製造業で取り扱う実データを対象としたRAGの改善

柴田健吾[1] 梶田久貴[1] 杉本崚[1] 森田克明[1]
[1]三菱重工業株式会社 {kengo.shibata.ku, hisataka.kajita.u4, ryo.sugimoto.ne, katsuaki.morita.py}@mhi.com

概要

企業における大規模言語モデルの活用が進む中、社内固有の知識を大規模言語モデルと結び付けるRetrieval Augmented Generation（RAG）は重要な技術である。多くの企業でRAGの改善が試みられているが、製造業でRAGを活用する場合、使用されるドキュメント群の特殊性から、一般的なデータセットで検証されたRAG改善手法が常に効果を発揮するとは限らない。本研究では、RAGの性能を改善するとされる要素（①テキストの正規化、②ハイブリッド検索、③リランキング、④検索ワードの拡張）について、製造業の設計現場で扱われるドキュメント群に対して、どの要素が効果的であるかを調査した。その結果、テキスト正規化、ハイブリッド検索、リランキングが検索性能向上に寄与した一方、検索ワードの拡張は検索性能向上に寄与しなかった。

1 はじめに

大規模言語モデル（Large Language Model: LLM）を企業の業務に活用する上で、LLMに社内固有の知識に基づいた応答を求めるニーズは高い。そのため、検索とLLMによる文章生成を組み合わせたRetrieval Augmented Generation（RAG）は重要な技術である。昨今、RAGの性能を向上させるための様々な研究が注目を集めている。

企業で扱う業務文書は、業界や用途によって分量や文章構造、表現形式が大きく異なる。そのため、RAGを実装する際は、実際に取り扱うドキュメント群の特性ごとにRAGの仕様を検討する必要がある。当社のような製造業（特に多品種少量生産型）の設計現場で扱うドキュメント群には以下の特徴がある。

1) 数千～数万規模の大量の文書ファイルを扱う。
2) 独自フォーマットのPDF, Wordファイルが多く、テキストを取り出す際にノイズが含まれやすい。
3) 専門用語や企業内独自の略語が多く含まれる。
4) 簡略化した表現が多い（例：「保守時は摩耗量が6 mm以内であることを確認する」ではなく「判定基準：摩耗6 mm以内」）。

このような特徴を持つドキュメント群に対し、単純な埋め込みベクトル検索によるRAG（文献[1]の”Naive RAG”と同じ構成）を実装しても、質問に対する回答根拠を含むコンテクストが検索でヒットしないことが多く、実用に十分な性能を発揮できない。この要因を前述の製造業におけるドキュメント群の特徴と対応づけると、1) 検索母数が非常に大きいこと、2) ノイズが多いこと、3) 専門用語が含まれること、4) ユーザークエリに含まれるトークンが引用すべきコンテクストには含まれないこと、などが挙げられる。

近年、RAGの検索性能を改善する様々な手法が提案されており、適切に取り入れることでこれらの課題を解決できる可能性がある。一方で、取り扱うドキュメント群の特性によって効果の程度は変わり得るため、各手法の適切な取捨選択が求められる。特に、製造業の設計現場で用いられるドキュメント群に対し、どのような手法が効果的であるかは十分に検証されていない。そこで本研究では、代表的なRAG改善手法である①テキストの正規化、②ハイブリッド検索、③リランキング、④検索ワードの拡張の４点について、業務における実文書データを対象とした実験を行い、各手法がRAGの検索性能を向上させることができるかを検証した。

2 関連研究

これまでにRAGの性能を向上させるための様々な手法が提案されている。検索エンジンの高度化という観点では、複数の検索アルゴリズムを組み合わせて検索を行うハイブリッド検索が挙げられる。具体的には、BM25 [2]に加え、BERT [3]などのテキスト埋め込みモデルによるベクトル検索を同時に行う方法がある [4]。ハイブリッド検索ではBM25によるキーワードの一致と、テキスト埋め込みモデルによる意味的な類似性の両方の観点からクエリとコンテクストの類似度を評価できるため、多様なクエリに対応可能な検索システムを実現することが期待さ

れる。他には、一度取得した検索結果を再度順位付けするリランキング [5]が提案されている。リランキングを行うことで検索プロセスが多段化されるため、低速だが高精度の意味的類似度評価手法（例：クロスエンコーダ型の意味的類似度評価モデル [6]）を活用できるなど検索の自由度が向上する。

検索エンジンを改善する方法の他に、検索ワードを最適化する方法も提案されている。通常の RAG ではユーザーが入力したクエリを検索ワードとするが、より検索ワードとしてふさわしい形式に変換することで、検索性能が向上する可能性がある。近年では、LLM を用いて検索ワードを拡張する方法が提案されている。例えば、LLM にユーザーが入力したクエリの言い換えをさせ、多様な表現で検索を行うクエリ拡張 [7]が提案されている。他には、ユーザークエリに対する仮想的な回答（間違っていてもよい）を LLM に回答させ、ユーザークエリとその仮想的な回答のセットを検索ワードにする手法である Hypothetical Document Embeddings（HyDE） [8]が提案されている。

本研究では、上記で列挙したそれぞれの手法について、製造業の設計現場で扱うドキュメント群に対して効果を発揮できるかという観点で、各手法の有効性評価を行う。

3　検索システム

3.1 データの前処理

PDF や Word 形式の生データからプレーンテキストを取り出し、検索時にスコア付けが可能な形式（ベクトルデータ等）に変換する処理をインデックス化と呼ぶ。インデックス化ステップでは、最初に LangChain [9]のファイルローダーを用いて生データからプレーンテキストを抽出した。次に、半角・全角の違いや Unicode の違いを統一する処理（NFKC 正規化）を行った。さらに、テキストに含まれるノイズ（空白、改行、タブ、縦棒、アスタリスク）をすべて半角スペースに置換した。本研究では NFKC 正規化とノイズ除去を合わせて「テキスト正規化」と呼ぶこととし、この有無による検索精度の違いを検証する。

テキスト正規化後、テキストを検索やプロンプト入力で処理する単位（以下チャンクと表記）に分割する処理を行った。本研究では、Microsoft の調査 [4]を参考に、チャンクサイズを 500 トークン、オーバーラップを 100 トークンとしてチャンク分割を行った。

最後に、チャンク単位でベクトルインデックスデータ、及び、BM25 インデックスデータを作成した。ベクトルインデックスデータの作成には、漢字に強い埋め込みモデルである bge-m3 [10]を用いた。

3.2 チャンクの検索手法

検索システムの全体像を図 1 に示す。検索ステップでは二段階のステップでチャンクの絞り込みを行う。一段階目は、埋め込みモデルによる検索と BM25 検索を組み合わせたハイブリッド検索で、ユーザークエリとの類似度が高い上位 100 件のチャンクに絞り込む。埋め込みモデルによる検索の結果と BM25 検索の結果に対して Reciprocal Rank Fusion (RRF) [11]スコアを算出することで合算し、総合的な順位を決定した。チャンクの RRF スコアS_{RRF}は、埋め込

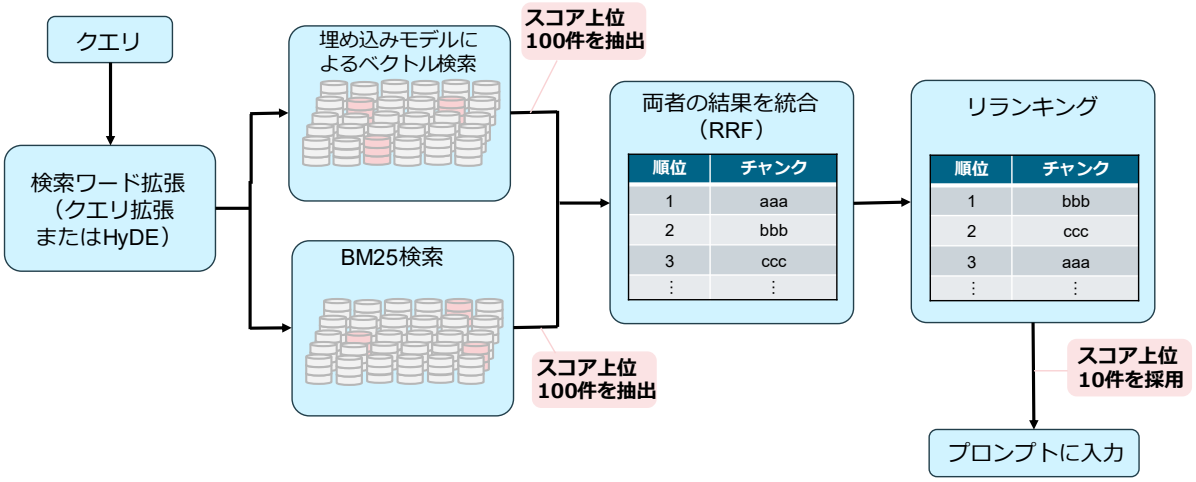

図 1 検索の全体像

みモデルによる検索の順位r_{EMB}と BM25 検索の順位r_{BM25}を用いて

$$S_{\mathrm{RRF}} = \frac{1}{r_{\mathrm{EMB}} + 60} + \frac{1}{r_{\mathrm{BM25}} + 60} \quad (1)$$

と算出される。二段階目では、一段階目で取得した 100 件のチャンクに対し、クロスエンコーダ型意味的類似度計算モデルを用いて再度類似度の順位付けを行う（リランキング）。リランキングモデルには、埋め込みモデルと同じ系列であり漢字に強い bge-m3-reranker [10]を用いた。リランキングモデルが出力したスコアが高い上位 10 件のチャンクを最終的な検索結果とする。高速な埋め込みモデルによる検索・BM25 検索と、低速だが高精度な意味的類似度の評価が可能なリランキングモデルを組み合わせた二段構成とすることで、検索スピードと性能を両立させる。

検索を行う前に LLM によって検索ワード（通常はユーザークエリ）を拡張する処理（クエリ拡張および HyDE）を入れたケースも試行した。クエリ拡張では、LLM を用いて検索ワードを多様な表現に変換することで、表現のゆらぎに強い検索を行えることが期待できる。HyDE では、LLM が生成したクエリに対する仮の回答を検索ワードに含めることで、より質問に対する回答に近い文脈を探し出すことが期待される。いずれも前述の課題 4)を解決することを狙ったものである。

3.3 RAG による回答生成

回答生成時は、ユーザークエリ、参照コンテクスト、指示文（「以下の参考資料をもとに質問に回答してください」等の単純なもの）をプロンプトに入力する。参考コンテクストは、検索で取得した 10 件のチャンクをリランキング後の順位のまま与える。付録 A の図 2 に回答生成の画面イメージを示す。

4　評価実験

4.1 データセット

本研究では当社グループ会社である三菱重工機械システム株式会社の協力のもと、油圧機械、加速器、精密メカニクス製品に関する、客先仕様書や機器の取扱説明書、見積仕様書、社内技術資料などを対象としたドキュメント（ファイル数は約 15,500 件、総トークン数は約 1 億）を用意した。これらのドキュメントは、日々の業務で実際に参照される実データであり、製造業の設計現場で使用されるドキュメントとして典型的な前述の特徴 1) – 4) を含む。

4.2 評価用データセットの構築

RAG の精度を評価するためのテスト用問題を作成した。実務担当者の協力のもと、実務で想定される closed QA 形式の問題を 58 問用意した。問題の例を表 1 に示す。

表 1 テスト用問題の例

問題	模範解答
エアヒータの熱間試運転時、排ガス温度は何℃以下とする必要があるか？	燃焼排ガス入口温度が 450℃、かつ燃焼排ガス入口温度と燃焼用空気出口温度の算術平均値が 400℃以下
エアヒータロータ本体の保守管理基準において、腐食・減肉の判定基準は？	・溶接部 突合せ：板厚に準じる すみ肉（脚長に対し） 脚長 9㎜ 以下：30%以内 脚長 9㎜ 以上：3㎜ 以内 ・母材：40%以内

4.3 評価指標

RAG の評価では、一般に検索パート（クエリに対する回答が含まれるチャンクが得られるか）と生成パート（LLM が生成した内容は適切か）に分けて評価を行う。本論文では、生成の前提となる検索部分の改善に着目していることから、検索パートの評価を実施する。

検索パートでは、テスト用問題 58 ケースのうち、問題の回答に相当する情報が含まれるチャンクが上位 10 件以内にヒットしたケースの割合（以下 Hit Rate@10 と表記）を評価指標とした。問題の回答に相当する情報が複数チャンクに存在し、どれか一つを参照すれば良い場合は、そのいずれかのチャンクが上位 10 件に含まれればよいとした。

5　実験結果

テキスト正規化の有無、検索方法（埋め込みモデルによる検索のみまたはハイブリッド検索）、リラ

ンキングの有無、クエリ拡張の有無、HyDE の有無を変化させた全 6 ケースについてテスト用問題の検索性能を評価した。

検索パートについて、評価結果を表 2 に示す。埋め込みモデルによる検索だけを用いた単純な RAG をベースラインとし、RAG 改善手法を一つ以上取り入れたものを手法 1 - 5 としている。

表 2 検索の評価結果

テストケース	テキスト正規化	ハイブリッド検索	リランキング	クエリ拡張	HyDE	Hit Rate @10
ベースライン	–	–	–	–	–	40%
手法 1	✓	–	–	–	–	71%
手法 2	✓	✓	–	–	–	74%
手法 3	✓	✓	✓	–	–	87%
手法 4	✓	✓	–	✓	–	71%
手法 5	✓	✓	–	–	✓	66%

テキスト正規化、ハイブリッド検索、リランキングは検索性能の向上に寄与した。一方でクエリ拡張および HyDE は検索性能の向上に寄与しなかった。

テキスト正規化は、前述した課題「2) ノイズが多い」の解決に対応し、検索の改善に寄与したと考えられる。ハイブリッド検索は「3) 専門用語が含まれること」の緩和に一定程度寄与したと考えられる。これは、BM25 検索が加わったことで特定のキーワードが含まれたチャンクが高く評価され、言葉の意味からチャンクを評価する埋め込みモデルによる検索単体よりも、専門用語を含む検索に強くなったためと考えられる。リランキングは「1) 検索母数が非常に大きい」という課題に対して有効であったと考えられる。真に重要なチャンクの特定はリランキングに任せ、埋め込みモデルによる検索および BM25 検索ではチャンクを広く取得することで、検索母数の拡大に伴う検索性能の低下を軽減できると考えられる。

LLM による検索ワード拡張手法であるクエリ拡張および HyDE は「4) ユーザークエリに含まれるトークンが引用すべきコンテクストには含まれない」に対する解決策となることが期待されたが、検索性能向上には寄与しなかった。

クエリ拡張は LLM がユーザークエリを言い換えるため、一般的知識に基づいた言い換え表現にとどまり、社内固有の風習・暗黙知に基づく言い換えができなかったことが性能向上に寄与しなかった要因と考えられる。例えば、製品の設計を変更した時期を調べる際は「設計変更書の発行年月日」を見る必要があるが、「設計変更の時期を教えてください」というクエリから「設計変更書の発行年月日はいつですか」という表現を導き出すことはできなかった。このような言い換えを実現するにはクエリ拡張をする LLM をファインチューニングするなどの工夫が必要になると考えられる。

HyDE はクエリに対する仮の回答が全く関係ないものになり、真に必要なチャンクの類似度評価を高めることができなかった。HyDE が検索性能改善に寄与するためには、仮の回答の時点である程度は関連性のある回答を生成する必要があるが、実務で扱うような専門性が高いクエリに対しては、文脈レベルで異なる回答を生成するケースが散見された。

6 おわりに

本研究では、製造業の設計現場で取り扱われるドキュメント群に対し、有効な RAG 改善手法（特に検索部分）を明らかにした。テキスト正規化、ハイブリッド検索、リランキングが検索性能向上に有効であることが示された。一方で、今回扱ったような専門性の高いドキュメント群に対しては、クエリ拡張や HyDE といった検索ワードを拡大する方法は有効ではなかった。

今後、企業における RAG 活用を推進する上で、文書には書かれていない社内知識・暗黙知をどのように咀嚼して LLM と接続させるかが課題である。

謝辞

本研究の推進にあたり、株式会社 Lightblue に協力をいただきました。また、本原稿の執筆にあたり東京科学大学の荒瀬由紀教授に指導を賜りました。ここに感謝を申し上げます。

参考文献

1. Wang, Gao and Yun Xiong and Xinyu Gao and Kangxiang Jia and Jinliu Pan and Yuxi Bi and Yi Dai and Jiawei Sun and Qianyu Guo and Meng Wang and Haofen Yunfan. Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv:2312.10997, 2023
2. Zaragoza, Robertson and Hugo, Stephen. The Probabilistic Relevance Framework: BM25 and Beyond, Now Publishers Inc, Trends Inf. Retr. 3, 4, 2009
3. Jacob Devlin, Chang, Kenton Lee, and Kristina Toutanova, Ming-Wei. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Association for Computational Linguistics, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2023
4. Microsoft, Azure AI Search: Outperforming vector search with hybrid retrieval and reranking, https://techcommunity.microsoft.com/blog/azure-ai-services-blog/azure-ai-search-outperforming-vector-search-with-hybrid-retrieval-and-reranking/3929167, 2023
5. Michael Glass, Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, Alfio Gliozzo, Gaetano. Re2G: Retrieve, Rerank, Generate, Association for Computational Linguistics, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2701–2715, 2022
6. Gurevych, Reimers and Iryna, Nils. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Association for Computational Linguistics, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, 2019
7. Bendersky, Jagerman and Honglei Zhuang and Zhen Qin and Xuanhui Wang and Michael, Rolf. Query Expansion by Prompting Large Language Models, arXiv:2305.03653, 2023
8. Luyu Gao, Ma, Jimmy Lin, Jamie Callan, Xueguang. Precise Zero-Shot Dense Retrieval without Relevance Labels, Association for Computational Linguistics, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023
9. GitHub - langchain, https://github.com/langchain-ai/langchain
10. Jianlyu Chen, Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu, Shitao. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, Association for Computational Linguistics, Findings of the Association for Computational Linguistics: ACL 2024, pages 2318–233, 2024
11. Buttcher, V. Cormack and Charles L. A. Clarke and Stefan, Gordon. Reciprocal rank fusion outperforms condorcet and individual rank learning methods, Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009
12. Mikolov, Corrado, G.s and Chen, Kai and Dean, Jeffrey, Tomas. Efficient Estimation of Word Representations in Vector Space, International Conference on Learning Representations, 2013
13. Cho, Nogueira and Kyunghyun, Rodrigo. Passage Re-ranking with BERT, arXiv:1901.04085, 2019

A 付録

本研究の実施と並行して、RAG を活用した社内向けチャットボットアプリケーションを試作した。画面イメージを図 2 に示す。画面中央部の赤いアイコンがついたテキストがユーザーが入力したクエリであり、オレンジのアイコンのテキストが LLM が生成した回答である。画面右のテキストは、LLM が回答を生成する際に参照したチャンクのテキストを表す（画面の設定では 10 チャンク）。

図 2 RAG を活用したチャットボットアプリケーションの画面イメージ