微软开源用于专业领域问题的RAG系统：PIKE-RAG

GitHubStore GitHubStore 2025年02月19日 08:40 湖北

项目简介

微软开源的一个用于专业领域问题的RAG系统：PIKE-RAG，它解决了传统RAG处理专业领域知识时的局限性，比较适合处理深度领域知识和多步逻辑推理的场景

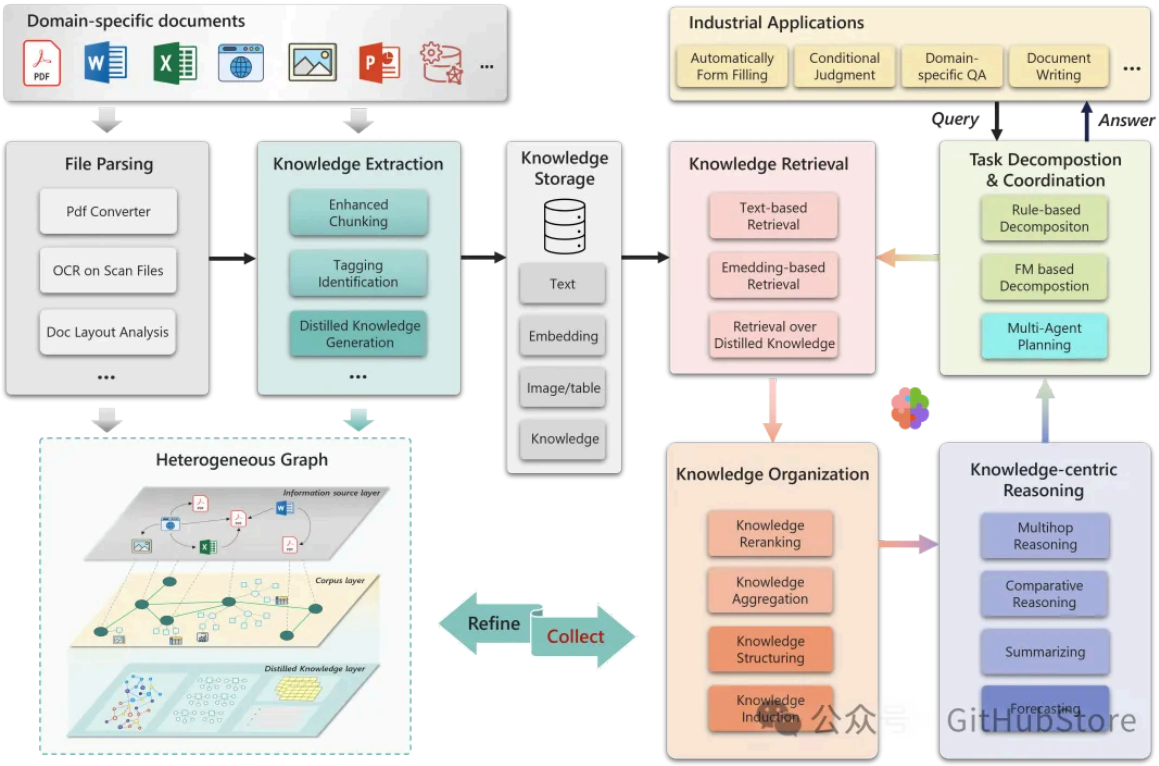
它通过提取、理解和应用领域特定知识，并构建连贯的推理逻辑，逐步引导LLM)生成准确答案

PIKE-RAG包含文档解析、知识提取、知识存储、知识检索、知识组织、以知识为中心的推理以及任务分解和协调多个基本模块，通过调整子模块，来构建针对不同功能的RAG 系统

已在医疗、工业制造、矿业等领域测试，显著提高了问答准确性

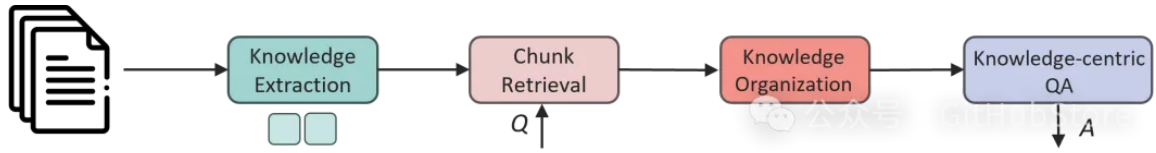
性能表现，在HotpotQA数据集上准确率87.6%；在2WikiMultiHopQA上准确率82.0%；在MuSiQue上准确率59.6%

近年来，检索增强生成（RAG）系统通过外部检索扩展了大型语言模型（LLM）的能力取得了显著进展。然而，这些系统在满足现实世界工业应用的复杂和多样化需求方面仍面临挑战。仅依靠直接检索不足以从专业语料库中提取深层领域特定知识并进行逻辑推理。为了解决这个问题，我们提出了 PIKE-RAG （sPecIalized KnowledgE and Rationale Augmented Generation）方法，该方法专注于提取、理解和应用领域特定知识，同时在构建连贯推理逻辑，逐步引导LLMs向准确响应。

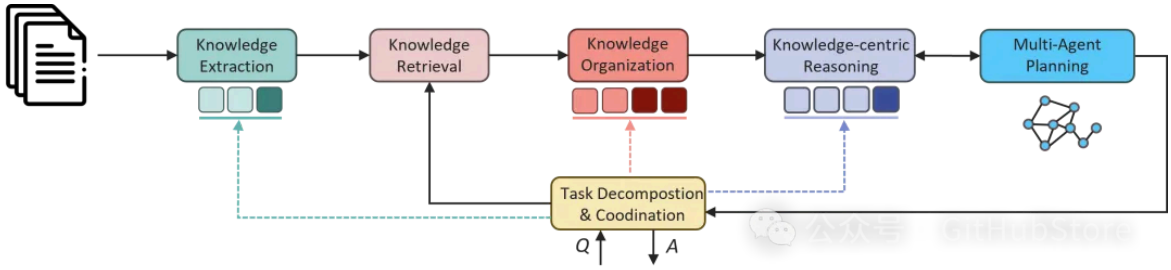


PIKE-RAG 框架主要由几个基本模块组成，包括文档解析、知识提取、知识存储、知识检索、知识组织、以知识为中心的推理以及任务分解和协调。通过调整主模块内的子模块，可以实现专注于不同能力的 RAG 系统，以满足现实场景的多样化需求。

例如，在患者历史病历搜索的情况下，它侧重于事实信息检索能力。主要挑战是：(1) 知识理解和提取常因不恰当的知识分割而受阻，破坏语义连贯性，导致检索过程复杂且效率低下；(2) 基于嵌入的知识检索受限于嵌入模型对专业术语和别名的对齐能力，降低了系统精度。通过 PIKE-RAG，我们可以在知识提取过程中使用上下文感知分割技术、自动术语标签对齐技术和多粒度知识提取方法，从而提高知识提取和检索的准确性，如以下流程图所示。



对于像为患者制定合理的治疗方案和应对措施建议这样的复杂任务，需要更高级的能力：需要强大的领域专业知识来准确理解任务，有时还需要合理分解它；还需要高级的数据检索、处理和组织技术来进行潜在趋势预测；而多智能体规划也将有助于考虑创造性和依赖性。在这种情况下，可以初始化一个更丰富的流程来实现这一点。



在公开基准测试中，PIKE-RAG 在多个多跳问答数据集上表现出色，例如 HotpotQA、2WikiMultiHopQA 和 MuSiQue。与现有基准方法相比，PIKE-RAG 在准确率和 F1 分数等指标上表现出色。在 HotpotQA 数据集上，PIKE-RAG 达到了 87.6%的准确率，在 2WikiMultiHopQA 上达到了 82.0%，而在更具挑战性的 MuSiQue 数据集上，它实现了 59.6%。这些结果表明，PIKE-RAG 在处理复杂推理任务方面具有显著优势，尤其是在需要整合多源信息和执行多步推理的场景中。

PIKE-RAG 经过测试，在工业制造、采矿和制药等领域显著提高了问答准确率。未来，我们将继续探索其在更多领域的应用。此外，我们还将继续探索其他形式的知识和逻辑，以及它们在特定场景中的最佳适应。

快速开始

1. 克隆此仓库并设置 Python 环境，请参阅此文档
2. 创建一个 `.env` 文件以保存您的端点信息（以及如果需要的话，一些其他环境变量），请参阅此文档
3. 修改 yaml 配置文件，并在 examples/目录下尝试脚本，参考此文档；
4. 构建您自己的管道和/或添加您自己的组件！

项目链接

> https://github.com/microsoft/PIKE-RAG

人工智能 658　AI应用 603　RAG 26

人工智能 · 目录

上一篇 · 开源本地化 RAG 系统Minima