Étude de l'utilisation des modèles de langages pour l'interrogation en langue naturelle des graphes de connaissances

Alexandra Épiphanie Padonou, Peggy Cellier, Sébastien Ferré

Univ Rennes, INSA, CNRS, Inria, IRISA - UMR 6074
Campus de Beaulieu, 35042 Rennes cedex, France
alexandra-epiphanie.padonou@etudiant.univ-rennes.fr - ferre@irisa.fr - cellier@irisa.fr

Résumé. Dans cet article, nous présentons les résultats d'une étude approfondie des performances des grands modèles de langage (LLM) dans le contexte de l'interrogation des graphes de connaissances en langue naturelle (KGQA). La méthodologie de l'expérimentation a été structurée en deux approches distinctes : la génération de requêtes SPARQL et l'interrogation directe. Les résultats sur le benchmark QALD-10 ont révélé des performances très faibles dans la première approche et des performances correctes dans la deuxième, avec des variations importantes selon le type des questions-réponses.

1 Introduction

L'émergence des grands modèles de langages (LLM) (Vaswani et al., 2017), tels que ChatGPT, a considérablement transformé le domaine du traitement automatique de la langue, ouvrant de nouvelles perspectives pour toutes sortes de tâches. La tâche à laquelle nous nous intéressons dans cette étude est l'interrogation en langue naturelle des graphes de connaissances (ex., Wikidata), appelée en anglais *Question-Answering over Knowledge Graphs* (KGQA) (Diefenbach et al., 2017). Cette tâche est cruciale pour l'accès à ces sources de connaissances et particulièrement difficile car elle nécessite de faire le pont entre la langue naturelle et les langages formels du Web sémantique : RDF, et SPARQL. Le benchmark QALD (*Question Answering over Linked Data*) (Usbeck et al., 2023) est couramment utilisé pour comparer différentes approches. Dans sa 10e et dernière édition (2022), la meilleure approche a obtenu un score F1 de 45.4%. Depuis l'arrivée des LLM et de ChatGPT, peu de travaux ont à ce jour étudié les performances des LLM pour la tâche KGQA (Klager et Polleres, 2023; Faria et al., 2023). Ces travaux utilisent entre autres les questions de QALD et ils comparent deux formes d'utilisation d'un LLM : (a) par génération d'une requête SPARQL puis évaluation de cette requête sur le graphe de connaissance et (b) par interrogation directe du LLM, sans passer par le graphe de connaissances.

Comparé à ces travaux, notre étude porte sur les 393 questions test de la dernière version de QALD, plutôt que sur d'anciennes versions et/ou de petits échantillons de questions. Pour cela, nous automatisons la comparaison des réponses obtenues et des réponses attendues en