ユーザ属性を考慮した検索拡張生成による学部教育課程相談チャットボット

竹内新 目良和也 梶山朋子
広島市立大学 情報科学部
g20116@e.hiroshima-cu.ac.jp
{mera, kajiyama}@hiroshima-cu.ac.jp

概要

本研究では，ユーザ属性を考慮した検索拡張生成（RAG）を活用し，学部教育課程に関する質問に応答するチャットボットを構築した．本提案では，学部や入学年度などのユーザ属性に基づき，外部 DB に含まれる全ドキュメントからユーザ属性に合ったドキュメント群を抽出する．そしてこのドキュメント群から質問に適合するパッセージ集合を生成し，質問とともに LLM に入力することで，質問への回答を生成する．提案手法により生成した回答は，全ドキュメントから生成した適合パッセージ集合による回答と比べて，質問内容に関係のない記載を含まない高評価の回答であることを確認した．

1 背景と目的

学生が教育課程に関する規則について質問したい場合，大学の公式ウェブサイトや配布された規則集などを閲覧する．これらの情報は分散して存在するだけでなく，在籍するすべての学生に対応できるように記載されているため，必要な情報を見つけるまでに時間がかかる．この問題を解決する 1 つの方法として，チャットボットの活用が挙げられる．

対話シナリオや FAQ 集を作成し，チャットボットを構築した研究が存在する[1, 2]．情報を追加するためには，これらの再構築が必要となるため，管理コストが高い．また，全ユーザに向けて共通の回答を提示する形となるため，ユーザ属性を考慮して回答を変化させることが難しい．

これらの問題を解決する 1 つの手法として，検索拡張生成（Retrieval-Augmented Generation, 以下, RAG と呼ぶ）[3]が挙げられる．RAG とは，外部 DB 内の全ドキュメントにある特定の文や段落（以下，パッセージ）から，質問に適合するパッセージ集合（以下，適合パッセージ集合）を生成し，質問とともに LLM に入力することで, 質問への回答を生成する手法である．情報を追加する場合，外部 DB を更新するだけで対応できるため, LLM を再度学習する必要がない．また，質問文にユーザ属性を明記することにより，そのユーザに適した回答を生成し提示することが可能である. RAG を用いたチャットボット[4]も提案されている．しかし，適合パッセージ集合を生成する際に，外部 DB に格納された全ドキュメントを対象とするため，不必要な情報が回答に含まれるという問題点があった．

本研究では，ユーザ属性を考慮した RAG を活用し，学部教育課程に関する相談に対応するチャットボットを構築することを目的とする．各ユーザに合った適合パッセージ集合を生成することで，質問内容に関係のない記載を含まない回答の生成を目指す．

2 提案システムの概要

図 1 は，提案システムの構成を示している．本提案では, 学部や入学年度などのユーザ属性に基づき, 外部 DB に含まれる全ドキュメントからユーザ属性に合ったドキュメント群を抽出する．そしてこのドキュメント群から質問に適合するパッセージ集合を生成し, 質問とともに LLM に入力することで, 質問への回答を生成する．

研究の対象は, 広島市立大学（以下, 本学）の 2021～2024 年度入学生の各学部（情報科学部, 国際学部, 芸術学部）の教育課程である．ユーザ属性は，教育課程に影響を与える 3 要素として, 入学年度, 学部, イノベーション人材育成プログラム（以下, イノベ）取得の有無とした．

2.1 外部 DB の作成

外部 DB の作成にあたり，学修の手引き，公式ウェブサイト，公立大学法人広島市立大学規定集の 3 つの情報源を活用した．これらの情報源から，(a)各入

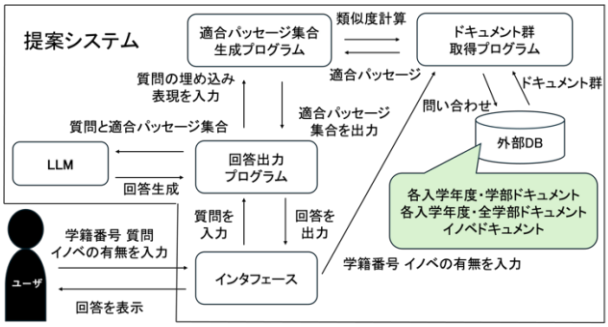

図1 提案システムの構成

学年度に対し全学部に共通する内容（4 種類），(b)各入学年度に対し各学部に関する内容（12 種類），(c)イノベに関する内容（1 種類）の計 17 種類のドキュメントを作成した．各ドキュメントには，3 つの情報源から抽出したパッセージが記載されている．

学修の手引きは，学部生を対象とし，入学年度ごとに発行されている．目次に掲載されているタイトルと該当する本文を連結したテキストを1つのパッセージとして定義した．タイトル内に学部名が含まれる場合は(b)，パッセージ内に「イノベーション人材育成プログラム」が含まれる場合は(c)，その他は(a)としドキュメントを作成した. 表はcsv形式に変換し，同様の手順でドキュメントに追記した．

公式ウェブサイトは，学修の手引きの章タイトルから抽出したキーワード（授業，履修，試験，成績，教育課程，教育課程表，教育職員免許，受領資格取得関係科目表，学芸，資格取得関係科目表，資格）が，ページタイトルに含まれるテキストを対象とした. タイトルと本文を連結させたテキストを1つのパッセージとして定義し，パッセージ内に年度の記載があれば，学修の手引きと同様の手順でドキュメントに追記した．ただし，博士課程に関する文言（大学院，研究科，博士）を含むパッセージは追記対象として除外した．

規定集は，上述のキーワードが，規定名，章名，節名，条文タイトルのいずれかに含まれるテキストを対象とした．規定名，各章，各節，各条文のタイトル, および本文を連結させたテキストを1つのパッセージとして定義し，公式ウェブサイトと同様の手順でドキュメントに追記した．

2.2 回答の生成

2.1節より作成した外部DBに含まれる全ドキュメントからユーザ属性に合ったドキュメント群の抽出は，ユーザの学籍番号により行った．入力された学籍番号によって，入学年度と所属学部を特定し，外部DB内の全ドキュメントから全学部適用とその学部のみに適用されるドキュメント群を抽出した．

そして, ユーザから入力された質問とのcos類似度計算によるスコアが，0.5以上のテキストからなるパッセージを適合パッセージ集合とした．なお，適合パッセージ集合とするcsvファイルは，スコアが0.5以上の「科目表タイトル」を持つcsvファイルとした．それらとLLMへの指示文を加えてプロンプトとし，OpenAIのAssistant APIを用いてLLM（gpt4o）に入力することで，質問への回答を生成した．なお，類似度計算を行う際の埋め込み表現の取得には，OpenAIのtext-embedding-3-largeを使用した．

2.3 インタフェース

図2は，提案システムの概観を示している．画面上の入力欄にユーザ属性と質問を入力し，「実行」ボタンを押すと，その下に回答が提示される．本学では学籍番号から入学年度と学部を判別できるため，ユーザ属性の入力として，学籍番号を活用した．

図2 提案システムの概観

3 評価

提案手法の評価にあたり，外部 DB 内の全ドキュメントから生成した適合パッセージ集合による回答（以下，既存手法）と比較する実験を行った．実験協力者は本学学部生 104 名で，計 140 件のデータを収集した．実験協力者に回答を提示する際に，既存

手法と提案手法を左右に並べて表示し，一対比較による評価を行った．表示位置による影響を排除するために，左右の表示は適宜変更するとともに，どちらの手法を用いたかは明示せず回答を提示した．実験協力者は，提示されたそれぞれの回答に対し，質問が解決できたか（Q1），回答に関係のない質問が含まれていなかったか（Q2）について 5 段階で回答した後，最終的にどちらの回答が良かったか 3 段階で回答した．

図3は，各手法に対するQ1，Q2の平均評価値を示している．いずれの質問も高い平均評価値となった． Q2の有意差が認められなかったが，Q1については，提案手法の方が問題を解決できることを確認した（$p=.015$）．

図 4 は Q3 に対する各選択肢の選択割合を示している．半数以上が提案手法による回答の方が良かったと回答し，既存手法より提案手法の方が高評価となった．既存手法の方が良かったと回答した人は，1 年生や 2 年生が多かった．4 年次進級条件など，まだ十分に把握できていない内容を提示された場合，良し悪しの判断が難しくなったためと考えられる．今後の課題は，回答の根拠を提示する機能などを追加し，利用者が情報の信憑性を判断しやすい仕組みを整備することである．

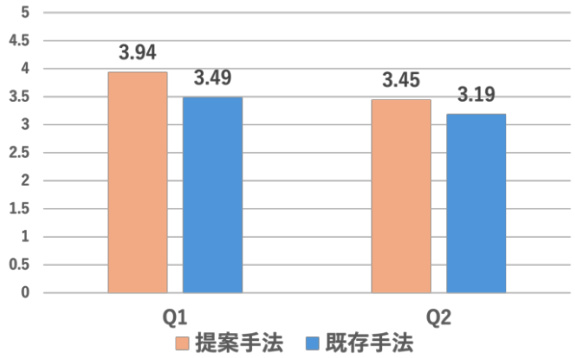

図 3　Q1 と Q2 の平均評価値

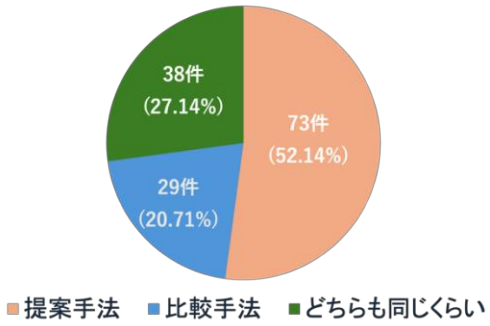

図 4　Q3 の回答結果

4 参考文献

1. 坂田亘，田中リベカ，黒橋 禎夫. 公式ウェブサイトをベースにした QA チャットボットの自動構築，言語処理学会 第 26 回年次大会発表論文集，pp. 327-330, 2020.

2. 吉田尚水ら. FAQ 集から自動で構築可能な QA 対話システム，人工知能学会研究会資料，SIG-SLUD-B902-26, pp. 113-114, 2019.

3. P. Lewis, et al., “Retrieval- augmented generation for knowledge-intensive NLP tasks,” In Proc. of the 34th International Conference on Neural Information Processing System , pp. 9459-9474, 2020.

4. Y. Mao, et al., "FIT-RAG: Black-Box RAG with Factual Information and Token Reduction," ACM Transactions on Information Systems, 26 pages, 2024. https://doi.org/10.1145/3676957