유지보수 처리내역(SR) 기반 RAG 시스템 구축 및 성능평가에 관한 연구

백승윤
숭실대학교

yoon604@soongsil.ac.kr

A Study on Building and Evaluating a RAG System Using Maintenance Service Requests (SR)

Seungyoon Baek
Soongsil Univ.

요 약

본 연구는 유지보수 서비스요청(Service Request, SR) 처리이력(요청내용-처리내용)을 지식베이스로 구축하고, SR 기반 검색증강생성(RAG) 파이프라인의 핵심인 검색(retrieval) 모듈 성능을 정량 평가하였다. SR 을 요청구분/처리유형/요청내용/처리내용/완료여부로 템플릿화한 뒤, BM25, SBERT 기반 밀집검색, 그리고 BM25 상위 100 후보에 대해 BM25 점수와 밀집 유사도를 가중합(Weighted Sum, α=0.98)으로 결합한 하이브리드를 비교하였다(전체 SR 1,140 건, 완료 673 건, 평가 질의 100 건; easy/paraphrase/id_drop). easy 및 id_drop 에서는 BM25 가 Hit@10=1.00 으로 가장 안정적이었으며 하이브리드/선택적 하이브리드도 성능을 유지하였다. paraphrase 에서는 하이브리드가 Hit@10 0.75→0.77, MRR@10 0.619→0.632 로 개선하였고, 특히 저신뢰 구간(bin0, n=33)에서 Hit@1 0.212→0.242, MRR@10 0.358→0.398 로 향상되었다. 또한 BM25 margin 하위 20%에만 하이브리드를 적용하는 신뢰도 라우팅은 전체 성능 저하를 방지하면서 저신뢰 질의의 개선을 가능하게 했다.

I. 서 론

공공·금융 업무 시스템의 유지보수에서는 반복적으로 유사한 SR 이 접수되며, 기존 처리이력의 재활용이 생산성과 품질을 좌우한다. SR 데이터는 자연어 요청과 식별자(기업번호, 화면/전문코드 등)가 혼재되어 검색 난이도가 높다. 본 연구는 SR 처리이력 기반 RAG 지원 모듈을 제안하고, 검색 성능 및 운영 정책(신뢰도 라우팅)을 정량 평가한다.

II. 본론

BM25 는 실무 검색의 강력한 기준선이다[1]. 문장 임베딩 기반 밀집 검색(SBERT[2], DPR[3])은 재서술 질의에서 의미 유사도를 반영할 수 있다. RAG 는 검색 결과를 LLM 입력에 결합하여 생성 품질과 사실성을 개선한다[4]. 본 연구는 하이브리드 검색의 효과가 구간별로 상이할 수 있음을 전제로 한다.

1. 제안 방법

파이프라인은 (1) 문서 템플릿화(필드 결합), (2) 비식별 처리, (3) 희소/밀집 인덱스 구축, (4) 하이브리드 결합, (5) 신뢰도 라우팅으로 구성된다. 하이브리드는 BM25 상위 N(100) 후보에 대해 BM25 점수와 SBERT 코사인 유사도를 정규화한 뒤 가중합(α=0.98)으로 결합한다. 라우팅은 BM25 margin 을 신뢰도로 사용하여 하위 20% 저신뢰 질의에만 하이브리드를 적용한다.

2. 실험 설계

데이터는 SR 1,140 건(완료 673 건)이며, 완료 SR 에서 100 건을 표본 추출하였다. 질의는 (a) 원문(easy), (b) 재서술(paraphrase), (c) 식별자 제거(id_drop)로 구성하였다. 지표는 Hit@K 및 MRR@K(K=1,3,10)이며, 정답 문서는 동일 SR(self-retrieval)로 정의하였다.

[표 1] 데이터 및 평가 설정 요약]

구분	항목	값
데이터	전체 SR	1,140
데이터	완료 SR	673 (59.0%)
평가	질의 표본(완료 SR)	100

3. 결과 및 논의

원문(easy) 및 id_drop 조건에서는 BM25 가 Hit@10=1.00 의 높은 성능을 보여 기준선으로 충분했으며, WSum 하이브리드와 신뢰도 라우팅(Selective Hybrid)도 성능을 유지하였다. 반면 paraphrase 조건에서는 WSum 하이브리드가 BM25 대비 Hit@10 0.75→0.77, MRR@10 0.619→0.632 로 소폭 개선하였다. 개선은 저신뢰 구간(bin0)에서 더 두드러졌으며, 이는 하이브리드의 효과가 질의 난이도(신뢰도)에 따라 달라짐을 시사한다.

[표 2] Variant 별 검색 성능(Hit@K, MRR@10). SelHybrid 는 BM25 margin 기반 신뢰도 라우팅으로 paraphrase 에서 하위 20% 질의에만 하이브리드를 적용한 결과이다.

Variant	방법	Hit@1	Hit@3	Hit@10	MRR@10
easy	BM25	0.980	1.000	1.000	0.990
easy	Dense	0.510	0.600	0.650	0.560
easy	Hybrid	0.980	1.000	1.000	0.990
paraphrase	BM25	0.550	0.650	0.750	0.619
paraphrase	Dense	0.260	0.430	0.540	0.355
paraphrase	Hybrid	0.560	0.670	0.770	0.632
id_drop	BM25	0.970	1.000	1.000	0.985
id_drop	Dense	0.320	0.380	0.420	0.354
id_drop	Hybrid	0.970	1.000	1.000	0.985
paraphrase	SelHybrid	0.560	0.670	0.770	0.632

재서술(paraphrase) 질의에서는 외국어/숫자/식별자 혼입 등 생성 품질 편차가 발생할 수 있으며, 이는 밀집 검색 및 하이브리드 성능 분산과 안정성에 영향을 줄 수 있다. 따라서 실무 적용 시에는 재서술 품질 검증(QC)과 재생성/폴백 정책을 함께 운영하는 것이 바람직하다.

[표 3] paraphrase 저신뢰 구간(bin0, margin 하위 33%(3 분위))에서 BM25 대비 하이브리드(WSum) 개선

구간	n	방법	Hit@1	MRR@10
저신뢰(bin0)	33	BM25	0.212	0.358
저신뢰(bin0)	33	Hybrid	0.242	0.398

III. 결론

운영 적용 시에는 (i) 검색 결과(Top-k SR)의 처리내용을 근거로 요약/가이드를 생성하고, (ii) 비식별 및 정책 검증(민감정보 차단, 근거 미제공 답변 금지), (iii) 로그/피드백 기반 개선을 포함해야 한다. 본 평가는 동일 SR 을 정답으로 하는 self-retrieval 로 실제 '유사하지만 다른 SR'을 찾는 문제를 완전히 대변하지 못한다. 향후에는 유사 SR 라벨링(또는 세미-자동 실버 라벨) 기반 평가와 생성 단계(근거 인용 포함)의 엔드투엔드 평가가 필요하다.

구현 로드맵(요약)은 1) SR 정규화/비식별 배치, 2) BM25+벡터 인덱스 구축, 3) 검색 API 및 라우팅 로직 적용, 4) 오프라인 평가 자동화(variant/구간별 리포트), 5) 파일럿 적용 및 사용자 피드백 수집 순으로 제안한다.

본 연구의 정량 평가는 동일 SR 을 정답으로 정의한 self-retrieval 설정으로, 실제 운영에서 요구되는 "유사하지만 다른 SR" 검색 과제를 완전히 대변하지 못한다. 향후에는 유사 SR 라벨(Top-k 다중정답) 구축 또는 저신뢰 구간 표본에 대한 인적평가를 병행하여 실무 유효성을 보완 검증할 필요가 있다. 또한 재서술(paraphrase)과 같이 표현 변형이 큰 질의에서는 밀집 모델의 도메인 적합도가 성능을 좌우한다. 범용 SBERT 대신 한국어·업무도메인 임베딩 모델로 교체하거나, SR 쌍을 이용한 약식 대조학습[9]을 적용하면 하이브리드의 개선 구간을 확대할 수 있다. 더 나아가 BM25 Top-N 후보에 대해 BERT 계열 재랭커를 적용한 2 단 검색은 의미적 정합성을 높이는 대안이다[5][6].

생성 단계는 처리내용 근거를 요약해 제시하는 수준에서 시작하되, 근거 문장 제공, 민감정보 차단, 근거 미제공 답변 금지 등 운영 통제 규칙을 결합해야 한다. 검색-생성 결합이 지식집약 과제에서 효과적이라는 선행 연구[4][7]를 기반으로, 유지보수 도메인에서도 근거 중심 생성 평가를 포함한 엔드투엔드 검증이 요구된다.

실험 재현성 측면에서, 희소/밀집 및 하이브리드 검색을 동일 프레임워크에서 비교할 수 있는 도구를 활용하면 데이터 갱신 주기마다 자동 리포트를 생성해 운영 의사결정을 지원할 수 있다[8].

추가적으로, 저신뢰 질의만 선별적으로 고비용 모듈(밀집/재랭킹/LLM)로 라우팅하면 지연시간과 비용을 관리할 수 있으며, 본 연구에서 확인한 '저신뢰 구간 개선' 결과와 직접 연결된다. 순위 기반 평가는 nDCG 와 같은 누적 이득 기반 지표로 보완될 수 있으며[10], 신경망 기반 정보검색의 설계·평가 논의는 [11]에 정리되어 있다. 또한 relevance feedback 기반 질의 확장[12]은 식별자 누락(id_drop)과 같이 단서가 부족한 질의에서 추가 개선 여지를 제공할 수 있다.

참 고 문 헌

[1] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, 2009.

[2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," EMNLP-IJCNLP, 2019.

[3] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," EMNLP, 2020.

[4] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.

[5] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," arXiv preprint arXiv:1901.04085, 2019.

[6] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," SIGIR, 2020.

[7] P. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain QA," arXiv, 2020.

[8] J. Lin et al., "Pyserini: Reproducible Sparse and Dense IR Toolkit," SIGIR, 2021.

[9] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," EMNLP, 2021.

[10] K. Järvelin and J. Kekäläinen, "Cumulated Gain-based Evaluation of IR Techniques," ACM TOIS, 2002.

[11] B. Mitra and N. Craswell, "An Introduction to Neural Information Retrieval," Foundations and Trends in Information Retrieval, 2018.

[12] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," Journal of the American Society for Information Science, 1990.