実用的な品質観点に基づく RAG 性能評価用 QA データセットの自動生成

寺井孝則　田原英一　大石悠河　湯浅晃
株式会社 NTT データグループ
{takanori.terai,eiichi.tahara,yuga.oishi,akira.yuasa}@nttdata.com

概要

RAG システムの評価には, 高品質な QA データセットが不可欠である. しかし, 現行の手動作成方法は多大な時間的・人的コストを要し, また既存の自動生成ツールでは生成されたデータセットの品質に課題があるため実案件での利用が難しいという問題がある. そこで本論文では, RAG の評価データセットに求められる品質観点を整理した上で高品質なデータセットの自動生成手法を提案する. 実験の結果, 一定程度の品質を持つデータセットの生成が可能であり, また従来の実案件データセットと同様に RAG の精度向上を適切に捉えられることを確認した.

1　はじめに

情報検索と生成を統合した RAG（Retrieval Augmented Generation）システムは, 外部データソースを活用しながら正確かつ柔軟な回答を生成できる点で注目を集めており [1], LLM による生成のみではカバーしきれない事実性の担保や, ドメイン知識の補完といった課題に対して有効なアプローチとして知られている [2]. この RAG システムの性能評価においては, 高品質な QA データセットの存在が不可欠であるとされている [3] が, 従来の人手による QA データセット作成手法は以下の課題を抱えている.

作成コストの高さ:　手動による QA データセットの構築は一般的に時間的・人的リソースの消費が大きく, 特に, ドメイン知識を要する実案件では, 専門家の確保や品質を確認のためにさらに多大なコストが必要となる.

ドメイン依存性:　汎用的な性能評価のためには, 複数のドメインにおける多様なデータセットが必要になるが, 人手では作成コストが莫大になるため現実的ではない.

データの更新容易性:　実運用では QA の対象や内容が時間の経過とともに変化するため, 評価データセットも随時更新を行う必要があるが, 人手更新の運用コストは大きな負担となる.

これらの課題を解決するため, 本研究では LLM を用いた高品質な QA データセットの自動生成に取り組む.

2　関連研究

LLM を用いた RAG 評価データセット生成ツールとして, RAGAS フレームワーク [4] 内で提供される TestsetGenerator が知られており, ドキュメントから効率的に QA データセットを生成し, 手動作業を大幅に削減できる点で注目されている.

しかし, TestsetGenerator が用いている QA 生成手法は与えられたドキュメントに対して単純に QA ペアの生成を指示するだけのものとなっており, この手法の主な問題として, 生成されるデータセットにおける回答の正確性を高めるための仕組みが用意されていない. また, 質問者の視点が固定されていないため, 生成される質問が実際のユースケースと乖離する場合があり, 結果として, 実在する QA とは程遠いデータが生成され, データセット全体の品質が低下することが懸念される. さらに, どのような質問を生成するのかという質問タイプの指定がないために, データセットが特定の質問観点に偏ってしまうことも考えられる.

以上の課題に対して, 本研究では, RAG 評価のためのデータセットに求められる品質観点に基づくデータセット生成手法を開発し, 生成したデータセットの品質検証および実際の RAG 評価における有効性検証を実施した.

3 提案手法

3.1 RAG 評価用 QA データセットに求められる品質観点の整理

Xiao らは, RAG システムを評価するための優れた QA データセットとは次の品質観点を満たすたものであると述べている [5]. 以下は Xiao らによる品質観点の定義を著者らによって日本語訳したものである.

現実性（Realism）： データセットは現実のユースケースを反映したものである必要がある.

信頼性（Reliability）： ベンチマークの正答のデータは正確でなければならず, 質問と回答の妥当性が担保されている必要がある.

多様性（Richness）： ベンチマークとなるデータセットは, 様々な質問タイプやユースケースをカバーしている必要がある.

洞察性（Insightfulness）： ベンチマークとなるデータセットは, 種類や難易度への対応などソリューションの持つ性能を多角的に評価できるものである必要がある.

持続性（Longevity）： ベンチマークのシナリオとデータはすぐに陳腐化することなく, 定期的に更新されている必要がある.

本研究では, これらの観点を満たす QA データセットを生成するため, QA ペアデータ生成, 品質判定によるフィルタリング, パラフレーズによるペアの拡張の三段階のプロセスに基づく QA データセット作成を実施し, LLM を用いた QA データセット生成の有効性検証を行った.

3.2 提案手法のプロセス

LLM を活用して QA データセットを効率的に作成するための処理プロセスを提案する. 本手法では, まず入力となる pdf ドキュメントから多様性観点に配慮した QA ペアの生成を行い, その後, 質問者の観点と QA ペアの正当性に基づく品質チェックを行うことで, 高品質なデータセット生成を実現する. 以下に各プロセスについて詳細を述べる.

step1: QA ペアデータ生成

第一ステップでは, LLM を活用して QA ペアデータを生成する. 生成する際に 5W1H の質問観点を用いたオープンクエスチョンや, 「はい」または「いいえ」で回答可能なクローズドクエスチョンのいずれかを生成する指示をプロンプトに与え, 多様性の観点に配慮した質問生成を行う. またここで, 質問者の立場に関する説明文を与え, 実際に想定される質問と回答が生成されやすくすることで, 現実性のより高いデータセット生成を実現する. さらに, 生成する質問は RAG の検索に使用する入力ドキュメントから回答可能なものにプロンプト上で限定することで, 回答がドキュメント内容と整合しないケースや回答不能な質問が生成されるリスクの低減を行う.

step2: 品質判定によるフィルタリング

このステップでは, 第一ステップで生成された QA ペアの品質を LLM に判定させ, データセットとして不適切なものを除するフィルタリング処理を行う. このフィルタリングでは, データセットの現実性と信頼性の向上を目的とし, 質問が想定する質問者からの質問として妥当であるか, 回答が質問に対して正しいものであり, 与えられているドキュメントから回答可能なものであるかを品質判定の要素とした.

step3: パラフレーズによる QA ペアの拡張

フィルタリング処理の終わったデータセットに対して, 各質問のパラフレーズ化によるデータセット拡張を行う. 質問の意図そのものは同じであるが, ドキュメント中での直接的な表現とは言い回しが異なる質問を本処理において生成することで, RAG 検索における表記揺れへの対応という多様性観点を踏まえたデータセットを作成することを目的としている.

4 実験

提案手法を用いた QA データセットの生成および評価を実施し, 提案手法の有効性を検証した.

4.1 実験概要

QA データ生成の対象として, マイナンバーに関連する公開 PDF ドキュメント（合計 130 ページ）を使用し, 想定する質問者を「マイナンバーシステムの利害関係者」と定めた. LLM にはマルチモーダル機能を持つ GPT-4o（モデルバージョン:2024-10-21）を採用し, ドキュメントの各ページを画像形式で入力することで図表情報も用いた QA 生成を行った.

有効性検証では, 提案手法を用いて生成した QA データからランダムに 40 問を抽出したデータセットを作成し, このデータセットを用いて 2 つの評価を実施した. 1 つはデータセット自体に対する品質検証, もう 1 つは RAG 性能評価のための適合性検証

であり, 生成したデータセットを用いて実際の RAG システムの性能評価を行い, その結果が著者らの所属組織における人手で作成されたデータセットを用いた場合と同様の傾向を示すか確認することによって評価した. これらの評価を通じて, 提案手法の有効性を検証した.

4.1.1 品質検証

生成したデータセットの品質検証では, RAG 評価用データセットに求められる観点のうち現実性, 多様性および信頼性について, 著者による主観評価に基づき評価を行った. ここで, 持続性は評価データとして使用する入力文書が適切に更新され陳腐化していないかに関わるものであり, 質問回答の生成プロセスとは関係がないため評価対象外とした. また, 洞察性を評価するためには, 様々な種類や難易度の QA を生成し, システムの応答を洞察可能か評価する必要があるが, 難易度の定義および手法検討は今後の課題としたため, こちらも評価の対象外とした. 今回の評価対象となる各観点の具体的な評価内容は以下の通りである.

- 現実性：質問の視点が想定される質問者の立場に基づき, 実際に発生し得る質問であるかを各 QA データについて二値評価により判定し, その適合割合を現実性の評価とする.
- 信頼性：各 QA データにおいて, 質問に対する回答が正確であり, 入力ドキュメントに基づいて導出可能かを二値評価により判定し, その適合割合を信頼性の評価とする.
- 多様性：各 QA データが 5W1H およびクローズドクエスチョンのいずれに該当するかを分類することで, データセット全体として多様な質問タイプが含まれているかを検証し, 割合をもとに多様性を評価する.

4.1.2 RAG 性能評価のための適合性検証

続いて, 本手法を用いて生成したデータセットが RAG システムの精度評価に適用可能であるかを検証するため従来型 RAG（NaiveRAG）と高度化された RAG（AdvancedRAG）を対象に本データセットを用いた性能評価を実施した. NaiveRAG は単純な検索機能と回答生成機能を持つ基本的な構成である. 一方, AdvancedRAG は RAGFusion[6] および ParentDocumentRetriever[7] といった手法を統合することで, 情報検索および回答生成の精度を向上させている. これら 2 種類の RAG に対し, 提案手法で生成したデータセット（以下, 生成データセット）と, 実際の業務で使用されている人手で作成された評価用データセット（以下, 実案件データセット）を適用し, 両データセットで得られた精度評価結果を比較することで, 生成データセットが実案件データセットと同様に RAG システムの性能を適切に反映できているかを検証した. ここで, 検索結果として取得するドキュメント数は 3 個に限定し, RAG 検索処理を実行した.

評価方法としては, RAGAS を用いた定量評価と人手による主観評価の 2 つを採用した. RAGAS に基づく定量評価では, Context Recall, Context Precision, Answer Semantic Similarity, Answer Correctness の 4 つの指標を用い, 各 RAG システムが生成した回答の精度を測定した. ここで, RAGAS はバージョン v0.2.3 を使用した. 一方, 人手による主観評価では, 各システムの回答内容を目視で詳細に確認し, 質問の意図に適合した回答がドキュメントの内容に基づいて正確に生成されているかについて, 検索結果の網羅性, 生成回答の網羅性および生成内容の事実性（ハルシネーションが起きていないか）の 3 つの観点について, それぞれ二値で評価を実施し, それぞれ検索網羅率, 回答網羅率, 事実性適合率とした. 併せて, 3 項目の基準を全て満たし完全に正しく回答しているものの割合を完全正解率として算出した.

4.2 実験結果

4.2.1 品質検証

品質検証の結果を以下の表 1, 表 2 に示す. まず現実性および信頼性の観点においては, そのどちらを評価する指標でも 85%以上の適合率が得られた. ここで代表的な誤りとしては, 現実性の部分ではマイナンバー制度と直接関係がない QA の生成, 信頼性の部分では図表の読み取りの失敗および回答情報の不足などがあった. また, 多様性の観点においては,「what」の観点に基づく質問が全体の 37.5%と多く含まれてはいるが, それ以外の質問タイプも生成されている様子が確認できた. さらに, 5W1H の How に関しては, 手段に関する質問の他, 件数や人数を尋ねる”How many”や「どの程度増加したか」といった”How much”の質問も含まれており, 幅広い質問形式が含まれていることが確認された.

表 1　データセット品質検証結果（現実性, 信頼性）

評価観点	評価内容	適合率 (%)
現実性	実際に発生し得る質問であるか	87.5
信頼性	質問への回答が正確であるか	85.0

表 2　データセット品質検証結果（多様性）

質問タイプ	該当件数/全件数	割合 (%)
What	15/40	37.5
When	6/40	15.0
Who	3/40	7.5
Why	2/40	5.0
Where	1/40	2.5
How	6/40	15.0
クローズドクエスチョン	2/40	5.0
その他	5/40	12.5

4.2.2　RAG 性能評価のための適合性検証

RAGAS による精度評価結果を図 1 に, 人手評価による評価結果を図 2 に示す.

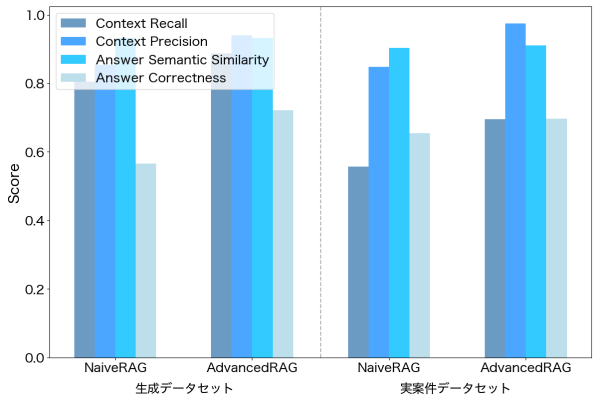

図 1　データセットごとの RAG 性能比較 (RAGA 評価)

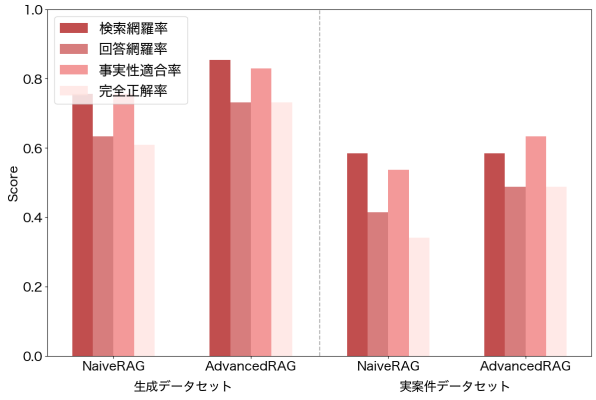

図 2　データセットごとの RAG 性能比較 (人手評価)

NaiveRAG と AdvancedRAG を比較すると, 生成データセット・実案件データセットのいずれにおいても, RAGAS および人手評価で同等の精度向上が確認された. したがって, 生成データセットも実案件データセット同様に RAG システムの精度向上を測定できることが示された.

5　考察

品質検証の結果, 現実性は 87.5%, 信頼性は 85%という適合率が得られ, 本手法を用いて一定の品質を持つデータセットを生成できることが確認された. また, RAGAS と人手評価の双方で, 生成データセットを用いて RAG の性能向上を追跡することが可能であり, RAG の改良前後を相対的に評価する基盤として一定の有用性を持つことが示された. 一方で, 評価指標の絶対値を実際に想定される質問への性能指標として信頼性の高いものとするためには, 現実性のさらなる向上が求められる. RAG システムの改善前後の性能差をより正確に測定するためには, 信頼性の向上も目指していく必要がある. さらに多様性の観点では, 様々な質問タイプを持つデータが提案手法で生成できることが確認できたが, 苦手な質問タイプの分析や, 実運用を想定した割合での評価を行うために, 質問タイプの割合をコントロールできる仕組みの導入が望ましい.

加えて今回の結果では, 生成データセットが実案件データセットに比べ, ほぼ全ての評価指標で高い結果を示した. これは, QA 生成に際してドキュメントをページ単位で処理していたため, 複数ページやドキュメントを統合した複雑な QA を生成できていないことが要因と考えられる. 現在の手法だと生成されている QA ペアは構造的に単純化され, RAG システムの高度な性能を評価するには不十分な可能性があるため, 今後は複数ドキュメントを統合して QA を生成する設計を導入し, 洞察性の観点も抑えた難易度の高いデータセットの構築を目指す.

6　おわりに

LLM を用いた QA データセットの自動生成手法を提案し, 作成されたデータセットの品質と RAG 性能評価のための適合性を検証した. 結果として, 一定程度の品質を持つデータセットを生成でき, 従来の実案件データセットと同様に RAG の精度向上を適切に捉えられることが示唆された. 一方で, 提案手法を用いて生成されたデータセットは, 比較対象とした実案件のデータセットよりも難易度が相対的に低かった点が課題として挙げられる. この原因は, 現在の生成手法が単一ページ内で完結する QA に限定されていることにあり, 複雑な QA を構築するためのさらなる改良が必要である.

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[3] Kunlun Zhu, Yifan Luo, Dingling Xu, Ruobing Wang, Shi Yu, Shuo Wang, Yukun Yan, Zhenghao Liu, Xu Han, Zhiyuan Liu, and Maosong Sun. Rageval: Scenario specific rag evaluation dataset generation framework, 2024.

[4] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.

[5] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. Crag – comprehensive rag benchmark, 2024.

[6] Zackary Rackauckas. Rag-fusion: A new take on retrieval augmented generation. **International Journal on Natural Language Computing**, Vol. 13, No. 1, p. 37–47, February 2024.

[7] LangChain Inc. Parentdocumentretriever, 2023. https://python.langchain.com/api_reference/langchain/retrievers/langchain.retrievers.parent_document_retriever.ParentDocumentRetriever.html.