 对于同义词或表达方式不同的描述, 请进行同义词替换, 确保没有重复的或者相同含义的三元组; 确保输出的三元组具有清晰的语义, 避免歧义。
2. 请参考以上抽取出的三元组数据, 重新提取并修复任何可能的错误或遗漏。
限制条件

图5 第二步提示词模板

Fig.5 Second-step prompt template

2.2.2 验证机制

对第二步生成的三元组(头实体-关系-尾实体)进行逻辑性、领域一致性和事实准确性的多维度校验,确保其符合文物领域知识图谱的构建标准。其具体步骤如下所示:

步骤1 定义验证规则,预定义合法关系列表[“出土于”“所属年代”“收藏于”“关联文化”“...”]。限定实体类型匹配规则:关系两端实体需符合类型约束,如“出土于”关系的尾实体必须为地点类实体。

步骤2 构建验证提示词,具体的提示如下:假如你是一名文物知识图谱质检专家,需严格验证三元组的格式、领域一致性、事实性和逻辑性。

步骤3 自动化验证流程。针对提取的每个三元组,大模型需执行以下操作:判断三元组中的关系是否属于预定义的合法关系列表;判断实体类型是否匹配;基于原文内容及外部知识库(权威文物数据库),核查三元组是否得到原文直接或逻辑隐含支持。若三元组通过全部检验,则标记为有效;否则标记为无效,并对无效三元组进行修正。

输出示例:

有效案例:输入:("圆釜宽叶倒钩铜矛","出土于","西宁市沈那遗址")。LLM分析:关系“出土于”合法,尾实体为地点,原文明确提及。输出:{"is_valid": true, "error_type": "无", "correction": ""}。

无效案例:输入:("圆釜宽叶倒钩铜矛","材质","青铜")。LLM分析:原文未明确提及材质,属幻觉推断。输出:{"is_valid": false, "error_type": "事实缺失", "correction": "删除"}。

步骤4 人工复核。对修改后的三元组进行人工检查,看是否合理。

步骤5 存储三元组。将最终获取到的三元组数据存储到Neo4j数据库中。

算法伪代码:

1. 构建提示模板、相关提示词、相关案例和关系列表
2. 编写连接Neo4j图数据库的函数
3. for i=0 to N do//遍历文物数据集
4. 使用文物数据填充第一步提示词模板,生成问题
5. 将问题提交给LLM,获得初步抽取结果
6. 清洗初步结果
7. for j=0 to M do//遍历初步抽取结果
8. 基于第一步抽取结果填充第二步提示词模板,生成补充完善问题
9. 将补充问题提交给LLM,获得三元组结果
10. 再次清洗和标准化结果
11. 构建当前文物、材质、出土地点、收藏机构等唯一id
12. for k=0 to P do//遍历三元组集合

13. 对三元组集合多维度校验,获得有效三元组与修改后的无效三元组,人工复核

14. 写入图谱

15. end for

16. end for

17.end for

3 实验

3.1 实验数据

由于目前尚缺乏针对唐蕃古道沿线文物的公开数据集,根据研究需求,自主收集并构建了相应的文物数据集。该数据集来源广泛,主要包括以下几个方面:一是通过网络爬虫技术获取互联网上的开放数据;二是整理各大博物馆官方网站及其数据库中发布的馆藏文物信息;三是参考权威文物专著与学术论文中的研究成果;四是依据地方文物志及地方文化遗产档案资料。

为保障所构建数据集的准确性、可靠性与完整性,对原始数据进行了系统化的预处理。首先,去除重复条目并消除冗余信息,以确保数据的一致性。其次,针对部分数据中存在的错误信息,采用人工校验结合多源交叉验证的方式进行修正;对于缺失的关键属性信息,则通过查阅权威文献和相关资料予以补全。此外,在处理地方文物志和馆藏数据中出现的模糊描述时,实施了语义标准化处理。例如,将“青海省海西州都兰县热水墓群”统一简化为标准地理标识“都兰县热水墓群”。经过清洗、去重、校验、补全以及规范化等流程后,最终构建形成了结构清晰、语义明确的唐蕃古道沿线文物数据集。

为进一步提高标注结果的准确度与一致性,本研究从可靠文献来源中选取了396条样本数据用于标注实验。多名具备专业背景的标注人员在不互通结果的情况下,分别独立完成样本标注。在此基础上,通过计算科恩卡帕系数(Cohen's Kappa)以评估标注者间一致性,并对标注不一致的案例引入领域专家进行第三方仲裁,从而形成最终一致的标注样本集合。

3.2 实验参数

本文所有实验均在一台配备64 GB内存的GPU服务器上运行,操作系统为Linux(Ubuntu 20.04)。实验采用Python 3.10作为编程语言,并基于PyTorch 2.1.0框架完成模型的构建与训练。基准模型选用BERT-BiLSTM-CRF,其关键参数配置如下:批处理大小(batch_size)设为1,输入序列最大长度(max_seq_len)为512,标签长度(num_p)为30,学习率(learning_rate)为1E-5,训练轮数(Epoch)为10。

为适应实验需求,对实验所需模型相关生成参数进行了统一调整:将生成过程中的随机性参数(tempera-

ture)设定为0.9,以增强输出结果的多样性。针对不同模型采用了差异化的采样策略:在调用ChatGLM模型时,采用top_p采样策略,设置top_p=0.7;而在使用DeepKE模型时,则采用top_k采样策略,设定k=5。其中,top_k策略从概率最高的k个候选词中进行采样,top_p策略则依据累积概率分布选择词汇。上述参数与策略的设置,旨在有效控制生成文本的随机性与多样性,从而提升模型输出的灵活性与质量,优化整体实验效果。

3.3 评价指标

本文采用了最常用的评价指标F1值、精确率P(Precision)、召回率R(Recall)3个指标对唐蕃古道文物知识抽取结果进行评估。

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%$$
 (1)

$$P = \frac{TP}{TP + FP} \times 100\%$$
 (2)

$$R = \frac{TP}{TP + FN} \times 100\%$$
 (3)

式中,TP表示模型正确提取的三元组数量,FP表示模型错误提取的三元组数量,FN表示模型未能提取的正确三元组数量。精确率反映模型提取出的三元组中正确结果所占的比例;召回率则是模型正确提取的三元组占全部正确三元组的比例;F1值是精确率和召回率的加权平均值。本文以F1值作为综合性评价指标,对模型性能进行整体判断。

3.4 实验结果

通过利用概念层中的本体结构及分步知识抽取方法,本文完成了对实体及其关系的有效抽取,具体结果如表5所示。模型在实体识别与关系抽取任务中分别取得了88.16%和84.78%的F1值,表明本文所设计的方法能够高效、准确地完成文物领域知识三元组的提取。

表5 实体关系抽取结果

Table 5 Entity-relation extraction results 单位:%			
类型	P	R	F1
实体	89.56	86.78	88.16
关系	85.83	83.77	84.78

为全面评估不同模型在知识抽取任务中的表现,实验中,以BERT-BiLSTM-CRF模型作为基准模型,开展实体与关系的联合抽取任务,训练集与验证集按照7:3的比例随机划分。为进一步比较不同方法的抽取效果,本文还引入多种大语言模型进行对比实验,包括ChatGLM3-6B、Qwen2.5-7B、Llama3-8B以及DeepSeek-R1-7B。针对这些大模型,采用分步式知识抽取策略,即通过设计合理的提示(prompt)引导模型依次完成实体识别与关系抽取,获取其原生的抽取结果,以全面评估各类模型在文物领域知识抽取任务中的表现。此外,为增

强实验结果的可比性与说服力,还引入了被广泛认可的UIE模型和DeepKE模型作为重要对比对象。

实验结果如表6所示。从整体性能来看,DeepSeek-R1与Qwen2.5显著优于基线模型,展现出更强的语义理解能力与结构化信息提取能力。相比之下,Llama3与ChatGLM3的表现略低于基线模型。进一步分析其信息抽取结果发现,部分文物文本描述具有较强的口语化与形象化特征,例如在描述纹饰时频繁使用“螭龙盘绕”“云雷纹交叠”等非标准化表达。此类表述虽在考古文献中较为常见,但对部分大语言模型而言仍存在语义解析与实体识别上的困难,从而影响了抽取效果。

表6 不同模型抽取结果对比

Table 6 Comparison of extraction results by different models 单位:%			
模型	F1	P	R
BERT-BiLSTM-CRF	83.12	84.98	81.34
ChatGLM3	72.31	71.53	73.16
Qwen2.5	84.66	85.10	84.24
Llama3	77.55	78.23	76.89
UIE	80.75	82.10	79.43
DeepKE	83.07	82.20	83.95
DeepSeek-R1	86.25	85.93	86.58

此外,通用信息抽取框架UIE与DeepKE的表现亦略逊于基线模型。这并非源于其框架设计存在根本性缺陷,而是反映出在面对高度专业化、表达非标准化的领域文本时,模型对领域适配性的要求极为严格。UIE与DeepKE在资源丰富、标注规范的通用场景下表现优异,具备良好的“开箱即用”能力,然而在文物这一特定垂直领域中,由于术语特殊、表达多样且缺乏大规模标注数据,其直接应用效果受限,凸显了通用模型向专业领域迁移时所面临的挑战。

值得指出的是,尽管各模型在参数规模、算力需求及算法架构等方面存在差异,但在本文提出的“分步抽取+验证机制”策略引导下,即通过任务导向型问答方式将复杂语料拆解为可处理单元,逐步解析并辅以精修去噪,即便是性能相对较弱的模型(如ChatGLM3-6B),依然能够达到72.31%的F1值。这一结果充分说明,本文方法在结构化设计与任务分解机制方面具有良好的通用性与鲁棒性。进一步分析表明,该分步抽取方法不仅能在高性能模型(如DeepSeek-R1、Qwen2.5)中发挥极致性能,亦可在中小规模模型或资源受限环境中实现稳定运行,从而维持高效、精准的知识获取能力。这种不依赖“巨模型、高能耗”的推理方式,而是通过结构化任务引导驱动抽取过程的设计思路,有效规避了大模型在开放语境中易出现的信息冗余、逻辑错位及抽取边界模

糊等问题。

综上所述,本文提出的知识抽取策略在实际应用场景中展现出显著优势。对于地方性文物普查、区域文化遗产数字化建设、基层博物馆知识管理等资源有限及环境复杂的场景而言,该方法提供了一条兼具高质量与低开销、稳定性与普适性的技术路径,为未来文物知识图谱的低成本、规模化构建提供了坚实支撑。

3.5 消融实验

为了验证本文提出的分步知识抽取策略的可行性和必要性,基于DeepSeek-R1模型进行了**消融实验**。实验设计了3种不同的三元组抽取方式:直接给定任务需求进行三元组抽取、仅采用分步抽取(未使用验证机制)以及结合“分步抽取+验证机制”的方法进行抽取。所有实验均使用相同的数据集以确保对比的有效性。实验结果汇总于表7。

表7 消融实验结果

Table 7 Results of ablation experiments 单位: %

分步抽取策略	P	R	F1
只给定任务需求	68.41	66.24	67.30
分步抽取(无验证策略)	82.19	80.31	81.24
分步抽取+验证机制	87.58	84.73	86.15

从消融实验的结果可以看出,本文提出的“分步抽取+验证机制”策略显著提升了三元组抽取的整体性能。具体而言,相较于仅给定任务需求直接进行三元组抽取的方式,采用分步抽取策略(但未引入验证机制)时,F1值提升了13.94个百分点。这一显著提升证明了分步处理框架在降低复杂文本理解难度方面的有效性。

进一步地,在分步抽取的基础上加入验证机制后,模型性能得到了持续改进,F1值达到了86.15%,相比未使用验证机制的情况提升了4.91个百分点。这表明,验证模块能够有效地识别并纠正分步抽取过程中产生的错误累积,从而提高最终抽取结果的准确性。实验数据证实,“分步抽取+验证机制”通过将复杂的抽取任务分解为一系列可管理的子任务,并利用验证机制对每个步骤的结果进行纠偏,共同构成了本文方法性能提升的关键因素。通过“任务拆分+验证加工”的双重机制,该方法不仅减轻了模型对复杂文物文本的理解负担,还大幅提升了信息抽取的准确性和完整性。

4 错误类型分析

为深入理解大语言模型在唐蕃古道文物知识图谱构建过程中生成三元组的错误机制,本文对抽取结果中的错误类型进行了系统分类与成因分析,旨在为后续分步知识抽取策略的优化提供数据支撑。本文系统构建了错误三元组的分类体系,如表8所示。

表8 错误三元组分类体系

Table 8 Error triple classification system

错误类别	子类	典型案例
实体识别错误	边界错误	将“贴金彩绘围屏石榻”错误切分为“贴金”“彩绘”“围屏石榻”3个实体
	文化术语歧义	“来通”(酒器)被误识别为动词短语
模型幻觉	缺乏原文本支持	“圆釜宽叶倒钩铜矛”未提及材质时断言“材质为青铜”
关系错误	类型错配	“三彩载货驼俑”与“唐代”的关系误标为“制作于”(应为“所属年代”)
时空错误	年代冲突	将魏晋时期“胡人牵驼画像砖”关联到唐代
其他	数据标准化缺失	“湟源”未规范化为“青海省湟源县”导致地点实体识别失败

错误统计分析结果显示(见图6),实体识别错误与模型幻觉合计占比达78.4%,是影响知识图谱构建质量的核心因素。通过对错误样本的深入挖掘,本文识别出3类高频错误类型,具体如下:

边界错误:主要源于复杂修饰结构解析失败。例如,“贴金彩绘围屏石榻”被错误切分为“贴金”“彩绘”“围屏石榻”3个实体,而实际上应作为一个复合实体整体识别。**实验表明,当名词短语中包含3个及以上修饰成分时,实体边界识别错误率显著上升。**

文化术语歧义错误:由专业术语的跨文化歧义引发。例如,“来通”在文物语境中特指一种古代酒器,而非动作性动词;**此类错误在涉及少数民族文化背景的文物描述中尤为常见,反映出模型对领域特定术语的文化语境理解不足。**

模型幻觉:模型对文本理解不充分、不准确,导致其基于错误的理解“脑补”出不存在于源文本中的信息,需要进行人工验证。例如“圆釜宽叶倒钩铜矛”未提及材质时断言“材质为青铜”。

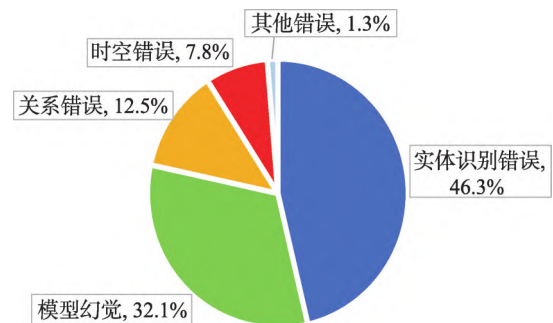


图6 错误类型比例图

Fig.6 Proportion of error types

分析揭示了当前模型在复杂实体识别、文化语境理解和生成可靠性方面的关键缺陷,后续研究可基于此进行改进,以系统提升领域知识图谱的构建质量。

5 知识图谱展示与分析

基于前述本体构建方法与分步知识抽取策略,本文利用大语言模型对唐蕃古道文物数据进行了系统化的知识抽取,最终生成结构化的知识三元组(实体-关系-实体)。所有提取的知识以统一格式进行标准化处理后,被存储至Neo4j图数据库中,完成了唐蕃古道文物知识图谱的构建。

该知识图谱共包含7 866个实体节点、2 138种关系类型,形成共计16 709对三元组的知识网络。依托Neo4j强大的图数据可视化能力,可实现对文物信息的多维度展示与交互式查询。例如,图7展示了“圆釜宽叶倒钩铜矛”的可视化效果,直观呈现了该文物的基本信息。

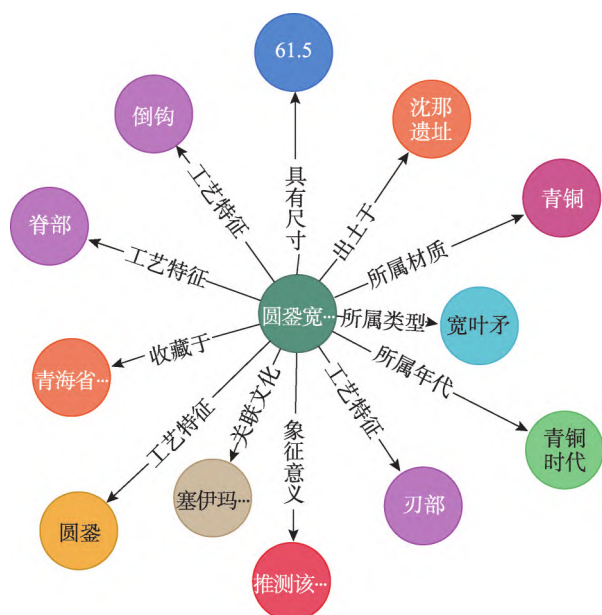


图7 “圆釜宽叶倒钩铜矛”可视化展示

Fig.7 Visualization of barbed bronze spear

此外,知识图谱还支持复杂查询与初步的数据分析功能。为深入了解唐蕃古道文物中唐代金器常见的纹饰特征,本文设计并执行了一条Cypher查询语句如下:

```
MATCH (a:Artifact {era:"唐代", type:"金器"})-[:HAS_PATTERN]->(p:Pattern)
WITH p.name AS pattern, COUNT(*) AS freq
WHERE freq > 5
RETURN pattern, freq
ORDER BY freq DESC LIMIT 5
```

查询显示,出现频率前5的纹饰分别为联珠纹、宝相花纹、动物纹、缠枝花卉纹、摩羯纹。进一步统计其占比情况如图8所示。结果表明,唐代金器纹饰呈现出“以几何纹为主导,植物纹与神兽纹并重”的分布特征,反映出当时工艺美术风格的多元融合趋势。

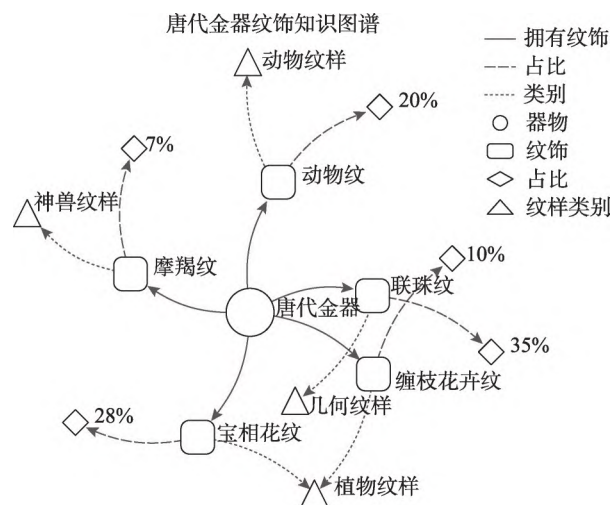


图8 唐代金器纹饰占比图

Fig.8 Proportion of decorative patterns on Tang dynasty gold artifacts

6 结束语

针对唐蕃古道文物领域知识抽取中面临的多源异构数据处理困难、标注成本高昂以及领域知识缺失等核心挑战,提出一种融合大语言模型与分步知识抽取策略的文物知识图谱构建方法。通过设计“任务型问答粗抽取-精修加工细抽取”的分层抽取架构,有效缓解了大语言模型在处理复杂物文本时的理解压力,降低了对大规模人工标注数据与超参数调优模型的依赖,在提升抽取精度的同时兼顾了计算效率。实验结果表明,本文方法在实体与关系联合抽取任务中取得了显著成效,F1值达到86.25%,相较传统深度学习方法(BERT-BiLSTM-CRF)提升了3.13个百分点,为文物领域知识图谱的轻量化、高效构建提供了新的技术路径和方法范式。本文在知识图谱概念层的设计上仍存在一定局限,表现为扩展性不足、应用领域受限等问题。在后续研究中,将进一步拓展数据来源与规模,完善知识图谱的整体结构。

参考文献:

- [1] PENG C Y, XIA F, NASERIPARSA M, et al. Knowledge graphs: opportunities and challenges[J]. Artificial Intelligence Review, 2023, 56(11): 13071-13102.
- [2] LEE Y, CHUNG J J Y, KIM T S, et al. Promptiverse: scalable generation of scaffolding prompts through human-AI hybrid knowledge graph annotation[C]//Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2022: 1-18.
- [3] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a web of open data[C]//Proceedings of the 2007 International Semantic Web Conference. Berlin: Springer, 2007:

- 722-735.
- [4] VRANDEČIĆ D, KRÖTZSCH M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.
- [5] MARCHAND E, GAGNON M, ZOUAQ A. Extraction of a knowledge graph from French cultural heritage documents [C]//ADBIS, TPD and EDA 2020 Common Workshops and Doctoral Consortium. Cham: Springer, 2020: 23-35.
- [6] CARRIERO V A, GANGEMI A, MANCINELLI M L, et al. Pattern-based design applied to cultural heritage knowledge graphs[J]. Semantic Web, 2021, 12(2): 313-357.
- [7] PELLEGRINO M A, SCARANO V, SPAGNUOLO C. Move cultural heritage knowledge graphs in everyone's pocket[J]. Semantic Web, 2023, 14(2): 323-359.
- [8] 李永卉, 刘沁芃, 卢章平. 基于知识图谱的南朝陵墓石刻信息资源开发研究[J]. 图书馆杂志, 2023, 42(9): 86-93.
- LI Y H, LIU Q P, LU Z P. Research on information development of Southern Dynasties' mausoleum stone carvings based on knowledge graph[J]. Library Journal, 2023, 42(9): 86-93.
- [9] 周冬艳, 高鲁放, 吴艳芳, 等. 基于本体的青瓷知识图谱构建方法研究[J]. 图书馆杂志, 2024, 43(3): 91-100.
- ZHOU D Y, GAO L F, WU Y F, et al. Research on celadon knowledge map construction based on ontology[J]. Library Journal, 2024, 43(3): 91-100.
- [10] XU D R, CHEN W, PENG W J, et al. Large language models for generative information extraction: a survey[J]. Frontiers of Computer Science, 2024, 18(6): 186357.
- [11] LIU Y R, ZHOU J P, SANG G B, et al. The journey of language models in understanding natural language[C]//Proceedings of the 2024 International Conference on Web Information Systems and Applications. Singapore: Springer, 2024: 331-363.
- [12] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute-optimal large language models[EB/OL]. [2025-07-12]. <https://arxiv.org/abs/2203.15556>.
- [13] WU T Y, HE S Z, LIU J P, et al. A brief overview of ChatGPT: the history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122-1136.
- [14] CHEN B H, BERTOZZI A L. AutoKG: efficient automated knowledge graph generation for language models[C]//Proceedings of the 2023 IEEE International Conference on Big Data. Piscataway: IEEE, 2023: 3117-3126.
- [15] CHEN X, ZHANG N Y, XIE X, et al. KnowPrompt: knowledge-aware prompt-tuning with synergistic optimization for relation extraction[C]//Proceedings of the ACM Web Conference 2022. New York: ACM, 2022: 2778-2788.
- [16] YU S, HUANG T, LIU M Y, et al. BEAR: revolutionizing service domain knowledge graph construction with LLM [C]//Proceedings of the 2023 International Conference on Service-Oriented Computing. Cham: Springer, 2023: 339-346.
- [17] SHANAHAN M. Talking about large language models[J]. Communications of the ACM, 2024, 67(2): 68-79.
- [18] ZHONG L F, WU J, LI Q, et al. A comprehensive survey on automatic knowledge graph construction[J]. ACM Computing Surveys, 2023, 56(4): 1-62.
- [19] MA X G. Knowledge graph construction and application in geosciences: a review[J]. Computers & Geosciences, 2022, 161: 105082.
- [20] 高劲松, 张强, 李帅珂. 可移动文物的知识图谱构建及关联数据存储: 以湖北省博物馆为例[J]. 现代情报, 2022, 42(4): 88-98.
- GAO J S, ZHANG Q, LI S K. Construction of knowledge graph and linked data storage of movable cultural relics: taking Hubei provincial museum as an example[J]. Modern Information, 2022, 42(4): 88-98.
- [21] DOU J H, QIN J Y, JIN Z X, et al. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage[J]. Journal of Visual Languages & Computing, 2018, 48: 19-28.
- [22] TANG X M, FENG Z Q, XIAO Y T, et al. Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development[J]. Geoscience Frontiers, 2023, 14(5): 101426.
- [23] 王娜, 蒋智慧. 动态本体构建的国内外研究现状综述[J]. 现代情报, 2020, 40(4): 159-166.
- WANG N, JIANG Z H. Research overview on dynamic ontology construction at home and abroad[J]. Modern Information, 2020, 40(4): 159-166.



刘华瑞(2000—),男,硕士研究生,主要研究方向为知识图谱、大语言模型等。

LIU Huarui, born in 2000, M.S. candidate. His research interests include knowledge graph, large language model, etc.



雒伟群(1969—),男,硕士研究生,教授,主要研究方向为自然语言处理、知识图谱、大语言模型等。

LUO Weiqun, born in 1969, M.S. candidate, professor. His research interests include natural language processing, knowledge graph, large language model, etc.



张兆基(1978—),男,硕士研究生,教授,主要研究方向为知识图谱、计算机应用等。

ZHANG Zhaoji, born in 1978, M.S. candidate, professor. His research interests include knowledge graph, computer applications, etc.