SRA の技術情報誌
GS LetterNeo
Vol.182　2023 年 9 月

Azure 環境で大規模言語モデルを用いたチャットに独自データを組み込む

土田 拓実

技術本部 技術開発室

はじめに

生成系 AI 技術の進歩により、この技術を活用したいというニーズが高まっています。自然言語処理の分野では、OpenAI 社の ChatGPT の台頭もあり、人間らしい回答を返す高性能な Q&A Chatbot の開発が注目されています。

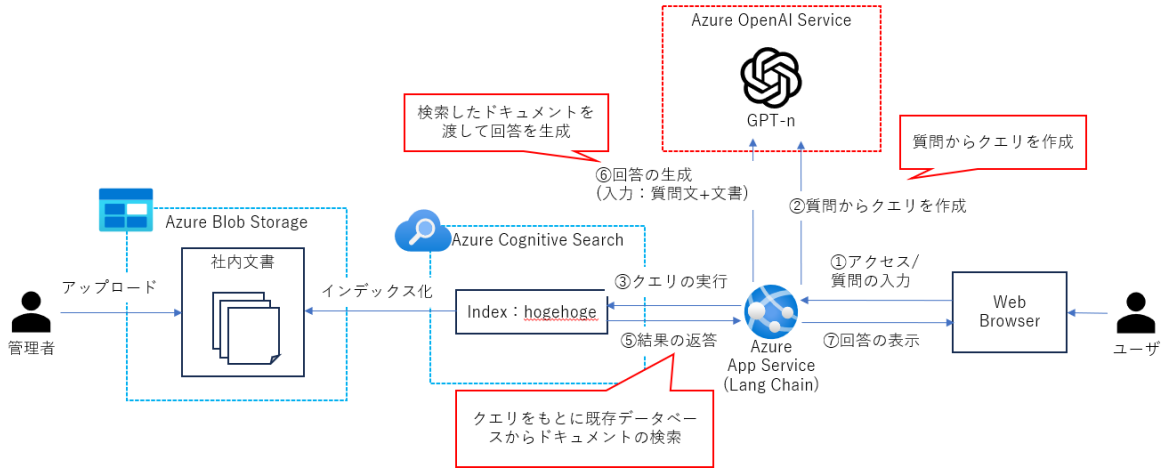
しかし、課題となるのがハルシネーションと呼ばれる現象への対策です。ハルシネーションとは，AI からの回答に事実に基づかない，いわゆる嘘が混じることです。このような事実かどうか不明瞭な回答を出力することがあるにもかかわらず、ユーザが文面上尤もらしいというだけの理由で AI の回答を信じ込む恐れがあります。また、独自サービスのデータで学習した Q&A Chatbot などでは、学習したデータで回答できない質問には対応できず、新しいデータを追加する際に学習が必要となることも懸念されます。

特定のドメイン知識を必要とする場合において、このような問題に対処する方法として、ドメイン知識に関するドキュメントを蓄積したデータベースを組み込むアーキテクチャが考えられます。質問をもとにデータベースを検索し、得られた検索結果をもとに回答を生成するので、回答を検索したドキュメントの範囲に限定することができます。これにより、ハル

シネーションが発生した場合には回答に利用したドキュメントを特定し精査することができます。また、モデルに影響を与えずに，ドキュメントを後から追加して回答を改善することができます。本稿では本構成を Azure 上で実装した場合をご紹介します。

Azure 環境の実装例

例として Azure 環境で社内ドキュメントに関する Q&A Chatbot の構築を考えます。簡単に構成すると以下のようになります。



Azure OpenAI Service では、GPT-n 等の大規模言語モデルを Azure 内の Private 空間で利用できます。検索には Azure Cognitive Search を利用します。Azure Cognitive Search は、フルマネージドの検索サービスで、クラウドストレージに保管されているデータを検索できるサービスです。データをインデックス内に格納し、インデックス内のデータを検索します。検索対象のファイルはキーと値のペアで保持され、様々なフィールドを持つことができます。上の例では Blob ストレージのコンテナに社内ドキュメントをアップロードし、インデックスを定義します。その後インデクサーと呼ばれるクローラを用いてインデックスにコンテンツをマッピングします。アプリケーションには App Service を利用します。

ユーザの入力を受け取った App Service は、質問文とプロンプトをモデルに送り、モデルが Cognitive Search Service の入力となるクエリを生成します。App Service はモデルが生成したクエリを用いて、Cognitive Search Service でドキュメントを検索します。最後に検索結果を受け取った App Service がドキュメント、質問文、プロンプトをモデルに送り、モデルが回答を生成します。

精度向上へのアプローチ

上記の構成では、ドキュメント検索の精度向上が回答の精度向上につながります。そのため、Cognitive Search においてより高精度の検索を行う必要があります。技術的な対策として、例えば、ベクトル検索[1]を実装することが考えられます。ベクトル検索は、テキストを数値データに変換し、その類似性に基づいて検索を行う方法です。これにより検索語句がドキュメントと一致しない場合でも検索クエリに似た内容のドキュメントを検索結果に含めることが期待できます。

また、回答の質を向上させるアプローチとして、Azure OpenAI Service のモデルに対するプロンプトをチューニングすること、Q&A のやり取り履歴などのデータを用いて大規模言語モデルを学習(ファインチューニング)したモデルを利用すること、GPT-4 のような、より大量のデータで事前学習したモデルを利用することなどが考えられます。

もっと気軽に試すには

Azure OpenAI Service には On Your Data(Preview)[2]という機能があり、気軽に自前のデータを入れて試すことができます。On Your Data(Preview)では、Cognitive Search Service の検索クエリのチューニングや，複数データソースを対象とした検索などは利用できませんが、簡単に試すことができるので少しだけ触ってみたいという方にはお勧めです。

おわりに

Azure 環境で大規模言語モデルを用いた AI チャットに独自データの組み込みを行う構成例を紹介しました。大規模言語モデルにドキュメントの検索サービスを接続することで、検索結果を利用した回答を生成することができます。特定のドメイン知識に関する Chatbot の構築を考えている方の一助となれば幸いです。

[1] https://learn.microsoft.com/ja-jp/azure/search/vector-search-overview
[2] https://learn.microsoft.com/ja-jp/azure/ai-services/openai/use-your-data-quickstart?tabs=command-line&pivots=programming-language-studio

GSLetterNeo Vol.182
2023 年 9 月 20 日発行
発行者　株式会社 SRA 技術本部 先端技術研究室

編集者　熊澤努　方 学芬
バックナンバー https://www.sra.co.jp/public/sra/gsletter/
お問い合わせ　gsneo@sra.co.jp

株式会社SRA
〒171-8513 東京都豊島区南池袋 2-32-8

夢を。Yawaraka Innovation
やわらかいのべーしょん