TCRMQ: An Application for Generating CIDOC-CRM SPARQL Queries from Text by Using Large Language Models

by Michalis Mountantonakis and Yannis Tzitzikas (FORTH-ICS and University of Crete)

TCRMQ (Text-2-CIDOC-CRM Query) is a research prototype that can generate SPARQL queries over CIDOC-CRM Knowledge Graphs using a novel two-stage ontology path pattern method and Large Language Models (LLMs). The generated SPARQL queries can include in the filtering conditions information that are not even part of the KG, by exploiting knowledge from the LLM.

There is a strong need for exploiting Artificial Intelligence (AI) techniques, to enhance the accessibility and reusability of Knowledge Graphs (KGs) of Cultural Heritage domain, which have been modelled through the CIDOC-CRM ISO Standard [1]. Hence, a relevant task is to enable the answering of natural questions over CIDOC-CRM based KGs, by generating SPARQL queries from a given natural question. This is a highly challenging task, especially for sophisticated (event-based) ontologies like CIDOC-CRM, since in most cases one has to derive queries with long property path expressions to answer a given question [2]. To tackle this difficulty, we present the research prototype TCRMQ (i.e., Text-2-CIDOC-CRM Query) [L1], which is an application that can generate SPARQL queries over CIDOC-CRM KGs, by exploiting a novel two-stage ontology path pattern method that combines knowledge from a given KG and from Large Language Models (LLMs) [2].

Regarding the functionality, as we can see in Figure 1, TCRMQ receives a question from the user in natural language, such as "Give me the number of artworks for each artist born after the end of the first World War" and then the user selects a method and the desired KG to answer the given question. The application uses a novel two-stage method combining Ontology Path Patterns and Knowledge from LLMs [2] for generating the SPARQL query. First, the method sends a prompt with candidate ontology path patterns and receives a prediction with the most relevant path patterns for answering the question. Then, a second prompt is sent to the LLM, for generating the SPARQL query for the given question, based on the predictions. For instance, in Fig. 1, TCRMQ managed to predict the long property path that connects an artwork with the corresponding artist and his/her birth date, and then it generated the corresponding multi-hop SPARQL query. Subsequently, the generated SPARQL query (see the bottom-left part of Figure 1) is sent to the SPARQL endpoint of the desired KG and the application retrieves the answer and presents the results of the query to the user (see the right part of Figure 1).

A very interesting finding regarding the synergy between LLMs and KGs is that LLMs "know" background knowledge (dates, names, etc.) that is not present in the KG, and this synergy can yield queries with the right filtering conditions, even if such information is not present in the KG. For example, in Figure 1, the KG did not contain any information about the dates of the First World War. However, LLMs (such as ChatGPT) include this relevant information and, in our case, the LLM added the desired date in the filtering condition of the SPARQL Query (i.e., "1918-11-11"). Another functionality of the TCRMQ is that the user can (i) compare several methods (based on different prompts) for generating the SPARQL

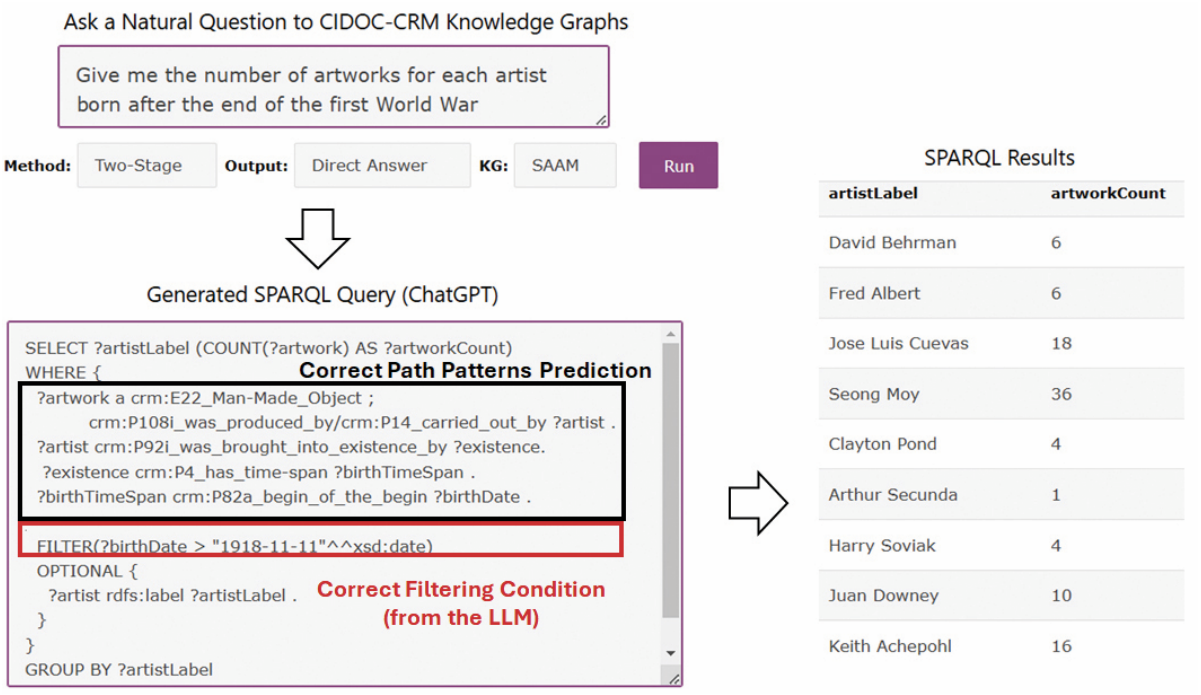

Figure 1: The functionality of the TCRMQ research prototype.

query, (ii) use different ChatGPT versions (currently it supports versions 3.5, 4-mini and 4), (iii) ask a question in languages other than English, and iv) edit the generated SPARQL query, e.g., to restrict the results and/or to correct possible errors in the generated query.

Concerning evaluation results, we have tested the methods of TCRMQ by using a benchmark [L2] with 100 questions over two real CIDOC-CRM based KGs including data about artworks. The benchmark consists of five categories of questions, i.e., each category includes 20 questions requiring a path with length a) 1, b) 2, c) 3, d) 4 and e) mixed paths (combining paths of different length). The evaluation was performed by using the ChatGPT version 3.5. Regarding the results, we managed to generate the correct SPARQL query for 66 out of the 100 questions [2], by using the two-stage method, whereas the corresponding accuracy for the baseline method was 19 out of 100 questions. As regards the different categories of questions, the accuracy was lower for questions requiring following either a path of length 4 or a mixed path, compared to the other categories. In the most erroneous cases, either a totally incorrect path or a longer path was selected. All the details for the benchmark and the evaluation can be found in the paper [2] and on the GitHub page [L2].

As future work, we plan to evaluate the methods by also using newer versions of ChatGPT, by also expressing the questions in other languages, to add more KGs in an automatic way (including also other KGs using event-based models) and to perform an analysis of the path patterns of all the available CIDOC-CRM KGs.

Links:

[L1] https://demos.isl.ics.forth.gr/Text2CIDOC/

[L2] https://github.com/mountanton/CIDOC-QA-using-LLMs

References:

[1] Y. Tzitzikas, et al., "CIDOC-CRM and machine learning: a survey and future research", Heritage, 5(3), pp.1612-1636, 2022.

[2] M. Mountantonakis and Y. Tzitzikas, "Generating SPARQL Queries over CIDOC-CRM using a Two-Stage Ontology Path Patterns Method in LLM Prompts", ACM Journal on Computing and Cultural Heritage, 18, 1, Article 12, March 2025, 20 pages. https://doi.org/10.1145/3708326

Please contact:

Michalis Mountantonakis
FORTH-ICS and University of Crete, Greece
mountant@ics.forth.gr

Yannis Tzitzikas
FORTH-ICS and University of Crete, Greece
tzitzik@ics.forth.gr

AI Multilingual Search Platform for EU Audiovisual Media Archives

by Pilar Orero, Chiara Gunella, and Sarah McDonagh (Universitat Autònoma de Barcelona)

Culture and linguistic diversity in Europe is both a jewel and a barrier to communication. Broadcaster's archives are part of the EU cultural heritage –inaccessible for many reasons. The MOSAIC project aims to develop tools for multilingual translation, automatic subtitling, and AI-driven content adaptation. MOSAIC seeks to empower broadcasters and media producers by making their content available to a wider audience, enhancing cultural exchange and unity across Europe.

European broadcasters, news agencies, and corporations are known for high-qulitymedia content and information production, distribution, and consumption and for the richness of their content archives, containing a wealth of cultural, political, historical, and artistic content in various formats, including film/video, television, radio, and digital media. Nevertheless, digital platforms outside Europe already exert control over the media landscape, leading to a fragmentation of accessible knowledge within the European Union. The potential of Europe's richness remains underused with a serious risk that Europe will lose valuable resources and archival material. In this scenario, the adoption of artificial intelligence (AI) technologies by organisations, and thus the creation of media content using advanced techniques, is essential not only for the cultural sovereignty of Europe but also for the preservation and presentation of heritage, education, and entertainment across a European context. Creating a more unified European media identity, based on a fragmented but common cultural heritage, is a major challenge: on one side, both public and private broadcasters have to ensure that national regulatory authorities are independent of political or commercial influence to ensure media freedom and pluralism, on the other side, the common challenge is the constant need to adapt to new technologies to stay relevant and to have the chance to make the European media market competitive worldwide and European media content archives an unparalleled source of cultural richness.

Artificial Intelligence (AI), Natural Language Processing (NLP), Natural Language Understanding (NLU), Language Technologies (LTs), and Speech Technologies(STs) have the potential to enable multilingualism technologically but, according to the META-NET White Paper Series "Europe's Languages in the Digital Age" [1] published in 2012, our languages suffer from an extreme imbalance in terms of technological support: English is well supported through technologies, tools, and datasets, but languages such as Maltese, Estonian or Icelandic still have very poor support.

The goal is to enable multilingualism technologically since "the EU and its institutions have a duty to enhance, promote and uphold linguistic diversity in Europe" (European Parliament 2018) [2]. Today, Generative AI (GenAI) can cre-