The Development of Algorithm for Knowledge Graphs Construction through Knowledge Extraction from Large Language Models

Author: Thin Prabhong Advisor: Assistant Professor Dr. Areerat Trongratsameethong
Mentor: Assistant Professor Dr. Natthawut Kertkeidkachorn
Company: Japan Advanced Institute of Science and Technology (JAIST)

Abstract

Large language models have gained attention for their ability to generate human-like text and provide extensive knowledge from vast datasets. Knowledge extraction from these models is a key area of interest for generating insights. Meanwhile, knowledge graphs offer a flexible way to organize and interpret complex information. This study develops an algorithm to construct knowledge graphs through knowledge extraction from large language models, using ISWC-2024 LM-KBC Challenge dataset.

Our task is to predict objects from given subjects and relations in a Subject-Predicate/Relate-Object format. The ISWC-2024 LM-KBC Challenge provides five distinct relations: awardWonBy, companyTradesAtStockExchange, countryLand BordersCountry, personHasCityOfDeath, and seriesHas NumberOfEpisodes. Our approach leverages large language models with retrieval-augmented generation, web scraping, and web crawling, while also using large language model to filter unrelated data from web scraping.

Our study evaluates the algorithm's performance on both the validation set and the test set. When compared to the baseline, the developed algorithm demonstrates superior performance, achieving Macro-F1 scores of 0.695 and 0.698 on the validation set and test set, respectively, indicating its effectiveness and consistency.

Introduction

A pre-trained large language model is a large-scale language model capable of generating human-like conversations and information. It possesses extensive knowledge as it has been trained on vast amounts of textual data. Knowledge extraction from large language models is an intriguing topic, as it enables the transformation of knowledge from LLMs into structured information for further use.

Traditionally, data and knowledge are stored in relational databases, which are efficient for structured data. However, for semantic tasks, relational databases face limitations in handling highly flexible data, contextual variations in meaning, complex relationships, and diverse semantics that are difficult for computers to process. To address these issues, data is increasingly stored in the form of a knowledge graph, which represents information as subject-predicate-object triples. This structure allows for better semantic representation and improved machine understanding.

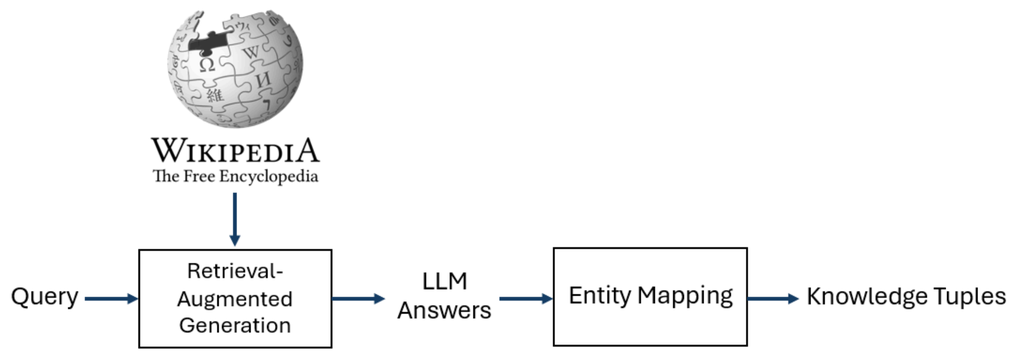
This study integrates both approaches, aiming to extract new knowledge from large language models and store it in a knowledge graph. The study is conducted using data from ISWC 2024 to explore methods for knowledge extraction and representation.

Technology

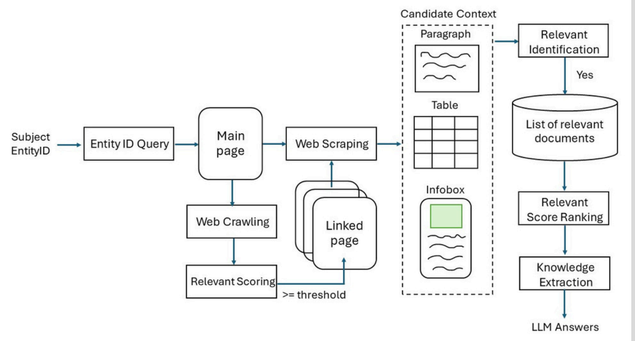
Huggingface Colab VS Code CodaLab

Methodology

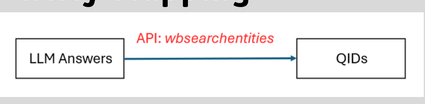
Overview



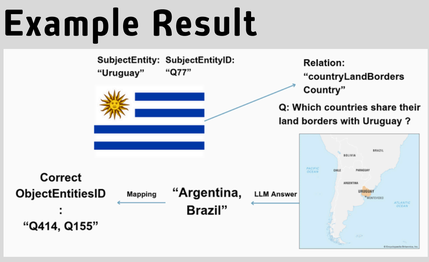
Retrieval-Augmented Generation



Entity Mapping



Example Result



Results

Data Set	Relation	Statistics					
		Macro-P		Macro-R		Macro-F1	
		Baseline	Ours	Baseline	Ours	Baseline	Ours
Validation Set	awardWonBy	0.238	0.771	0.028	0.103	0.045	0.125
	companyTradesAtStockExchange	0.540	0.842	0.703	0.795	0.474	0.678
	countryLandBordersCountry	0.961	0.921	0.912	0.900	0.919	0.850
	personHasCityOfDeath	0.700	0.750	0.600	0.910	0.460	0.660
	seriesHasNumberOfEpisodes	0.493	0.725	0.160	0.700	0.155	0.697
	Average	0.638	**0.799**	0.552	**0.801**	0.455	**0.695**
Test set	awardWonBy	-	0.825	-	0.021	-	0.032
	companyTradesAtStockExchange	-	0.797	-	0.755	-	0.638
	countryLandBordersCountry	-	0.910	-	0.903	-	0.854
	personHasCityOfDeath	-	0.695	-	0.920	-	0.647
	seriesHasNumberOfEpisodes	-	0.800	-	0.770	-	0.770
	Average	-	**0.792**	-	**0.810**	-	**0.698**

*No baseline is available for comparison in test set.

Conclusion

This study aims to develop an algorithm for constructing a knowledge graph by extracting knowledge from large language models using retrieval-augmented generation, web scraping and web crawling to enhance the question-answering capabilities of large language models. Additionally, large language models are utilized to filter out irrelevant information from web scraping.

The study experiments with data from ISWC-2024 LM-KBC Challenge, evaluating the algorithm's performance on CodaLab using both validation and test sets. The results show that our algorithm outperforms the baseline in most relations, except for *countryLandBordersCountry*. This exception highlights that while more information can enhance performance, it can also introduce noise, which may mislead the large language model. Furthermore, large language models can be effectively utilized for filtering irrelevant data, enhancing the overall performance of knowledge extraction and representation.

For future investigations, it is recommended to explore the implementation of automatic relevant document retrieval instead of relying solely on question-answering combined with relevant scores.

References

1. **Farrelly, T., & Baker, N. (2023). Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice. Educ. Sci., 13(11), 1109. https://doi.org/10.3390/educsci13111109.**
2. **CodaLab Competitions. Available at: URL: https://codalab.lisn.upsaclay.fr/ competitions/19136. Accessed July 20,2024.**
3. **Hugging Face. meta-llama/Meta-Llama-3-8B-Instruct. Available at: URL: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct. Accessed Aug 2, 2024.**
4. **lm-kbc/dataset2024. Available at: URL: https://github.com/lm-kbc/ dataset2024. Accessed Aug 2, 2024.**
5. **MediaWiki API Documentation. Wikibase/API. Available at: URL:https://www.mediawiki.org/wiki/Wikibase/API/en. Accessed Aug 2, 2024.**
6. **Wikipedia. Uruguay. Available at: URL: https://en.wikipedia.org/wiki/Uruguay. Accessed Aug 28,2024.**

204496 Cooperative Education, Academic Year 2024
Department of Computer Science, Faculty of Science, Chiang Mai University