基于开源大语言模型的知识库问答系统实践

Practice of the Knowledge-based Question Answering System Based on Open Source LLM

周绍彬 肖铮

厦门大学，厦门，361005

传统的基于规则库的问答系统主要是利用检索技术在知识库中查找相关的信息，缺乏对语义的理解和推理的能力，在回答复杂或开放式的问题时会存在准确率较低的问题。基于大语言模型的问答系统在理解和生成自然语言方面具有显著优势，能够有效克服传统问答系统的局限，但是大语言模型在回答问题的时候会不稳定，甚至编造答案。因此我们采用传统“ FAQ（基于常见问题）问答系统 ”与“ AI 问答系统 ”结合的方式来建立知识库问答系统。在选择大语言模型的时候，线上大语言模型会存在数据隐私性差，模型不可控的问题，所以我们采用本地部署的大语言模型来建立知识库问答系统。

本实践通过开源框架 Ollama 部署了“ Qwen:7b ”开源大语言模型，向量模型使用了 M3E（Moka Massive Mixed Embedding）模型。问答系统使用了开源框架 FastGPT 来搭建。

该问答系统采用流程化的处理的方式来回答问题，使得问答系统的扩展性更灵活。该问答系统的基础流程是用户提出问题首先经过“ FAQ 问答系统 ”进行匹配处理，简单的问题可以由 FAQ 问答系统进行回答。没有被 FAQ 问答系统处理的问题则进入“ AI 问答系统 ”处理，根据语义从知识库中检索相关的内容，同问题一起提供给本地部署的大语言模型，由大语言模型来给出回答。系统定期统计经 AI 问答系统处理的问题和回答内容，再将常用的问题补充到 FAQ 问答系统中，以帮助 FAQ 问答系统提升问题命中率。

基于本地部署的开源大语言模型可以避免敏感数据泄漏的问题，同时可以通过微调等方式使得模型更符合应用场景。本实践将 FAQ 问答系统和 AI 问答系统结合到一起，弥补了传统问答系统在语义理解和推理能力方面较弱的缺点，也对具有明确回答的问题提供了更精确地回答。采用流程处理的方式，使得系统具有很强的扩展性，例如可以在流程入口和出口增加关键词过滤，可以过滤掉敏感的内容，后期可以通过接口调用与其他系统进行集成。

关键词：知识库问答，大语言模型，知识库

中图分类号： G251

作者简介及联系方式:

周绍彬，工程师，研究方向：数字图书馆

中国知网 https://www.cnki.net

肖铮，高级工程师，研究方向：数字图书馆

中国知网 https://www.cnki.net