Facilitating Search of the Virtual Record Treasury of Ireland Knowledge Graph using ChatGPT

Alex Randles[1]*, Lucy McKenna[1], Lynn Kilgallon[2], Beyza Yaman[1], Peter Crooks[2] and Declan O'Sullivan[1]

[1] *ADAPT Centre for Digital Content, Trinity College Dublin, Dublin, Ireland*
[2] *Department of History, Trinity College Dublin, Dublin, Ireland*

Abstract

The Virtual Record Treasury of Ireland (VRTI) is an initiative to digitally recreate the contents of the Irish central archive which was destroyed during the Civil War. The project has created a Knowledge Graph (KG) to facilitate information discovery and reasoning over the recovered items. However, complex queries must be created to retrieve data in the KG, which require a high level of technical expertise. In this paper, we explore the application of Large Language Models (LLMs) to facilitate searching of the VRTI-KG by users who lack this technical expertise and to decrease workload for those who do not. The VRTI-ChatGPT framework is proposed which uses ChatGPT to interpret requests from users and to facilitate the creation of queries which can be executed on the KG.

Keywords

KG Search, User Interface, ChatGPT

1. Introduction

The Virtual Record Treasury of Ireland (VRTI) [1–3] is a state-funded programme hosted at Trinity College Dublin. The VRTI began with the objective of digitally reconstructing archival records destroyed during the 1922 Irish Civil war [2]. A fire during the war destroyed the Public Record Office of Ireland, which damaged records dating back more than 700 years. The staff at the time spent months to recover documents which were recreated a century later. The initial VRTI Knowledge Graph (KG) was created as a result of the lead project named Beyond 2022. The VRTI-KG contains notable information about people, places, roles, organisations and their interconnections from Irish history. Representing this information using a KG allows the integration of heterogenous source data formats and supports reasoning and inference of the data and applied to a range

of scenarios already from event based [4] networks to more recently climate action related applications [5]. The KG was implemented using RDF, which means data must be retrieved using SPARQL [6] queries. Creating these queries is time-consuming and requires an understanding of the SPARQL query language and structure of relevant schemas. Many of the historians who would interact with the VRTI-KG do not possess the technical expertise to create these complex queries. Large Language Models (LLMs) such as ChatGPT [7] provide functionality which could allow the data in the VRTI-KG to be easily searched and the results presented using natural language. It was decided to use ChatGPT in the proposed approach as it provided the best results in early experimentation. With the emergence of generative AI, we are interested in exploring what benefits it can have for the VRTI-KG system [1]. However, it is important to ensure that the proposed application of generative AI to the VRTI-KG is constrained to information only contained in the VRTI and does not pollute responses with external information on the requested topic. In order to explore how generative AI could be applied, we propose the **VRTI-ChatGPT framework** which was designed to facilitate searching of the VRTI-KG through natural language questions and answers. The framework uses strict prompt templates to interact with ChatGPT in order to process the users input and form sentences from KG query results. A recent survey [8] has highlighted the importance of providing straightforward interaction between semantic interfaces and respective domain experts. The survey compared 28 interfaces based on interaction paradigm, information being displayed, and strategies used to improve the understanding of information. The survey concluded that many of these approaches still require some level of technical expertise to be used effectively, which some domain experts may lack. It is hoped natural language interaction can bridge the gap between domain experts (historians) and diverse data held in VRTI-KG. An existing tool designed for KG natural language querying by Ontotext[1] was experimented with before deciding to create a bespoke solution. The tool uses LLMs to create SPARQL queries from a provided ontology and natural language question. The endpoint of the VRTI-KG and ontology were provided to the tool, however, it was observed that it struggled to create syntactically correct queries for most test cases. The incorrect queries could be a result of the complex structure of the VRTI ontology. Using an approach involving query templates ensures that the query created from the natural language is syntactically correct and retrieves all of the required information to provide a sufficient response. The query templates used in the framework are configurable which is hoped to allow the approach to be customised for other KGs in future. Early observations from the historians in the VRTI has been positive when the involved prompts are strictly constrained so that ChatGPT does not make inference on the provided information.

2. VRTI-ChatGPT framework

This section discusses the design and implementation of the VRTI-ChatGPT framework. The implementation of the framework is **available online**[2]. The framework is configurable to

---

[1] https://www.ontotext.com/blog/natural-language-querying-of-graphdb-in-langchain/
[2] https://vrti-graph.adaptcentre.ie/gpt-search

allow changes in the VRTI-KG to be easily synchronised with the involved prompts and queries. **Figure 1** presents an overview of the activities of the framework.
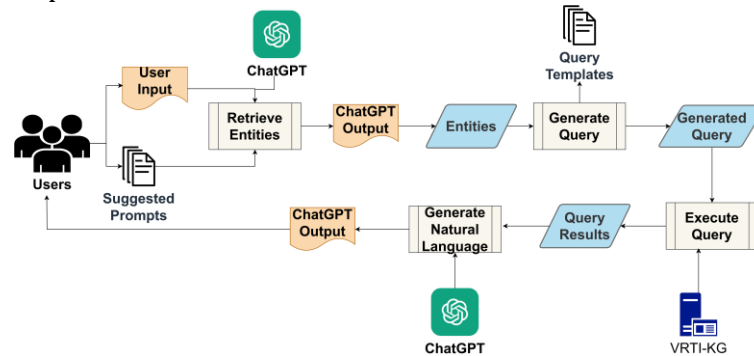

Figure 1: Workflow of the VRTI-ChatGPT framework

Several Python libraries[3] were used to implement the framework. Flask is a customizable web framework which was used to create the web application. SPARQLWrapper is used to execute queries on the endpoint of the VRTI-KG. The Open-AI library is used to communicate with ChatGPT 4.0[4]. Python string formatting are used to create prompts and queries from the templates. **Figure 2** presents search results displayed on the implementation.

Figure 2: Screenshot of implementation displaying results for "Tell me about Michael Collins"

Initial Processing of Users Input. First, the user enters a question into the search bar (A –Figure 2) or selects a suggested question. For instance, a question could ask about a specific (<person>) such as "Tell me about <person>?" ", "Where and when was <person> born", "Was <person> in the army?" or "What job did <person> have?". For the **running example**, the user inputs "Tell me about Michael Collins". Michael Collins[5] is a notable Irish person who was involved in the Irish civil war. The question is inserted in a prompt template which extracts the name of people and places from the user's input. The generated prompt is "Extract the names of people and places in this text 'Tell me about Michael Collins' and output the result into a JSON dictionary". Then, the prompt is fed into ChatGPT 4.0 using a request carried out by the OpenAI library. The result is a dictionary containing key-value pairs of names of people and places which is stored in memory.

Creation of SPARQL query. The extracted entity ("Michael Collins") is inserted into a SPARQL [6] query template[6] defined in the configuration file. The insertion involves translating the key-value pairs from the created JSON dictionary into FILTER conditions (FILTER CONTAINS(?Name, "Michael Collins")) using string formatting methods. The query retrieves resources with a matching name along with their related properties, such as birth date and place. Thereafter, the query is executed on the VRTI-KG using the SPARQLWrapper library to retrieve matching resources. The query results are represented in dictionary format.

Creation of Natural Language Response. The dictionary containing the query result is inserted into a prompt template to generate the natural language answer. For this example, the generated prompt is "Answer this question 'Tell me about Michael Collins' using only the information in this dictionary '{Person: <....>, Occupation: <...>, BirthDate: "..."}'. Do not include any external information in the answer."). The prompt template is designed to constrain ChatGPT to use only the information from the query results from the VRTI-KG rather than external information it has on the topic. The response (B – Figure 2) from ChatGPT is then displayed on the interface. In addition, the URI of each resource returned from the query are presented in a tabular format (C – Figure 2), which allows further exploration with the application.

3. Future Work and Conclusion

Future work includes usability testing of the framework with a cohort of historians. The testing will allow the user requirements to be refined and validated. The testing will involve the participants interacting with the framework to complete several tasks which mimic the expected user interaction. In addition, it is hoped to configure the framework to answer questions from information stored in other KGs.

The VRTI-ChatGPT framework proposed in this paper provides possible direction for the integration of generative AI, such as LLMs in the VRTI-KG system [1]. It is hoped the proposed approach can facilitate searching by users who lack relevant technical expertise. Thus, reducing workload and improving the uptake of information by domain experts.

---

[5] https://kb.virtualtreasury.ie/person/Collins_Michael_c20_dib_a1860
[6] https://github.com/alex-randles/VRTI-ChatGPT/blob/main/sample-query.rq

Finally, the prompts used by the framework are hoped to provide guidance for researchers who propose similar approaches.

Acknowledgements

References

[1]   B. Yaman, L. McKenna, A. Randles, L. Kilgallon, P. Crooks, D. O'Sullivan, Digital Prosopography Information in Virtual Record Treasury of Ireland Knowledge Graph, in: Proceedings of the 1st International Workshop of Semantic Digital Humanities (SemDH) Co-Located with the 21st Extended Semantic Web Conference, 2024. https://ceur-ws.org/Vol-3724/paper2.pdf.

[2]   P. Crooks, Z. Reid, C. Wallace, The Virtual Record Treasury of Ireland: A century of Recovery from the 1922 Four Courts Blaze - and Beyond, Hist Irel 30 (2022) 38–41.

[3]   C. Debruyne, G. Munnelly, L. Kilgallon, D. O'Sullivan, P. Crooks, Creating a Knowledge Graph for Ireland's Lost History: Knowledge Engineering and Curation in the Beyond 2022 Project, J. Comput. Cult. Herit. 15 (2022). https://doi.org/10.1145/3474829.

[4]   J. Keeney, D. Roblek, D. Jones, D. Lewis, D. O'Sullivan, Extending Siena to support more expressive and flexible subscriptions, in: Proceedings of the Second International Conference on Distributed Event-Based Systems, Association for Computing Machinery, New York, NY, USA, 2008: pp. 35–46. https://doi.org/10.1145/1385989.1385995.

[5]   J. Wu, F. Orlandi, D. O'Sullivan, S. Dev, An ontology model for climatic data analysis, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021: pp. 5739–5742.

[6]   S. Harris, A. Seaborne, E. Prud'hommeaux, SPARQL 1.1 Query Language, World Wide Web Consortium (W3C) Recommendation 21 (2013) 778. https://www.w3.org/TR/sparql11-query/ (accessed April 1, 2023).

[7]   T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, Y. Tang, A brief overview of ChatGPT: The history, status quo and potential future development, IEEE/CAA Journal of Automatica Sinica 10 (2023) 1122–1136.

[8]   E. Bernasconi, C. Miguel, M. Mecella, Linked Data interfaces: a survey, in: 19th Conference on Information and Research Science Connecting to Digital and Library Science, 2023: pp. 1–16.