LLMs in a Knowledge Graph Pipeline: From Knowledge Extraction to SPARQL Querying
The Case of FashionDB

Teresa Liberatore[1][0009−0001−8602−8957]

INDELab, University of Amsterdam t.liberatore@uva.nl

Keywords: Knowledge Graph Construction · Knowledge Extraction · Text to SPARQL · Fashion Knowledge Graph · Knowledge Graphs Question Answering · Large Language Model

1 Introduction

The web hosts an abundance of textual information, with platforms like Wikipedia indexing knowledge through article titles and hyperlinks. While text is a rich source of information, extracting specific, structured, and aggregated insights from it remains a challenge. Traditional keyword-based search and manual extraction processes are time-consuming and lack the flexibility of structured queries. On the other hand, structured data formats, such as knowledge graphs (KGs), provide an efficient means of organizing and retrieving information, supporting advanced query capabilities that enable complex and aggregated retrieval.

However, two main challenges often hinder the usability of knowledge graphs:

1. Construction – Extracting structured data from unstructured text requires a scalable and automated approach.
2. Interaction – Querying structured databases typically requires knowledge of SPARQL, a barrier for many domain experts unfamiliar with knowledge graph query languages.

In this work, we present FashionDB, a domain-specific knowledge graph for fashion, encompassing data on fashion designers, fashion houses, and fashion collections. While information about fashion is widely available in sources like Wikipedia, Vogue, and The Fashion Model Directory, it is underrepresented in structured datasets. Building FashionDB required overcoming the two major challenges of knowledge graph construction and interaction, which we addressed by integrating large language models (LLMs) into both processes.

We describe a pipeline for automated KG construction and interaction, demonstrating how LLMs can bridge the gap between unstructured textual data and structured knowledge retrieval. We believe this framework can be generalized to other domains where text alone is insufficient for complex and aggregated retrieval, and structured representations are essential for accurate and meaningful insights.

2 Constructing FashionDB

The fashion industry is inherently dynamic and temporally structured, with designers moving across different fashion houses, collections evolving over time, and trends influencing multiple creative directions. To model these aspects, we develop a specialized ontology focusing on temporal relations, that extends Wikidata properties to represent fashion designer career trajectories, the evolution of fashion houses and characteristics of fashion collections.

To automatically populate FashionDB, we leverage LLMs for knowledge extraction, structuring unstructured textual data according to our ontology. Recent advances in pre-trained language models have demonstrated state-of-the-art performance in knowledge extraction [7]. In particular, in-context learning and instruction-following capabilities in LLMs have improved performance on ontology-guided extraction and temporal information retrieval [3, 4].

Inspired by these advancements, we use LLMs to extract structured facts along with their temporal attributes, ensuring that FashionDB not only stores factual information but also captures historical evolution and context.

3 Interacting with FashionDB

While FashionDB enables powerful SPARQL queries, formulating queries remains a significant barrier for users unfamiliar with SPARQL or knowledge graph structures. Recent research in text-to-SPARQL translation has explored ways to bridge this gap by leveraging LLMs.

For instance, Auto-KGQAGPT [6] proposes a fragment selection mechanism, where relevant subgraphs of the knowledge base are retrieved and provided to an LLM to generate the corresponding SPARQL query. Similarly, SPARQLGEN [2] introduces a multi-source prompting strategy, incorporating the user's natural language question, an RDF subgraph, and an example SPARQL query from a different question to improve accuracy. However, these approaches rely on large, closed-source LLMs, which pose challenges in terms of scalability, interpretability, and accessibility. As suggested in [1], smaller, open-source models can offer a more efficient and domain-adapted solution for text-to-SPARQL tasks.

Building on these insights, we propose a domain-specific text-to-SPARQL pipeline based on Gemma 2 [5], an open-source model by Google. Instead of relying on knowledge graph fragments in the prompt—an approach suited for general-purpose knowledge bases—our method leverages the structured nature of FashionDB's ontology. By providing the model with a controlled vocabulary of FashionDB properties (IDs and labels) and relevant classes, we ensure that queries remain accurate and schema-compliant. To further refine the translation process, we annotate a small dataset of natural language questions paired with their corresponding SPARQL queries. When a user submits a new query, we retrieve the most semantically similar question-query pairs using cosine similarity over text embeddings, incorporating them into the prompt to improve translation accuracy.

References

1. Brei, F., Frey, J., Meyer, L.P.: Leveraging small language models for text2sparql tasks to improve the resilience of ai assistance (5 2024), http://arxiv.org/abs/2405.17076
2. Kovriguina, L., Teucher, R., Radyush, D., Mouromtsev, D.: Sparqlgen: One-shot prompt-based approach for sparql query generation (2023), https://github.com/danrd/sparqlgen
3. Mihindukulasooriya, N., Tiwari, S., Enguix, C.F., Lata, K.: Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In: International Semantic Web Conference. pp. 247–265. Springer (2023)
4. Polat, F., Tiddi, I., Groth, P.: Testing prompt engineering methods for knowledge extraction from text. Semantic Web. Under Review (2024)
5. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C.L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C.A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J.P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L.L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L.B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R.A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S.M., Cogan, S., Perrin, S., Arnold, S.M.R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., Andreev, A.: Gemma 2: Improving open language models at a practical size (2024), https://arxiv.org/abs/2408.00118
6. Viktor, C., Avila, S., Vidal, V.M.P., Franco, W., Casanova, M.A.: Experiments with text-to-sparql based on chatgpt . https://doi.org/10.1109/ICSC59802.2024.00050, https://bit.ly/41qfjGM
7. Whitehouse, C., Vania, C., Aji, A.F., Christodoulopoulos, C., Pierleoni, A.: WebIE: Faithful and robust information extraction on the web. In: Rogers, A., Boyd-

Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7734–7755. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.428, https://aclanthology.org/2023.acl-long.428