Enhancing Requirements Engineering QA with Retrieval-Augmented Generation and Knowledge Graphs

Darius Sabaliauskas, Jolanta Miliauskaitė

Institute of Data Science and Digital Technologies, Vilnius University, darius.sabaliauskas@mif.stud.vu.lt, jolanta.miliauskaite@mif.vu.lt

ABSTRACT

Automated question answering (QA) systems in requirements engineering (RE) can greatly speed up the processes of specification analysis, validation, and decision-making. While traditional QA models, such as those based on BERT architectures, excel at extracting specific spans of text, they often encounter difficulties with long, fragmented, and domain-specific requirement texts [1-2].

Retrieval-Augmented Generation (RAG) improves the precision of answers by using external knowledge during the inference stage [3]. In this research, we assess five RAG strategies for RE QA, including sparse lexical retrieval with BM25 [4], dense vector-based retrieval using semantic embeddings [2], a hybrid semantic reranking process utilising cross-encoders [1], and graph-enhanced retrieval leveraging concept-based knowledge expansion [5]. we implement a multi-hop retrieval extension that incorporates entity-level reasoning to uncover contextual evidence spanning multiple segments [6]. These methods are designed to alleviate information overload by pinpointing potentially relevant parts of requirement documents, potentially enhancing answer precision and interpretability in subsequent reasoning activities. We assess the models using a domain-specific RE dataset with four standard QA metrics: Exact Match (EM) and F1 score for span-level accuracy, ROUGE-L for lexical similarity, and BERTScore for semantic alignment.

RQ/METHOD/MODEL

RQ1: Do RAG methods improve QA performance in RE?
RQ2: Which retrieval strategy performs best on RE data?
RQ3: Do graph-based and multi-hop retrieval methods improve semantic and contextual accuracy?

Table 1. Retrieval methods applied to two QA models (BERT+LSM, DistilBERT QA)

Method/Models	Description
Baseline	**BERT+LSTM span prediction (no retrieval)**
DistilBERT QA	**Pre-trained distilbert-base-uncased-distilled-squad model for span-based QA (no retrieval)**
BM25 (Lexical RAG)	Lexical retrieval (Top-3)
Dense RAG	Dense semantic retrieval (Top-8)
Semantic Reranking RAG	BM25 → Cross-Encoder reranking
Graph-Enhanced RAG	Concept expansion (dense) + reranking
Multi-Hop	Graph-based hop → Entity hop → Dense → Reranker

BERT+LSTM BASELINE MODEL

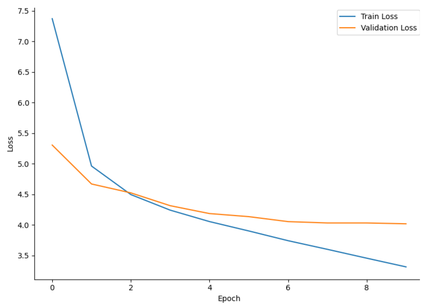

Figure 1. Training vs validation

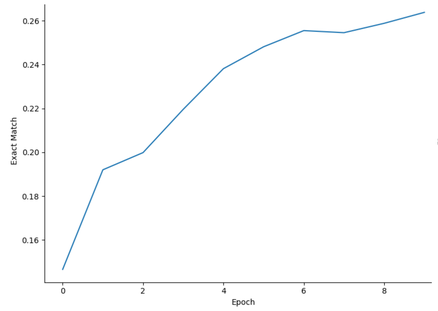

Figure 2. Validation EM per Epoch

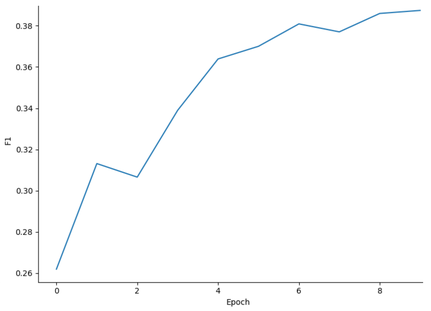

Figure 3. Validation F1 per Epoch

- The baseline QA model combines a frozen BERT encoder with a custom three-layer LSTM network for span prediction.
- BERT is used exclusively for contextual token embedding generation, while a stacked LSTM (1024 → 512 → 512 units) learns to model sequential dependencies, predict start and end positions of the answer span (see Figure 1-3).

RESULTS

Table 2. Results for Experiment 2 (BERT+LSTM + RAG)

Method/Model	F1	EM	ROUGE-L	BERTScore
Baseline – BERT+LSTM	**0.573**	**0.233**	**0.569**	**0.509**
BM25 RAG	0.314	0.133	0.315	0.332
Dense RAG	0.358	0.167	0.355	0.376
Semantic RAG	0.337	0.167	0.344	0.369
Graph-Enhanced RAG	0.335	0.167	0.332	0.365
Multi-Hop RAG	**0.423**	0.2	0.419	**0.424**

Table 3. Results for Experiment 1 (DistilBERT + RAG)

Method/Model	F1	EM	ROUGE-L	BERTScore
Distilbert QA	**0.796029**	**0.333333**	**0.791085**	**0.66833**
BM25 RAG Top-3	0.469373	0.2	0.472882	0.428694
Dense RAG Top-8	**0.742438**	0.366667	0.742255	**0.655656**
Semantic RAG Top-8	0.595246	0.266667	0.60662	0.537645
Graph+Semantic RAG Top-3	0.651106	0.3	0.646772	0.571804
Multi-hop Retrieval	**0.720514**	0.3	0.719078	**0.618059**

Our experiments reveal that Retrieval-Augmented Generation does not reliably surpass a strong BERT+LSTM baseline within the Requirements Engineering domain (see Table 2).

Although certain retrieval approaches—particularly Dense and Multi-Hop retrieval—yield gains in specific scenarios with fragmented or widely distributed information, the overall performance on the complete dataset is frequently similar to, or even worse than, that of the baseline (see Table 3).

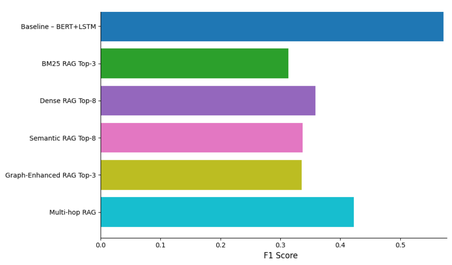

Figure 4. RAG on BERT+LSTM –F1 Score

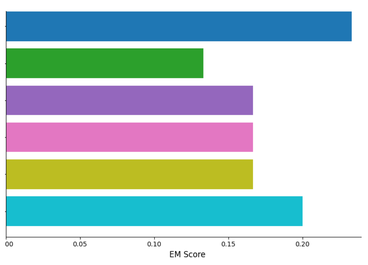

Figure 5. RAG on BERT+LSTM – Exact Match

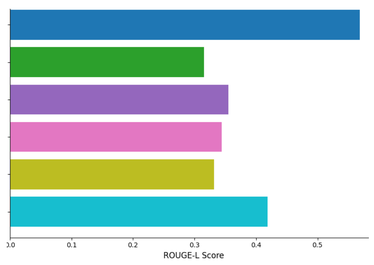

Figure 6. RAG on BERT+LSTM – ROUGE-L

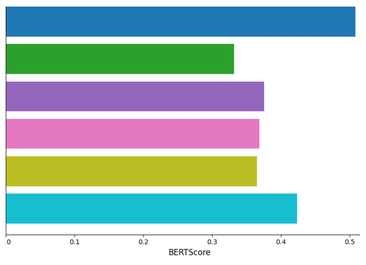

Figure 7. RAG on BERT+LSTM – BERTScore

CONCLUSIONS

- Our experiments indicate that retrieval-augmented methods do not consistently outperform a strong BERT+LSTM baseline or a DistilBERT QA model in the Requirements Engineering domain.
- Of all the approaches assessed, Dense and Multi-Hop retrieval deliver the best overall F1, ROUGE-L, and BERTScore, highlighting their stronger capability to identify semantically relevant evidence across multiple contexts.
- Future work should focus on adaptive RAG strategies, as our findings reveal that more advanced retrieval is not inherently advantageous for domain-specific QA and may even harm performance by adding noise when robust, pertinent context is already present.

REFERENCES

[1] Fu, H., Yao, Y., Lin, J., & Liu, Z. (2023). S-BERT: A semantic information selecting approach for open-domain question answering. arXiv:2305.17413.

[2] Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. ACL.

[3] Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge Intensive NLP Tasks. NeurIPS.

[4] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in IR, 3(4), 333–389.

[5] Xu, X., et al. (2024). GraphTrace: A modular retrieval framework combining knowledge graphs and LLMs for multi-hop question answering. NeurIPS.

[6] Zhao, K., et al. (2024). Advancing Large Language Models with Enhanced RAG: Evidence from Biological UAV Swarm Control. Aerospace Science and Technology.