摘要

在现代医疗领域中，数据的作用日益凸显，它不仅是医学研究的基石，也是临床决策的重要依据。医疗数据的类型多样，包括但不限于电子健康记录、医学影像、基因组数据、病理报告以及临床试验结果。这些数据的深度挖掘和分析对于医学研究和临床实践具有重要价值，综合利用可以显著提升疾病的诊断准确性，优化治疗方案，预测疾病的发展趋势。本文旨在围绕着大语言模型在临床医学领域的能力，构建了医学文档知识提取框架解决真实临床病案中的知识挖掘任务。并构建了中文大语言模型临床医学能力评估基准对大语言模型的医学能力进行了多方面多角度的评估。最终设计了一个基于大语言模型和智能体的框架。

为了解决传统文档智能方法在处理多样化和复杂的真实临床医疗病案时的局限性，本文提出了一个创新的自适应医学文档知识提取框架 CliniDocParser。该框架整合了对病案文档的文本信息感知、版面特征感知，并引入了大语言模型模块，利用其具有的优异的语义理解能力和多任务适应性，有效实现了医疗病案文档的知识提取。本研究使用了来自国内 24 家医院的原始住院病案文档，构建了一个庞大的临床医疗知识数据集。该数据集包含 480 份病案和约 28800 个表单项，覆盖了大约 400 个人口学和临床检验指标。基于该数据集，本文设计了针对医学指标提取的指令微调方法，对医学文档知识提取框架进行不同策略的训练。实验结果表明，相比传统的自然语言处理方法，依赖于更少量数据训练的 CliniDocParser 在提取数值型和语义型医学指标方面展现出了更高的准确率和查全率，实现了低资源自适应地针对多样化的医疗病案进行知识提取。

为了探究大语言模型在医疗领域任务的能力，弥补现有的评估基准无法全面权威地评估模型在临床医学上的能力的问题，本文构建了中文大语言模型临床医学能力评估基准 PKUCliniEval。本工作构建旨在模拟一位临床医生从高等教育学生时期到最终临床出诊中每个阶段需要掌握的知识和能力。评测数据集包含 9453 道题目，来源于北大课程测试题目、国家执业医师、临床三基等权威考试题目，包含三个板块：知识部分、推理部分和临床诊疗部分，覆盖了医学基础学科、临床诊疗科室、人体八大系统等 53 个科目。对国内外的开源和闭源大语言模型的评测实验结果表明，相对于英文大语言模型、小规模垂直领域医学大模型，更大规模的通用中文大语言模型在各个模块的能力更优。同时，模型在不同模块等不同科目下的表现有较为显著的差异。本文针对以上实验结果进行了原因探究与分析，证明了大语言模型在临床医学领域的能力，为未来的模型优化方向提供了依据。

医疗病案文档知识抽取方法 CliniDocParser 为智能体提供了对文档及医疗数据的感知能力，中文大模型临床医学能力评估基准 PKUCliniEval 验证了大模型对临床学科具有的知识储备。故而基于以上两部分工作，本文构建了一个基于大语言模型的医学科研智能体系统，具备医学科研能力。智能体由大脑、感知、行动三个部分组成，利用大型语言模型的强大处理能力，在海量医疗病案文档中自动挖掘关键医疗指标，构建相关的临床医学课题，并利用统计分析方法对其进行验证，形成有效的医学研究成果。这一智能体系统能够生成医学真实世界研究中具有实际应用价值的科研成果，为深度挖掘医疗数据的潜在价值和加速医疗工作者的真实世界研究做出了实质性贡献。

总体而言，本文工作对充分利用医学数据、提高医疗质量的创新性架构做出了贡献，并为未来医学领域大模型的发展提供了测试基准和指导。未来可以进一步深化对模型性能和实际应用的理解，以及探索新的研究方向。

关键词：预训练语言模型，大语言模型，信息提取，智能体

Study on medical knowledge extraction and application based on large language model

Hanwen Du (Computer Science and Technology(Intelligence Science and Technology))
Directed by: Prof. Shiliang Zhang

ABSTRACT

In the modern medical domain, the role of data is increasingly prominent, serving not only as the cornerstone of medical research but also as a crucial basis for clinical decision-making. Medical data is diverse, encompassing electronic health records, medical imaging, genomic data, pathology reports, and clinical trial results. Deep mining and analysis of these data types are invaluable for medical research and clinical practice, enabling significant enhancements in disease diagnosis accuracy, treatment optimization, and disease progression prediction. this thesis focuses on leveraging large language models (LLMs) in clinical medicine, proposing a medical document knowledge extraction framework, CliniDocParser, to address the challenges of knowledge mining in real clinical case documents. Furthermore, this thesis establish a Chinese LLM clinical medicine capability assessment benchmark, PKUCliniEval, to evaluate the medical capabilities of LLMs from multiple perspectives comprehensively. Innovatively, a framework based on large-scale language models and intelligent agents has been developed.

To address the limitations of traditional intelligent document processing methods in handling diverse and complex real clinical medical cases, this thesis propose an innovative adaptive medical document knowledge extraction framework, CliniDocParser. This framework integrates text information perception and layout feature awareness with a large language model module, leveraging its exceptional semantic understanding capabilities and multi-task adaptability to effectively perform knowledge extraction from medical case documents. Utilizing original inpatient case documents from 24 hospitals across the country, this thesis have constructed a vast clinical medical knowledge dataset. This dataset comprises 480 cases and approximately 28,800 form items, covering about 400 demographic and clinical diagnostic indicators. Based on this dataset, this thesis have designed a command fine-tuning method for medical indicator extraction and trained the knowledge extraction framework using various strategies. Experimental results demonstrate that CliniDocParser, trained on less data, ex-

hibits higher accuracy and recall rates in extracting numerical and semantic medical indicators compared to traditional natural language processing methods, achieving adaptive knowledge extraction for diverse medical cases in resource-constrained settings.

To explore the capabilities of LLMs in medical domain tasks and address the lack of comprehensive and authoritative benchmarks for evaluating LLMs in clinical medicine, this thesis have developed the PKUCliniEval, a Chinese LLM clinical medicine capability assessment benchmark. This benchmark simulates the knowledge and skills a clinical physician needs from medical school education to clinical practice, including 9,453 questions from Peking University course exams, national medical licensing exams, and other authoritative tests, covering knowledge, reasoning, and clinical treatment across 53 medical subjects. Comparative testing of domestic and international open-source and proprietary LLMs reveals that larger-scale general-purpose Chinese LLMs outperform their English counterparts and smaller specialized medical models in various modules. this thesis investigates and analyzes the reasons behind these performance differences, affirming the capability of LLMs in clinical medicine and providing insights for future model optimization.

The medical document knowledge extraction method, CliniDocParser, equips the intelligent agent with the ability to perceive documents and medical data, while the Chinese large model clinical medicine capability benchmark, PKUCliniEval, verifies the knowledge reservoir of LLMs in clinical disciplines. Thus, based on these components, this thesis has constructed an automated medical research intelligent agent system pothis thesisred by large language models. This agent system, composed of brain, perception, and action modules, utilizes the processing pothis thesisr of large models to automatically mine key medical indicators from massive medical case documents, construct related clinical medical topics, and validate them using statistical analysis methods to produce effective medical research findings. This intelligent agent system can generate research outcomes with practical value for real-world medical studies, making a substantive contribution to unlocking the potential value of medical data and accelerating real-world research by medical professionals.

Overall, this thesis contributes to the innovative architecture for fully leveraging medical data and enhancing medical quality and provides benchmarks and guidance for the future development of large models in the medical field. Future research could further deepen understanding of model performance and real-world applicability and explore new research directions.

KEY WORDS: Pre-trained Language Models， Large Language Models， Information Extraction， Agent