·情报方法与技术创新·

融合知识图谱与大语言模型的科技文献复杂知识对象抽取研究

陈文杰[1,2,3] 胡正银[1,2,3*] 石 栖[1,2] 卢 颖[1,3]

（1. 中国科学院成都文献情报中心，四川 成都 610299；
2. 中国科学院大学经济与管理学院信息资源管理系，北京 100190；
3. 四川省科技信息智能挖掘与应用工程研究中心，四川 成都 610299）

摘 要：［目的/意义］科技文献复杂知识对象对科技文献中的深度知识内容进行细粒度、全面的知识表示，可有效支撑数智驱动的科学发现与知识发现，是重要的科技创新要素。［方法/过程］首先，通过轻量级本体构建、BRAT 知识标注和 Neo4j 知识存储等步骤实现领域知识图谱构建，其次，本地化部署大语言模型 ChatGLM2-6B 并通过低秩适应（Low-Rank Adaptation，LoRA）技术微调模型，最后基于思维记忆（Memory of Thoughts，MOT）机制将知识图谱中的复杂知识注入提示中，通过与大语言模型的多轮问答从科技文献中抽取出复杂知识对象。［结果/结论］以有机太阳能电池（Organic Solar Cells，OSC）为例验证方法的有效性，结果表明融合知识图谱与大语言模型的抽取方法优于大语言模型单独支撑的抽取方法，在准确率 P、召回率 R 和 F1 值 3 个指标上分别提升 14.1%、10.3%和 12.3%。知识图谱能够增强大语言模型对科技文献的复杂知识对象抽取能力，提升 OSC 领域的科技文献挖掘效率与准确性。

关键词：知识图谱；大语言模型；科技文献；太阳能电池；知识抽取；提示构建

DOI：10.3969/j.issn.1008-0821.2025.07.002

〔中图分类号〕G254 〔文献标识码〕A 〔文章编号〕1008-0821（2025）07-0014-12

Research on Scientific and Technological Literature Complex Knowledge Object Extraction Fusing Knowledge Graph and Large Language Model

Chen Wenjie[1,2,3] Hu Zhengyin[1,2,3*] Shi Xi[1,2] Lu Ying[1,3]

（1. National Science Library（Chengdu），Chinese Academy of Sciences，Chengdu 610299，China；
2. Department of Information Resources Management，School of Economics and Management，University of Chinese Academy of Sciences，Beijing 100190，China；
3. Science & Technology Information Intelligent Mining and Application Engineering Research Center of Sichuan，Chengdu 610299，China）

Abstract：［Purpose/Significance］The complex knowledge objects in scientific and technological literature provide fine-grained and comprehensive knowledge representation of the deep knowledge content in scientific and technological literature，which can effectively support data-driven scientific and knowledge discovery and is an important element of techno-

收稿日期：2024-09-09

基金项目：国家社会科学基金青年项目“基于超网络的关键核心技术识别与技术机会发现研究”（项目编号：24CTQ044）；国家重点研发计划项目“从生物医学和流行病学研究数据中自动生成因果图的系列工具研发”（项目编号：2022YFF0712000）；中国科学院文献情报能力建设专项“实验方法自动抽取与建模研究”（项目编号：E2C0003008）。

作者简介：陈文杰（1990-），男，工程师，研究方向：智能情报分析。石栖（1998-），女，硕士，研究方向：知识图谱与知识演化。卢颖（1996-），女，助理工程师，研究方向：知识抽取。

通信作者：胡正银（1979-），男，正高级工程师，研究方向：学科领域知识发现。

logical innovation. [Method/Process] Firstly, the domain knowledge graph was constructed through steps such as lightweight ontology construction, BRAT knowledge annotation, and Neo4j knowledge storage. Next, the large language model ChatGLM2-6B was locally deployed and fine tuned through LoRA technology. Finally, based on the MOT mechanism, the knowledge in the knowledge graph was injected into the prompts, and complex knowledge objects were extracted from scientific literature through multiple rounds of Q&A with the large language model. [Result/Conclusion] Taking organic solar cells(OSCs) as an example to verify the effectiveness of the method, the results show that the extraction method intergrating knowledge graph and large language model is superior to the extraction method supported by large language model alone, with improvements of 14.1%, 10.3%, and 12.3% in accuracy P, recall R, and F1 score, respectively. Knowledge graph can enhance the ability of large language models to extract complex knowledge objects from scientific literature, and improve the efficiency and accuracy of scientific literature mining in the OSC field.

Key words: knowledge graph; large language model; knowledge extraction; prompt building

科技文献中蕴含大量“可信、专业、规范”的领域知识与科学数据组成的复杂知识对象，是重要的科技创新要素[1]。知识单元是揭示文献知识内容的基本元素，通常以三元组、特征向量和属性—值对等形式描述文献的研究问题、实验原理和研究主题等特征[2]。而科技文献复杂知识对象是由若干知识单元关联、组织形成的统一知识结构，以面向对象的视角对科技文献中的深度知识内容进行细粒度、全面的知识表示。例如，科技文献中实验方案通常包括实验原理、实验元素、实验步骤、实验结果等不同类型的知识。其中，实验原理可由简单的知识单元进行表示，实验元素通常是由实验材料、实验试剂、科学仪器等知识单元组成复合型知识对象，而实验步骤包含科学实验流程等时序性知识对象，实验结果则是一种知识与数据融合性的知识对象。上述由实验元素、实验步骤、实验结果构成的实验方案就是一种典型的科技文献复杂知识对象[3]。通过对这些复杂知识对象进行抽取与分析，能够发现不同学科领域内潜在的、深层次的知识关联与传递，可有效支撑数智驱动的科学发现与知识发现[1]。传统的知识对象抽取关注从文本中提取简单知识结构，如实体、关系和简单事实，这些信息通常是扁平的，不需要复杂的结构化处理。而科技文献复杂知识对象抽取旨在识别和抽取科技文献中的细粒度、结构化的知识单元，并将它们组织成更高层次的知识结构，侧重于揭示深层次的知识关联和传递。然而，以领域专家为核心的科技文献知识抽取模式存在效率低下和主观性强等缺陷，难以支撑大规模科技文献的挖掘与建模。因此，如何从科技文献中高效抽取复杂知识对象成为一个困难却有价值的问题。

以 ChatGPT[4] 为代表的大语言模型是一类使用大量文本数据训练的深度神经网络模型，凭借其强大的涌现能力和零样本迁移能力，在知识抽取任务中得到广泛应用[5]。但大语言模型无法覆盖所有领域的知识，尤其是更新速度快、专业性强的领域知识，在适配特定研究领域时存在性能损失和推理能力不足的问题。知识图谱是一种描述客观世界中各类实体和关系的大规模语义网络，通过本体描述实体的层次结构和关联关系，以三元组的形式表示具体知识，可有效对科技文献复杂知识对象进行精准的知识表示。将知识图谱蕴含的形式化知识作为先验知识注入大语言模型，可以有效提升大语言模型在专业领域知识的理解、抽取方面的能力，并增强模型推理结果的可解释性，如将知识三元组转换为指令数据来微调预训练模型以契合下游抽取任务，将领域知识或本体注入提示模版来引导模型对实体和关系类型的识别，利用已有三元组辅助新三元组生成等[6]。

综上，本文旨在利用知识图谱增强大语言模型对科技文献复杂知识对象的知识抽取能力。首先，利用轻量级本体建模方法完成知识图谱模式层构建，通过知识标注工具和图数据库完成知识图谱实例层构建；然后，针对调用 ChatGPT 在线服务接口存在数据与隐私泄露问题，本地化部署了大语言模型 ChatGLM2-6B[7] 并利用 LoRA[8] 技术微调模型，使其更适用于专业领域复杂知识对象抽取任务；最后，为了缓解 ChatGLM2-6B 输出结果不稳定和存在幻觉的问题，利用思维记忆(Memory of Thoughts, MOT)[9] 和知识图谱实现提示(Prompt)的领域知识注入，以增强模型抽取结果的稳定性和有效性。

1 研究现状

现有的科技文献知识抽取方法可以分为传统抽取方法、基于知识图谱的抽取方法和基于大语言模型的抽取方法。传统抽取方法采用人工抽取、规则构建和机器学习等手段，对文献内容进行分类和标注。人工抽取通常利用专业的标注工具，虽然准确率较高，但要求标注者具有专业领域知识并且主观性较强，无法在短时间内完成大批量文献数据的标引。基于规则的抽取方法利用领域专家制定的线索词来构建抽取规则或模版，再通过模式匹配的方式从文本中抽取知识[10]。叶光辉等[11]通过“人工标注—构建规则—模式识别—补充规则”的流程构建知识规则库，从文献中抽取出不同类型的知识单元并对其分布特征进行了分析。郑梦悦等[12]先构建了知识元本体模型，然后统计句子的类型、位置和线索词建立规则库，最后基于本体模型和规则库实现非结构化摘要的知识抽取。这类方法在保证一定准确率的情况下提高了知识抽取的效率，但需要人工参与规则的制定和维护，难以移植到其他学科领域。为了提高知识抽取的效率和可移植性，部分学者开始利用机器学习技术来实现知识对象的自动抽取，涵盖了支持向量机、条件随机场和人工神经网络等模型。Liu X H 等[13]采用半监督方式抽取实体，先利用 k 邻近算法进行实体分类，然后通过条件随机场模型标注实体边界。Lample G 等[14]通过两种神经网络架构实现知识抽取，一种结合双向 LSTM（Long Short-Term Memory）和条件随机场标注实体，另一种基于转移分块模型为句子分段和打标签。这类方法能够极大缓解领域依赖性，但仍存在需要大量标注样本和准确率不高的问题。

除了科技文献固有的文本信息外，外部的知识图谱可以为知识对象的抽取提供一些额外的先验知识，以帮助抽取模型更好地分析、理解文献中蕴含的深层语义信息。Mintz M 等[15]最早通过远程监督的方式为知识抽取任务生成高质量训练数据集，首先对输入文本进行分词、词性标注和依存分析，然后利用知识图谱匹配出现的实体并提取其词法特征和句法特征，最后结合三元组信息和实体特征对文本中的句子进行自动标注以得到远程监督训练数据。Han X 等[16]提出，知识图谱与文本之间的相互注意力机制，在一个统一的语义空间中为知识图谱和文本生成表示向量，从而在知识抽取中能够更好地区分噪声数据和筛选有价值的三元组。为了消除文本中的噪声影响，Hu L M 等[17]采用门控机制，从实体描述信息和知识图谱的结构信息中生成标签，再结合注意力机制筛选有效样例以实现关系的分类。为解决长尾关系问题，Zhang N Y 等[18]利用句子编码器和知识图谱嵌入模型分别学习关系的隐性和显性特征，然后通过知识感知注意力机制增强长尾关系的预测能力。这类方法存在错误传播和长尾关系问题，如何避免将知识图谱中的错误或偏差传播到知识抽取中并有效抽取低频关系仍然是一个难题。

随着大语言模型的出现与发展，加上 BERT 和 ChatGPT 等模型在自然语言任务上展现出的优越性，有学者开始尝试利用大语言模型进行知识对象抽取。Tang X Y 等[19]构建了一个多任务 BERT-BiLSTM-AM-CRF 模型，利用 BERT 提取上下文信息中的动态词向量，接着将 BiLSTM 模块训练后的结果输入 CRF 进行解码，最后利用 CRF 对观测标注序列进行分类和提取得到知识抽取结果。Wei X 等[20]通过与 ChatGPT 的多轮问答实现零样本知识抽取，在第一轮问答识别句子中实体、关系和事件的类型，在后续几轮问答中利用链式抽取模版识别句子中的细粒度知识。Yuan C 等[21]设计了零样本提示、事件排序提示和思考链提示三类提示模版，通过与 ChatGPT 的三轮问答实现零样本时序关系抽取。张颖怡等[5]利用 ChatGPT 通过实体识别、训练集生成和伪标签生成等流程实现学术论文实体识别，并从性能、价格和时间 3 个维度进行了可行性分析。苏杭等[22]提出了一个基于提示调优的两段式知识抽取方法，第一阶段微调预训练模型进行关系分类，第二阶段复用微调后的模型进行实体识别。王震宇等[23]通过计算多模态样本间的相似度生成高质量辅助知识，然后将原始输入与辅助知识输入到大语言模型中实现关系抽取。这类方法在零样本和少样本的知识抽取任务上取得了较优性能，但相关研究仅用于通用领域，对于专业领域科技文献挖掘的研究与探索较少。

以上方法只能从文本中抽取出实体、关系和事件等细粒度、离散化的简单知识对象，而复杂知识

对象的抽取需要对领域知识有深入理解，处理更加复杂的知识结构和语义关系，如时序性、层次性和多维度关系。当前，针对复杂知识对象抽取的研究较少，仍处于探索阶段。对此，本文旨在将知识图谱中有效的专业领域知识注入大语言模型中，通过与大模型的多轮问答实现复杂知识对象的抽取。

2 科技文献复杂知识对象抽取

传统的知识对象抽取关注从文本中提取实体、关系等简单知识结构，这些信息通常是扁平的，不需要复杂的结构化处理。而科技文献复杂知识对象抽取旨在识别和抽取科技文献中的细粒度、结构化的知识单元，并通过语义组织形成更高层次的知识结构，侧重于揭示深层次的知识关联和传递。本节描述了融合知识图谱与大语言模型的科技文献复杂知识对象抽取方法，包括领域知识图谱构建、模型微调和复杂知识对象抽取 3 个阶段，如图 1 所示。其中，第一阶段通过本体构建、BRAT[24] 标注和 Neo4j[25] 存储完成领域知识图谱模式层与实例层构建；第二阶段基于实例层三元组构建指令数据集，利用 LoRA 技术实现大语言模型微调；第三阶段通过 MOT 机制选择 Top-k 的问题答案(Question-Answer，QA)对作为领域知识整合到提示中，经过与模型的多轮问答实现复杂知识对象抽取。

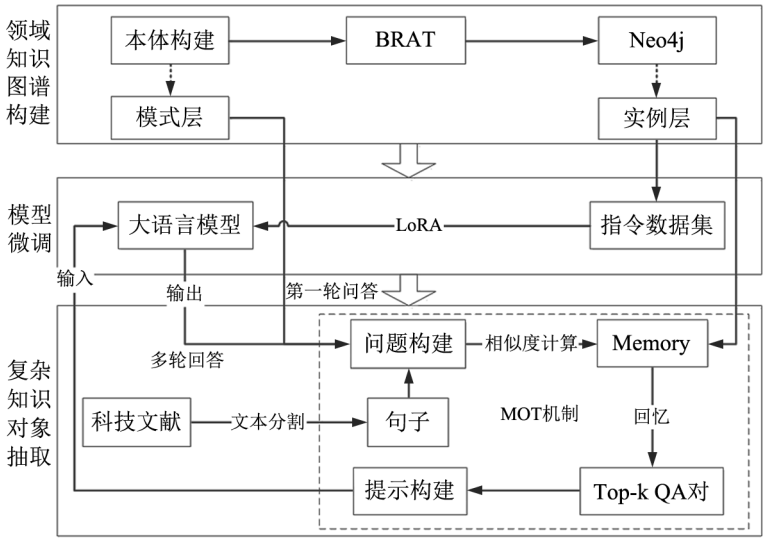

图 1 科技文献复杂知识对象抽取流程

Fig. 1 Process of Complex Knowledge Object Extraction for Technological Literature

2.1 领域知识图谱构建

知识图谱包括模式层和实例层两部分，前者定义实体、关系和属性的层次结构与语义关系，后者以三元组的形式存储具体的领域知识。领域知识图谱的构建分为模式层设计和实例层数据填充两个步骤。本体是概念体系的明确化和规范化的描述说明，将其作为模式层能够更有效地支撑领域知识图谱融合与复用，以形成结构合理、冗余度低和覆盖全面的知识结构。综合研究国内外已有的本体构建方法，发现资源消耗低、轻量化的本体建模方式更适用于特定学科领域的本体快速建立[26]。因此，本文首先利用两阶段式轻量级本体建模方法[26] 实现知识图谱模式层的设计，在准备阶段确定特定学科领域的范围与界限，在构建阶段通过基本框架搭建、知识结构完善和知识结构对齐三轮循环完成各种实体类型、实体之间的语义关系以及实体属性的定义。在本体模型中，将知识实体分为语句级、词汇级和科学数据级 3 种类型，语句级实体是具有特定语义的核心句，词汇级实体是领域术语或关键词，科学数据级实体是特定的评价指标、统计量等。

在完成模式层设计后，通过数据预处理、知识抽取、知识审核和知识存储等步骤实现实例层数据填充。其中，数据预处理从专利和论文等科技文献中提取出文本内容并分割成不同长度的句子。知识抽取阶段基于本体模型构建标签体系结构，利用标注工具 BRAT 实现科技文献中知识实体及其属性关系的标注。BRAT 是一个基于 Web 的快速标注工具，具有高质量可视化页面、多功能标注支持和自然语

言处理(Natural Language Processing，NLP)模型集成等特性，广泛应用于知识抽取任务中。知识审核阶段先由标引人员对标注结果进行交叉验证，对标注出的实体进行统一、规范化的表示，再由领域专家完成最终审定。知识存储阶段将标注结果转换为三元组形式导入图数据库 Neo4j 完成实例层数据填充，Neo4j 内置的 Cypher 语句和图数据科学模块可以方便实现知识的关联查询与深度挖掘。

2.2 大语言模型微调

以 ChatGPT 为代表的大语言模型在零样本和少样本信息抽取任务上表现优异，但是这些模型仅能通过在线的 API 使用，存在数据泄露和不可重复等问题[27]。受限于实验室硬件条件，本文选择能够在消费级显卡上部署的 ChatGLM2-6B 作为基础模型。相较于其他开源模型，ChatGLM2-6B 具备上下文理解和指令遵循能力，可以更好地理解长文本和执行用户的微调指令，灵活适配于不同的下游任务场景。在数据规模较小时，ChatGLM2-6B 内置的微调模块 P-Tuning v2 容易产生“灾难性遗忘”现象，并且会占用下游任务输入序列的空间，而 LoRA 通过低秩适配器能够简单高效地微调大语言模型。基于此，本文本地化部署 ChatGLM2-6B 模型进行科技文献挖掘，利用知识图谱实例层的三元组构建指令数据集，再通过 LoRA 技术微调模型以适配特定学科领域的知识抽取任务。

知识抽取任务包括实体类型识别和实体抽取两部分。对于知识图谱中的三元组(实体，属性，属性值)，首先，利用属性值所在文本和属性构建 QA 对(问题答案对)，用于微调实体类型识别任务；随后，将属性值所在文本和属性构建为问题，属性值构建为答案，用于微调实体抽取任务。具体的微调流程如下：

1）冻结 ChatGLM2-6B 的全部参数。大语言模型的参数量巨大，直接调整其参数效率低下且资源消耗高。

2）通过构建秩分解矩阵 $\boldsymbol{BA}$ 来模拟 ChatGLM2-6B 参数的更新量 $\boldsymbol{\Delta W}$，使得 $\boldsymbol{\Delta W}=\boldsymbol{BA}$。将微调指令数据集作为模型输入，模型训练时仅更新降维矩阵 $\boldsymbol{A}$ 和升维矩阵 $\boldsymbol{B}$，在极大压缩所需更新参数量的同时达到模型微调的效果。

3）多次调整批量大小、迭代次数和训练的学习率以取得最优微调效果。

2.3 复杂知识对象抽取

2.3.1 基于多轮问答的知识抽取

对科技文献 d 进行文本分割得到句子集 S，依次为每个句子 $s\in S$ 构建多个提示，通过与大语言模型的多轮问答得到知识实体，该过程形式化表示如式（1）所示：

$$P(e|s,o)=P(type_1|p_1(s,o))\cdots\cdots P(type_n|p_n(s,o,type_{n-1}))P(e|p_{n+1}(s,o,type_n)) \quad (1)$$

其中，$p_i(i\in[1,n])$ 是第 i 轮问答中构建的提示，o 是领域本体。复杂知识对象由若干知识实体构成，其本体通常被定义为树状层次结构，因而实体类型识别可以视作一个文本层次分类任务。在第 1 到 n 轮问答基于领域本体的树状结构判断句子 s 对应实体类型，对 s 进行自上而下的迭代式分类以得到精确的实体类型。在第 $n+1$ 轮问答基于已识别出的实体类型抽取具体知识实体 e。最后，通过领域本体将从不同句子中抽取出的知识实体组织关联起来形成复杂知识对象。

2.3.2 MOT 机制的提示构建

由于大语言模型生成的内容具有不稳定性、偏见性和存在幻觉[28]，本文在每一轮问答通过 MOT 机制将知识图谱中的领域知识注入提示模版中，以增强模型在特定领域的事实推理和知识抽取能力。MOT 机制基于外部知识构建了一个记忆模块，在设计提示时会从记忆模块中读取最相关且高度置信的知识作为问答样例，通过上下文学习(In-Context Learning)[29]引导模型完成下游任务。相关研究表明，上下文学习对问答样例的顺序、质量和多样性高度敏感[29]。因此，基于 MOT 机制构建提示的流程为：

1）利用领域知识图谱构建 QA 对(与模型微调阶段相同)以形成外部记忆模块 Memory。

2）利用 LDA[30] 主题模型将记忆模块中的全部记忆分为 N 个不同主题的记忆簇 $\{M_1,M_2,\cdots,M_N\}$。主题模型的输出分别为：QA 对—主题分布矩阵 $\boldsymbol{\theta}$ 和主题—单词分布矩阵 $\boldsymbol{\vartheta}$。$\boldsymbol{\theta}(qa)$ 表示问题答案对 qa 在不同主题的分布概率，若 $\boldsymbol{\theta}(qa)_i>\alpha$，则将 qa 划分到 M_i 记忆簇中，超参数 $\alpha\in[0,1]$。

3）从 Memory 中检索出 k 个与目标问题高度

语义相似的 QA 对作为上下文学习的问答样例。对于输入句子 s，构建目标问题 q，依次从每个记忆簇中选择一个与 q 最相关记忆，将得到 N 个有效记忆$\{M'_1,M'_2,\cdots,M'_N\}$作为候选问答样例。若直接用 N 个候选问答样例构建提示，提示的长度会超过 ChatGLM2-6B 输入序列的最大长度限制，因此，还需进一步从候选样例中筛选出相似度 Top-k 的样例。其中，有效记忆计算如式（2）、式（3）所示：

$$M'_i=Max_{m\in M_i}sim(q,m) \tag{2}$$

$$sim(q,m)=\cos(doc2vec(q),doc2vec(m)) \tag{3}$$

$sim(q,m)$采用向量间的余弦相似度来评估目标问题 q 和记忆 m 的语义相似性；doc2vec[31]是一种大型文本向量化模型，能够从不同长度的句子中学习到固定长度的低维向量。

4）将问答样例注入提示，其内容组织为“问答样例：$[M'_1,M'_2,\cdots,M'_k]$目标问题：$[q]$”。提示输入大语言模型后得到回答结果，如式（4）所示：

$$answer=LLM(M'_1,M'_2,\cdots,M'_q) \tag{4}$$

2.3.3 融合多轮问答和 *MOT* 机制的复杂知识对象抽取

在第 1 到 n 轮问答中，基于领域本体的层次结构获取实体类型列表，接着利用大语言模型判断句子文本特征与不同实体类型的匹配度，并基于 MOT 机制获取 Top-k 有效问答样例，生成提示模板如下：

现在你是一个文本分类器，文本类型包括[实体类型 1，实体类型 2，……]

Question：判断句子“文本 1……”的文本类型

Answer：实体类型 1

……

Question：判断句子“文本 k……”的文本类型

Answer：实体类型 k

Question：判断句子“文本 s……”的文本类型

该阶段将实体类型识别任务分解为若干个局部分类任务，从根节点出发自上而下地判断知识实体类型。

得到句子对应的知识实体类型后，判断其粒度，若属于语句级，则直接生成三元组(科技文献 d，知识实体类型，句子 s)；若属于词汇级和科学数据级实体，则基于 MOT 机制构建提示以开启第 n+1 轮问答，进一步提取句子中出现的细粒度知识实体，提示模板如下：

Question：“文本 1……”提取上述句子中[实体类型 1]类型的实体

Answer：知识实体 1

……

Question：“文本 k……”提取上述句子中[实体类型 k]类型的实体

Answer：知识实体 k

Question：“文本 s……”提取上述句子中[实体类型 s]类型的实体

抽取出知识实体后，得到三元组(科技文献 d，知识实体类型，知识实体)，然后进一步判断是否满足本体中定义的约束条件，去除不满足约束的三元组，如科学数据级实体通常被限定为一定范围内的数字。复杂知识对象抽取算法的详细步骤如表 1 所示。

3 实 验

3.1 OSC 领域知识图谱

有机太阳能电池凭借成本低、重量轻、可溶解加工和柔韧性好等特点，成为全球能源领域的研究热点[32]。因此，本文将 Web of Science 论文数据库和 IncoPat 专利数据库作为数据源，搜集 OSC 领域的科技文献，通过领域专家制定检索式共获取论文 3 369篇、专利 421 篇，并从中遴选出高质量的论文与专利作为核心数据集，该数据集详细的统计结果如表 2 所示。在 OSC 领域知识图谱构建中，首先，通过两阶段式轻量级本体建模方法得到 OSC 本体模型，如图 2 所示。然后，基于本体模型对 OSC 核心数据集进行人工标注，共得到4 700个知识实体和15 377个三元组，组织关联后形成 328 个复杂知识对象，部分知识实体如表 3 所示。为验证领域知识图谱的质量，邀请了中国科学院福建物质结构研究所、化学所多名领域专家对 OSC 的本体层和实例层进行完备性、一致性和准确性评估。评估结果表明，OSC 领域知识图谱涵盖了有机太阳能电池领域内的关键概念，能够准确反映概念的属性与关系，并且没有明显的逻辑矛盾或冲突。

3.2 ChatGLM2-6B 模型的部署与微调

模型部署与微调的硬件环境为 GPU NVIDIA Tesla A40，CPU AMD EPYC 7543 32-Core Processor，内存 80G，操作系统 Ubuntu 18.04。首先，安装 CUDA Toolkit、PyTorch 及其他依赖库，其次，从 Huggingface

表 1 复杂知识对象抽取算法

Tab. 1 Complex Knowledge Object Extraction Algorithm

算法 1：复杂知识对象抽取算法
输入：科技文献 d，知识图谱 G=(ontology,triples)，大语言模型 LLM
输出：复杂知识对象
1：令三元组集 triplet_set ={}
2：基于知识图谱的三元组 triples 构建外部记忆模块 Memory={$M_1,M_2,\cdots,M_N$}
3：将科技文献 d 分割为句子集合 sentence_list
4：for each s∈sentence_list
5： 基于句子 s 和本体 ontology 构建提问 q
6： 令候选问答样例 candidate={}
7： for each M_i∈Memory
8： 从记忆簇 M_i 中选择与 q 最相似的记忆作为有效记忆M_i'并将其添加到 candidate
9： end for
10：从 candidate 中选择相似度 Top-k 记忆[$M_1',M_2',\cdots,M_k'$]作为问答样例 qa_list
11：基于 qa_list 和 q 构建提示模版 prompt1，将其输入 LLM，解析得到实体类型 entity_type
12：基于 ontology 获取 entity_type 的子类型 subtypes
13： if subtypes!=Ø
14： 利用 s 和 subtypes 构建提问 q，返回第六步
15： if entity_type∈语句级
16： 构建三元组(d，entity_type，s)，并将其添加到 triplet_set
17： *else*
18： 基于 s 和 entity_type 构建提问 q2，并获取问答样例 qa_list
19： 基于 q2 和 qa_list 构建提示模版 prompt2，通过大模型得到细粒度实体 entity
20： 生成三元组(d，entity_type，entity)，并将其添加到 triplet_set
21：基于 triplet_set 和 ontology 生成复杂知识对象

表 2 OSC 核心数据集

Tab. 2 Core Dataset of OSC

文献类型	研究类型	文献数量
论文	有机太阳能电池材料研究	125
专利	有机太阳能电池材料研究	97
论文	有机太阳能电池制备方法研究	21
专利	有机太阳能电池制备方法研究	16
论文	有机太阳能电池机理研究	18
专利	有机太阳能电池机理研究	14
论文	有机太阳能电池结构研究	18
专利	有机太阳能电池结构研究	19

下载模型参数权重实现 ChatGLM2-6B 部署，最后按照 5∶1 的比例将领域知识图谱中三元组集分为训练集和测试集，生成对应的问答对进行模型的微调训练与测试。ChatGLM2-6B 模型使用 LoRA 微调的参数设置，如表 4 所示。为了分析模型微调的收敛性和收敛速度，统计了模型损失函数值 Loss 随着 Epoch 的变化情况，如图 3 所示。可以发现，模型在初期 Loss 随着 Epoch 的增大而快速下降，在中期 Epoch 为 0.4 时 Loss 下降速度放缓，在后期 Epoch 大于 1 时逐渐收敛，最终 Loss 值稳定在 0.08 左右。

3.3 MOT 构建

领域知识抽取基于 MOT 机制和知识图谱来构建提示，通过与大语言模型的多轮问答得到知识三元组。MOT 利用 LDA 主题模型实现外部记忆的划分，每个主题表示一个记忆簇，主题或记忆簇的最佳个数由困惑度来评估，如式（5）所示：

$$Perplexity=\exp\left(-\frac{\sum_{i=1}^{M}\ln p(w_m)}{\sum_{i=1}^{M}N_m}\right) \tag{5}$$

其中，M 表示外部记忆中 QA 对的个数，N_m 表示第 m 个 QA 对中单词的数量，w_m 是第 m 个 QA 对的单词集，$p(w_m)$表示 w_m 生成的概率。困惑度越低，说明在当前主题个数下 LDA 的预测和泛化能力越强。实验在训练集上统计了主题个数 5~50 时困惑

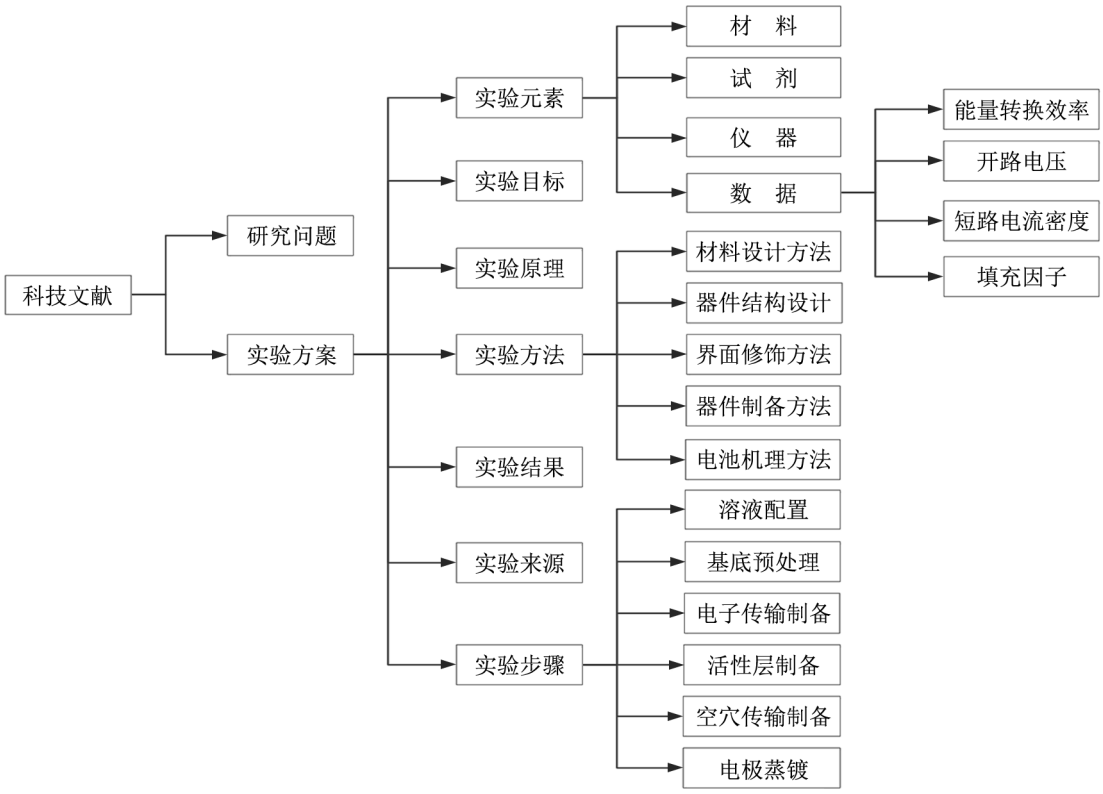

图 2 OSC 本体模型

Fig. 2 OSC Ontology Model

表 3 部分知识实体

Tab. 3 Part of Knowledge Entity

实体名称	实体类型	数量	示 例
试 剂	词汇级实体	143	Chlorinated benzenes，o-DCB，DCM，DIM…
仪 器	词汇级实体	351	Keithley 4200，XRD，AFM，UV-vis…
给体材料	词汇级实体	353	BPAPF，BTP-eC9，BTP-S2，BTPT-4F…
受体材料	词汇级实体	407	A4T-16，AQx-1，ATT-9，BCP，Bis-PC61BM…
能量转换效率	科学数据级实体	315	2.5%，0.77%，13.42%，17.6%…
开路电压	科学数据级实体	308	0.57，0.51，0.61，0.63…
短路电流密度	科学数据级实体	298	7.1，4.24，-2.68，18.28…
填充因子	科学数据级实体	313	63%，46.4%，74.6%，77.8%…
研究问题	语句级实体	297	To date，there are rare studies that report the use of flexible graphene films…
实验目标	语句级实体	324	In order to permit an attractive（photovoltaic） performance whilst…
实验原理	语句级实体	327	The separated charges（hole or electron） from the exciton must also be able…
实验结果	语句级实体	309	The performance of organic photovoltaic(OPV) bulk-heterojunction blends…

度的变化情况，如图 4 所示。其中，横轴和纵轴分别表示主题个数与困惑度。可以看出，困惑度随主题个数的增加呈现先急剧下降后波动上升的趋势，在主题个数为 30 时困惑度达到最低值。基于 t-SNE（T-Distribution Stochastic Neighbour Embedding）[33] 将每个记忆(QA 对)的主题概率分布特征映射到二维空间中，如图 5 所示。可以发现，图中节点空间分布呈现明显的社区结构，每一个社区表征一个记忆簇；两个 QA 对的主题分布越相似，在空间中的距离就越接近。

表 4 模型微调参数

Tab. 4 Model Fine-Tuning Parameters

参 数	参数值
lora_rank	8
lora_dropout	0. 1
batch_size	4
gradient_accumulation_steps	1
learning_rate	1e-4
num_train_epochs	1

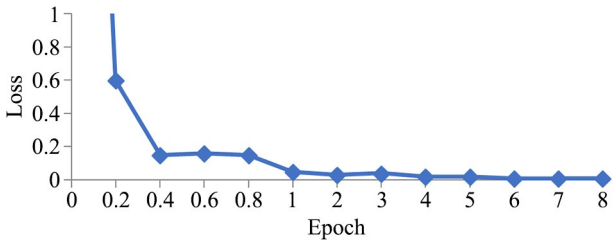

图 3 微调模型损失函数值变化情况

Fig. 3 Change of Loss Function Value for Fine-Tuning Model

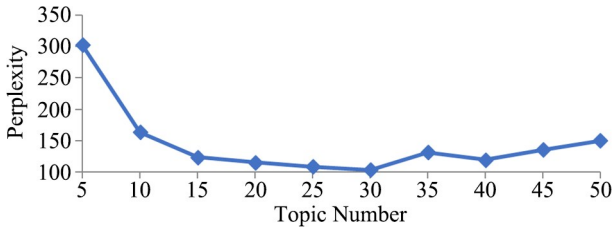

图 4 困惑度变化情况

Fig. 4 Change of Perplexity

MOT 利用 Doc2vec 模型将不同长度的文本转换为固定长度的向量，以便计算与目标问题最近似的 QA 对。Doc2vec 模型的基本原理就是已知文本 d 和

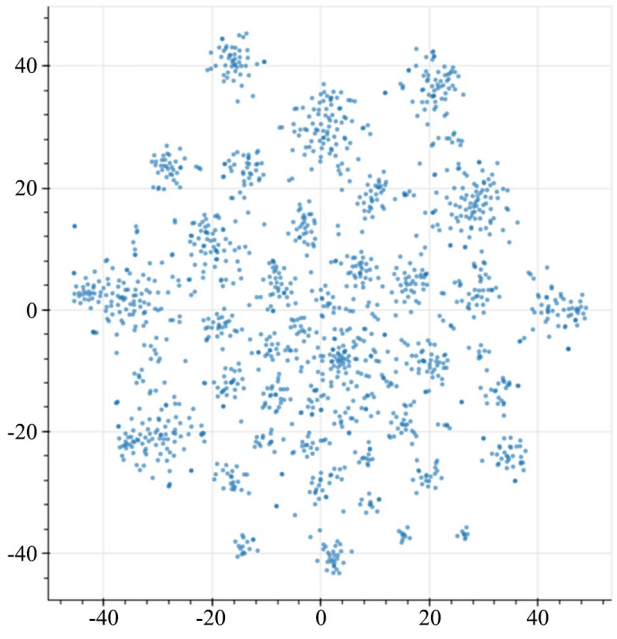

图 5 记忆簇

Fig. 5 Memory Cluster

单词 w 的上下文 $context(w)$，最大化单词 w 的预测值，目标函数如式（6）所示：

$$L=\sum_{d\in D}\sum_{w\in sampling(d)}\log p(w\mid v,contex(w)) \quad (6)$$

其中，D 是由训练集中 QA 对和测试集中问题 Q 整合而成的文本集，$sampling(d)$ 是从文本 d 中采样得到的单词集。经过多次实验得到 Doc2vec 的最优参数配置：向量维数设为 20，迭代次数设为 10，学习率为 0. 025，上下文窗口大小为 10。Doc2vec 模型的输出结果如表 5 所示，每一行表示一个 QA 对或 Q 的文本向量。QA 对和 Q 的语义相似性由余弦相似度来评估，其值越大则两个文本越相似，每个 Q 与 Top-k 个 QA 对的相似度如表 6 所示。

表 5 文本向量

Tab. 5 Text Vectors

文本	1	2	……	20
文本 1	-0. 02832495	-1. 3030231	……	-0. 08972332
文本 2	-0. 27721792	-0. 24999158	……	-0. 2448817
……	……	……	……	……
文本 3219	1. 5489659	-0. 06562421	……	0. 10191738

3. 4 领域复杂知识对象抽取

3. 4. 1 实验结果

实验选取准确率 P、召回率 R 和 F1 值作为知识抽取结果的评价指标，P 表示正确预测样本在所有预测为正例的样本的占比，R 表示正确预测样本在所有正例样本中的比例，F1 是 P 和 R 的调和平

表 6 文本相似度

Tab. 6 Text Similarity Value

Q	QA	相似度
A structured polymer solar cell ……	Polymer solar cells(PSCs) with……	0.96
A structured polymer solar cell ……	Polynitrogen heterocyclic polymers……	0.92
A structured polymer solar cell ……	Preferably a polymer solar cell……	0.84
……	……	……
Organic photovoltaics(OPVs) have……	Nonfullerene organic photovoltaics……	0.93
Organic photovoltaics(OPVs) have……	Improving semi-transparent organic……	0.86
Organic photovoltaics(OPVs) have……	Upscale the flexible organic……	0.78

均值。3 种评价指标的定义如式（7）~（9）所示：

$$P=\frac{M}{M+N} \tag{7}$$

$$R=\frac{M}{M+T} \tag{8}$$

$$F1=\frac{2PR}{P+R} \tag{9}$$

其中，M 是正确预测为正例的样本数，N 是将反例预测为正例的样本数，T 是将正例预测为反例的样本数。通过比较 ChatGLM2-6B、ChatGLM2-6B+LoRA1、ChatGLM2-6B+LoRA2、ChatGLM2-6B+LoRA2+MOT 4 种不同模式在测试集上的抽取结果来验证本文所提方法的有效性。其中，LoRA1 利用 ChatGLM 基于科技文献的文本内容自动生成的微调指令集，其提示模版为“把文本中关键的内容提取出来，制作成对话数据集的格式”，LoRA2 利用知识图谱生成微调指令集。通过多次实验获取超参数的最优配置：问答样例个数 $k=3$，超参数 $\alpha=0.7$，阈值 $\beta=0.8$。知识抽取结果如表 7 所示，可以看出 ChatGLM2-6B+LoRA 模式优于 ChatGLM2-6B，在 3 个指标上均有所提升，说明通过 LoRA 微调能够改善模型在特定领域任务上的处理能力。相较于 ChatGLM2-6B+LoRA1，ChatGLM2-6B+LoRA2 的提升效果更明显，验证了知识图谱数据微调大语言模型的有效性。ChatGLM2-6B+LoRA2+MOT 模式优势明显，在 3 个指标上均取得了最优值，这说明充分利用知识图谱来微调模型和构建提示可以有效提升大语言模型的知识抽取的性能。此外，比较 ChatGLM2-6B+LoRA2+MOT 模式在不同类型实体上的抽取结果如表 8 所示，发现语句级实体预测的准确率高于词汇级实体和科学数据级实体。这是由于这两类实体通常处于本体结构的底层，需要通过更多轮的问答才能得到抽取结果，存在错误传播的问题，即把上一轮的文本分类产生的错误传播到下一轮实体提取问答中，降低了预测结果的准确性。

表 7 知识抽取结果

Tab. 7 The Results of Knowledge Extraction

模 式	P (%)	R (%)	F1 (%)
ChatGLM2-6B	47.3	48.4	47.8
ChatGLM2-6B+LoRA1	48.5	49.7	49.1
ChatGLM2-6B+LoRA2	54.5	53.9	54.2
ChatGLM2-6B+LoRA2+MOT	61.4	58.7	60.1

表 8 不同类型实体抽取结果

Tab. 8 The Results of Different Types of Entities

实体类型	P (%)	R (%)	F1 (%)
语句级	67.6	64.8	0.66
词汇级	57	53.4	55.1
科学数据级	58.5	56.6	57.6

3.4.2 实例分析

在 WOS 号为 000280276900013 的科技文献 d 中，句子 s1“The influence of molecular weight on various optical and physico-chemical properties of poly…”，通过一轮问答得到实体类型为“研究问题”，生成三元组(d，研究问题，s1)。句子 s2“in order to permit an attractive（photovoltaic）performance…”，通过两轮问答得到实体类型路径为“实验方案→

实验目标”，生成三元组(d，实验目标，s2)。句子 s3“Interest in fluorene-based organic semiconductors has surged thanks to their ability to display high…”，通过两轮问答获取实体类型路径为“实验方案→实验原理”，形成三元组(d，实验目标，s3)。句子 s4“F8TBT batches 1-14 were prepared according to previously reported procedures…”，通过三轮问答获取实体类型路径为“实验方案→实验方法→材料的设计方法”，形成三元组(d，材料的设计方法，s4)。句子 s5“A LiF electron-blocking layer (thickness 6A°) and aluminum top electrodes…”，通过三轮问答得到实体类型路径为“实验方案→实验步骤→电极蒸镀”，生成三元组(d，电极蒸镀，s5)。句子 s6“Electronic characterization of FETs was conducted with a Keithley 4200 parameter analyzer…”，通过三轮问答获取实体类型路径为“实验方案→实验元素→试剂”，经第四轮问答抽取出具体的实体 Keithley 4200，形成三元组(d，试剂，Keithley 4200)。处理完所有句子后，利用本体模型将三元组进行组织关联得到复杂知识对象，如图 6 所示。

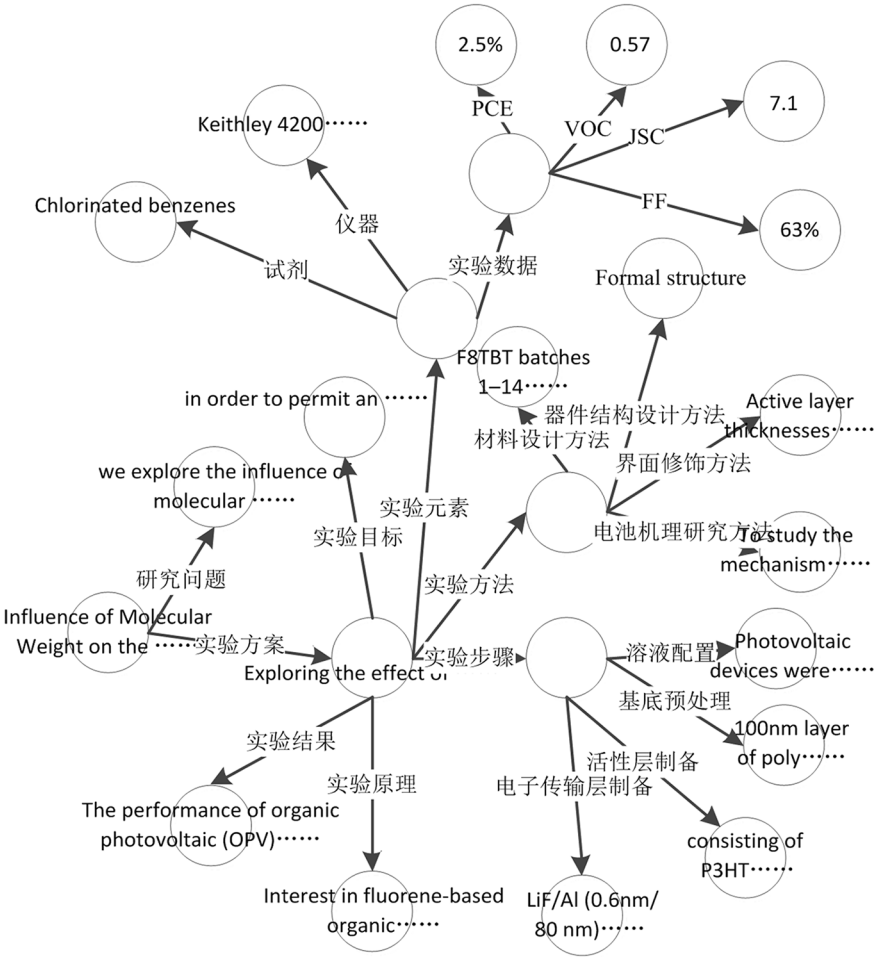

图 6 复杂知识对象

Fig. 6 Complex Knowledge Object

4 结束语

本文提出了一种融合知识图谱和大语言模型的科技文献复杂知识对象抽取方法，首先构建了轻量级本体并通过 BRAT、Neo4j 来标注科技文献和构建领域知识图谱，然后基于知识图谱数据通过 LoRA 微调大语言模型 ChatGLM2-6B，最后在构建提示时注入高度置信的问答样例，通过多轮问答实现科技文献细粒度知识挖掘。本研究贡献包括：①利用知识图谱对多源异构的科技文献信息进行细粒度、多维度的知识组织与融合，通过模型微调、MOT 机制融合了领域知识和大语言模型内置知识，实现细分学科领域科技文献的复杂知识对象高效精准抽取。②能够通过循环迭代的方式提升知识抽取的效果，对大语言模型输出的复杂知识对象进行过滤和审核

后，可进一步用于补充和完善领域知识图谱。③当标注数据不足或没有时，可不微调模型，直接基于领域本体和大语言模型实现小样本或零样本的领域知识抽取。本研究存在的局限：①知识抽取需要多轮迭代，对于科技文献中每一个句子，需要与大语言模型进行多轮问答才能从中抽取出对应知识。②科技文献类型仅包含论文和专利，还需进一步引入领域专著、项目、研究报告、标准和技术档案等多源异构数据。③难以快速应用于其他学科，需要领域专家深度参与构建领域知识本体。

知识图谱中存在关于实体和关系的丰富信息，如层次结构信息和描述文本，若能有效融合这些信息，将会进一步提升大语言模型的知识抽取能力。实体和关系的类型通常被组织成树状结构，标签路径是从根节点到叶子节点的完整路径，能够有效表征类型树的结构特征。在未来，考虑融合标签路径信息到大语言模型中，以提高知识实体的识别效果。目前，向大语言模型注入的知识形式为三元组格式，更为复杂的知识路径、知识结构、推理信息等没能得到充分利用[34]，后续计划设计更好的知识注入方法将不同类型和结构的复杂知识引入大语言模型中。

参 考 文 献

[1] 代冰，胡正银. 基于文献的知识发现新近研究综述 [J]. 数据分析与知识发现，2021，5（4）：1-12.

[2] 李广建，袁钺. 基于深度学习的科技文献知识单元抽取研究综述 [J]. 数据分析与知识发现，2023，7（7）：1-17.

[3] 石栖，陈文杰，胡正银，等. 面向知识发现的科学实验知识图谱构建研究 [J]. 数据分析与知识发现，2025，9（3）：1-15.

[4] Open AI. ChatGPT：Optimizing Language Models for Dialogue [EB/OL]. [2023-03-12]. https://openai.com/blog/chatgpt/.

[5] 张颖怡，章成志，周毅，等. 基于 ChatGPT 的多视角学术论文实体识别：性能测评与可用性研究 [J]. 数据分析与知识发现，2023，7（9）：12-24.

[6] Chen X，Zhang N，Xie X，et al. KnowPrompt：Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction [EB/OL]. [2021-04-15]. https://arxiv.org/abs/2104.07650.

[7] Zeng A H，Liu X，Du Z X，et al. Glm-130b：An Open Bilingual Pre-trained Model [EB/OL]. [2022-10-05]. https://arxiv.org/abs/2210.02414.

[8] Edward H，Yelong S，Phillip W，et al. Lora：Low-Rank Adaptation of Large Language Models [EB/OL]. [2021-06-17]. https://arxiv.org/abs/2106.09685v2.

[9] Li X N，Qiu X P. MoT：Pre-thinking and Recalling Enable ChatGPT to Self-Improve with Memory-of-Thoughts [EB/OL]. [2023-05-10]. https://arxiv.org/abs/2305.05181v1.

[10] 沈雪莹，欧石燕. 科学文献知识单元抽取及应用研究：梳理与展望 [J]. 情报理论与实践，2022，45（12）：195-207.

[11] 叶光辉，彭泽，陈国梁，等. 学术文献中的知识单元抽取及其分布特征识别研究 [J]. 情报理论与实践，2023，46（4）：90-98.

[12] 郑梦悦，秦春秀，马续补. 面向中文科技文献非结构化摘要的知识元表示与抽取研究——基于知识元本体理论 [J]. 情报理论与实践，2020，43（2）：157-163.

[13] Liu X H，Zhang S D，Wei F R，et al. Recognizing Named Entities in Tweets [C] //The 49th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies，Proceedings of the Conference，19-24 June，2011，Portland，Oregon，USA，2011.

[14] Lample G，Ballesteros M，Subramanian S，et al. Neural Architectures for Named Entity Recognition [EB/OL]. [2016-06-12]. https://arxiv.org/abs/1603.01360v3.

[15] 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：Volume 2. Morristown，USA，2009：1003-1011.

[16] Han X，Liu Z Y，Sun M S. Neural Knowledge Acquisition via Mutual Attention between Knowledge Graph and Text [C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto：AAAI Press，2018：4832-4839.

[17] Hu L M，Zhang L H，Shi C，et al. Improving Distantly-Supervised Relation Extraction With Joint Label Embedding [C] //Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). Hong Kong，China. Stroudsburg，PA，USA：Association for Computational Linguistics，2019：3819-3827.

[18] Zhang N Y，Deng S M，Sun Z L，et al. Long-Tail Relation Extraction via Knowledge Graph Embeddings and Graph Convolution Networks [EB/OL]. [2020-08-13]. https://arxiv.org/pdf/1903.01306.pdf.

[19] Tang X Y，Huang Y，Xia M，et al. A Multi-Task BERT-BiLSTM-AM-CRF Strategy for Chinese Named Entity Recognition [J]. Neural Processing Letters，2023，55（2）：1209-1229.

[20] Wei X，Cui X，Cheng N，et al. Zero-Shot Information Extraction via Chatting with ChatGPT [EB/OL]. [2023-02-21]. https://arxiv.org/abs/2302.10205.

（下转第 63 页）

[20] 裴佳音，单鹏. 基于粗糙集约简的公共事件微博舆情趋势影响因素提取 [J]. 情报科学，2019，37 (3)：112-118.
[21] 赵栋祥. 大数据环境下个人的数字囤积行为研究——基于扎根理论的探索 [J]. 中国图书馆学报，2024，50 (1)：96-114.
[22] 刘天元，贾煜. 青年群体数字囤积行为的形成机理及其结果效应——基于内部动因视角的分析 [J]. 中国青年研究，2023 (2)：93-100.
[23] 张成虎，张一凡，刘建新. 数字囤积行为的影响因素组态路径研究——基于 fsQCA 的分析 [J]. 珞珈管理评论，2023 (4)：90-111.
[24] Wang H X，Miao P，Jia H Y，et al. The Dark Side of Upward Social Comparison for Social Media Users：An Investigation of Fear of Missing Out and Digital Hoarding Behavior [J]. Social Media+ Society，2023，9 (1)：20563051221150420.
[25] Tugtekin E B. Investigation of the Relationship between Digital Hoarding，Information Technologies Self-Efficacy，and Anxiety [J]. European Journal of Education Studies，2022，9 (1)：157-177.
[26] 代宝，张鹏金，王玥倩. 研究生学术性数字囤积行为的影响因素及后果——基于扎根理论的探索性研究 [J]. 现代情报，2024，44 (4)：53-65.
[27] 赵栋祥. 个人的数字囤积行为研究进展与展望 [J]. 情报杂志，2022，41 (8)：194-200，184.
[28] Oravec J A. Digital(or Virtual) Hoarding [J]. International Journal of Computers in Clinical Practice，2018，3 (1)：27-39.
[29] Zaremohzzabieh Z，Abdullah H，Ahrari S，et al. Exploration of Vulnerability Factors of Digital Hoarding Behavior Among University Students and the Moderating Role of Maladaptive Perfectionism [J]. Digital Health，2024，10：20552076241226962.
[30] 谢笑，高鑫圆，汪卓赟，等. 数字囤积：概念溯源、活动分析和研究展望 [J]. 图书情报工作，2023，67 (23)：124-137.
[31] 张艳丰，刘敏. 移动社交媒体用户数字囤积行为生成机理与引导策略研究 [J]. 情报资料工作，2024，45 (2)：48-56.
[32] Thorpe S，Bolster A，Neave N. Exploring Aspects of the Cognitive Behavioural Model of Physical Hoarding in Relation to Digital Hoarding Behaviours [J]. Digital Health，2019，5：2055207619882172.
[33] Deng J L. Introduction to Grey System Theory [J]. Journal of Grey System，1989，1 (1)：1-24.
[34] 洪闯，李贺，祝琳琳，等. 活动理论视角下社会化问答平台用户知识协同模型与关键影响因素研究——基于模糊 DANP 方法 [J]. 情报理论与实践，2019，42 (11)：100-106.
[35] 王国凯. 基于模糊 DANP 的公共图书馆健康信息服务质量影响因素识别研究 [J]. 农业图书情报学报，2022，34 (10)：33-43.
[36] 曹玉昆，张亚芳，陈雨荷，等. 基于 Grey-DEMATEL 模型的生态文化产业发展关键影响因素识别研究——以中国重点国有林区为例 [J]. 生态经济，2023，39 (11)：195-202.

(责任编辑：杨丰侨)

(上接第 25 页)

[21] Yuan C，Xie Q，Ananiaidou S. Zero-shot Temporal Relation Extraction with ChatGPT [EB/OL]. [2023-04-05]. https://arxiv.org/abs/2304.05454.
[22] 苏杭，胡亚豪，谢艺菲，等. 利用提示调优实现两阶段模型复用的关系实体抽取方法 [J]. 计算机应用研究，2022，39 (12)：3598-3604.
[23] 王震宇，朱学芳，杨睿. 基于多模态大语言模型的关系抽取研究 [J]. 数据分析与知识发现，2025，9 (1)：90-99.
[24] Stenetorp P，Pyysalo S，TopiC G，et al. BRAT：A Web-Based Tool for NLP-Assisted Text Annotation [C] //Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics，2012.
[25] Neo4j. High-Speed Graph Database with Unbounded Scale，Security，and Data Integrity for Mission-Critical Intelligent Applications [EB/OL]. [2023-06-12]. https://neo4j.com/docs/.
[26] 周毅，刘峥，粟小青，等. 融合多层次数据的问答知识图谱本体模型构建 [J]. 图书情报工作，2022，66 (5)：125-132.
[27] Ji B. VicunaNER：Zero/Few-shot Named Entity Recognition using Vicuna [EB/OL]. [2023-05-05]. https://arxiv.org/abs/2305.03253v1.
[28] 舒洪水，彭鹏. ChatGPT 场景下虚假信息的法律风险与对策 [J]. 新疆师范大学学报 (哲学社会科学版)，2023，44 (5)：124-129.
[29] Liu J，Shen D，Zhang Y，et al. What Makes Good In-Context Examples for GPT-3? [EB/OL]. [2021-01-17]. https://arxiv.org/abs/2101.06804.
[30] Blei D M，Ng A Y，Jordan M I，et al. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research，2003，3 (4/5)：993-1022.
[31] Le Q，Mikolov T. Distributed Representations of Sentences and Documents [C] //Proceedings of International Conference on Machine Learning，2014：1188-1196.
[32] 沈伊凡，刘耀虎，张建齐，等. 大面积有机太阳能电池的研究进展 [J]. 中国科学：化学，2022，52 (11)：2001-2026.
[33] Laurens V D M，Hinton G. Visualizing Data Using t-SNE [J]. Journal of Machine Learning Research，2008，9 (86)：2579-2605.
[34] 鲍琛龙，吕明阳，唐晋韬，等. 与知识相结合的提示学习研究综述 [J]. 中文信息学报，2023，37 (7)：1-12.

(责任编辑：杨丰侨)