RAGモデルにおける情報ソース参照範囲の最適化

Optimization of information source reference ranges in RAG models

横井　大将
指導教員　細野　繁

東京工科大学　コンピュータサイエンス学部　コンピュータサイエンス学科
サービスシステムデザイン研究室

本研究では，外部情報を検索し，それを基に応答を生成するRAGモデルが外部情報を検索し応答を生成する際に，参照する情報ソースの範囲を変更することによって，応答の精度と速度のバランスを保つ方法を探る．これにより，RAGモデルの有用性とパフォーマンス向上を目指す．

キーワード：LLM，RAG，情報検索，自然言語処理

１．はじめに

近年，LLM は日常的な対話から専門的なアシスタントまで，様々な領域で活用されており，OpenAI 社の Chat-GPT が代表的である．さらに，外部情報の検索を組み合わせて応答を生成する RAG モデルは，高い柔軟性と効果的な知識利用能力により注目されている．RAG を活用した応答生成システムは，外部の豊富な情報ソースを参照することで，正確で適切な応答を生成する．しかし，参照する情報ソースの違いや分量による応答のドリフトや速度の低下といった課題がある．そのため，精度と速度のバランスを維持した応答の生成が求められる．

２．LLM

LLM（大規模言語モデル）とは，膨大なテキストデータを基に訓練された，自然言語処理モデルである．これらのモデルは多様な文脈やトピックに対する深い理解と生成能力を持っている．LLM は，文章の生成や質問応答，対話システムの構築など，広範囲のタスクに応用されており，人間に近い自然で一貫した応答を提供する．これにより，日常的な対話から専門的なアシスタントまで，多岐にわたるアプリケーションが実現可能となっている．

３．RAG

RAG（Retrieval-Augmented Generation）とは，情報検索（Retrieval）とテキスト生成（Generation）の技術を組み合わせた自然言語処理における手法の一つである．この手法では，外部の情報を検索し，その情報を生成プロセスに統合することで，従来の生成モデルよりも多様かつ詳細な応答が可能になる．RAG では，従来の生成モデルの限界を超えた情報提供を実現するため，外部知識を効果的に活用することを目指している．

４．先行研究

Lewis et al.（2021）の「Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks」[4] では，RAG の基本概念とその優れた性能が示されている．この研究では，情報検索とテキスト生成の技術を組み合わせることにより，知識集約型タスクにおける情報提供の多様性と詳細性が大幅に向上することが示されている．RAG のアプローチにより，データベースの情報更新頻度が低下し，運用コストの削減が実現される．また情報の充実によるハルシネーション（虚偽情報の生成）問題の軽減が期待されている．しかし，参照する情報ソースの違いや情報量による応答のドリフトや生成速度の低下といった課題も指摘されている．これらの課題に対処するためには，情報検索範囲の最適化が重要である．

５．研究目的

本研究の目的は，RAG が参照する情報ソースの範囲変更により，応答の精度と速度のバランスを保つ方法を探ることである．情報ソースの検索範囲を段階的に変更し，その影響の評価により最適なバランスの発見を行う．これにより，RAG モデルにおける情報ソース参照範囲の最適化を目指す．

６．提案手法

本研究では，RAG モデルにおいて参照する情報ソースの変更が応答の精度と速度に与える影響を評価し，検索範囲を最適化するために以下の手法を提案する．

６．１　RAGの流れ

本研究で実装する RAG モデルにおける処理の流れについて，図 1 に示す．

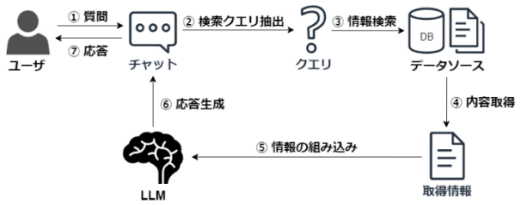

図 1　RAG モデルにおける処理の流れ

RAG を用いた応答生成は全 7 工程で行われる．1 では，ユーザがテキストベースでチャットなどに質問を送信する．2 では，ユーザの質問から検索クエリを抽出する．3 では，抽出した検索クエリをもとに適切な情報ソースから情報検索を行う．4 では，データソースからユーザの質問に関連する情報を取得する．5 では，取得した情報を LLM に組み込む．6 では，LLM は取得した情報を活用して応答生成を行う．7 では，生成された応答をユーザに返す．このプロセスによって，外部情報を効果的に活用した応答の生成が可能となる．

6．2　情報ソースの分類と選定

情報ソースの選定は，RAG システムのパフォーマンス向上において重要なプロセスである．対象の領域に応じて，信頼性が高く最新の情報を提供する情報ソースの選定を行う必要がある．情報ソースを一般的情報源と専門的情報源に分類する．一般的情報源は，広範囲の知識を提供する情報源である．例として，Wikipedia や百科事典などが挙げられる．これらの情報源は，多様なトピックに関する基本的な情報を提供するため，応答の基礎となる情報として有用である．専門的情報源は，特定の分野の知識を提供する情報源である．例として，Google Scholar や arXiv などが挙げられる．これらの情報源は，特定の分野やトピックに特化した情報を提供するため，特定の専門知識が必要な応答を生成する際に有用である．これらの要素を考慮し，RAG システムにおいて情報ソースの選定と分類を行うことで，質問に沿った応答生成が可能となる．これらを踏まえて情報ソースの選定を行う仕組みを形成することにより，RAG モデルのパフォーマンス向上を目指す．

6．3　参照範囲の設定

外部情報の参照範囲設定は，応答の精度と速度に直接影響を及ぼす．情報ソースのサイズやトピックの広さに応じて検索範囲の設定を行う．範囲を広げることで情報の網羅性は向上するが，処理データ増加に伴う応答速度の低下に繋がる可能性があるため，バランスを考慮して範囲設定を行う必要がある．範囲設定は段階的に行うものとし，各ステップにおいて精度と速度を評価する．検索範囲の最適化後に再度評価を行い，範囲設定が適切であるか確認を行う．

6．4　評価指標

精度の評価では，RAG パイプライン全体に対して信頼性，適切さ，総合的なパフォーマンスなどを評価する．以下に本研究で使用する，Ragas の評価指標を図 2 に示す[3]．

評価指標	説明
忠実性	生成された回答とコンテキストの一貫性
回答の関連性	質問に対する回答の適切性
コンテキスト精度	参照した情報の正確性
コンテキストリコール	情報の網羅性
意味的類似性	応答と正解の意味的類似性
回答の正確さ	情報の正確性

図 2　使用する Ragas の評価指標

6．5　範囲設定の影響評価

情報ソースの検索範囲変更が応答の精度と速度に与える影響を評価するため，最初にモデルの精度評価を行う。先に挙げた評価指標を用いて定量的な評価を行い，範囲変更前後における応答の正確性を比較する．次に，応答生成の速度を測定し，検索処理や応答時間の変化を分析する．これにより，速度の低下または向上の定量的な評価を行う．最後に，精度と速度のバランスと取るためトレードオフ分析を行い，最適なバランス点を特定する．

7．　これまでの調査

これまでの調査では，Web ページから必要な情報を抽出する BeautifulSoup や RAG モデルの評価を行うフレームワークの Ragas について調査を行った．また Wikipedia の情報を基に，ユーザの質問に対して応答生成を行う RAG を用いた LLM の作成を行った．

8．　今後の展望

本研究では，RAG モデルの参照する情報ソース範囲の最適化により RAG モデルの有用性とパフォーマンス向上を図った．今後の展望として，情報ソースの多様化や検索範囲の動的な調整を行うアルゴリズムの開発を進める．これにより，RAG モデルの実用性とパフォーマンスのさらなる向上を目指す．

参考文献

[1] 大規模言語モデル（LLM）の概要について，https://book.st-hakky.com/data-science/llms-overview/，2024 年 5 月 9 日閲覧
[2] Welcome to Kedro’s award-winning documentation! https://docs.kedro.org/en/stable/，2024 年 6 月 13 日閲覧
[3] Ragas，https://docs.ragas.io/en/stable/，2024 年 6 月 27 日閲覧
[4] Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2021)
[5] Shahul Es et al. RAGAS:Automated Evaluation of Re-trieval Augmented Generation (2023)