大规模语言模型驱动的场景化知识图谱构建研究

程楠楠[1] 魏璐露[1]
CHENG Nannan WEI Lulu

摘 要 针对传统知识抽取成本高，且现有医疗知识图谱在知识类型与表达能力受限的问题，文章提出利用大规模语言模型进行知识抽取，并构建场景化知识图谱，用于提升医疗知识图谱在实际应用中的适用性。具体而言，用大模型驱动场景化知识图谱的融合式构建思路，设计一种基于场景程序语言模板的Prompt，通过实验对比基准模型与传统文本Prompt的指标表现，并观察输入示例数量对数据指标的影响。结果显示，大语言模型在知识抽取方面 F_1 指标较高，场景程序语言模板代码 Prompt 表现尤为突出，且随着输入示例的增加，三元组完整抽取的效果越好。

关键词 大规模语言模型；场景化知识图谱；知识抽取；知识图谱构建；Prompt

doi：10.3969/j.issn.1672-9528.2025.06.004

0 引言

知识图谱（knowledge graph, KG）[1] 利用图形结构建模、识别和分析推断事物之间复杂的关联关系和沉淀领域知识，已经被广泛应用于语义搜索、智能问答、推荐系统、数据分析与决策引擎等领域。目前，随着人工智能技术的发展，医疗领域作为知识图谱应用最广的垂直领域之一，在疾病评估、临床辅助决策、智能推荐、医疗问答系统等智能医疗领域都有着良好的发展前景 [2]。

知识图谱虽然具有强大的知识建模与表征能力，但更强调知识的存在性而非准确无误的适用性 [3]。而医学知识的特点是高度的专业性、表达多样性、复杂性，同一概念可能蕴含不同的专业术语，如糖尿病亦被称作消渴症 [4]。此外，其在业务实践中要求极低的容错率，这对医学知识的应用提出了更为严苛的标准。鉴于医疗领域信息的多维度特性，当前的知识图谱在知识类型和表达能力上受到限制，难以满足真实场景中的临床诊疗需求。例如，在智能问答系统中，除需要涵盖专业医学词汇外，还需纳入大众常用的口语化表达，系统不仅需理解概念实体间的关系，还应具备处理经验性关系的能力。因此，医学知识图谱的构建需进一步延伸至基于具体场景的深度医学知识图谱，以期更好地提升在疾病评估、临床辅助决策、诊疗推荐、用户推荐等智能医疗领域任务的效果。

传统的知识图谱构建方法主要依赖于人工标注和规则匹配，数据获取难度高，知识表示不灵活、效率低下且难以扩展。近年来，将大语言模型（large language model, LLM）与知识图谱融合的研究工作引起了众多研究人员和从业者的关注 [5-7]。LLM 凭借强大的语言理解能力、高效的语义增强技术和持续的优化学习能力，可以与知识图谱的嵌入技术相结合，识别、提取和推导出更为丰富的实体及其关系，从而自动化地构建知识图谱 [8]，降低人力和时间成本外，进一步丰富和完善医学知识图谱的质量，更好地落地具体应用场景。

1 相关工作

1.1 场景化医疗知识图谱

目前国内外关于医学知识图谱的构建已有一定成果，如2018 年，美国德克萨斯大学健康科学中心制定了标准 UMLS 术语，并构建了医学知识图谱 BMKG[9]。2022 年，Huo 等人 [10] 基于 PubMed 进行信息抽取构建医疗百科知识图谱 BKG。国内医学知识图谱 CUMLS[11] 是中国医学科学院医学信息研究所基于 UMLS 开发的中文一体化医学语言系统。除此之外，中医药领域的 TCMLS[12] 知识图谱共有 9 种类型，包含“基于中医药学语言系统的知识图谱”“中医美容知识图谱”“中医养生知识图谱”“中医临床知识图谱”等。在工业领域，阿里云于 2021 年推出的糖尿病知识图谱 DiaKG[13]。但当前知识图谱受限于存储的知识类型和表达能力，难以满足真实场景的临床诊疗需求。

受文献 [3,14] 启发，本文提出场景化医疗知识图谱，将场景化医疗知识图谱的场景属性按照真实的医疗活动给出具体的场景要素，包括知识的产生场景和知识的应用场景，其

1. 江西科技学院 江西南昌 330098

[基金项目] 江西省教育厅科学技术研究项目：场景化知识图谱的智能医疗推荐研究（GJJ2202609）

中知识的产生场景包括疾病知识、知识的依据及来源等，应用场景是知识的应用目的和价值、应用对象的年龄、性别、职业、病史等，即可以补充实际医疗场景的知识维度，提高医疗知识图谱在实际应用中的针对性和适用性。

1.2 大语言模型在知识图谱构建中的应用

在大规模语言模型驱动的场景化知识图谱构建中，信息抽取（实体、关系、属性、事件等抽取）是关键环节之一。通过 LLM 对文本进行深入理解，可以准确识别实体、属性和关系，为知识图谱的构建提供丰富的知识来源。

抽取技术主要分为两大类，一类是基于自然语言的大模型抽取方法，如通过结构化专家编写的指令来微调 LLM 进行信息抽取的 InstructUIE[15]、实现零样本自动识别和提取功能 ChatIE[16]；另一类是通过生成具有统一编程模式的代码进行信息抽取和知识图谱的构建，如 Code4UI[17] 和联合模式感知提示的 CodeKGC[18]。第一类方法更能理解和生成人类的语言，且简洁易于设计，但对于隐藏复杂的关系捕捉难度较大。第二类方法可以基于代码天然的结构属性，以编码的形式以保持图谱内在的语义结构，更适用于复杂推理和结构预测任务。

2 大语言模型驱动的场景化知识图谱的融合式构建

大语言模型驱动的场景化知识图谱的融合式构建框架如图 1 所示，场景化知识图谱的构建流程包括数据收集与预处理、场景化知识图谱构建与融合、存储与应用阶段，其中本文实验重点在场景化知识图谱阶段的各类型知识抽取中，还包括按照数据结构化程度的分类。

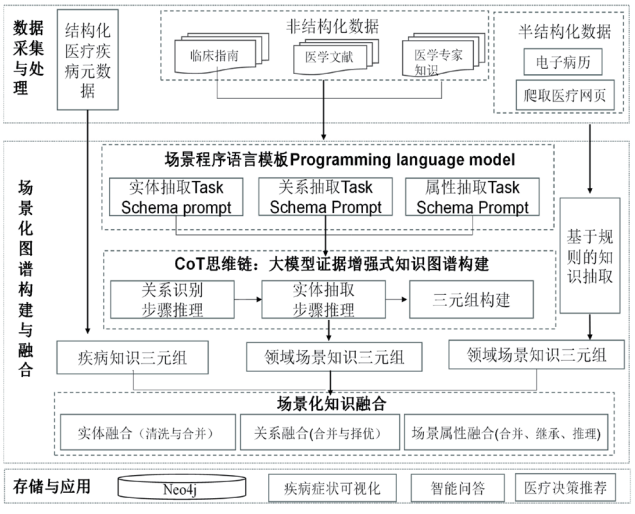

图 1 大语言模型驱动的场景化知识图谱的融合式构建框架图

在数据收集与预处理阶段，需要收集与场景相关的文本数据并进行清洗和格式化；在知识抽取阶段，利用 LLMs 从文本中提取结构化知识；在实体识别与关系建立阶段，对提取的知识进行进一步处理和分析；在图谱构建与评估阶段，将处理后的知识以图的形式表示并评估其质量和准确性。

2.1 数据采集与处理

因本研究主要关注医疗健康和导诊问诊场景，因此数据来源主要为结构化和半结构化的文本数据，包括 PubMed、Medline 等数据库中的研究论文、国家临床指南、电子病历、问诊和导诊数据等。

采集到的数据需要使用 Pandas 等工具进行数据清洗，确保数据的准确性和一致性。将疾病名称、药物名称、单位等进行标准化处理，单位统一成“g”，疾病名称全部映射到国家标准 ICD-10 的编码。

2.2 场景化图谱构建

图谱构建主要包含知识抽取、知识融合步骤。在构建过程中，本研究将 Transformers 作为 LLM 模型技术框架，将开源 LLM 大模型作为模型底座。基于有向属性图模型构建了“实体 - 关系 - 实体”和“实体 - 场景属性 - 属性值”两种核心语义关系，不同实体之间的语义关系列举如表 1 所示。另外基于文献 [3] 中对医疗各类场景属性的抽取效果，本文场景属性主要使用时间属性、地点属性、患者病史属性、患者年龄属性、患者性别属性、患者职业属性。

表 1 “实体 - 关系 - 实体”语义关系表

头实体	关系	尾实体
患者	患有	疾病
患者	发病部位	部位
患者	相关疾病	疾病
疾病	症状表现	症状
疾病	就诊	科室
疾病	检查	相关检查
疾病	采取	治疗方式
疾病	采用	常用药品
疾病	护理	护理与营养
疾病	原因	发病原因
疾病	学术	相关学术

2.2.1 场景化知识抽取

在构建知识图谱的过程中，知识抽取的效率和人力成本一直是非常重要的。为有效地解决以上问题，本文的场景化知识抽取方法如下。

首先，由于疾病、症状、科室、治疗等结构化图谱知识已存储在关系数据库中，因此可直接转换为三元组，即图中的疾病三元组。对于以网页形式存在的互联网问诊等半结构化数据，可先用 Python 的 Scrapy 框架对数据进行爬取，然后通过正则等技术获取问诊场景化知识三元组，并将其作为知识图谱的补充和非结构化知识抽取任务的训练数据，这部分半结构化的数据包括疾病、别名、发病部位、传染性、多发人群、并发疾病、是否医保、就诊科室、治疗方法、治疗周期、治愈率、治疗费用、相关检查、常用药品等。

对其临床指南和医学文献等非结构化文本，本文构造面向各类抽取任务的场景程序语言模板场景程序语言模板（programming language model），利用 Python 类代码语言结构化表达，再以大语言模型为驱动，以实现知识图谱构建中涉及的实体抽取、关系抽取和场景属性抽取任务。本文采用结构化的代码 Prompt 设计。具体的 Prompt 示例如图 2 所示。

```
Schema Prompt
class Entity:
    def __init__(self, name: str):
        self.name = name
class Relation:
    def __init__(self, name: str):
        self.name = name
class Attribute:
    def __init__(self, name: str):
        self.name = name
……
class Disease(Entity):
""" 疾病实体 """
    def __init__(self, name: str):
        super().__init__(name = name)
class Dise_Symb (Relation):
""" 疾病症状 """
    def __init__(self, name: str):
        super().__init__(name = name)
……
class Triple:
""" 元组 """
    def __init__(self, head, relation, tail):
      self.head = head
        self.relation = relation
        self.tail = tail
class Extract:
""" 知识抽取 """
    def __init__(self, triples):
        self.triples = triples
```

图 2 实体类的定义示例

受思维链和文献 [18] 的启发，本文还采用一种证据增强的生成方法提高推理能力，从而优化任务结果，具体来说是将复杂的知识图谱推理任务分解成多个步骤。对于正常提示，给定文本输入 T_c、提示 P 和概率模型 plm，最大化生成的事实图谱 G 的可能性 p 用公式表示为：

$$p(G_c|T_c,P)=p(G_c|T_c,P,R)p(R|T_C,P) \quad (1)$$

其中，

$$p(G_c|T_c,P,R)=\prod_i^{|G_c|}\mathrm{plm}(g_i|T_c,P,R,g<i) \quad (2)$$

$$p(\mathrm{R}|T_c,P)=\prod_i^{|R|}\mathrm{plm}(r_i|T_c,P,r<i) \quad (3)$$

式中：r_i 是 $|R|$ 推理总步数的一步。

另外对于知识图谱的构建，关键在于识别关系并提取其对应的实体。因此本文的推理链包括 3 个步骤：

步骤 1：识别文中的候选关系。

步骤 2：抽取文中的候选实体。

步骤 3：使用 Extract 方法生成所有关系的三元组。

2.2.2 场景化知识融合

由于不同数据源中实体具有不同的术语表达，因此需对数据进行清洗。对实体的 name 属性和标签类别做相似性计算，判断是否为同一个实体，进行属性的去重和合并。对齐的实体仍然包括多源的关系，对同一对实体对的多源关系进行融合，包括冗余关系去重、关系属性合并。

3 实验

3.1 实验设置

本文的基准模型是 Bert-BiLSTM-CRF，其在知识抽取中凭借其强大的文本表示能力、序列建模能力、标签序列优化能力、易于集成与扩展等优势，成为该领域的基准模型之一。大模型的方面，本实验采用的是 OpenAI 提供的 GPT-3 模型系列中的 Text-davinci-003，在回答问题、文本完成和风格多样性方面表现出色且比之前的版本更善于遵循用户意图。Text-davinci-003 的训练语料中也包含大量的代码数据。在利用 OpenAI 提供的 API 接口时将 temperature 设置为 0.5，max_tokens 设置为 512。当出现 """, class 或者 # 等特殊模式时，停止代码生成。

3.2 评估指标

本实验使用严格的三元组评估，也就是头实体和尾实体均要一致才算预测正确。性能指标使用 F_1 值（F_1 score），其计算公式为：

$$F_1 = 2 \times (\mathrm{Precision} \times \mathrm{Recall}) / (\mathrm{Precision} + \mathrm{Recall}) \quad (4)$$

其中，

$$\mathrm{Precision} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP}) \quad (5)$$

$$\mathrm{Recall} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \quad (6)$$

式中：TP 表示模型提取出正确的三元组个数；FP 表示模型提取出错误的三元组个数；FN 表示模型未提取出的正确三元组个数。因此，Precision 精确率越高表示在模型提取出的三元组中，正确的三元组比例越高。Recall 召回率表示模型提取出的正确三元组个数与样本正确的三元组总数的概率，其值越高，表示其识别正确的三元组方面的能力越强。F_1 分数是精确率和召回率的调和平均数，用于综合评估模型的性能。

3.3 结果与分析

利用大模型技术进行知识抽取时，本实验还对比了普通的文本提示 textPrompt 和本实验采用的代码提示 codePrompt 类型，在提示过程中均会给出 2 个样本示例进行学习。实验结果如表 2 所示。实验结果表明，在本数据集中，大语言模型相比需要大量数据训练的 Bert-BiLSTM-CRF 模型来说，其

知识抽取的性能较好。

表 2 模型指标表现

模型	F_1 值 /%
Bert-BiLSTM-CRF	60.3
Text-davinci-003（textPrompt）	64.8
Text-davinci-003（codePrompt）	**65.2**

对比传统的文本提示 textPrompt，本实验中使用的场景程序语言模板 codePrompt 又进一步提升了三元组抽取的 F_1 综合性能，但提升效果相对受限。考虑到医疗术语的专业性以及大模型自动化抽取过程中的复杂性，本文新增 Few-shot 实验，在 Prompt 工程部分提供 2 个、5 个、10 个、15 个示例供模型学习，实验结果如图 3 所示。

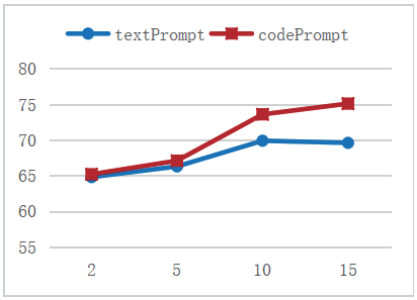

图 3 Few-shot 设置 [2,5,10,15] 下的对比分析

从图 3 可以看出，随着输入样本示例的增多，对比文本提示，本实验采用的结构化的代码提示 codePrompt 性能指标越来越高，这也进一步表明代码语言与文本语言在模型的结构化信息表达上更有优势。

4 总结与展望

本文针对传统知识抽取高成本问题，探索如何利用大规模语言模型进行知识图谱的抽取与构建。另外鉴于医疗领域信息的多维度特性，当前的知识图谱在知识类型和表达能力上受到限制，本文提出场景化知识图谱，以便补充实际医疗场景的知识维度，提高医疗知识图谱在实际应用中的针对性和适用性。与此同时，提出了大模型驱动场景化知识图谱的融合式构建思路，实验结果表明，大语言模型在知识抽取方面展现了较高的准确率和召回率，其中基于场景程序语言模板代码 Prompt 相比传统的文本 Prompt 表现尤为突出，且随着输入示例的数量增加综合 F_1 数据指标更好。

本文为以应用场景为目标的知识图谱构建提供了新的思路和方法，但在探索研究的过程中，需要考虑数据集的多样性和大语言模型接口成本的问题，未来的研究中会尝试探索成本效益更高的替代方案，为医疗知识图谱的场景化落地应用发展做出贡献。

参考文献：

[1] 陈华钧 . 知识图谱导论 [M]. 北京 : 电子工业出版社 , 2021.

[2] YE Q, HSIEH C Y, YANG Z Y, et al. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system[J]. Nature communications, 2021, 12(1): 6775.

[3] 陆泉 , 陈静宇 , 陈帅朴 , 等 . 场景化知识图谱及构建方法 [J]. 情报科学 , 2024,42(3):1-9.

[4] 王楚童 , 李明达 , 孙孟轩 , 等 . 融合大规模医学事实的跨语言双层知识图谱 [J]. 软件学报 , 2025,36(3):1240-1253.

[5] PAN S R, LUO L H, WANG Y F, et al. Unifying large language models and knowledge graphs: a roadmap[J]. IEEE transactions on knowledge and data engineering, 2024, 36(7): 3580-3599.

[6] 冯拓宇 , 李伟平 , 郭庆浪 , 等 . 大语言模型增强的知识图谱问答研究进展综述 [J]. 计算机科学与探索 , 2024, 18(11): 2887-2900.

[7] 张学飞 , 张丽萍 , 闫盛 , 等 . 知识图谱与大语言模型协同的个性化学习推荐 [J]. 计算机应用 ,2024,45(3):773-784.

[8] JIANG J H, ZHOU K, WEN J R,et al. UniKGQA: unified retrieval and reasoning for solving multi-hop question answering over knowledge graph[EB/OL] .(2023-03-01) [2024-09-26].https://doi.org/10.48550/arXiv.2212.00959.

[9] CONG Q, FENG Z Y, LI F, et al. Constructing biomedical knowledge graph based on semmedDB and linked open data[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway: IEEE, 2018: 1628-1631.

[10] HUO C G, MA S T, LIU X Z. Hotness prediction of scientific topics based on a bibliographic knowledge graph[J]. Information processing & management, 2022, 59(4): 102980.

[11] 李丹亚 , 胡铁军 , 李军莲 , 等 . 中文一体化医学语言系统的构建与应用 [J]. 情报杂志 ,2011,30(2):147-151.

[12] LONG H, ZHU Y, JIA L R. et al. An ontological framework for the formalization, organization and usage of TCM-knowledge[J].BMC medical informatics and decision making, 2019,19:53.

[13] CHANG D J, CHEN M S, LIU C Z, et al. DiaKG: an annotated diabetes dataset for medical knowledge graph construction[C]//Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction. Berlin: Springer, 2021: 308-314.

[14] DING Y, WU B, ZHOU E Q, et al. The scenario-oriented method for recording and playing-back healthcare information[C]//International Conference on Enterprise Information Systems. Berlin: Springer, 2011:175-184.

[15] WANG X, ZHOU W K, ZU C, et al. InstructUIE: multi-task instruction tuning for unified information extraction[EB/

一种基于 EOG 的睡眠分期算法研究

梁文虎[1] 杨其宇[1] 王振学[1]
LIANG Wenhu YANG Qiyu WANG Zhenxue

摘 要 睡眠质量影响着人类的身体健康与生活质量，准确的睡眠分期是睡眠质量评估的重要依据。针对眼电信号中的规律性，文章提出了一种用于实现睡眠分期的基于 EOG 的深度学习网络 -AMSTNet。首先，采用 3 个不同尺度的一维卷积网络进行 EOG 信号特征提取；其次，构建一个自适应特征校准模块，动态进行特征权重的选取校准；接着，采用双向门控循环单元学习各睡眠阶段间的时间序列规则；最后，通过公开数据集 Sleep-EDF 验证模型实现睡眠分期的效果。实验结果表明，所提模型在利用 EOG 信号进行睡眠分期中有较好的性能，用 EOG 做睡眠分期是可行的。

关键词 眼电信号（EOG）；睡眠分期；自适应特征提取；卷积神经网络；双向门控循环单元

doi：10.3969/j.issn.1672-9528.2025.06.005

0 引言

睡眠是人类一项基本的生理活动。近年来，利用脑电信号进行自动睡眠分期的算法取得了较好的性能。如 Goshtasbi 等人 [1] 采用全卷积神经网络对单通道 EEG 进行睡眠分期，准确率为 84.8%。Gabriel-alexandru 等人 [2] 采用 CNN-LSTM 网络同样对单通道 EEG 进行睡眠分类，准确率为 88.8%。Seifpour 等人 [3] 采用时域特征统计 SBLE 网络对单通道 EEG 数据进行睡眠分类，准确率达到 90.6%。然而，脑电信号的采集过程相对专业与繁琐。相比之下，眼电信号的采集相对简单方便，并且不需要经过专业培训也能进行信号的正确采集。这为设计便携、可穿戴的自动睡眠分期系统提供了可能性 [4]。研究表明，眼电信号（EOG）是一种潜在的睡眠分期方式 [5]。这是由于 EOG 信号能反映眼睛的活动情况，并且眼电信号与脑电信号具有高度的相似性 [6]。Fan 等人 [7] 采用 CNN-RNN 网络对 EOG 信号进行特征提取与序列学习以进行睡眠分类，在 MASS 数据集和 SleepEDF 数据集上分别获得准确率为 81.2% 和 76.3%。利用眼电信号做睡眠分期在便携性以及经济性方面优势较为明显。

本文提出一种深度学习网络模型（adaptive multi-scale temporal network, AMSTNet），用于对眼电信号进行特征提取与学习，实现自动睡眠分期。首先，针对眼电信号多层次特征提取问题，本文采用 3 种尺度的一维卷积神经网络进行提取不同维度的特征。其次，构建一个自适应特征校准模块，通过全局特征信息对各通道特征进行动态加权，有效提升模型对关键特征的关注。最后，使用双向门控循环单元（bidirectional gate recurrent unit, Bi-GRU）捕捉学习各睡眠

1. 广东工业大学 广东广州 511400

OL].(2023-04-17)[2024-10-03].https://doi.org/10.48550/arXiv.2304.08085.

[16] WEI X, CUI X Y, CHENG N, et al. ChatIE: zero-shot information extraction via chatting with ChatGPT[EB/OL]. (2024-05-27)[2024-08-19].https://doi.org/10.48550/arXiv.2302.10205.

[17] GUO Y C, LI Z X, JIN X L, et al. Retrieval-augmented code generation for universal information extraction[C]//Natural Language Processing and Chinese Computing. Berlin: Springer, 2024: 30-42.

[18] BI Z, CHEN J, JIANG Y N, et al. CodeKGC: code language model for generative knowledge graph construction[J].ACM transactions on asian and low-resource language information processing, 2024, 23(3): 1-16.

【作者简介】

程楠楠 (1987—)，女，江苏南通人，博士研究生，高级工程师、讲师，研究方向：知识图谱与大语言模型、智能医疗，email:chengnaluedu@126.com。

魏璐露 (1995—)，女，江西南昌人，硕士，高级工程师、助教，研究方向：AI 应用研究、生信大数据。

（收稿日期：2024-12-30 修回日期：2025-04-29）