専門家と非専門家はどう問うか
歴史調査における対話システムへの質問の定量分析

佐原恭平 [1]　関根聡 [2]
[1] 株式会社 COTEN　[2] 株式会社いちから
kyohei.sahara@coten.co.jp　satoshi.sekine@ichikara.ai

概要

歴史調査において，専門家と非専門家の立てる問いにはどんな違いがあるのか．本研究では歴史調査で使用されている対話システム（Leonardo）のログを用いて，専門家と非専門家による質問の差異を探索的・定量的に分析した．本システムは約 1,800 冊の歴史書の情報をベクトル DB に保持し，特定のユーザーからの質問に対する回答を RAG によって生成するものである．ここで収集された 840 件の質問を構造・語彙・品詞の 3 つの観点から比較したところ，専門家は多様な表現で抽象概念や関係性を問う傾向が見られた．一方，非専門家は具体的な事柄を表す語彙を多用するパターンが確認された．

1　はじめに

大規模言語モデル（LLM）の発展により，歴史研究をはじめとする人文学分野でも調査支援システムが実用化されつつある．特に書籍や史料を対象とした Retrieval-Augmented Generation（RAG）[1] は，膨大な文献を効率的に探索することで，研究者の調査効率を大幅に向上させる可能性がある．しかし RAG を用いる対話システムでは，同じ要求であっても質問の表現によって得られる結果が大きく異なることが知られている．

歴史調査における問いは，単なる情報検索のみならず，専門知識と調査能力を駆使して行われる総合的な営みである [2]．こうした質問のスタイルの違いは長年の経験を通じて習得される暗黙知として存在しており，明示的に教えられることは少ない．そこで本研究では我々が先に提案し [3]，実際に歴史調査で使用している Leonard の対話ログを用いて，専門家と非専門家の質問の違いを解析する．本システムは約 1,800 冊の歴史に関する書籍の情報をベクトル DB に保持して，ユーザーからの質問に RAG を用いて回答するもので，COTEN RADIO[4] のコンテンツ制作時などに利用されている．図 1 に Leonardo の概要を示す．

本研究の目的は実運用されている対話システムのログを用いて，歴史調査における専門家と非専門家の質問の差異を構造・語彙・品詞の観点から定量的に明らかにすることである．これにより，これまで暗黙知として存在していた「問いの立て方」の違いを可視化し，初学者への教育支援や RAG をベースとした対話システムの設計にも応用可能な知見を提供する．

2　関連研究

専門性と検索行動に関する研究は Web[5]，歴史データベース [6]，医療情報 [7] など複数の文脈で行われてきた．これらの研究から明らかになっている傾向として，専門家は非専門家に比べて長いクエリを使用し，ドメイン固有の概念や高度な検索機能を多用することがわかっている．一方，非専門家や初心者は広いトピックに基づいてクエリを作成する傾向がある．たとえば Hölscher らはインターネット専門家が初心者に比べてより長いクエリや Boolean 演算子などの高度な検索機能を多用することを示し [5]，また Tamine らは臨床領域において専門家が非専門家よりも多くの医学概念を使用して階層的にクエリを作成することを報告している [7]．

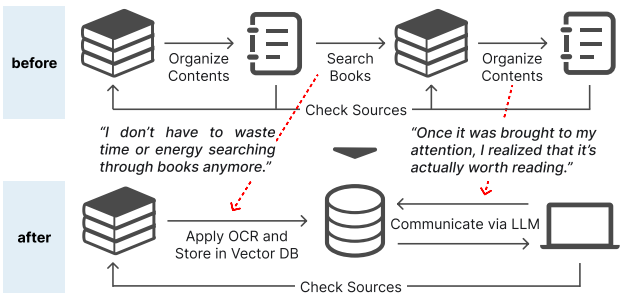

図 1　Leonardo の概要 [3]

件数	専門家	非専門家	全体
質問	797	43	840
ユニークユーザー	17	4	21
セッション	288	13	301
質問数の平均値	47	11	40
質問数の中央値	26	10	22

表 1　データセットの基本情報

こうした先行研究は専門性と質問との関係性に重要な示唆を与える．しかしこれらの対象は従来の Information Retrieval（IR）システムに限定されており，RAG をベースとした対話システムにおいて専門性が与える影響については未解明なところも多い．RAG を用いた情報検索は同じ要求であっても質問の表現によって得られる回答の質が大きく変わるため，専門家と非専門家がどのように問いを立てるかを明らかにすることは，システムの効果的な利用において重要な意味を持つ．本研究では歴史調査という専門知識を要する領域の実運用ログを用いて，質問のスタイルの差異を構造・語彙・品詞の観点から定量的に比較する．

3　データセット

本研究で使用するのは，RAG をベースとした対話システム（Leonardo）から取得した実運用ログである．

分析の前処理としてまず歴史調査に従事している者，すなわちその業務を行うチームに所属しているユーザーを専門家，それ以外を非専門家とした．専門家には歴史コンテンツの制作を職務とする研究者や編集者が含まれる．一方，非専門家は社内の他部門のスタッフや歴史調査を主業務としないユーザーである．ただし，Leonardo の開発やテストに関わるユーザーは本システムの挙動に精通しているうえ，動作確認を目的とする質問を行っている可能性があるため，分析対象から除外した．

対話ログのうち，質問が空白のレコードは取り除き，また完全に同一の質問文は重複とみなして 1 件のみを残した．さらに質問数が 3 件以上のユーザーだけを分析の対象とした．質問数が極端に少ないユーザーは，単にテストのために本システムを利用した可能性が高いためである．これらの前処理によって得られたデータセットの基本情報を表 1 に示す．データセットは専門家による質問が約 94.9%を占めるが，この不均衡は Leonardo が主に専門家向けの調査支援ツールとして運用されていることに起因する．

4　方法

本研究では専門家と非専門家の質問の差を，構造・語彙・品詞の 3 軸の特徴量で比較する．期間やユーザーによって調査内容が異なるため，トピックそのものを極力特徴量として扱わないような設計とした．

まず各質問文の構造的な特徴を解析した．特徴量の一覧は表 2 に示すとおりである．このうち，`has` の接頭辞がついているものは真偽値，`len_chars` は離散量，`uniq_ratio` は連続量である．次に質問文に含まれる語彙を対象に，分類語彙表（WLSP）[8] の中分類で分析を行った．ここでは MeCab[9] の形態素解析により質問文から単語を抽出し，各語に WLSP の中分類を付与した．ただし，特定のトピックによる寄与を取り除くため固有名詞は分析の対象外とした．最後に質問文で使用される品詞を比較した．ここでも形態素解析には MeCab を用いたが，「フィラー」「その他」と分類される語は本研究とは無関係である可能性が高いので，分析対象から除外した．語彙および品詞は出現数に基づく離散量として扱った．

特徴量	説明	専門家	非専門家
`has_constraint`	回答方法の指定の有無	0.083 ± 0.055	0.132 ± 0.160
`has_list`	箇条書きの有無	0.133 ± 0.150	0.211 ± 0.366
`has_number`	数値表現の有無	0.092 ± 0.111	0.125 ± 0.221
`has_qmark`	「？」の有無	0.394 ± 0.231	0.391 ± 0.251
`has_quote`	引用符の有無	0.091 ± 0.117	0.063 ± 0.110
`has_wh`	疑問詞の有無	0.247 ± 0.168	0.216 ± 0.149
`len_chars`	文字数（空白を含む）	34.716 ± 13.270	24.442 ± 8.006
`uniq_ratio`	全語に対するユニーク語の比率	0.955 ± 0.025	0.885 ± 0.098

表 2　構造解析における特徴量と平均値（± 標準偏差）

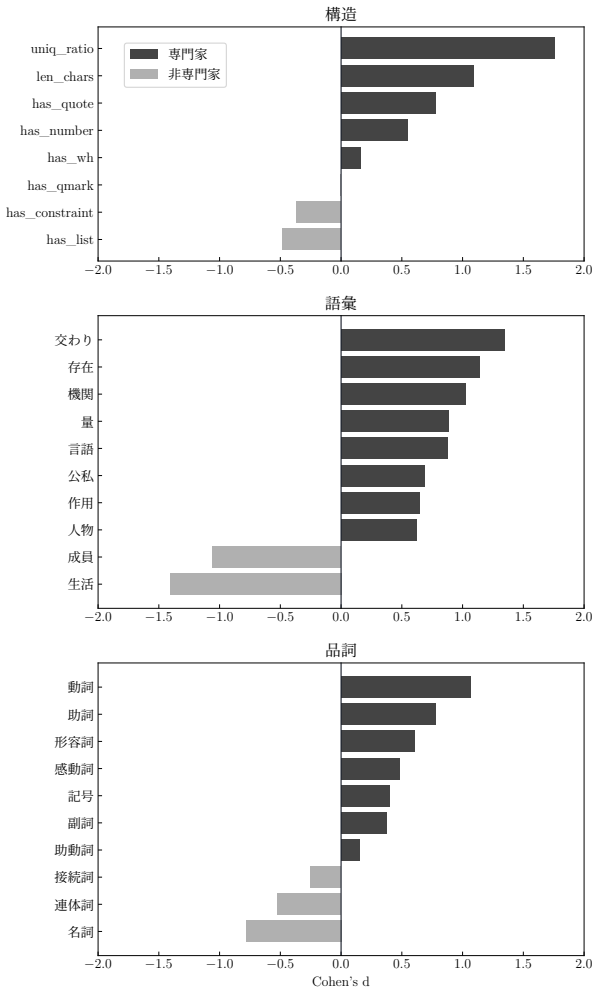

図 2　解析結果

これらの特徴量をユーザー単位で集計し，各ユーザーの平均を代表値とした。専門家・非専門家の差異は，このユーザー平均に対して Cohen's d[10] を算出することで評価した。サンプルサイズが限られることから効果量の解釈を重視する．

5　結果

すべての解析結果を図 2 に示す．語彙に関しては d の絶対値が大きいもの 10 件を抜粋して表示している．構造に関する特徴量の平均値と標準偏差は表 2 のとおりである．専門家については，構造で `uniq_ratio` や `len_chars`，語彙で「交わり」や「存在」，品詞で動詞や助詞が特徴として見られた．非専門家に関しては，構造で `has_list` や `has_constraint`，語彙で「生活」や「成員」，品詞で名詞や連体詞の使用が多いことが観察された．

区分	質問
専門家	戦争と平和のメカニズムを調べている。アクターが置かれている”環境”を分類する仮説を持ちたい。過去の戦争書籍において、戦争するアクターあるいは戦争を回避したアクターが置かれていた環境として言及されている物を列挙、分類して。
専門家	東ティモールの歴史を時系列で詳細にまとめて。特に東ティモールの独立に関わる紛争の部分を詳細にしてほしい。その際年代の下のレイヤーでは、関係していたアクター毎の振る舞いを軸にしてほしい。
専門家	イラン・イラク戦争が終了後、イラクから債務の帳消しの要請に対して、サウジアラビアはその要請を受け入れ、一方クウェートは受け入れを拒否したことには、どのような背景がありますか？
非専門家	桶狭間の戦いでの戦略とはどのようなものですか？
非専門家	女性の役割の変遷を教えて
非専門家	圧倒的なカリスマで統治した統治者は？

表 3　実際になされた代表的な質問

6　考察

構造では表現の多様性や文章量，疑問詞の有無などに大きな差異があった．また語彙において，専門家は関係性や抽象概念を表す単語の寄与が大きく，非専門家は具体的な事象を指すものが相対的に多い傾向が見られた．品詞の使い方に関して，専門家の質問は動詞・助詞といった単語の比率が高く，複雑な関係性や動的なプロセスを表現していると考えられる．一方，非専門家の質問は名詞や連体詞の割合が比較的大きく，具体的な事物が結びついている傾向を示している．これらを総合すると，専門家が多様な表現で関係性や抽象概念を問う一方，非専門家は具体的な情報を求める質問が多いと解釈できる．

定量分析で観察された傾向を具体的に示すため，表 3 に代表的な質問を掲載した．専門家の例では前提や目的を明示したうえで関係性や因果を問う記述が多く，長い文章で構成されていることが確認でき

る．一方，非専門家の例は対象や主題を端的に指示するものが中心であり，単一の問いとして表現されることが多い．

7 限界

本研究にはいくつかの限界がある．第 1 にサンプルサイズが小さく（専門家 17 名，非専門家 4 名），観察された差異が統計的に有意な水準に達していない可能性がある．第 2 にデータが単一システム（Leonardo）のログに限定されており，歴史調査の質問全般に一般化できるかどうか不明である．第 3 に本研究は質問の差異を記述することに焦点を置いており，質問のスタイルが RAG の回答品質にどう影響するかという因果関係までは検討していない．

8 おわりに

本研究では歴史調査で利用される対話システムのログを用いて，専門家と非専門家の質問の差異を定量的に分析した．その結果，専門家は抽象概念を表す語彙で関係性を問う傾向があり，非専門家は名詞中心で具体的な事象を質問する傾向があることが明らかになった．本研究はこれまで暗黙知として存在していた，人文学における「問いの立て方」の差異を可視化した探索的研究として位置づけられる．

より信頼度の高い結果を得るためには，複数の対話システムや異なる領域（文学，哲学など）へ対象を拡張することで，より大規模かつ精緻に差異を検証する必要がある．また質問のスタイルが RAG の回答に与える影響の調査や，専門家の表現を組み込んだ教育などへの応用も今後の展開としてありうる．これらはこれまで暗黙知とされてきた経験やスキルの定量的な評価を明らかにし，LLM を活用した人文学研究の発展に寄与するものと考えられる．

謝辞

本研究はいちから Ichikara TRY 2025 の支援を受けて実施されました。ここに深く感謝の意を表明いたします．

参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474. Curran Associates, Inc., 2020.

[2] M. Kainulainen, M. Puurtinen, and C. A. Chinn. Regrounding inquiry-based learning in history: A study of historians' epistemic processes. **Cognition and Instruction**, Vol. 43, pp. 291–315, 2025.

[3] Taisei Klasen, Hinatsu Kusano, and Kyohei Sahara. Leonardo: Book-based Q&A system accelerating digital humanities research with source attribution. **Proceedings of the 14th Conference of the Japanese Association for Digital Humanities**, pp. 113–115, 2025.

[4] https://coten.co.jp/services/cotenradio/.

[5] Christoph Holscher and Gerhard Strube. Web search behavior of internet experts and newbies. **Computer Networks**, Vol. 33, pp. 337–346, 2000.

[6] Charles Cole, John E. Leide, Emeka Nwakamma, Jamshid Beheshti, and Andrew Large. Structure of domain novice users' queries to a history database. **Proceedings of the 66th Annual Meeting of the American Society for Information Science and Technology (ASIST 2003)**, 2003.

[7] L. Tamine and C. Chouquet. On the impact of domain expertise on query formulation, relevance assessment and retrieval performance in clinical settings. **Information Processing & Management**, Vol. 53, pp. 332–350, 2017.

[8] https://clrd.ninjal.ac.jp/goihyo.html.

[9] https://taku910.github.io/mecab/.

[10] Jacob Cohen. **Statistical Power Analysis for the Behavioral Sciences**. Routledge, New York, 2nd edition, 1988.