J-033

行政 DX のための検索拡張生成システムの開発
Development of RAG System for Digital Transformation of Local Government

川島 壮生† 白松 俊† 水本 武志‡
Soki Kawashima Shun Shiramatu Takeshi Mizumoto

1. はじめに

現代の行政は複雑かつ迅速な意思決定を要求される場であり，情報アクセスの効率化は不可欠である．愛知県庁都市交通局航空空港課では，管理・運営に関わる業務のルールや法規・規定等の文書集は 2TB 以上のデータ量で蓄積されている．県幹部職員に提出する資料作成や，民間事業者等からの問い合わせへの回答作成などにおいては，過去の事例などを参照する必要がある．現状では OS のキーワード検索機能や，フォルダ名やファイル名から当たりをつけてファイルを一つずつ閲覧していくなどの手段を使っている．こういった既存の情報検索手法では蓄積された文書から適切な情報を見つけ出すのに多大な時間と労力が必要である．特に文書ファイルの保存形態が統一されていない現状では，職員個人の経験や知識に依存した非効率的な作業となりがちである．そこで RAG (Retrieval Augmented Generation) [1]を用いて，情報検索及び回答作成システムを開発した．

数年ごとに部署を異動する愛知県庁職員にとって，過去の経緯や事例についての情報検索は特に大きな負担となっている．RAG が過去の経緯や事例に関する時系列を加味した出力を行うためには，日付などの時系列情報を文書から取得する必要がある．しかし，RAG システムは文書全体を一括で検索するのではなく，文書を分割したテキストセグメントごとに検索する．そのため，日付情報とユーザのクエリに関連する情報が文書内において離れて記述されているとき，日付情報が検索されないことがある．そこで文書全体から時系列情報を事前に抽出し，プロンプトに追加する手法を実装した．

本研究は AichiXTech という愛知県と企業等との連携による DX 推進事業の一環として実施されており，県庁内の各所属が抱える行政課題に対し，ICT を活用して解決することを目指す．[2]

2. 関連研究

RAG は，外部情報を検索することで，LLM に独自の文書を参照した回答を生成させることができる．一般的な RAG の手法は，入力クエリの埋め込みベクトルと類似度の高い外部情報を抽出し，それを LLM に与えるプロンプトに追加するものである．ゼロショットで外部の情報にアクセスできる優位性から法学[3]，農学[4]，数学[5]や物理学[6]など，多様な分野での応用が期待されている．本研究では，地方行政における RAG の有効性を検証する．

RAG の性能評価方法に関する研究も行われており，RAGS[7]や ARES[8]などの自動評価フレームワークが発表されている．本研究では，被験者に問題を解いてもらうことで，実際の行政業務に似た状況での評価を行った．

†名古屋工業大学 Nagoya Institute of Technology
‡ハイラブル株式会社 Hylable Inc.

3. 提案手法

3.1 システムの概要

RAG はまず，入力されたクエリに基づいて，関連性の高い文書をデータベースから検索する．ベクトルデータベースには Pinecone System 社が提供する Pinecone を用い，コサイン類似度に基づく近似最近傍探索によって最も類似度の高い文書を特定する．検索された文書に基づくプロンプトを使用して，LLM がクエリに対する回答を生成する．システム構成図を図 1 に示す．埋め込みベクトルモデルには OpenAI の text-embedding-ada-002 を用い，LLM には OpenAI の gpt-4-1106-preview を使用した．また分割文字数に対して 20%のオーバーラップを設けた．

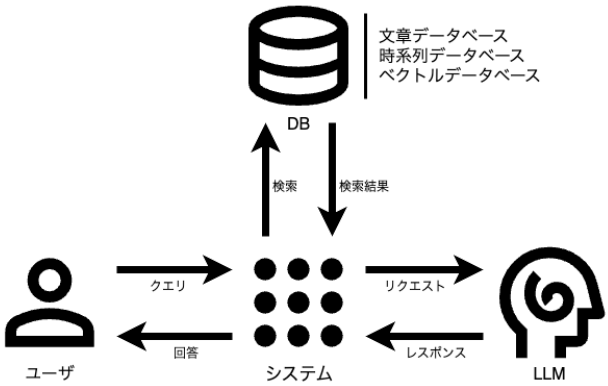

図 1 ：システム構成図

3.1.1 時系列情報

文書全体から時系列に関する情報を LLM で事前に抽出する．その際のプロンプトを図 2 に示す．RAG では一つの文書を複数セグメントに分割し検索を行う．文書内の 1 セグメント以上が検索されたとき，その文書から事前に抽出された日付情報を全てプロンプトに追加する．これによっ

```
あなたは空港航空課の行政ファイルから重要な日付に関する情報を抽出してタグ付けを行うスペシャリストです。
タグ付けを出力する際には、以下のフォーマットで回答する必要があります。
[{"year": "書類内容に関する西暦年",
"month": "書類内容に関する月",
"day": 書類内容に関する日",
"annotation": "日付の説明"}]
以下のテキストに対してタグ付けしてください。
【ファイルの内容が入る。】
```

図 2 ：時系列情報抽出プロンプトの一部

て未検索のテキストセグメントに記述された時系列情報を考慮することができる．

3.2 回答生成

ユーザのクエリから作成された埋め込みベクトルを用いてベクトルデータベース内でコサイン類似度に基づく類似ベクトルを検索する．検索されたセグメントと，そのセグメントが含まれる文書から事前に抽出した時系列情報を，類似度の高いものからプロンプトに加える．このようなプロセスにより生成されたプロンプトの例を図３に示す．

```
# 命令
あなたは選ばれたファイルを参照して、クエリに回答するスペシャリストです
クエリに的確に回答してください。HTML に挿入するので HTML タグを用いてください。
- 絶対に HTML タグを用いて内容にメリハリをつけること
- 過去の意思決定について問われた場合は、その経緯や根拠、因果関係が分かるように説明すること
- 過去の意思決定について賛否があった場合は、その賛否の意見と根拠を示すこと
- 意見を提示する際は、誰のどの立場の意見であるかを明示すること
- ファイルの絶対パスや、文脈を根拠に、時系列を認識すること
- 具体的な事例については、分かる限り時期を明示すること
- ファイル内容からの文脈を広く参照すること
- 参照元のファイル NO を（*1) のように注釈として明示すること
- 日本語で出力すること
- markdown 記法を使用しないこと
- 表で示せるものは HTML の表で出力すること
- HTML タグは h2, h3, p, ul, ol, table, td, th のみ使用してください。
# クエリ
## 状況
ミュージアム〇周年のセレモニー内容を考えています.
## 質問・命令
ミュージアム 100 万人達成セレモニーの内容を教えて下さい.
# 文脈
## ファイル NO1
### ファイル名
100 万人達成セレモニー.pptx
### ファイル内容
20xx 年 yy 月 zz 日:50 万人達成
――――――――
ミュージアムの来場者数が・・・
## ファイル NO2
以下同様
```

図３：回答生成プロンプトの例

3.3 ユーザインターフェース

ユーザインターフェース（図４）ではユーザはクエリと状況（任意）を入力する．ユーザのクエリでは具体的に質問したり命令したりする．ユーザの状況にはクエリの背景や経緯を入力する．状況の入力は任意であり，システムはクエリのみの入力でも動作する．しかし，ユーザの状況を入力することで，検索の精度が向上し，ユーザの認知不可の低減に繋がると考えた．

中央には RAG システムの出力が表示され，その右部に RAG システムが参照した文書のリストとその内容が表示される．

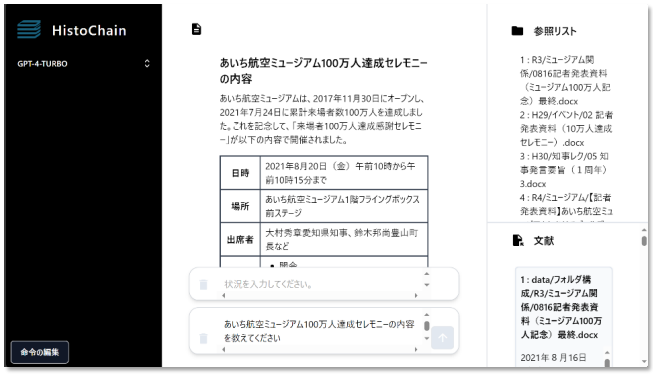

図４：ユーザーインターフェース

4. 回答精度

4.1 実験方法

一問一答形式を用いて，正答率，作業時間，及び回答への信頼度を調べる．被験者に行政に関する問題を与え，正解が記述されたファイルを探し，回答するという実験を行った．また，本実験は愛知県庁都市交通局航空空港課にて実施した．この実験で使用したファイルは実際の行政業務で作成されたものであり，ファイル数は 164 である．これらのファイルの形式は doc，docx，pptx，xls，xlsx，rtf となっている．

担当の行政職員（事前知識あり・行政経験あり）と他課の行政職員（事前知識なし・行政経験あり）と学生（事前知識なし・行政経験なし）の各カテゴリから３人ずつ，合計９人を被験者とし，３つのグループに分ける．各グループは異なるカテゴリの一人ずつから構成される．全９人という人数は統計的に充分とは言えないが，本実証実験事業に関与する部署の人数的制約のため，これ以上被験者を増やすことは不可能であった．問題セットは短答式問題５問で構成され，これを３セット用意する．各グループは手作業，100 文字分割（オーバーラップ 20 文字）による RAG システム，4000 文字分割（オーバーラップ 800 文字）による RAG システムの３つの異なる手法で問題セットに取り組む．グループと問題セットを振り分け，問題の難易度によるバイアスを考慮にいれる．問題セットは事前に被験者以外の担当課の行政職員が作成し，全ての問題は必ずいず

れかのファイルに解答が記述してある．問題内容は数値や名称など明確な答えがあるものになっており，正解か不正解かで評価する．

作業時間は問題を開始してから回答を終了するまでの時間で問題ごとに計測する．手作業では，OS のキーワード検索や，フォルダ構成などに基づき，実際にファイルを開いて問題に取り組む．システム使用時は，システムの精度を測定するために，被験者にはシステムの出力結果をそのまま信頼して回答してもらう．すらわち，ファイルを開いてシステムの回答の正当性を確認することはしない．また，作業後に自分の回答が信頼できるかどうかを 7 段階で評価してもらう．「全く信頼できない」が 0 であり，「非常に信頼できる」が 6 となっている．

4.2 結果

4.2.1 正答率：

表 1 に正答率を示す．担当課職員と他課行政職員において，手作業，100 文字分割システム，4000 文字分割システムを用いた正答率に大きな差は見られなかった．学生においては，手作業時の正答率 53.3%から 100 文字分割システム使用時に 86.7%，4000 文字分割システム使用時に 93.3%まで向上した．

表 1 ：正答率

	手作業	RAG(100-20)	RAG(4000-800)
担当課行政職員	0.933	0.933	0.867
他課行政職員	0.800	0.733	0.800
学生	0.533	0.867	0.933

4.2.2 作業時間：

表 2 に作業時間を示す．RAG システム使用によって手作業時の作業時間と比較して，担当課行政職員は 23.6%から 25.1%，他課行政職員は 75.9%から 77.6%，学生は 77.4%から 79.8%の作業時間削減を達成した．

表 2 ：短答式問題の作業時間（秒）

	手作業	RAG(100-20)	RAG(4000-800)
担当課行政職員	156.4	119.5	117.1
他課行政職員	368.6	83.1	89.0
学生	405.2	91.7	81.8

図 5 に正解時，図 6 に不正解時の作業時間を示す．担当課行政職員と学生の作業時間は手作業において，正解時より不正解時に長くなった．一方，システム使用においては，正解時より不正解時に短くなった．

4.2.3 信頼度：

図 7 に正解時，図 8 に不正解時の信頼度の結果を示す．この信頼度とは，被験者が回答した後に自分の回答がどれだけ信頼できるものであるかを見積もったものである．担当課行政職員は手作業時に置いて正解時の信頼度平均 4.93 から不正解時の信頼度平均 2.00 となっており，著しく低下した．学生も同様に手作用時において正解時の信頼度平均 5.63 から不正解時の信頼度平均 3.28 となっており，著しく低下した．しかし，担当課行政職員も学生も，システム使用時には正解時と不正解時の信頼度に大きな差は見られなかった．他課行政職員は手作業，システム使用時ともに不正解時の信頼度が著しく低下した．

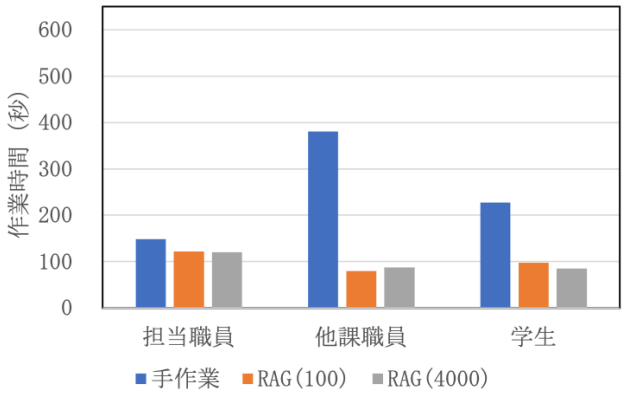

図 5 ：作業時間（正解時）

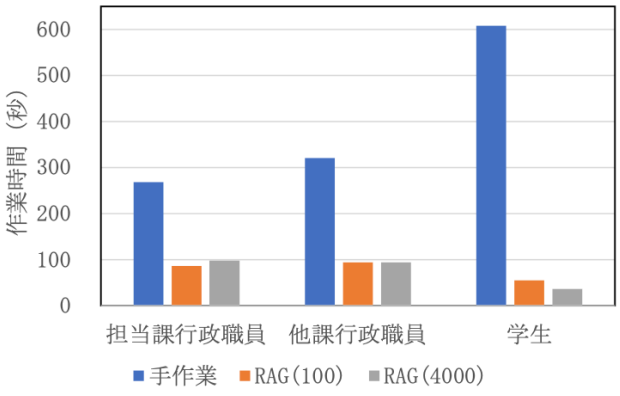

図 6 ：作業時間（不正解時）

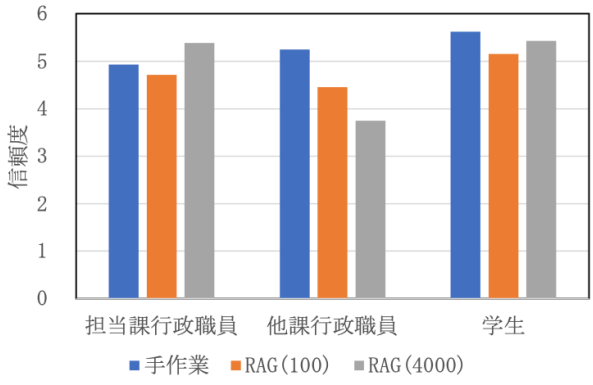

図 7 ：信頼度（正解時）

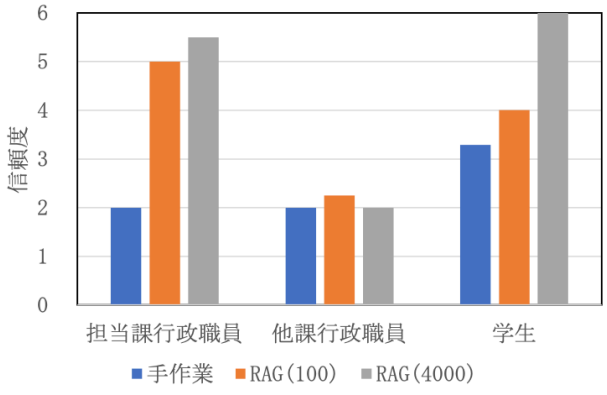

図 8 ：信頼度（不正解時）

4.3 考察

4.3.1 正答率

学生のみがシステムの使用により正答率を大幅に向上させたことが分かった．学生は行政に関する専門知識がほぼないため，システムを利用して必要な情報を得ることで回答できたと考えられる．一方で行政職員は事前知識や行政経験があるため，手作業でも高い正答率を達成し，システム使用による正答率向上は見られなかった．また，学生と行政職員の正答率を比較したときに，手作業では行政職員の正答率のほうが優れているが，システム使用時には正答率に大きな差が見られなかった．これはシステムが専門知識や行政経験の差を補い，効率的な問題解決を支援したためだと考えられる．作業時間を保持・削減したうえで回答精度を維持・向上させたことは，特に経験の浅い職員の働き方改革に繋がる可能性を示した．

4.3.2 正解・不正解時の作業時間と信頼度の比較

担当課行政職員は不正解時の際に手作業で作業時間が長くなり，システム使用時は作業時間が短縮された．また，手作業時には不正解時よりも不正解時の信頼度が大きく低下したが，システム使用時には正解時と不正解時の信頼度にほぼ差がなかった．これは，手作業においてはより正確に回答の信頼度を見積もることが出来ていることを示しており，ファイルを調べる過程で自分の回答の信頼度を見積もることができるからだと考える．一方で，システム使用時には回答がすぐに出力されるためその回答の信頼度を測ることが難しくなっていたと考える．特に，システムの回答が担当課行政職員の先入観や期待と合致するときに批判的な分析が欠ける可能性がある．

他課行政職員については作業時間に手作業とシステム使用で差は殆どなかった．また，手作業においてもシステム使用においても，正解時よりも不正解時の信頼度が著しく低下していた．これは，自身の回答の信頼度を正しく見積もることが出来ていることを示している．事前知識がない一方で行政経験があるため，不確実な情報に対して懐疑的であり，システムの回答を慎重に扱う傾向があると考えられる．また，システムの回答が根拠を示している場合に限りその信頼度を高く見積もる傾向がある．根拠を含むシステムの回答はより正確であると推察される．一方根拠が明記されていないシステムの回答は，LLM が推論に必要な情報を十分に得られておらず，回答が不十分な可能性がある．

学生は，手作業のときは正解時よりも不正解時に作業時間が長くなり，逆にシステム使用時には正解時よりも不正解時に作業時間が短縮された．これは，システムの回答を疑うための事前知識や行政経験がないため，システムの出力結果を容易に信じる傾向によるものと考えられる．また，手作業時において，不正解となったときに作業時間が長くなっていたのは，問題に対するクリティカルな根拠が示されたファイルを探すのに時間をかけたためであり，最終的にそのファイルを見つけられなかったときに信頼度が低くなると推察する．

5. おわりに

本研究では行政 DX のための検索拡張生成システムを開発し，その効果を実証実験にて検証した．手作業と比較してシステム使用によって，行政職員は同程度の正答率で作業時間を 23.6%から 77.6%短縮し，事前知識のない学生は作業時間を 79.8%削減したうえで正答率を 40.0 ポイント向上した．この結果は本 RAG システムが行政職員の業務を効率化させ，働き方改革に繋がる可能性を示した．

この研究では，164 ファイルのデータセットを使用して実証実験を行ったが，実業務における膨大なデータ量において情報検索の作業負担は更に増加すると予想される．今後はさらに多くの行政ファイルに対しても本 RAG システムを効率的に使用できるかを調査する．

謝辞

実証実験を行うにあたり，愛知県庁都市交通局航空空港課及び DX 推進課の職員の方々にご協力いただきました．また，実業務に本システムを導入するための意見や提案を数多くいただき，システム改良を進めることが出来ました．ここに心より感謝します．

また，本研究の一部は愛知県 AichiXTech および JST CREST (JPMJCR20D1)の支援を受けたものです．

参考文献

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., tau Yih, W., Rocktaschel, T., Riedle, S. and Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2021).

[2] AichiXTech: 蓄積された行政情報を簡単に発見できる庁内システムを構築したい！, https://aichixtech.jp/projects/project10.html.

[3] Ryu, C., Lee, S., Pang, S., Choi, C., Choi, H., Min, M. and Sohn, J.-Y.: Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA, Proceedings of the Natural Legal Language Processing Workshop 2023 (Preotiuc-Pietro, D., Gonanta, C., Chalkidis, Il, Barrett, L., Spanakis, G. J. and Aletras, N., eds.), Singapore, Associatation for Computational Linguistics, pp. 132-137(online), DOI: 10.18653/v1/2023.nllp-1.13(2023).

[4] Silva, B., Nunes, L., Estevao, R., Aski, V. and Chandra, R.: GPT4 as an Agronomist Assistant? Answering Agriculture Exams Using Large Language Models (2023).

[5] Levonian, Z., Li, C., Zhu, W., Gade, Al, Henkel, O., Postle, M.-E and Xing, W.: Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference (2023)

[6] Anand, A., Goel, Al, Hira, M., Buldeo, S., Kumar, J., Verma, A., Gupta, R. and Shah, R. R.: SciPhyRAG – Retrieval Augmentation to Improve LLMs on Physics Q&A, Big Data and Artificial Intelligence (Goyal, C., Kumar, N., Bhowmick, S. S., Goyal, P., N. and Kumar, D., eds.), Cham, Springer Nature Switzerland, pp. 50-63 (2023).

[7] Es, S., James, J., Espinosa-Anke, L. and Schockaert, S.: RAGAS: Automated Evaluation of Retrieval Augmented Generation (2023)

[8] Saad-Falcon, J., Khattab, O., Potts, C. and Zaharia, M.: ARES: An Automated Evaluation Framework for Retrieval-Augmented Generaton System (2023)