LLM による SPARQL 問合せ式作成支援

井坂 志穂

近年、インターネットの普及や IoT デバイスの進化によって膨大な量のデータが生成・蓄積されており、特にグラフデータはその中で重要な役割を果たしている。グラフデータは、データ間の関係性をノード（点）とエッジ（線）で直感的に表現できるという特徴を持つ。これにより、複雑な関係性を容易に表現することができるため、幅広い分野で利用が進んでいる。しかし、グラフデータを操作・検索するには、特有のデータ構造や問合せ言語に精通する必要がある。このため、初学者にとっては、グラフデータの利用難易度が高い現状がある。

このような背景から近年、Text-to-SPARQL という手法が注目されている。これは、大規模言語モデル(LLM: Large Language Model) を利用して自然言語からグラフデータに対する SPARQL 問合せ式を生成する手法である。LLM は文章だけではなく、プログラム生成にも優れた性能を持つ。一方で、Hallucination（幻覚）と呼ばれる現象が課題として挙げられる。これはモデルが学習していない情報や誤った情報をあたかも正確であるかのように生成してしまう問題である。

しかし、これまでの Text-to-SPARQL の関連研究ではファインチューニングを行わず In-context Learning のみの SPARQL 問合せ式生成は困難である、事前に与えるべき情報量が増加してしまう等の問題があった。そのため、本研究では少ない情報量の提示で正確な生成結果を獲得できることを目的に新たな Text-to-SPARQL 手法について考察する。より具体的には、SP^{2}Bench の提供データセットを対象とし、LLM の生成する Hallucination やデータ量に対応するべく以下の手法を採用した。まず、ShEx(Shape Expression Schema)スキーマを LLM を用いて生成した。ここで、ShEx はグラフデータのスキーマ言語であり、データの構造を簡潔に表現することができる。次に、生成した ShEx スキーマと手動作成した最低限のデータを用いて GPT-4o のファインチューニングを行い、GPT-4o の新たなモデルを作成した。このように作成したモデルを用いることにより、自然言語での質問から性能の良い SPARQL 問合せ式を生成できるようになる。評価実験の結果、ファインチューニングした対話例と類似性が低い問合せに対しての生成精度は低いものの、類似性が見られる問合せに対しては大きな精度向上が見られた。また、ShEx スキーマの導入、ファインチューニングの実行の両者でそれぞれ性能向上が見られた。

（指導教員　鈴木　伸崇）