Computer Science and Application 计算机科学与应用, 2024, 14(11), 21-27
Published Online November 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.1411212

Hans汉斯

一种基于RocketQA与ChatGLM模型的问答机器人系统研究与设计

宋丽萍*，王天与，宋丽华，孙虹飞

北方工业大学信息学院，北京

收稿日期：2024年10月3日；录用日期：2024年10月30日；发布日期：2024年11月8日

摘　要

随着信息技术的快速发展，用户对高效获取精准信息的需求日益增长。此外，为了支持国产操作系统软件生态，文章基于Deepin操作系统，集成RocketQA和ChatGLM模型，设计并实现了一个智能文档问答系统。系统首先使用RocketQA模型将问题向量化，利用Faiss进行检索和排序，然后由RocketQA模型二次搜索和排序，对结果进行增强，检索到的文档通过ChatGLM模型转化为自然语言答案，以“参考链接 + 答案”的格式呈现给用户。最后，在deepin wiki数据集上进行了广泛的测试，结果表明系统在问答准确性和响应速度方面均表现优异。

关键词

智能问答，RocketQA，ChatGLM，自然语言处理，Faiss，Tornado框架，Deepin操作系统

Research and Design of a Question-Answering Robot System Based on RocketQA and ChatGLM Model

Liping Song*, Tianyu Wang, Lihua Song, Hongfei Sun

College of Information, North China University of Technology, Beijing

Received: Oct. 3rd, 2024; accepted: Oct. 30th, 2024; published: Nov. 8th, 2024

Abstract

With the rapid development of information technology, users' demand for efficiently obtaining

*通讯作者。

文章引用: 宋丽萍, 王天与, 宋丽华, 孙虹飞. 一种基于 RocketQA 与 ChatGLM 模型的问答机器人系统研究与设计[J]. 计算机科学与应用, 2024, 14(11): 21-27. DOI: 10.12677/csa.2024.1411212

accurate information is growing. In addition, to support the ecosystem of domestic operating system software, this paper designs and implements an intelligent document question-answering system based on the Deepin operating system, integrating the RocketQA and ChatGLM models. The system first uses the RocketQA model to vectorize the question, utilizes Faiss for retrieval and sorting, then uses the RocketQA model for a second search and sorting to enhance the results. The retrieved documents are converted into natural language answers by the ChatGLM model and presented to the user in the format of "reference link + answer". Finally, extensive testing was conducted on the deepin wiki dataset, and the results show that the system performs excellently in both question-answering accuracy and response speed.

Keywords

Intelligent Question Answering, RocketQA, ChatGLM, Natural Language Processing, Faiss, Tornado Framework, Deepin Operating System

1. 引言

随着信息技术的快速发展，信息检索已成为人工智能领域的重要分支。在信息化的背景下，信息检索系统在帮助用户从海量数据中快速准确地找到所需信息方面发挥着至关重要的作用。国产操作系统，如 Deepin，作为推动信息技术自主可控的重要力量，其生态系统的建设和完善对于保障国家信息安全、促进软件产业发展具有重大意义[1]。然而，当前国产操作系统在信息检索和智能问答方面仍存在诸多挑战，尤其是在处理中文查询和提供准确答案方面。因此，开发适合国产操作系统的高效、智能问答系统对于提升用户体验、推动国产操作系统的广泛应用具有重要的现实意义和深远的战略意义。

在国际上，智能问答系统的研究已经取得了显著进展，在多个领域和场景中得到广泛应用，能够理解用户的自然语言查询并提供准确的答案。但它们主要基于国外开发的技术和平台，对于中文和其他多语言的支持尚未达到最佳状态，需要进一步优化以满足非英语用户的需求。

在国内，随着国产操作系统的快速发展，越来越多的研究者开始关注基于国产操作系统的智能问答系统。尽管如此，国产操作系统在智能问答领域的研究仍面临着数据资源、算法优化、系统部署等多方面的挑战。

针对现有的问答机器人在中文处理能力不足、数据资源有限等问题，本研究旨在开发一个基于 Deepin 国产操作系统的智能问答机器人，以提供高效、准确的问答服务。本研究利用了深度学习技术在语言理解、文本编码和自然语言生成方面的优势，集成了 RocketQA [2]和 ChatGLM [3]模型，设计并实现了一个智能文档问答系统。系统首先使用 RocketQA 模型将问题向量化，利用 Faiss [4]进行检索和排序，然后由 RocketQA 模型二次搜索和排序，对结果进行增强。检索到的文档通过 ChatGLM 模型转化为自然语言答案，以“参考链接 + 答案”的格式呈现给用户。这种集成方式不仅提高了对中文查询的理解能力，而且通过优化算法和简化系统部署，提升了用户体验。

本研究为 Deepin 平台专门打造的基于 Deepin 领域的问答机器人[5]，不仅为国产操作系统的使用和推广提供了有力的支持，而且为智能问答领域提供了新的解决方案，推动了国产操作系统的广泛应用和智能化发展。在 Deepin wiki 数据集上进行的广泛测试进一步验证了系统的有效性和实用性。

2. 相关研究

尽管已有多种问答模型，如 BERT [6]、RoBERTa [7]、T5 [8]等预训练语言模型在问答任务中表现出色，但它们通常需要大量的计算资源，并且在特定领域的适应性上有所不足，具体特性对比见表 1。

Table 1. Comparison of characteristics of existing Q&A models
表 1. 现有问答模型特性对比

特性模型	BERT	RoBERTa	T5	RocketQA
模型类型	预训练语言模型	预训练语言模型(BERT 的改进版)	文本到文本的转换器	检索增强的问答模型
问答能力	需要额外的训练进行问答任务	能够处理问答任务，性能较 BERT 有所提升	专为文本生成设计，适用于问答	集成检索和生成，专注于问答
检索能力	有限，需要与其他系统配合	有限，需要与其他系统配合	非其主要设计目标	强，专为检索设计
生成能力	能够生成流畅的文本	能够生成流畅的文本，性能较 BERT 增强	生成能力强，特别适用于文本生成任务	生成能力结合检索结果
上下文理解	强	强，较 BERT 更强	强	强，结合检索结果
适用场景	多种 NLP 任务，包括问答	多种 NLP 任务，包括问答	文本摘要、翻译、问答等文本生成任务	开放域问答，需要文档检索的场景
创新性	作为早期的预训练模型具有创新性	在 BERT 基础上的改进	T5 的文本到文本方法具有创新性	检索增强的问答方法结合了检索和生成

信息生成(Information Generation, IG)技术[9]通常指的是利用人工智能算法自动生成信息、文本或其他类型内容的技术。IG 技术在问答系统中的应用可以包括自动回答生成[10]、内容创作、数据增强等。该技术自动生成的信息可能不总是完全准确，尤其是在复杂或专业性很强的领域。对于一些复杂或模糊的查询，IG 技术可能难以准确理解用户的真正意图，从而生成不相关或不准确的回答。

为了解决这些问题，本研究探索了检索式和生成式模型[11]的集成，以期在保持高性能的同时，提高系统的灵活性和适应性。

RocketQA 是一个基于预训练的检索式问答模型，具有强大的语言理解和知识获取能力，是全世界首个面向中文的端到端搜索问答工具包。与此同时，ChatGLM 是一个生成式对话模型，旨在模仿自然语言对话，并能够生成连贯、流畅的文本回复，从而与用户进行交互。ChatGLM 模型的核心思想是利用大规模的对话数据来学习对话的模式和规律，从而生成对话文本。该模型在对话生成方面表现出色，能够模拟人类对话的特点，并且能够根据上下文信息作出合适的回复。

本研究创新性地将 RocketQA 和 ChatGLM 模型结合起来，这种架构允许系统在检索到的信息不足以生成完整答案时，通过生成式模型进行补充。同时，可以根据需要调整检索式和生成式模型的参数，以适应不同的应用场景和数据集，使系统具备良好的泛化能力。

3. 系统设计

系统设计包括以下几个关键组件：数据预处理、文档编码、索引构建、查询处理、相似性搜索、文档检索和答案生成。每个组件都针对 Deepin 国产操作系统的特定需求进行了优化：

1、数据预处理：本系统采用基于规则的文本清洗方法，结合自然语言处理(NLP)技术，对文档进行清洗和格式化，确保它们适合模型处理[12]。这一步骤包括去除文档中的噪声数据，如 HTML 标签、特

殊字符等，为后续的编码和索引构建打下基础。

2、文档编码算法：文档编码是将原始文本转换为机器可理解的向量表示。本系统采用了 RocketQA 模型，该模型基于深度双向编码器，能够捕捉文档的语义信息。具体步骤如下：

1) 文本清洗：使用正则表达式和 NLP 工具去除无用的标点符号、停用词等。

2) 分词：采用基于深度学习的分词技术，提高分词的准确性，将文本分解为单独的词汇单元。

3) 词嵌入：使用预训练的词向量模型，Word2Vec 来捕捉词汇的语义关系，将每个词汇转换为高维空间中的向量。

4) 编码：RocketQA 模型采用 Transformer 架构，利用自注意力机制对文档进行编码，将文本转换为高维空间中的向量表示。

3、索引构建算法：索引构建是提高检索效率的关键。采用了 Faiss (Facebook AI Similarity Search)库，该库专为高效相似性搜索和密集向量索引而设计。构建索引的步骤包括：

1) 向量归一化：对文档向量进行归一化处理，确保向量长度一致，以提高搜索的准确性。见公式 1：

$$normalized\_emd = \frac{x}{\sqrt{\sum_{i=1}^{d} x_i^2}} \tag{1}$$

2) 索引创建：使用 Faiss 创建索引结构，IndexFlatIP (Index Flat with Inner Product)用于内积搜索的索引类型，它适用于大规模数据集的内积搜索。见公式 2：

$$\text{similarity}(q,d) = q.d \tag{2}$$

3) 向量添加：采用批量添加技术，将归一化的文档向量添加到索引中，提高索引构建的效率。

4、查询处理算法：查询处理涉及将用户的自然语言查询转换为机器可处理的形式。采用了以下步骤：

1) 查询解析：使用依存句法分析来识别查询中的实体和关系，识别和提取查询中的关键词汇。

2) 查询编码：将查询编码为与文档编码相同的向量空间，以便进行有效的相似性比较。

5、相似性搜索算法：相似性搜索是问答系统的核心，采用了 Faiss 库中的向量搜索算法。具体步骤包括：

1) 向量搜索：采用局部敏感哈希(LSH)技术，提高搜索速度，找到与查询向量最相似的文档向量。

2) 结果排序：根据相似度得分对搜索结果进行排序。使用负对数似然损失(Negative Log Likelihood Loss)，最大化正例文档与问题之间的相似度，同时最小化负例文档与问题之间的相似度。见公式 3：

$$\mathrm{L}\left(q, p^{+}, \left\{p_j^{-}\right\}_{j=1}^{m}\right) = -\log \frac{e^{sim\left(q,p^{+}\right)}}{\sum_{j-1}^{m} e^{sim\left(q,p_j^{-}\right)}} \tag{3}$$

6、文档检索算法：文档检索是选择最相关文档的过程。利用 RocketQA 模型对搜索结果进行进一步的排序和筛选：

1) 相似度评分：采用余弦相似度，计算查询向量与文档向量之间的相似度。

2) 结果筛选：引入阈值机制，只选择高于特定阈值的文档作为候选答案。

7、答案生成算法：答案生成是将检索到的文档转换为自然语言答案的过程。采用了 ChatGLM 模型，该模型能够生成流畅自然的文本：

1) 文档理解：使用文本摘要和信息抽取技术，从文档中提取出最有价值的信息。

2) 答案生成：ChatGLM 模型采用序列到序列的架构，生成连贯且信息丰富的答案，对 RocketQA 模型检索到的答案进一步增强。

4. 系统实现

系统采用端到端的形式实现，服务端与客户端进行交互通信，系统实现见图 1。

1) 封装 API：服务端的 API 被封装为一个 HTTP POST 端点，客户端通过发送 HTTP POST 请求与之交互。请求体中包含了查询和所需的参数。

2) 调用 API：客户端使用 requests.post ()函数调用服务端 API。这个函数接受服务端地址、请求数据和其它配置(如 headers 等)。

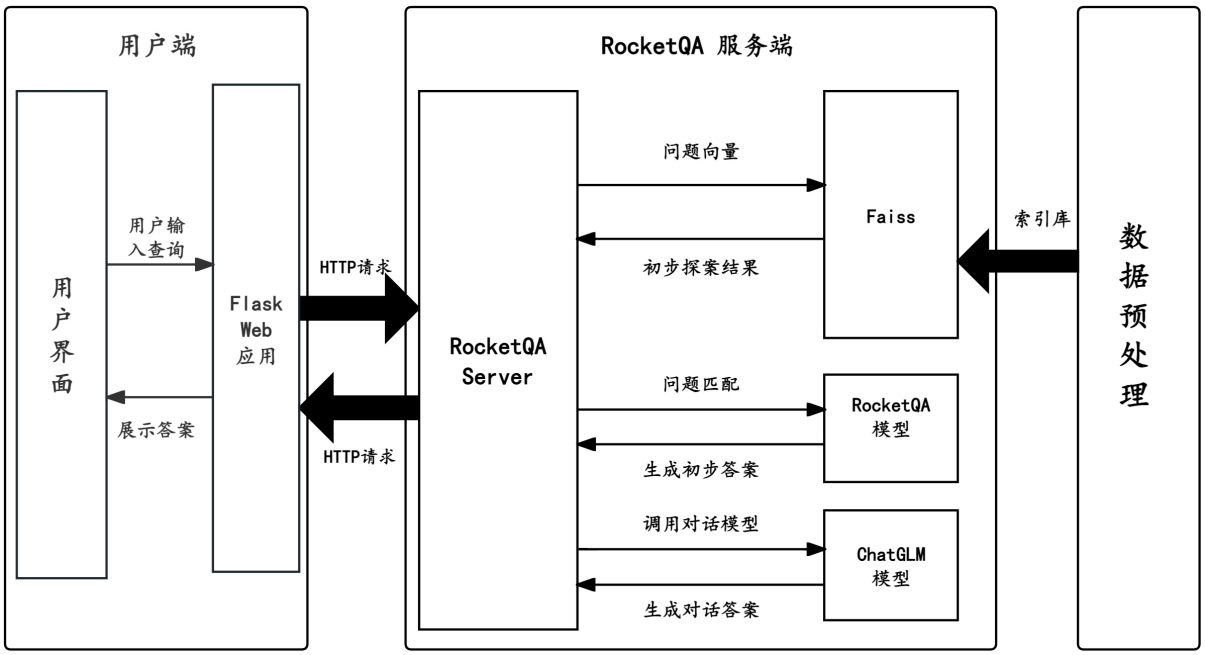

Figure 1. System framework diagram
图 1. 系统框架图

1) 输入阶段：用户输入的查询通过客户端的自然语言处理模块进行初步处理，以提高查询的质量和准确性。而后查询被发送到服务端。

2) 预处理：服务端接收到查询后，进行分词、去除停用词等预处理步骤。

3) 查询向量生成：使用 RocketQA 模型对查询进行编码，生成能够代表查询语义的向量。

4) 相似性搜索：在 Faiss 索引中进行向量搜索，找到与查询向量最相似的文档向量，并使用 LSH 技术提高搜索效率，找到 k 个最相似的文档向量。

5) 文档解码：检索到的向量通过 RocketQA 模型解码，以理解文档内容。

6) 答案提取：根据解码后的信息，提取可能的答案候选。

7) 答案优化：ChatGLM 模型对答案候选进行优化，生成自然、流畅的答案。

8) 结果封装：最终答案与相关文档的参考链接结合，形成完整的回答。

9) 输出阶段：系统将封装好的答案通过客户端展示给用户，提供直观、易理解的界面。

5. 实验分析

为了全面评估该智能文档问答系统的性能，在两个不同的数据集上进行了广泛的测试：Deepin wiki 数据集和玲珑使用数据集。此外，此系统不仅在 Windows 平台上进行了实现，同时成功部署在了 Deepin 操作系统上。

1、数据集和实验环境

Deepin wiki 数据集：包含 900 多条 Deepin 系统相关的中文教程和词条，是主要测试数据集。

玲珑使用数据集：提供了额外的文档资源，增加了数据的多样性和复杂性。

操作系统：系统在 Windows 和 Deepin 操作系统上均进行了部署和测试，确保了跨平台的兼容性。

硬件环境参数：GPU：1*NVIDIA V100，内存：32 GB。

2、为了能更好地体现本系统准确性和响应速度方面都具备良好的性能，我们采用 RocketQA 与 MacBERT 模型集成作为对照组，从以下几个方面对系统性能进行了评估：

1) 准确性：在 Deepin wiki 和玲珑使用数据集上进行了问答准确性的测试。测试结果显示，系统在两个数据集上均达到了 80%以上的答案相关性，证明了模型在理解用户查询和提供准确答案方面的能力。

2) 响应时间：系统的响应时间是衡量用户体验的关键指标。在测试中，RocketQA 与 ChatGLM 集成系统的平均响应时间为 13 秒，见表 2；RocketQA 与 MacBERT 集成系统的平均响应时间为 12.6 秒，见表 3。

Table 2. Response time of the integrated system of RocketQA and ChatGLM
表 2. RocketQA 与 ChatGLM 集成系统响应时间

试验次数	实验 1	实验 2	实验 3	实验 4	实验 5	……	实验	实验	平均响应时间
响应时间	19s	8s	20s	18s	12s	……	14s	13s	13 s

Table 3. Response time of the integrated system of RocketQA and MacBERT
表 3. RocketQA 与 MacBERT 集成系统响应时间

试验次数	实验 1	实验 2	实验 3	实验 4	实验 5	……	实验	实验	平均响应时间
响应时间	12 s	8 s	19 s	7 s	12 s	……	15 s	15 s	12.6 s

回答信息密度：旨在使用接收信息字数/响应时间进行表示。在测试中，RocketQA 与 ChatGLM 集成系统的平均回答信息密度为 12 字/秒，见表 4；RocketQA 与 MacBERT 集成系统的平均回答信息密度为 0.66 字/秒，见表 5。

Table 4. Answer information density of the integrated system of RocketQA and ChatGLM
表 4. RocketQA 与 ChatGLM 集成系统回答信息密度

试验次数	实验 1	实验 2	实验 3	实验 4	实验 5	……	实验	实验	平均回答信息密度
回答信息密度	12 字/秒	18 字/秒	9 字/秒	15 字/秒	17 字/秒	……	10 字/秒	9 字/秒	12 字/秒

Table 5. Answer information density of the integrated system of RocketQA and MacBERT
表 5. RocketQA 与 MacBERT 集成系统回答信息密度

试验次数	实验 1	实验 2	实验 3	实验 4	实验 5	……	实验	实验	平均回答信息密度
回答信息密度	0.5 字/秒	0.2 字/秒	1 字/秒	1 字/秒	0.4 字/秒	……	0.5 字/秒	1 字/秒	0.66 字/秒

Table 6. Comparison between ChatGLM model and MacBERT-large Model
表 6. ChatGLM 模型与 MacBERT-large 模型对比

特性/模型	MacBERT-large	ChatGLM
模型类型	基于 BERT 的预训练语言模型	基于 GPT 的生成式语言模型
回答风格	简短、直接、或过于生硬	自然连贯、更符合人类语言习惯
生成能力	主要依赖于输入文本，生成能力有限	能够自主生成流畅的文本
上下文理解	强，但可能缺乏创造性	强，并且能够创造性地生成回答
交互性	有限，主要基于已有文本	交互性好，根据上下文生成回答
模型集成	需要与检索模型(如 RocketQA)配合使用	可以独立生成回答，也可以与检索模型集成
部署难度	相对容易部署	部署可能更复杂，需要考虑生成文本的质量和多样性

结果分析：MacBERT-large 是基于大规模 MRC 数据再训练的模型，在阅读理解/分类等任务上均有大幅提高，将此模型放到下游任务微调可比直接使用预训练语言模型提高 2 个点/1 个点以上。以此作为对比，两者在响应时间上的性能相仿，但在回答信息密度上，RocketQA 与 ChatGLM 模型集成系统是 RocketQA 与 MacBERT 模型集成系统的 18 倍，可以充分的体现出本系统的优越性能。更多功能对比见表 6。

6. 结论

本研究为 Deepin 国产操作系统提供了定制化的智能问答解决方案，推动了国产操作系统的发展。将 RocketQA 和 ChatGLM 模型的集成，实现了从查询编码到答案生成的完整流程，提高了系统的智能化水平。通过深度学习模型的应用，提高了对中文查询的理解能力和答案的自然性。期待该系统能够扩展至更多领域，并在未来的工作进一步优化模型性能，提高系统的个性化和智能化水平。

基金项目

北京市大学生创新创业训练计划项目(2024)；北方工业大学教育教学改革项目。

参考文献

[1] 国产操作系统迎来新突破[J]. 市场瞭望, 2024(15): 1.

[2] Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., *et al*. (2021). RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. arXiv: 2010.08191.

[3] Zhang, X., Zhang, X. and Yu, Y. (2023). ChatGLM-6B Fine-Tuning for Cultural and Creative Products Advertising Words. 2023 *International Conference on Culture-Oriented Science and Technology* (*CoST*), Xi'an, 11-14 October 2023, 291-295. https://doi.org/10.1109/CoST60524.2023.00066

[4] Douze, M., Guzhva, A., Deng, C., *et al.* (2024) The Faiss Library. arXiv: 2401.08281.

[5] 关殿玺, 黄琨, 崔年治. 基于大模型、RAG 和智能体技术的勘察岩土问答机器人研究[J]. 中国勘察设计, 2024(8): 101-104.

[6] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805.

[7] Liu, Y., Ott, M., Goyal, N., *et al*. (2019) Roberta: A Robustly Optimized Bert Pretraining Approach. arXiv: 1907.11692.

[8] Ni, J., Abrego, G.H., Constant, N., *et al*. (2021) Sentence-t5: Scalable Sentence Encoders from Pre-Trained Text-To-Text Models. arXiv: 2108.08877.

[9] 王若佳, 范科鸣, 刘智锋, 等. 生成式人工智能环境下用户信息检索式行为研究[J/OL]. 数据分析与知识发现: 1-15. http://kns.cnki.net/kcms/detail/10.1478.G2.20240117.1057.008.html, 2024-09-09.

[10] 赵芸, 刘德喜, 万常选, 等. 检索式自动问答研究综述[J]. 计算机学报, 2021, 44(6): 1214-1232.

[11] 刘邦奇, 聂小林, 王士进, 等. 生成式人工智能与未来教育形态重塑: 技术框架、能力特征及应用趋势[J]. 电化教育研究, 2024, 45(1): 13-20.

[12] 黄施洋, 奚雪峰, 崔志明. 大模型时代下的汉语自然语言处理研究与探索[J/OL]. 计算机工程与应用: 1-19. http://kns.cnki.net/kcms/detail/11.2127.tp.20240925.1046.017.html, 2024-10-01.