Integrating SPARQL and LLMs for Question Answering over Scholarly Data Sources*

Fomubad Borista Fondi[1], Azanzi Jiomekong[1,2,*,†]

[1]Department of Computer Science, University of Yaounde I, Yaounde, Cameroon
[2]TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract
The Scholarly Hybrid Question Answering over Linked Data (QALD) Challenge at International Semantic Web Conference (ISWC) 2024 focuses on Question Answering (QA) over diverse scholarly sources: DBLP, SemOpenAlex, and Wikipedia-based texts. This paper describes a methodology that combines SPARQL queries, divide and conquer algorithms, and BERT-based-case-SQuad2 predictions. It starts with SPARQL queries to gather data, then applies divide and conquer to manage various question types and sources, and uses BERT to handle personal author questions. The approach, evaluated with Exact Match and F-score metrics, shows promise for improving QA accuracy and efficiency in scholarly contexts.
Keywords: Scholarly Question Answering, Large Language Models, Divide and conquer.

1. Introduction

The Scholarly Hybrid Question Answering over Linked Data (QALD) aims to answer questions in scholarly publications provided in natural language [6]. A challenge on Question Answering over Linked Data (QALD) which is hosted at the International Semantic Web Conference(ISWC) 2024 [5] since 2023. The 2024 edition is devoted to the development of question answering (QA) systems capable of integrating and querying information from three distinct but interconnected sources: DBLP Knowledge Graph[1], SemOpenAlex Knowledge Graph[2], and Wikipedia-based texts[3].

1. **DBLP Knowledge Graph** is a comprehensive dataset documenting research publications, authors, and affiliations.
2. **SemOpenAlex Knowledge Graph** is an extensive KG containing detailed information about authors, institutions, and publications.
3. **Wikipedia-Based Scholarly Text** is composed of textual data derived from Wikipedia, offering supplementary information on scholarly topics.

The primary objective of this paper is to describe the methodology employed in addressing the Scholarly Hybrid QALD Challenge. This includes detailing the integration of SPARQL queries across different KGs [2], the application of divide-and-conquer algorithms [9], and the utilization of BERT [10] to improve response accuracy. To assess this methodology, the dataset provided by the organisers was used. This dataset was composed of training set and test set. The training set was composed of 5000 questions along with their answers, while the test set was composed of 702 questions. The approach proposed in this paper shows promising results for this challenge.

The rest of the paper is organised as follows: Section 2 is the detailed methodology used, the Section 3 presents the results and the Section 4 conclude this work.

Woodstock'24: Symposium on the irreproducible science, June 07–11, 2024, Woodstock, NY
borista.fomubad@facsciences-uy1.cm (F. B. Fondi); fidel.jiomekong@facsciences-uy1.cm (A. Jiomekong)
0009-0005-9448-7722 (F. B. Fondi); 0000-0002-8005-2067 (A. Jiomekong)

[1]https://dblp-april24.skynet.coypu.org/sparql
[2]https://semoa.skynet.coypu.org/sparql
[3]https://drive.google.com/file/d/1ISxvb4q1TxcYRDWlyG-KalInSOeZqpyI/view?usp=drive_link

2. Methodology

To address the Scholarly Hybrid Question Answering over Linked Data (QALD) Challenge, we adopted a multi-step approach combining natural language processing techniques for data processing, SPARQL queries, divide and conquer algorithms, and LLM-based predictions. This methodology is designed to efficiently handle the complexity of integrating information from multiple sources and producing accurate answers for a given set of questions. Fig. 1 provides an overview of the methodology pipeline used in this work. This figure illustrates the main steps involved in processing the data, executing queries, applying LLM-based predictions, generating answers, and refining them.

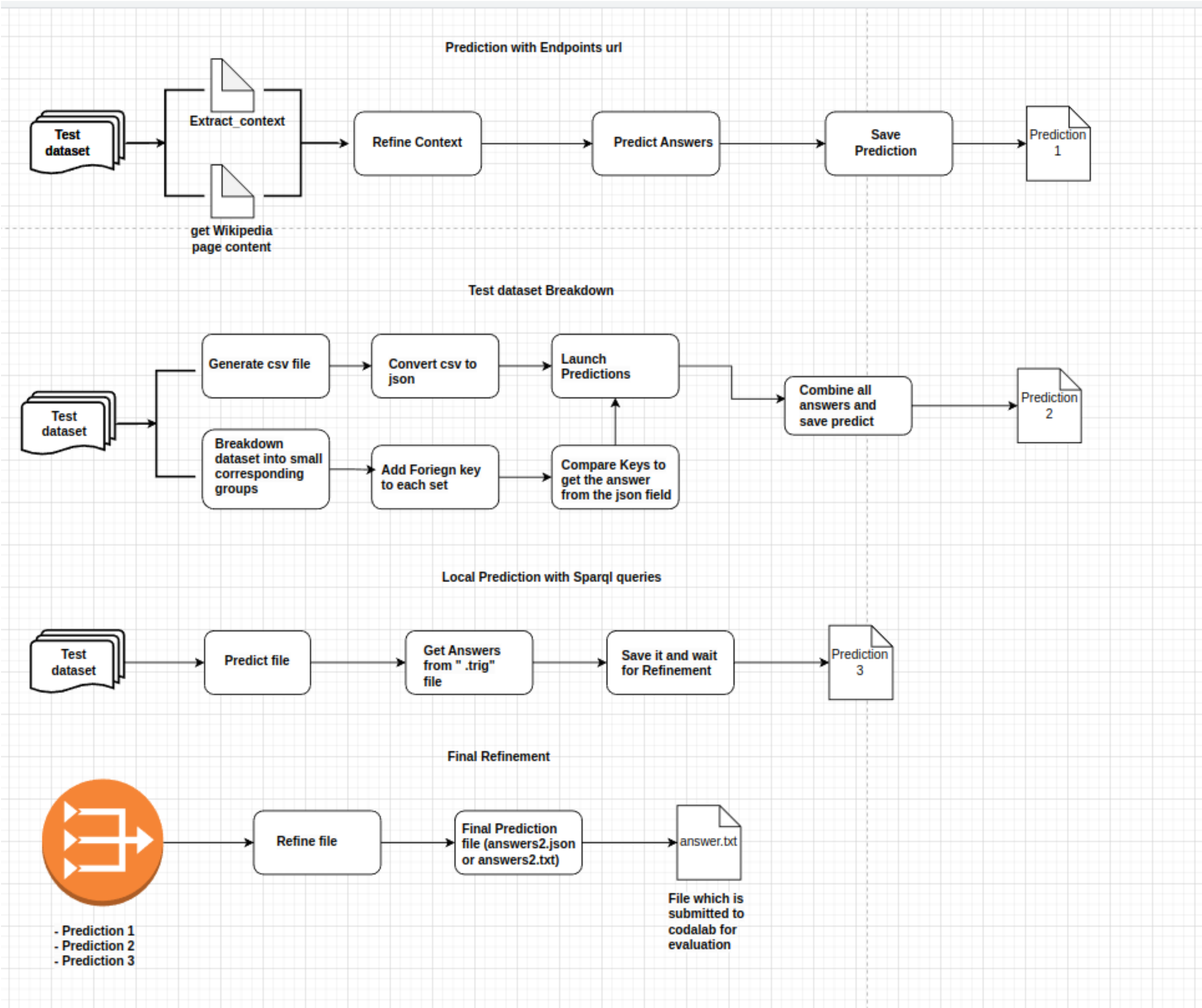

Figure 1: Methodology pipeline for the Scholarly Hybrid QALD Challenge.

2.1. Data Processing and Query Execution

The process began with executing a general script containing SPARQL queries against SemOpenAlex for both authors and institutions. This step involved querying the `semoa_authors.trig` file and the `institution-semopenalex.trig` dataset from October 2023 locally. The query execution took approximately 62-65 hours to complete due to the size and complexity of the datasets. In some cases, it filled up the RAM and returned "**Killed**".

Fig. 2 provides an overview of the data processing outputs from the the various KGs given. The data processing involved cleaning the dataset to remove noise such as un-added parts, mis-matched names, Names which are not well spelled, and alot more and irrelevant information such as None-useful links, . The questions were subjected to a thorough, individual examination , and keywords from the questions were noted. An example is provided by the Equation 1 . The following points presents the different steps of the data processing:

1. the datasets was transformed in alphabetical order of the questions. This allow us to see which questions had a similar structure/pattern. For questions that had a similar structure, we found that the responses were usually in the same position of the KG.
2. Look for the names returned by the *author_dblp_uri* and compare the results found in the endpoint and the results found in the Knowledge Graph (semoa_authors.trig file).

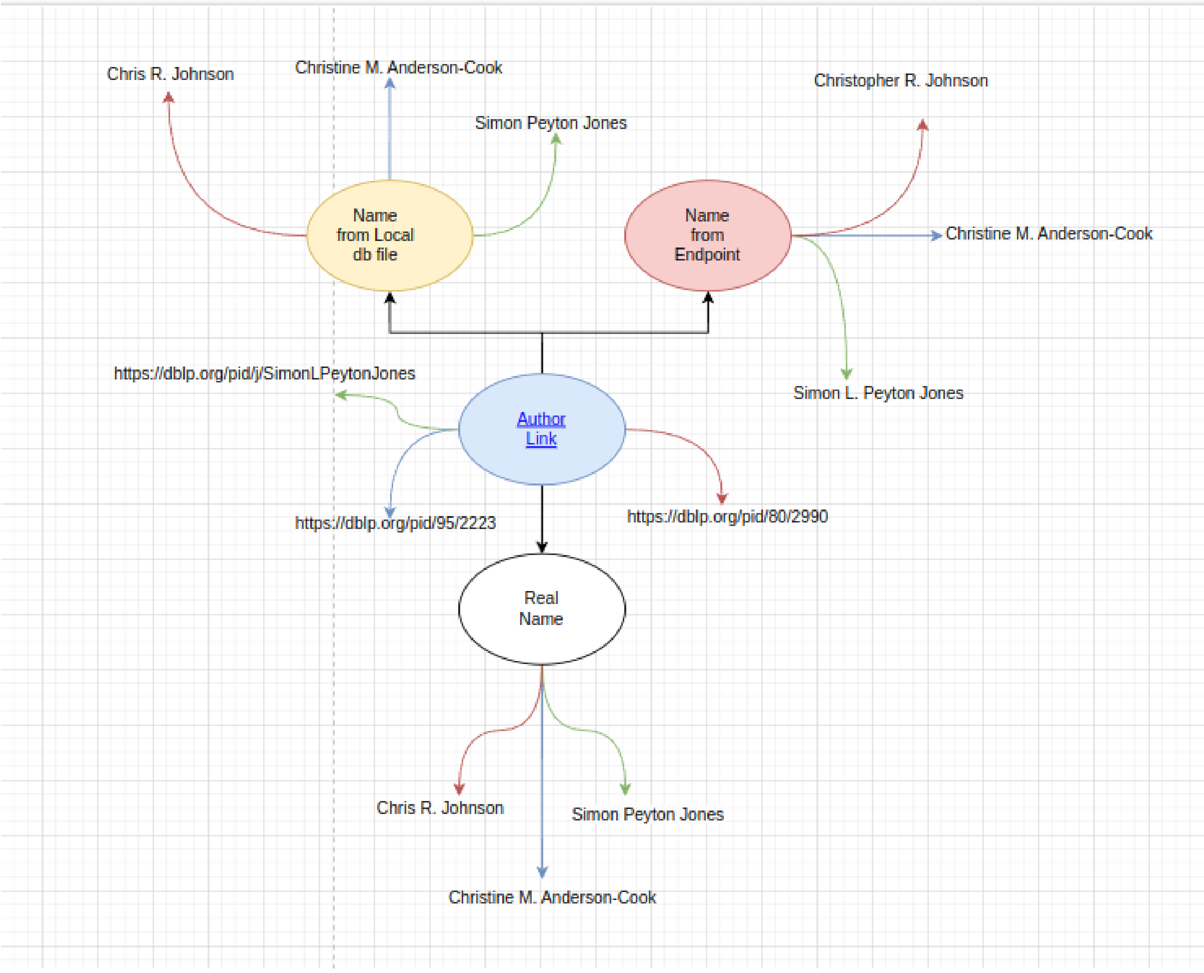

Figure 2: Data processing and cleaning.

$$M = \text{Breakdown sets}(DPR) = \begin{cases} \text{List\_author\_dblp\_uri}_{\text{set}} \\ \text{Authors}_{\text{set}} \\ \text{institution}_{\text{set}} \\ \text{hIndex}_{\text{set}} \\ \text{i10index}_{\text{set}} \\ \text{acronym}_{\text{set}} \\ \text{etc} \end{cases} \tag{1}$$

2.2. Divide and Conquer Approach

To manage the diverse nature of the questions and data, we implemented a divide and conquer strategy:

1. **Initial Data Breakdown:** We first segmented the test data based on whether the `author_dblp_uri` contained multiple links or a single link. This allowed us to address questions with multiple author identifiers separately from those with single identifiers.
2. **Further Segmentation:** Questions were further classified into those concerning individual authors and those about authors' institutions. Keywords like "Organizations," "Affiliations," "institution," etc. were used to automate this classification.
3. **Detailed Sub-Classification:** Questions about authors were subdivided into more specific queries based on the information sought (e.g., publication details, affiliations).

2.3. Data Retrieval and Aggregation

We employed a script to generate a CSV file containing all potential responses for each question by querying the endpoints provided on the challenge's website. This CSV file included detailed author information such as names, publication counts, and institutional affiliations.

1. **CSV to JSON Conversion:** The CSV file was converted to JSON format. Duplicate entries, resulting from multiple author names, were removed to ensure a clean dataset.
2. **Merging Results:** The JSON file was then used to cross-reference and extract answers for each specific question. The answers were aggregated and merged to create a comprehensive set of responses.
3. **Final Refinement:** The merged results were refined by integrating them with the initial general predictions and LLM-generated responses to ensure accuracy and completeness. This final step resulted in the creation of the `answers2.txt` file submitted for evaluation.

2.4. Large Language Model-Based Predictions

The LLM used in this challenge was BERT-base-cased-squad2, a pretrained non-finetuned model download from Hugging Face[4], which was used for Question Answering tasks.

After executing SPARQL queries, we used the BERT-based model `bert-base-cased-squad2` to predict responses to personal questions about authors. The context for these predictions was generated from the results of the SPARQL queries. This step was crucial for answering questions that required detailed and context-specific information. The overall LLM prediction steps are:

1. **Context Generation:** The context for each question was constructed from the data retrieved through SPARQL queries.
2. **LLM Inference:** Using the `bert-base-cased-squad2` model, we generated predictions based on the context. This model was trained on the SQuAD2 dataset to handle the intricacies of question answering with contextual information. This model was from Hugging Face[5].
3. **Integration:** The LLM-generated responses were integrated with the initial query-based results before the final refinement stage to enhance the accuracy and completeness of the answers.

2.5. Evaluation and Finalization

The evaluation of our approach was carried out by submitting the results obtained after the application on the test set to the codalab [6] provided by the organisers. The results was assessed based on Exact Match and F-score metrics.

2.6. Experimentation Environment

The experimentation was conducted using an HP EliteBook 745 G5 laptop equipped with an AMD Ryzen™ 5 PRO 2500U w/ Radeon™ Vega Mobile Gfx × 8 CPU, 24 GB of RAM, and a 512.0 GB SSD disk. The operating system used was Ubuntu 24.04.4 LTS.

[4]https://huggingface.co/deepset/bert-base-cased-squad2
[5]https://huggingface.co/deepset/bert-base-cased-squad2
[6]https://codalab.lisn.upsaclay.fr/competitions/19747

3. Results and Discussion

Fig. 3 presents the results obtained after applying the methodology presented in Section 2. It shows that the best results is obtained when the SPARQL queries are combined with the LLM for predicting responses.

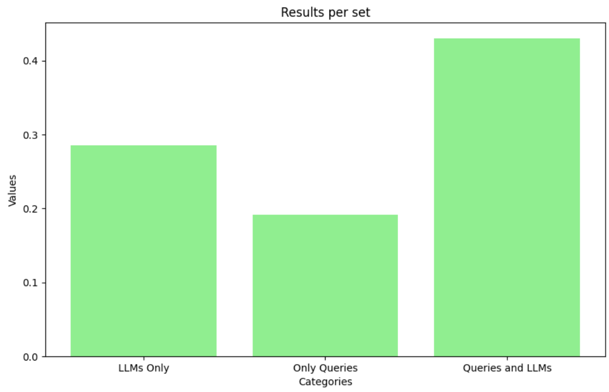

Figure 3: Performance comparism of the different(All) results of our model.

During this work, we found that:

- To manage complex queries on authors, institutions, affiliations and publications on the Semopenalex, we integrated SPARQL queries and LLMs prediction.
- the BERT-base-cased-squad2 model combined with DPR algorithm significantly improved the accuracy of entity and relation extraction on the DBLP KG. It should be noted that these information are needed to provide the context for the prediction by the LLMs.
- To handle the complete dataset, the DPR algorithm was employed, so as to be able to get through all the broken sets of the dataset.

4. Conclusion

In this paper, we presented a novel approach forHybrid Question answering over Linked Data. This approach was assessed on the training and test datasets of the Scholarly Hybrid Question Answering over Linked Data (QALD) Challenge 2024. We found that the integration of SPARQL queries with LLM-based predictions offers a robust solution for Question Answering over diverse scholarly data sources. Our approach demonstrated significant improvements in handling complex queries and providing accurate responses. Despite the results obtained, there were several challenges, particularly in handling the large and complex nature of the SemOpenAlex and DBLP datasets. Future work will focus on improving the model's ability to generalize across different types of scholarly data and incorporating more sophisticated rule-based systems on the one hand. On the other hand, we will focus on refining the methodology and exploring additional enhancements to further improve the system's performance.

References

[1] A. Jiomekong, V. Tsague, B. Foko, U. Melie and G. Camara, "Towards an approach based on knowledge graph refinement for answer type prediction," 2020.

[2] A. Jiomekong, G Camara and M. Tchuente "Extracting ontological knowledge from java source code using hidden markov models," *Open Computer Science,* Vol. 9, no. 1, pp. 181-199, 2019.

[3] Y. Mao et al., "HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5130-5141.
[4] J. Smith and A. Doe, *Scholarly QALD 2023: Enhancing Question Answering over Knowledge Graphs*, in *Proceedings of the Scholarly QALD 2023 Workshop*, 2023.
[5] International Semantic Web Conference (ISWC) 2024. Available at: https://iswc2024.semanticweb.org
[6] J. Smith and A. Doe, *Scholarly QALD 2023: Enhancing Question Answering over Knowledge Graphs*, in *Proceedings of the Scholarly QALD 2023 Workshop*, 2023.
[7] Wikipedia contributors, *Wikipedia-Based Scholarly Text*, retrieved from https://drive.google.com/file/d/1ISxvb4q1TxcYRDWlyG-KalInSOeZqpyI/view?usp=drive_link, 2024.
[8] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed., Prentice Hall, 1999.
[9] D. S. Johnson and C. H. Papadimitriou, "On the Performance of Divide and Conquer Algorithms," in *Journal of Computer and System Sciences*, vol. 19, no. 2, pp. 127-149, 1981.
[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171-4186, 2019.

5. Online Resources

The source code for this project is available via

- GitHub