様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２８ 日現在

機関番号：８２６５７
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ１２１３２

研究課題名（和文）Search-Oriented Dialog System for Data Science

研究課題名（英文）Search-Oriented Dialog System for Data Science

研究代表者

金　進東（Kim, Jin-Dong）

大学共同利用機関法人情報・システム研究機構（機構本部施設等）・データサイエンス共同利用基盤施設・特任准教授

研究者番号：４０５３６８９３

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究の主な成果としては、（1）人間解剖学3Dモデルを検索するために開発されたカスタマイズGPTであるAnatomy3DExplorer（現在GPTストアで公開中）、（2）RDFデータ検索システムであるLODQAの拡張機能として開発されたダイアログインターフェース、そして（3）テキストアノテーションのウェブインターフェースであるTextAEの拡張機能として開発されたダイアログアノテーション機能が挙げられる。研究成果を述べた論文はGenomics&Informatics誌に投稿され、現在審査中である。全体として、検索志向のダイアログシステムがLLMを活用して効果的に開発できることを示した。

研究成果の学術的意義や社会的意義
本研究成果は、LLMを活用することで自然言語ダイアログインターフェースを効果的に開発できることを示している。この成果は既に他のデータベース検索インターフェースの開発にも拡張されている。学術的には、LLMが人間の知能とデータに埋め込まれた知能を結びつける効果的なレイヤーを提供することを示している。社会的には、データベース検索スキルを持たないユーザーに対してデータベースにアクセスするための実用的な手段を提供できることを示している。これは、一般の人々が専門知識により容易にアクセスできるようになることを意味する。

研究成果の概要（英文）：The primary outputs of this research include Anatomy3DExplorer, a customized GPT developed for searching human anatomy 3D models (currently available on the GPT store), the dialog interface implemented in LODQA, and a web-based user interface for analyzing and annotating dialogs, integrated as a feature of TextAE. Currently, Anatomy3DExplorer is successfully being utilized by the developers of BodyParts3D, a database of anatomical 3D models. A paper detailing our findings has been submitted to the journal of Genomics & Informatics and is currently under review. Overall, we have demonstrated that search-oriented dialog systems can be effectively developed leveraging LLMs.

研究分野：Computational Linguistics

キーワード：Natural Language　Dialog Interface　User Interface　Database　Data Science　Large Language Models　Customized GPT

1. 研究開始当初の背景

近年、データ駆動型科学の発展は、遺伝学、ゲノミクス、創薬などの様々な分野に大きな影響を与え、データの利用を通じて大きな進展が見られる。しかし、これらの機会にもかかわらず、膨大なデータ資源にアクセスするためには、SQLやSPARQLなどの検索専用言語を筆頭とする専門知識が必要な場合が多く、多くの科学者がデータサイエンスを十分に活用できていない。

一方、人工知能(AI)の著しい進歩により、人間親和的な自然言語による対話インターフェースへの関心が急増している。本研究は、データサイエンスにおけるデータアクセス問題を解決できる自然言語対話インターフェースの開発が可能かどうかという疑問から始まった。つまり、データ検索を助ける対話システムを開発し、その有用性を確かめることが本研究の目的である。これをデータサイエンスにおける検索指向の対話システム(Search-oriented dialog system for data science)と称する。

2. 研究の目的

この研究の主な目的は、検索指向の対話システムのプロトタイプを開発し、実用性を評価することである。当初、ユーザーの一連の発話を通じて表現される検索意図を把握することを、検索指向の対話システムの開発の核心的な問題として想定し、研究の中心的な課題として全体の研究が設計された。

しかし、広範な人間の知識コーパスで訓練された大規模言語モデル(LLM)の出現に伴い、研究開発の状況が大きく変わった。LLMは既に高度な対話能力を備えているため、重複投資を避けて最先端の技術を本来の研究に取り組むために、研究の設計を転換し、データアクセスを支援するためのLLMの応用を探求することを目指すことにした。

3. 研究の方法

この研究の進展は、大規模言語モデル(LLM)の出現により、研究の状況が一変し、方法論の適応gが必要となった。これらの変化に対応するため、LLMの機能をプロジェクトに統合することを選択した。具体的には、最先端のチャットインターフェースで知られるChatGPT APIの利用と、GPTのカスタマイズという2つのアプローチを取った。

(1) ChatGPT APIの利用

検索向けの対話システム開発の一環として、既存のデータ検索インターフェースであるLODQAシステムにChatGPT APIを利用して対話インターフェースを実装した。LODQAはリンクドオープンデータ(Linked Open Data)を検索するための質疑応答形式のユーザーインターフェースである。図1は、LODQAシステムの元のバージョン(a)と対話インターフェースを実装した後(b)の様子を示す。元のバージョンは、受け取った自然言語クエリ(natural language query; NLQ)を解析し、その解析に基づいてデータ検索を行い、結果を回答として出力する。修正バージョンでは、受け取ったNLQをまずコンテキスト化してから解析する点が異なる。コンテキスト化モジュールは、受け取ったクエリとコンテキストの内容を含む新たなクエリを作成する。その過程でChatGPT APIが活用される。コンテキスト化されたクエリは、次に受け取るクエリのコンテキストとして機能する。図2は、コンテキスト化モジュールの動作の例を示す。最初にクエリが届くと、コンテキストは存在しないため、コンテキスト化は不要で、クエリは直接コンテキスト化されたクエリとして転送される。次のクエリに対しては、最初のクエリがコンテキストとして機能し、コンテキスト化されたクエリが作られる。クエリのグラフ表現は、LODQAによるクエリの内部的な解釈を示す。

このLODQAの拡張は、対話インターフェースを実装しているが、実装された対話はクエリの修正に限られる。LLMのチャットインターフェースの多用途な可能性を考えると、クエリの修正だけにその応用を制限することは、LLMの機能を十分に活用しているとは言えない。

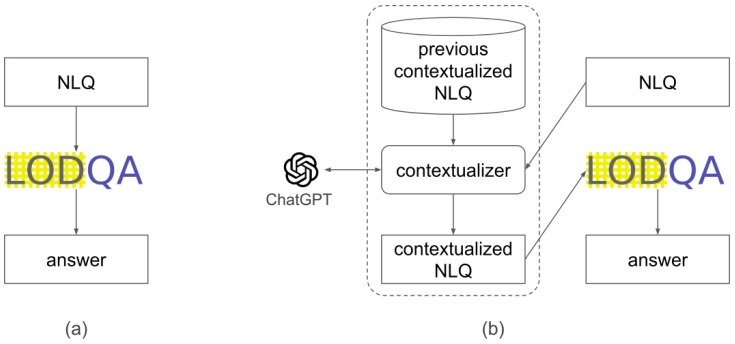


図1. LODQAシステム、(a) 対話インターフェース拡張前と (b) 拡張後

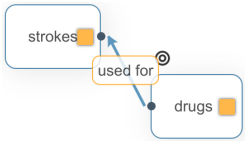
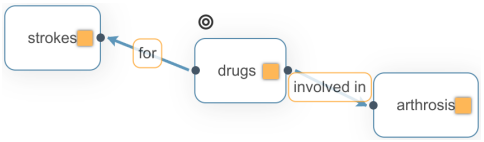
	初期クエリ	2番目のクエリ
コンテキスト	(empty)	What drugs are used for strokes?
クエリ	What drugs are used for strokes?	What are involved in arthrosis?
コンテキスト化されたクエリ	What drugs are used for strokes?	What drugs for strokes are involved in arthrosis?
クエリのグラフ表現		

図2. 2つの連続する例を用いたクエリのコンテキスト化の図解

(2) Customized GPT

言語処理における大規模言語モデル(LLM)の多用途な可能性を十分に活用するために、GPTをカスタマイズする方法での検索志向対話システムを開発を行った。その結果、人間の解剖学の3Dモデルのデータベースであ るBodyParts3Dにアクセスするための自然言語対話インタフェースである Anatomy3DExplorerを開発した。

BodyParts3D(https://lifesciencedb.jp/bp3d)は、Foundational Model of Anatomy(FMA)識別子を使用して解剖学の3Dモデルを取得するためのAPIを提供する。以下はリクエストの一例である:

/FMASearch_SegmentUI/latest/?id=FMA14385&id=FMA42352&id=FMA24922&expansion=all

上記のパスは、図5に示されているように、正中神経(FMA14385)、手根管(FMA42352)、手首(FMA24922)などの構造を含む3Dモデルを生成する。Anatomy3DExplorerは、このAPIとの通信のために、このAPIのOpenAPIスキーマに従って構成されたアクションを使用する。

このAPIを使用する際の大きな問題は、ユーザーがFMA識別子を提供する必要があることで、これは非常に非人間親和的な作業である。この問題はBodyParts3Dに限らず、さまざまなデータベースで共通している。ChatGPTや類似の大規模言語モデル(LLM)を使用してこれらの識別子を取得することが可能と思えるかもしれないが、これらのモデルは、誤った情報や作り上げられた情報を生成するリスク(「ハルシネーション」として知られる現象)があるため、直接的なデータ取得タスクには通常適していない。この問題の例は図6を参照してほしい。この課題を軽減するために、信頼できる識別子取得の代替手段としてPubDictionariesという外部サービスを利用する。

PubDictionaries(https://pubdictionaries.org/)は、さまざまな辞書をホストしている公共リポジトリであり、自然言語の用語から識別子を取得するためのAPIを提供する。以下は使用例である:

/find_ids.json?dictionary=FMA-PAE&labels=median+nerve|carpal+tunnel|wrist

図3はAnatomy3DExplorerでの典型的なワークフローを示す。まず、ChatGPTとの対話を通じて欲しい解剖学的用語を収集する。FMA IDが必要な場合は、PubDictionariesに相談して取得する。識別子が用意されると、BodyParts3Dから対応する3Dモデルを取得するのができる。

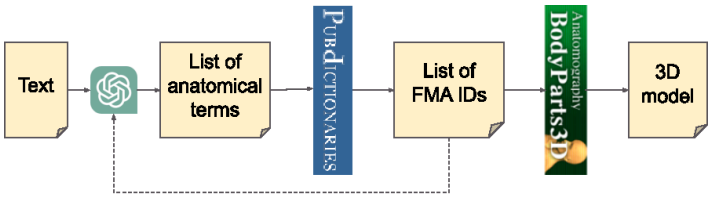


図3. Anatomy3DExplorerでの典型的なワークフロー

4. 研究成果

この研究プロジェクトの主な成果は以下の通りである: (1) Anatomy3DExplorerは、人間の解剖学3Dモデルを検索するためのカスタマイズされたGPTであり、GPTストアで利用可能である。(2) LODQA対話インターフェースはChatGPT APIを使用して開発された対話インターフェースで、https://lodqa.org で利用可能である。(3) TextAEウェブインターフェースは対話を分析および注釈するためのインターフェース実装していて、https://textae.pubannotation.org で利用可能である。特に、Anatomy3DExplorerはBodyParts3Dの開発者によって成功裏に採用されで実活用されている。この研究の結果は、大規模言語モデルを使用した検索指向の対話システムの効果的な開発が可能なことを示している。この内容は論文として纏まれて、Journal of Genomics & Informaticsに提出され、現在審査中である。継続中の研究の一環として、複数のデータベースにわたる応用を広げるために、追加のカスタマイズされたGPTを開発する予定である。

理解を助けるために、以下ではAnatomy3DExplorerのデモンストレーションを含め、その実際の応用と有用性を紹介する。

Anatomy3DExplorerのデモンストレーション

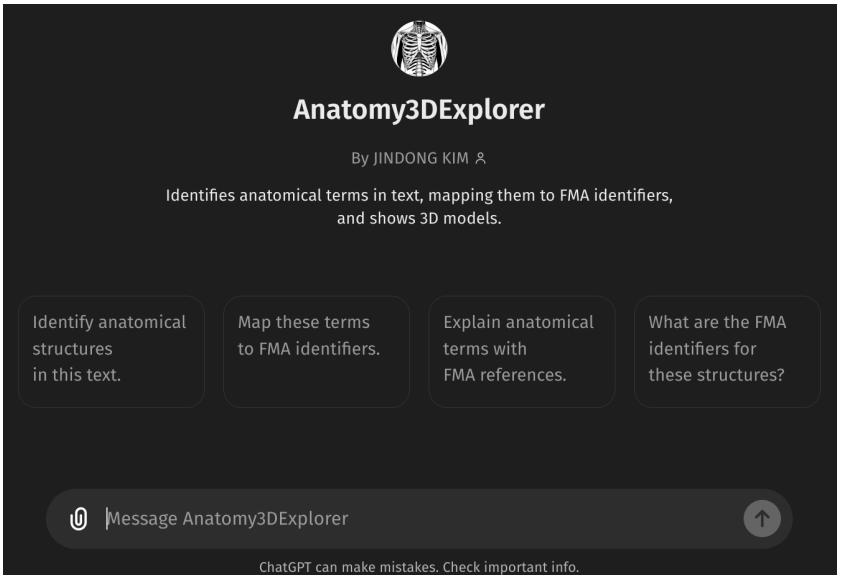


図4. Anatomy3DExplorerの初期画面

図4は、Anatomy3DExplorerの初期インターフェースを示す。テキスト内の解剖学用語を特定し、対応するFMA識別子にマッピングして、3Dを見せるツールとして説明されている。図5は、テキストが提供された際のAnatomy3DExplorerの応答を示す。提供されたテキスト内で見つかった解剖学用語が表示され、対応するFMA識別子が付随している。「*Talked to pubdictionaries.org*」というメッセージは、識別子の取得のためPubDictionariesに相談したことを示す。

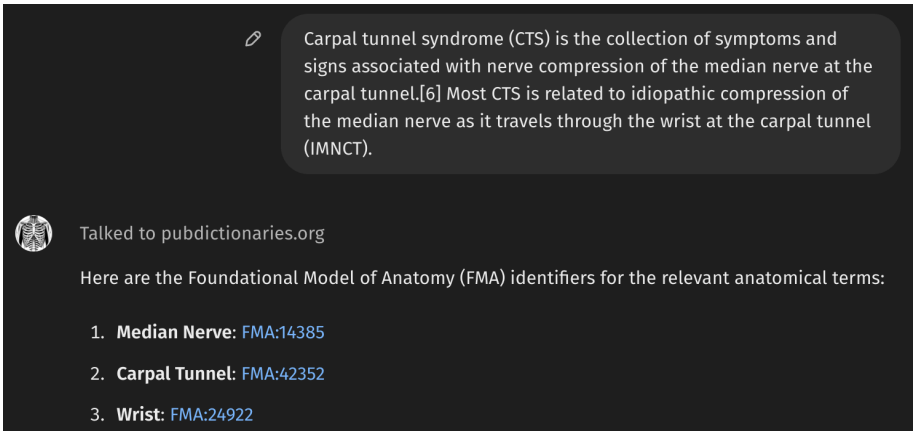


図5. テキストのブロックが提供された際のAnatomy3DExplorerの応答。

図6は、同じテキストの入力に対してPubDictionariesに相談せずにGPTが応答したFMA識別子を示す。すべてが誤っていて、ハルシネーションの一般的な例を示していて、FMA識別子の取得にPubDictionariesを使用することの理由を強調している。

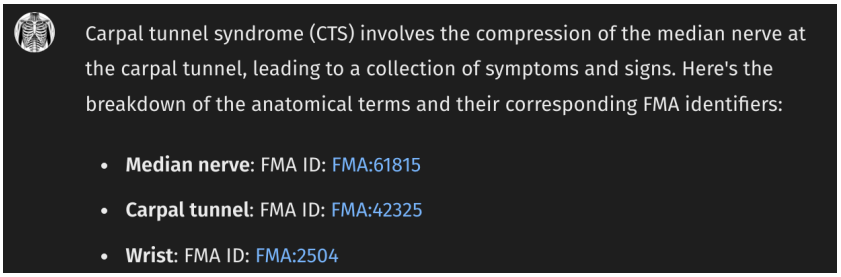


図6. PubDictionariesを参照しなかった場合にChatGPTが示した不正確なFMA識別子の応答。

図7では、ユーザーがこれまでに特定したすべての解剖学用語を含む3Dモデルを要求している。Anatomy3DExplorerはBodyParts3Dと通信し、手根管症候群に関連する解剖学構造を示す3Dモデルへのリンクを提供する(図8)。

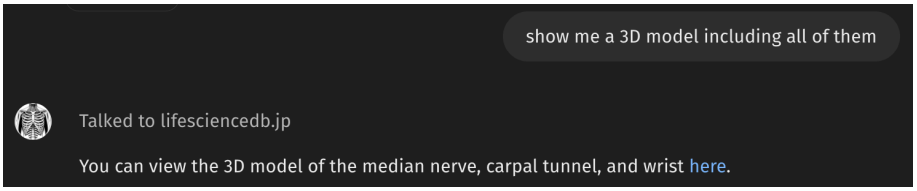


図7. Anatomy3DExplorerに解剖学構造の3Dモデルを作成するよう指示する例

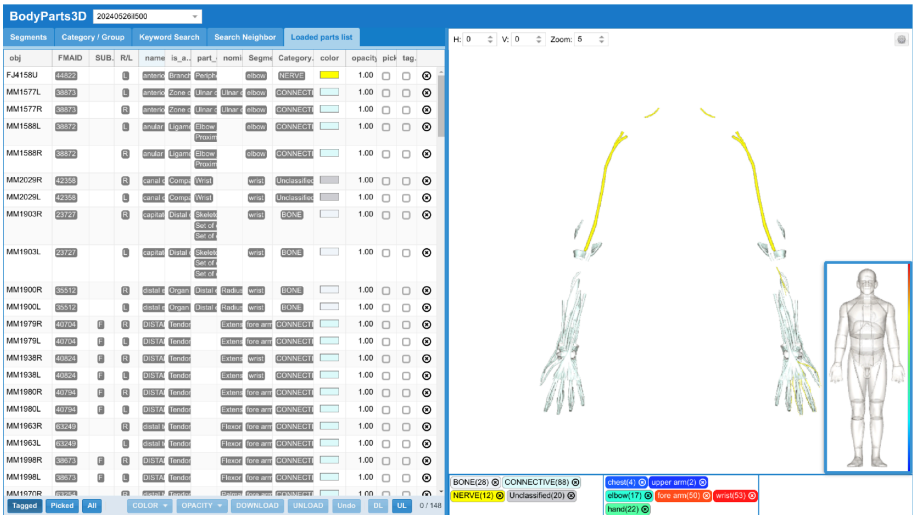


図8. BodyParts3Dから取得された手根管症候群に関連する解剖学的3Dモデル

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計1件

国際研究集会	開催年
Beyond QA Kickoff Meeting of XQA and DialoQ	2019年～2019年

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関