文章编号:1008-1534(2025)02-0103-08

一种基于领域知识的检索增强生成方法

张高飞[1],李 欢[1],池云仙[1],赵巧红[2],勾智楠[3],高 凯[1]

(1. 河北科技大学信息科学与工程学院,河北石家庄 050018;2. 河北轨道运输职业技术学院,河北石家庄 050801;3. 河北经贸大学管理科学与信息工程学院,河北石家庄 050061)

摘 要:为了提高当前大语言模型(large language model,LLM)在利用检索文档生成答案时的准确性,提出一种基于领域知识的检索增强生成(retrieval-augmented generation,RAG)方法。首先,在检索过程中通过问题和领域知识进行第 1 层的稀疏检索,为后续的稠密检索提供领域数据集;其次,在生成过程中采用零样本学习的方法,将领域知识拼接在问题之前或之后,并与检索文档结合,输入到大语言模型中;最后,在医疗领域和法律领域数据集上使用大语言模型 ChatGLM2-6B、Baichuan2-7B-chat 进行多次实验,并进行性能评估。结果表明:基于领域知识的检索增强生成方法能够有效提高大语言模型生成答案的领域相关度,并且零样本学习方法相较于微调方法表现出更好的效果;采用零样本学习方法时,融入领域知识的稀疏检索和领域知识前置方法在 ChatGLM2-6B 上取得了最佳提升效果,与基线方法相比,ROUGE-1、ROUGE-2 和 ROUGE-L 评分分别提高了 3.82、1.68、4.32 个百分点。所提方法能够提升大语言模型生成答案的准确性,为开放域问答的研究和应用提供重要参考。

关键词:自然语言处理;开放域问答;检索增强生成;大语言模型;零样本学习;领域知识

中图分类号:TP391.1 **文献标识码**:A **DOI**:10.7535/hbgykj.2025yx02001

A retrieval-augmented generation method based on domain knowledge

ZHANG Gaofei[1], LI Huan[1], CHI Yunxian[1], ZHAO Qiaohong[2], GOU Zhinan[3], GAO Kai[1]

(1. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China; 2. Hebei Vocational College of Rail Transportation, Shijiazhuang, Hebei 050801, China; 3. School of Management Science and Information Engineering, Hebei University of Economics and Business, Shijiazhuang, Hebei 050061, China)

Abstract: In order to enhance the accuracy of current large language model (LLM) in generating answers using retrieval documents, a retrieval-augmented generation method based on domain knowledge was proposed. Firstly, during the retrieval process, the first layer of sparse retrieval was conducted using both the question and domain knowledge, providing a domain-specific dataset for subsequent dense retrieval. Secondly, in the generation process, a zero-shot learning method was employed to concatenate domain knowledge before or after the question, and combined it with the retrieved documents to input into the large language model. Finally, extensive experiments were conducted on datasets in the medical and legal domains using ChatGLM2-6B and Baichuan2-7B-chat, and performance evaluations were conducted. The results indicate that the retrieval-augmented generation method based on domain knowledge can effectively improve the domain relevance of the answers generated by large language models, and the zero-shot learning method outperforms the fine-tuning method. When the zero-

收稿日期:2024-11-08;修回日期:2025-01-19;责任编辑:丁军苗

基金项目:河北省自然科学基金(F2022208006,F2023207003);河北省高等学校科学技术研究项目(QN2024196)

第一作者简介:张高飞(2000—),男,河北邯郸人,硕士研究生,主要从事自然语言处理方面的研究。

通信作者:高凯教授。E-mail:gaokai@hebust.edu.cn

张高飞,李欢,池云仙,等. 一种基于领域知识的检索增强生成方法[J]. 河北工业科技,2025,42(2):103-111.

ZHANG Gaofei, LI Huan, CHI Yunxian, et al. A retrieval-augmented generation method based on domain knowledge[J]. Hebei Journal of Industrial Science and Technology, 2025, 42(2): 103-111.

shot learning method is used, the sparse retrieval incorporating domain knowledge and the method of placing domain knowledge before the question achieve the best improvement on ChatGLM2-6B, with ROUGE-1, ROUGE-2 and ROUGE-L scores increasing by 3.82, 1.68 and 4.32 percentage points respectively compared to the baseline method. The proposed method can enhance the accuracy of the answers generated by large language models and provide an important reference for the research and application of open-domain question answering.

Keywords: natural language processing; open-domain question answering; retrieval-augmented generation; large language model; zero-shot learning; domain knowledge

在开放域问答任务中,大语言模型[1-3](large language model,LLM)利用检索增强生成(retrieval-augmented generation,RAG)系统提供的检索文档能够生成精确的答案[4-5]。尽管 RAG 方法展现出显著的优势,但仍然存在一些问题,如 LLM 利用检索文档生成答案的准确性不高,检索文档中可能会包含不相关的内容[6-7],这些问题都会导致 LLM 生成错误的答案[8-9]。因此,如何提高 LLM 利用检索文档生成答案的准确性成为一个亟待解决的问题。

现代开放域问答系统已普遍将传统的检索技术与神经阅读理解模型相结合[10]。检索器[11-12]检索文档后,提取式或生成式阅读器使用检索文档生成答案[13]。近年来,诸多研究大多数聚焦于 RAG 系统的鲁棒性[14-17]。ASAI 等[15]引入生成特定标记的机制,使 LLM 能够自我评估文档与问题的相关性,从而自我检查生成结果的准确性。YORAN 等[17]的研究要求 LLM 在生成答案前先评估文档的相关性,该方法仅依赖稀疏的二进制标签评估文档的相关性,难以精确捕捉细粒度的关联信息。ZHANG 等[18]通过微调 LLM,使其学习如何过滤干扰文档,从而优化特定领域的 RAG 系统,虽然效果显著,但是微调 LLM 非常耗时且需要大量的算力资源。GLASS 等[19]在检索阶段增加了重排器,对检索到的文档集进行重排,从而提升检索准确性,但要实现这一效果,重排前的检索文档的质量必须足够高。KIM 等[20]提出了新的 RAG 框架,该框架通过检索相关文档并生成摘要,进而提升 LLM 生成答案的准确性。尽管 RAG 领域已经取得了一系列卓越的研究成果,但现有研究在整合领域知识方面仍有待加强,这种局限性阻碍了对检索文档中领域知识的深入挖掘,同时也影响了大语言模型利用检索文档生成答案的准确性和可靠性。

针对上述问题,本文提出了一种基于领域知识的检索增强生成方法。在检索阶段,通过融入领域知识向 LLM 提供问题所属领域的相关检索文档;在生成阶段,将领域知识拼接于问题之前或之后,并结合检索阶段的检索文档,以零样本学习(zero-shot learning, ZSL)[21]方式输入给 LLM。该方法能够增强 LLM 对专业领域知识文档的理解,提升 LLM 对检索文档的适应性和选择性,提高 LLM 生成答案的准确性。

1 基于领域知识的检索增强生成方法提出

1.1 基本框架

在利用 RAG 方法解决开放域问答任务时,首先,针对给定查询 q,通过检索器在检索阶段进行稀疏检索和稠密检索,从文档集中提取出前 k 个相关文档(可选用重排器进一步优化结果)[22-23];然后,将查询 q 与检索到的参考文档 $D=\{d_i\}_{i=1}^{k}$ 一同输入到阅读器,由其生成一个答案 A。

$$A=\{\mathrm{LLM}(q,d_i)\mid d_i\in D\}\ 。\tag{1}$$

本文提出一种基于领域知识的检索增强生成方法,该方法在检索和生成过程中融入领域知识,从而提升大语言模型生成答案的准确性,其基本框架如图 1 所示。

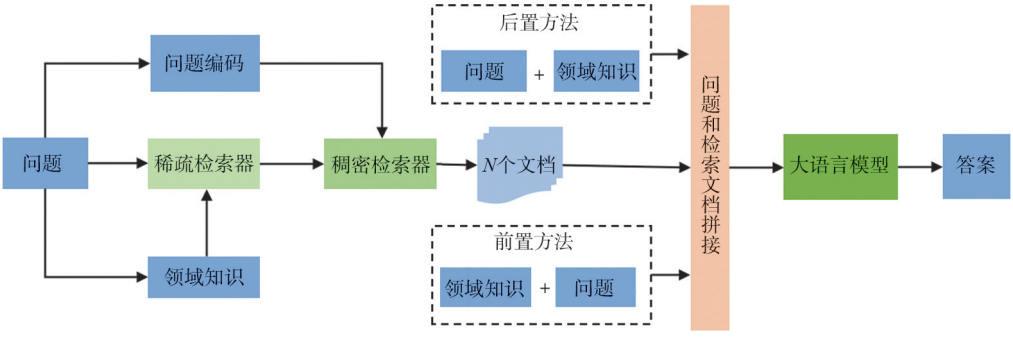


图 1 基于领域知识的检索增强生成方法基本框架

Fig. 1 Basic framework of retrieval-augmented generation based on domain knowledge

在检索阶段的稀疏检索部分，通过问题与领域知识的联合检索，为稠密检索提供训练数据集，从而更精准地获取高领域相关度的检索文档。在生成过程中，采用前置或后置拼接领域知识的方式，使 LLM 能够更有效地选择与问题领域相关度更高的检索文档，进而提高答案的准确性。

1.2　基于领域知识的检索

在 RAG 方法中，通过第 1 层的稀疏检索构建的文档集包含正样本（包含 1 条黄金样本和多条正样本，黄金样本与正确答案完全一致，而正样本则为其他正确的答案）、负样本（错误的样本）和强负样本（强负样本是指和黄金样本非常相似但是依旧错误的样本）。

在稀疏检索阶段中融入领域（domain，Dmn）知识，不仅可以为第 2 层的稠密检索器提供更好的训练集合，而且可以使检索到的文档和问题的领域相关度更加契合。在第 1 层检索中，采用 BM25[24] 算法进行联合检索，从而利用问题及其领域知识共同检索相关内容。具体公式如下

$$\mathrm{SCORE}(D_{\mathrm{mn}}+q,D)=\sum_{i=1}^{n}\mathrm{IDF}(D_{\mathrm{mn}i}+q_i)\times\frac{f(D_{\mathrm{mn}i}+q_i,D)\times(k_1+1)}{f(D_{\mathrm{mn}i}+q_i,D)+k_1\times(1-b+b\times\frac{|d|}{\overline{|d|}})},\tag{2}$$

式中：D_{mn} 为领域信息；d 表示检索的文档集；n 表示查询中包含的词项数；f 函数表示查询词项在样本集中出现的频率；b 和 k_1 是该算法的超参数；$|d|$ 表示文档集 d 的长度；$\overline{|d|}$ 表示所有文档集平均长度。

通过第 1 层稀疏检索获得 1 个按照相关度排行的文档集，用于为第 2 层的稠密检索构建训练文档集。该文档集正样本（以第 1 条正样本为黄金正样本）、负样本和强负样本的比例按照稠密段落检索（dense passage retriever，DPR）方法中的 1∶2∶2 设置[25]。负样本的选择借鉴了 RocketQA 的方法[26]，在其他文档的黄金正样本中随机选取多个样本，作为该文档集的负样本，这种方法有助于更好地区分正、负样本。强负样本是选取相关度排行文档集中排名靠后的样本。对检索阶段进行领域知识的融入，保证检索到的检索文档集都是在同一领域下的。

无论是在稀疏检索还是稠密检索阶段，文档集均基于训练集中的问答对构建，且不包含任何测试集中的问题和答案。

1.3　基于领域知识的生成

在特定领域，准确的领域知识对 LLM 处理检索文档起到了至关重要的作用。通过前置或后置方法，将领域知识融入到 RAG 的生成过程中，从而提升开放域问答任务的准确性和专业性。

1）前置方法　通过前置方法，将问题与该问题所对应的领域知识进行前置拼接，这样做的目的是，在问题理解阶段为 LLM 提供直接且明确的语境指引，促进 LLM 更快捕捉到问题的专业领域和特定背景，前置领域知识的方法可以使 LLM 给予领域知识更高的权重信息，提高生成文档和问题的领域相关度。具体公式如下。

$$a_{\mathrm{front}}=\mathrm{LLM}(D_{\mathrm{mn}}+Q,D),\tag{3}$$

式中：a_{front} 为前置方法的大模型生成的答案；Q 为查询的问题。

2）后置方法　后置方法通过将检索问题与其对应领域知识后置拼接，作为背景提示贯穿模型处理问题的全过程，影响 LLM 对检索文档领域知识的判断，并最终影响生成决策。具体公式如下。

$$a_{\mathrm{rear}}=\mathrm{LLM}(Q+D_{\mathrm{mn}},D),\tag{4}$$

式中：a_{rear} 为后置方法的大模型生成的答案。

2　实验与结果分析

2.1　数据集和实验设置

2.1.1　数据集

本文基于医疗领域和法律领域 2 个数据集进行实验。

1）医疗数据集　针对开放域问答任务，实验采用了一个包含 25 万条数据的数据集。该数据集是人工构建的数据集，主要包括详细的病人问题和医生科室信息、治疗建议，训练集和测试集的比例为 8∶2。医生所属科室信息为领域知识，如表 1 所示。

表 1　医疗数据集不同科室数量统计

Tab. 1　Statistics on the number of different departments in the medical data sets

医生所属科室	病人问题/个
内科	129 120
妇产科	93 502
外科	57 167
耳鼻喉科	38 750
皮肤科	24 109
儿科	15 618
肿瘤科	11 624
整形美容外科	8 476
精神病科	6 617
性病科	6 200
传染病	3 505

2)法律数据集　采用 LaWGPT[27] 中的法律知识问答数据集,共 2.3 万条法律问答数据。该数据集包括法律咨询问题及其答案和该问题所属的案件类型。案件类型包括房产纠纷、劳动纠纷、债权债务、交通事故、婚姻家庭 5 种,案件类型是领域知识。训练集和测试集的比例为 8∶2。数据集中案件类型的数量统计如表 2 所示。

表 2　法律数据集不同案件类型数量统计

Tab. 2　Statistics on the number of different case types in legal data sets

案件类型	法律咨询问题/个
房产纠纷	2 459
劳动纠纷	4 372
债权债务	3 570
交通事故	4 774
婚姻家庭	3 034

2.1.2　实验设置

实验使用基于 64 位 Linux 操作系统的环境,主机硬件配置如下:显卡为 NVIDIA RTX 4090(由英伟达公司提供);内存为 24 GB(由金士顿科技公司提供)。采用 Elasticsearch(ES)进行稀疏检索;采用 DPR 进行稠密检索。在生成阶段选择 ChatGLM2-6B[28] 和 Baichuan2-7B-chat[29] 大语言模型,初始参数使用 LLM 的默认值。

2.2　基线方法和评估指标

2.2.1　基线方法

本文通过 2 种基线方法验证领域知识和零样本学习在开放域问答任务中的重要性。首先,通过未融入领域知识的 RAG 方法证明领域知识的重要性;其次,通过 LoRA[30] 微调的 RAG 方法证明利用零样本学习的方式能更有效地利用检索文档,从而提升答案的准确性。

1)未融入领域知识的 RAG 方法 在该方法中,使用常规的混合检索,第 1 层检索使用 ES 进行稀疏检索,第 2 层检索使用训练好的 DPR,采用检索文档集前 k(取 $k=3$)个样本与测试集问题组成知识密集型数据集。将这种未融入领域知识的 RAG 方法标记为 no-Dmn,并将其作为基线方法。

2)LoRA 微调的 RAG 方法 在检索过程中将测试集检索改为训练集检索,并利用检索文档和训练集问题对 ChatGLM2-6B、Baichuan2-7B-chat 进行 LoRA 微调。

2.2.2　评估指标

开放域问答任务中,确保答案的准确性和完整性至关重要。本文使用 Rouge-N 指标(具体为 Rouge-1 和 Rouge-2)评估生成文本的准确性,使用 Rouge-L 通过最长公共子序列评估文本的完整性。

2.3　实验设计

通过 5 种实验设计,探究在检索阶段和生成阶段融入领域知识的方法。

1)方法 1　首先,在稀疏检索阶段添加领域知识的联合检索;然后,利用稠密检索作为第 2 层检索;最后,采用检索文档集前 k 个样本和测试集的问题作为一个新的知识密集型数据集。该方法为融入领域知识的稀疏检索方法(标记为 Dmn)。

2)方法 2 和方法 3　首先,使用常规的稀疏检索方法作为第 1 层检索;然后,通过训练好的 DPR 模型对测试集中的问答对进行第 2 层检索。利用领域知识与测试集的问题进行前置或者后置拼接,并与检索文档集的前 k 个样本结合,得到 2 个新的知识密集型数据集。这 2 个方法分别为领域知识前置方法(标记为 ques-front)和领域知识后置方法(标记为 ques-rear)。

3)方法 4 和方法 5　首先,在稀疏检索阶段添加领域知识的联合检索;然后,使用训练好的 DPR 模型在测试集上进行检索,得到检索文档集。将初始测试集中的问题与检索文档集中的前 k 个样本相结合,同时将初始测试集中的领域知识字段添加到问题的开头或结尾,得到 2 个新的知识密集型数据集。这 2 个方法分别为融入领域知识的稀疏检索和领域知识前置方法(标记为 Dmn-ques-front)、融入领域知识的稀疏检索和领域知识后置方法(标记为 Dmn-ques-rear)。

本文对所提出的方法在开放域问答任务上的表现进行了评估。融入领域知识的稀疏检索(Dmn)是为了证明在 RAG 中检索阶段提供的领域文档的重要性;领域知识前置方法(ques-front)、领域知识后置方法(ques-rear)是为了证明 RAG 在生成阶段中,领域知识对 LLM 提高检索文档的领域相关度识别的重要性;融入领域知识的稀疏检索和领域知识前置方法(Dmn-ques-front)、融入领域知识的稀疏检索和领域知识后置方法(Dmn-ques-rear)是为了证明在 RAG 中的 2 个阶段同时融入领域知识对生成结果的影响。以上方法在生成过程中均采用 ZSL 策略,此外,通过将生成阶段的 ZSL 策略替换为 LoRA 微调,验证了在 ZSL 条件下融入领域知识相比 LoRA 微调对生成答案准确性的提升效果更显著(相关源代码已在 GitHub 开源,网址为 https://github.com/zgf1005/dpr-llm.git)。表 3 和表 4 分别展示了大语言模型 ChatGLM2-6B、Baichun2-7B-chat 在 2 个开放域问答任务上的实验结果。

表 3　大语言模型 ChatGLM2-6B 的实验结果

Tab. 3　Experimental results of the large language model ChatGLM2-6B

%

策略	方　法	医疗数据集			法律数据集		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ZSL	no-Dmn	80.77	48.07	70.50	25.99	14.41	21.99
	Dmn	83.73	**49.75**	70.40	28.34	15.54	23.77
	ques-front	83.92	49.60	73.96	29.63	**16.56**	24.07
	ques-rear	83.28	49.08	71.80	27.63	15.56	22.07
	Dmn-ques-front	**84.59**	49.10	**74.82**	**30.76**	14.87	**24.82**
	Dmn-ques-rear	82.86	49.17	71.36	29.88	14.31	22.93
LoRA	no-Dmn	74.80	37.16	39.01	2.27	2.12	5.13
	Dmn	76.10	37.93	40.75	1.64	0.51	2.63
	ques-front	79.20	46.17	54.97	2.81	1.49	5.41
	ques-rear	78.98	40.37	43.90	3.63	2.31	6.93
	Dmn-ques-front	78.00	44.85	49.75	3.26	1.64	6.23
	Dmn-ques-rear	78.72	42.92	42.13	3.46	2.05	6.64

表 4　大语言模型 Baichun2-7B-chat 实验结果

Tab. 4　Experimental results of the large language model Baichun2-7B-chat

%

策略	方　法	医疗数据集			法律数据集		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ZSL	no-Dmn	87.58	55.07	47.48	19.22	6.36	10.19
	Dmn	87.55	55.01	48.10	19.68	7.23	10.30
	ques-front	**88.52**	55.63	**49.45**	**19.73**	6.97	**11.12**
	ques-rear	87.84	**55.85**	49.35	19.37	6.72	10.99
	Dmn-ques-front	87.84	55.09	48.65	19.38	**7.33**	10.90
	Dmn-ques-rear	87.77	55.27	48.52	19.32	7.03	10.23
LoRA	no-Dmn	83.73	43.25	41.42	5.85	2.01	8.00
	Dmn	78.82	40.22	44.51	2.90	0.82	5.51
	ques-front	84.20	46.17	47.97	6.09	1.92	8.38
	ques-rear	80.55	40.44	42.08	5.47	1.94	8.34
	Dmn-ques-front	82.07	40.09	47.37	3.29	1.17	6.24
	Dmn-ques-rear	84.72	42.92	42.13	2.87	0.91	5.48

首先，在 ZSL 方式下，于 LLM 的生成阶段将领域知识融入 RAG 中，其效果优于 LoRa 微调方法。这种优势在 ROUGE-1、ROUGE-2 和 ROUGE-L 等评估指标上表现得尤为突出。

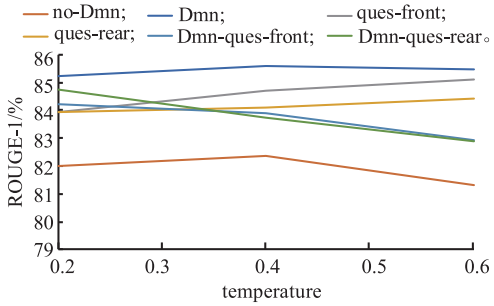
其次，在 ZSL 方式下，融入领域知识的稀疏检索和前置融入领域知识(Dmn-ques-front)，在 ChatGLM2-6B 模型上取得了最佳提升效果，与基线方法(no-Dmn)相比，ROUGE-1 提高了 3.82 个百分点，ROUGE-2 提高了 1.68 个百分点，ROUGE-L 提升了 4.32 个百分点。其他融入了领域知识的方法同样优于基线方法。

综上所述，基于领域知识的 RAG 方法普遍优于基线方法，更适用于开放域问答任务。

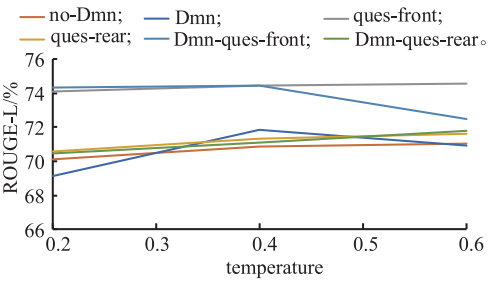
2.4　参数调整实验

为了全面评估本文所提方法的性能，通过调整生成阶段的参数配置，设计了一系列对比实验。在这些实验中，检索阶段的设置与之前保持一致，而生成阶段则针对 ChatGLM2-6B 和 Baichuan2-7B-Chat 模型进行了参数调整，主要涉及 temperature 和 top_k 参数值。通过调整这些参数值，并利用 ROUGE-1 和 ROUGE-L 指标对生成结果进行评估，以系统地分析不同参数配置对生成效果的影响。

1) 在 ChatGLM2-6B 和 Baichuan2-7B-chat 中调整 temperature 参数，当 temperature 值较高时，模型生成的文本会更加多样，但也可能会包含更多不相关或不连贯的内容；而较低的 temperature 值会使得模型更倾向于选择概率最高的词汇，从而生成更一致和连贯的文本，通过实验研究其在生成过程中的影响。如图 2 和图 3 所示，temperature 取值在[0.20，0.60]时，融入领域知识的方法生成的结果在完整性和准确性上优于未融入领域知识的方法。然而，随着 temperature 值的增加，某些方法的精度会有所下降。尽管融入领域知识稀疏检索的方法(Dmn)在 temperature＝0.4 时的 ROUGE-L 效果不及基线方法，但总体而言，其他方法表现优于基线方法。



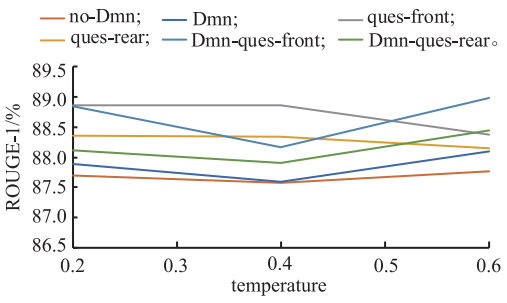
a）对ROUGE-1指标的影响



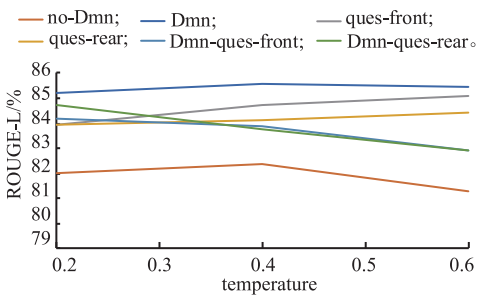
b）对ROUGE-L指标的影响

图 2 参数 temperature 对 ChatGLM2-6B 模型 ROUGE-1 和 ROUGE-L 指标的影响

Fig. 2 Impact of the parameter temperature on ROUGE-1 and ROUGE-L metrics in ChatGLM2-6B



a）对ROUGE-1指标的影响

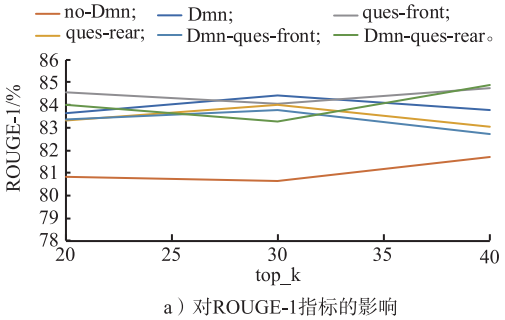


b）对ROUGE-L指标的影响

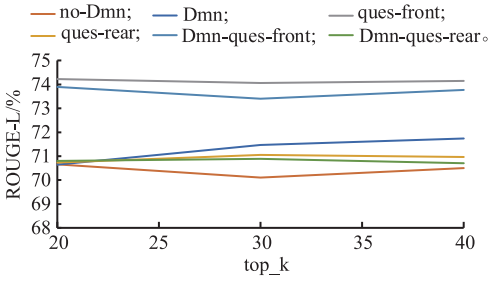
图 3 参数 temperature 对 Baichuan2-7B-chat 模型 ROUGE-1 和 ROUGE-L 指标的影响

Fig. 3 Impact of the parameter temperature on ROUGE-1 and ROUGE-L metrics in Baichuan2-7B-chat

2）在 ChatGLM2-6B 和 Baichuan2-7B-chat 中，通过调整采样参数 top_k 来优化生成效果。top_k 取值越大，生成文本的多样性越强；取值越小，生成文本多样性越弱。由图 4 和图 5 可知，top_k 取值在



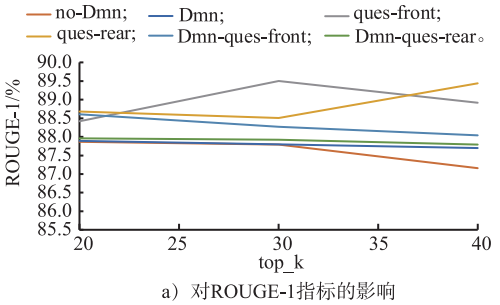
a）对ROUGE-1指标的影响



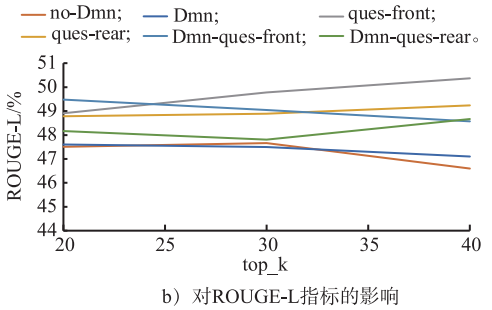
b）对ROUGE-L指标的影响

图 4 参数 top_k 对 ChatGLM2-6B 模型 ROUGE-1 和 ROUGE-L 指标的影响

Fig. 4 Impact of the parameter top_k on ROUGE-1 and ROUGE-L metrics in ChatGLM2-6B



a）对ROUGE-1指标的影响



b）对ROUGE-L指标的影响

图 5 参数 top_k 对 Baichuan2-7B-chat 模型 ROUGE-1 和 ROUGE-L 指标的影响

Fig. 5 Impact of the parameter top_k on ROUGE-1 and ROUGE-L metrics in Baichuan2-7B-chat

[20,40]时，融入领域知识的方法生成结果的完整性和准确性优于未融入领域知识的方法。尽管随着 top_k 值的增加，某些方法的性能会下降，但总体上融入领域知识的方法始终优于未融入领域知识的方法。

2.5 消融实验

为了验证各个模块在本文方法中的贡献，基于领域数据集设计了 9 个消融实验，分别对 RAG 框架的检索阶段中的稀疏检索和稠密检索 2 个模块进行了消融实验，实验设计组合包括：

1）ES 仅采用 ES 检索；

2）ES＋Dmn ES 检索与领域知识的联合检索；

3）ES(ques-front) ES 检索与前置领域知识；

4）ES(ques-rear) ES 检索与后置领域知识；

5）ES＋Dmn(ques-front) ES 与领域知识的联合检索与前置领域知识；

6）ES＋Dmn(ques-rear) ES 与领域知识的联合检索与后置领域知识；

7）DPR 仅采用 DPR 检索；

8）DPR(ques-front) DPR 检索与前置领域知识；

9）DPR(ques-rear) DPR 检索与后置领域知识。

DPR 的训练过程需要使用训练数据集。在 DPR 训练数据集的设计中，正样本为 1 条黄金正样本；使用 RocketQA 的方法，取其他文档的黄金样本作为该文档集的负样本；强负样本为空。通过训练数据集训练好 DPR 检索器后，再进行检索，从检索结果中获取排行较高的前 k 个检索文档作为强负样本，继续训练 DPR，找到一个最优配置，以上操作均基于 DPR 方法展开。生成阶段依旧是在 ChatGLM2-6B 和 Baichuan2-7B-chat 上利用 ZSL 进行实验。使用大语言模型 ChatGLM2-6B 和 Baichuan2-7B-chat 进行消融实验的结果分别如表 5、表 6 所示。

表 5　大语言模型 ChatGLM2-6B 的消融实验结果

Tab. 5　Ablation study results of the large language model ChatGLM2-6B

%

方　法	医疗数据集			法律数据集		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ES	89.93	58.95	46.67	28.04	14.36	22.97
ES＋Dmn	90.10	55.76	47.44	28.60	15.49	24.08
ES(ques-front)	90.06	57.05	50.03	30.14	16.51	24.44
ES(ques-rear)	89.26	56.30	48.66	29.63	15.74	23.50
ES＋Dmn(ques-front)	89.67	57.08	50.37	29.42	15.69	24.45
ES＋Dmn(ques-rear)	89.41	55.50	49.37	30.09	16.87	24.64
DPR	77.14	43.79	60.01	25.88	13.74	19.41
DPR(ques-front)	77.79	44.07	62.06	29.47	16.00	24.33
DPR(ques-rear)	78.05	44.68	60.16	30.19	15.95	23.92

表 6　大语言模型 Baichuan2-7B-chat 的消融实验结果

Tab. 6　Ablation study results of the large language model Baichuan2-7B-chat

%

方　法	医疗数据集			法律数据集		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ES	87.58	55.07	47.48	16.66	5.95	8.81
ES＋Dmn	87.55	55.01	48.10	19.22	7.03	10.12
ES(ques-front)	88.52	55.63	49.45	19.84	7.80	10.16
ES(ques-rear)	87.84	55.85	49.35	19.32	7.49	9.82
ES＋Dmn(ques-front)	87.84	55.09	48.65	19.99	7.64	10.22
ES＋Dmn(ques-rear)	87.77	55.27	48.52	18.30	7.03	9.52
DPR	82.32	49.69	45.04	19.89	8.36	11.28
DPR(ques-front)	82.32	48.07	47.39	19.91	8.56	10.59
DPR(ques-rear)	81.22	47.92	47.47	19.58	7.74	9.92

在消融实验中，ES 系列方法在 2 个模型上均表现优越，特别是将问题前置或后置的策略(ES(ques-front)和 ES(ques-rear))显著提升了性能。ES＋Dmn 方法在 2 个模型上的表现略有提升，但效果有限。由于缺少 ES 提供的联合检索训练数据集系列方法，DPR 系列方法在 2 个模型上的表现均低于 ES，但将问题前置或后置(DPR(ques-front)和 DPR(ques-rear))有助于提高性能。ChatGLM2-6B 和 Baichuan2-7B-Chat 在不同方法上的表现趋势相似，但 ChatGLM2-6B 提升差异更大。这表明，优化检索阶段的方法以及领域知识前置或后置拼接的方法，能够有效提升 LLM 利用检索文档生成答案的准确性。

3 结 语

本文提出了一种基于领域知识的检索增强生成方法，并在医学领域和法律领域开放域问答任务中进行了验证。该方法能够显著增强大语言模型对检索文档中领域知识的挖掘能力，从而提高开放域问答任务中答案的准确性和领域相关度，为特定领域的问答系统提供了一种有效的解决方案。主要结论如下。

1)基于领域知识的检索增强生成方法在生成阶段无需额外训练，可以直接通过零样本学习的方式生成高质量的答案。与微调方法相比，零样本学习方法在本研究中展现出了更好的效果。

2)采用零样本学习方法时，融入领域知识的稀疏检索和领域知识前置方法在 ChatGLM2-6B 上取得了最佳提升效果，与基线方法相比，ROUGE-1、ROUGE-2 和 ROUGE-L 评分分别提高了 3.82、1.68、4.32 个百分点。此外，其他不同的融入领域知识的方法也均优于基线方法，进一步证明了领域知识的重要性。

3)调整不同的大语言模型生成参数(temperature 和 top_k)，对生成结果有一定程度的影响，融入领域知识的方法均优于不融入领域知识的基线方法。

4)融入领域知识的稀疏检索提供的领域数据集对大型语言模型生成答案的准确性影响显著；融入领域知识的稀疏检索和领域知识前置方法能够有效提升大语言模型生成答案的准确性。

本文所提方法虽然在当前数据集上的预测效果有一定的提升，但是由于数据集结构形式的局限，对于其他自然语言处理任务的效果还需要进一步进行验证。未来的工作将聚焦于将本文方法应用到更细粒度(例如段落或句子级别)的数据处理，并探索该方法在其他知识密集型任务中的应用。

参考文献/References：

[1] 杜家驹，叶德铭，孙茂松. 中文开放域问答系统数据增广研究[J]. 中文信息学报，2022，36(11)：121-130.
DU Jiaju，YE Deming，SUN Maosong. Data augmentation in Chinese open-domain question answering[J]. Journal of Chinese Information Processing，2022，36(11)：121-130.

[2] BROWN T B. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems，2020，33：1877-1901.

[3] ZHAO W X，ZHOU Kun，LI Junyi，et al. A survey of large language models[J/OL]. (2024-10-13)[2025-01-19]. https://arxiv.org/abs/2303.18223.

[4] GAO Yunfan，XIONG Yun，GAO Xinyu，et al. Retrieval-augmented generation for large language models：A survey[J/OL]. (2024-03-27)[2025-01-19]. https://arxiv.org/abs/2312.10997.

[5] LEWIS P，PEREZ E，PIKTUS A，et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems，2020，33：9459-9474.

[6] LIU N F，LIN K，HEWITT J，et al. Lost in the middle：How language models use long contexts[J]. Transactions of the Association for Computational Linguistics，2024，12：157-173.

[7] SHI F，CHEN Xinyun，MISRA K，et al. Large language models can be easily distracted by irrelevant context[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu：ICML，2023：31210-31227.

[8] MALLEN A，ASAI A，ZHONG V，et al. When not to trust language models：Investigating effectiveness of parametric and non-parametric memories[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto：ACL，2023：9802-9822.

[9] REN Ruiyang，WANG Yuhao，QU Yingqi，et al. Investigating the factual knowledge boundary of large language models with retrieval augmentation[C]//In Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi：ACL，2025：3697-3715.

[10] CHEN Danqi，FISCH A，WESTON J，et al. Reading wikipedia to answer open-domain questions[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver：ACL，2017：1870-1879.

[11] REN Ruiyang，LYU Shangwen，QU Yingqi，et al. PAIR：Leveraging passage-centric similarity relation for improving dense passage retrieval[C]//Findings of the Association for Computational Linguistics：ACL-IJCNLP 2021. Online：ACL，2021：2173-2183.

[12] ZHANG Hang，GONG Yeyun，SHEN Yelong，et al. Adversarial retriever-ranker for dense text retrieval[J/OL]. (2022-10-30)[2025-01-19]. https://arxiv.org/abs/2110.03611.

[13] ZHU Fengbin，LEI Wenqiang，WANG Chao，et al. Retrieving and reading：A comprehensive survey on open-domain question answering[J/OL]. (2021-05-08)[2025-01-19]. https://arxiv.org/abs/2101.00774.

[14] ZHAO Penghao，ZHANG Hailin，YU Qinhan，et al. Retrieval-augmented generation for AI-generated content：A survey[J/OL]. (2024-06-21)[2025-01-19]. https://arxiv.org/abs/2402.19473.

[15] ASAI A，WU Zeqiu，WANG Yizhong，et al. Self-rag：Learning to retrieve，generate，and critique through self-reflection[J/OL]. (2023-10-17)[2025-01-19]. https://arxiv.org/abs/2310.11511.

[16] LUO Hongyin，CHUANG Y S，GONG Yuan，et al. Sail：Search-augmented instruction learning[J/OL]. (2023-06-25)[2025-01-19]. https://arxiv.org/abs/2305.15225.

[17] YORAN O，WOLFSON T，RAM O，et al. Making retrieval-augmented language models robust to irrelevant contex[J/OL]. (2024-05-05)[2025-01-19]. https://arxiv.org/abs/2310.01558.

(下转第 196 页)