アーカイブ資料専用の対話型システムの構築

板垣　光樹

アーカイブズとは、組織または個人がその活動に伴って生み出す記録のうち、重要なものを将来のために保存する施設であり、同時に資料そのものも指す。アーカイブズには今後に活かすことのできる貴重なアーカイブ資料が数多く眠っているが、アーカイブズの利用は普及しておらず、アーカイブ資料の活用も積極的には行われていない。その理由としてアーカイブ資料は保存されることに特化していて必要な資料を探し出すのに時間がかかってしまうことが挙げられる。また、アーカイブ資料固有の特徴として資料同士の結びつきが強いとうことがある。資料同士に内容の補完性があり、結びつきがある。つまり複数の資料から得られた情報同士から新たな知識が発見できるのである。アーキビストがいないと資料同士の結び付きを発見することが困難であることもアーカイブ資料の活用が積極的に行われていない理由として挙げられる。そこでアクセス性が高く、人とシステムが対話を繰り返すことによってアーカイブ資料の資料同士の結び付きを発見できる可能性がある対話型システムが有効ではないかと考えた。本研究の目的は、アーカイブ資料専用の対話型システムのプロトタイプの構築を行うことで、アーカイブ資料資料同士の結び付きを発見しアーカイブ資料を組み合わせた知識発見が可能であることを示すことである。

本研究では、図書館情報学アーカイブを対象とし、図書館情報大学と筑波大学の統合に関係する資料と関係しそうなアーカイブ資料のみ電子化を行った。対話型システムを構築する際には、RAG（Retrieval-Augmented Generation）を用いた。RAG は、LLM（Large Language Model）の回答の際に外部情報のデータベースから取得した情報を、直接 LLM のプロンプトに追加して推論させる方法である。つまり LLM は外部情報として与えられている、電子化されたアーカイブ資料を参照にして回答を生成する。

アーカイブ資料を外部情報とする際に、大きなデータを小さなセグメントにわけるチャンキングを行う。長すぎるデータなどを適切なサイズに分割することで、検索を行う際に必要な箇所のみの読み込みを行うことができるため処理の効率があがり、関連性の高い情報を含むチャンクを抽出することによる検索性能の向上も期待できる。アーカイブ資料専用の対話型システムに適したチャンクサイズを検証するためにシステム試験を行った。それぞれのチャンクサイズに強み、弱みがあり、知りたいことに応じて適切なチャンクサイズをその場で柔軟に設定できるようにシステムの設計を行うべきであることが示された。また、資料同士の結び付きを発見するには、回答を生成する際に複数の資料を参照することと、システムが対話履歴すべてを保持し必要に応じて参照しながら回答を生成することが有効であることが示され、アーカイブ資料を組み合わせた知識発見が可能であることを示した。

（指導教員　宇陀　則彦）