

Exploring SPARQL query types to improve Ontology Mapping and Retrieval Augmented Modelling for auto-generated questions

Tharaniya sairaj R^{1a}, Balasundaram S R^{1b}

¹Department of Computer Applications, National Institute of Technology, Tiruchirappalli, Tamilnadu,

^atharaniyasairaj@gmail.com, ^bblsundar@nitt.edu

Abstract

The use of SPARQL is evergreen for querying, extracting and expanding named entities from RDF datasets, significantly enhancing text processing applications such as Automatic Question Generation (AQG). Recent research employs pretrained Large Language Models (LLMs) for AQG to reduce manual costs, but these models are limited by their dependence on training data. This in turn makes LLMs to counterfeit the answer-revealingness challenge in automatic Multi-hop Question Generation. This challenge occurs in the process of Multi-hop Question Generation as it requires integration of named entities from multiple sources and deep comprehension of these interconnected concepts, which is quite challenging. In this context, Retrieval Augmented Models (RAM) have gained attention in NLP for improving text processing through enhanced information extraction, yet their application in AQG is limited. This research addresses this gap by RAM's workflow with attention to its first phase - Input Text Enrichment via Named Entity Expansion—is crucial for generating diverse, comprehensive questions. But, Effective named entity expansion is facilitated by ontology mapping to align entities to various ontologies, which is more demanding. To address this requirement, SPARQL querying techniques such as multi-querying, step-back querying, and sub-querying are examined to enhance named entity expansion accuracy, thereby improving RAM's efficacy, leading to well-formed auto-generated questions.

Keywords: Retrieval Augmented Model, Ontology Mapping, Named Entity Set Expansion, Semantic Similarity, Automatic Question Generation.

1. Introduction

The SPARQL Protocol and RDF Query Language remains crucial need for querying, extracting, and expanding relevant named entities from RDF (Resource Description Framework) datasets [7-9]. Notably, the expansion of named entities significantly enhances various text processing applications, including Automatic Question Generation (AQG) [7,8]. In this regard, the recent literature leverages the utilization of pretrained Large Language Models (LLMs) for this generative process to reduce the manual cost and time in synthesizing hand curated question templates. However, the outcomes of LLMs is limited to the phenomenon of training data dependence, which in turn may not cover the domain-specific knowledge required for certain AQG tasks (such as Multi-hop Question Generation) [10,11]. A multi-hop question requires combining information from multiple passages or sources to arrive at the correct answer. For example, consider the text “An equilateral triangle is a geometric shape characterized by three equal sides”. The auto-generated multi-hop question is “Discuss the angles of the equilateral triangle characterized by three equal sides.”. Here the named entity “equilateral triangle” should not be generated as a part of question as it is a part of the answer (**Figure 1**). In this context, named entities which seem to be the answer candidates should be replaced with answer non-revealing named entity by integrating information from multiple sections of text or diverse sources [5,6]. This generates an answer non-revealing multi-hop question as “Discuss the angles of the geometric shape characterized by three equal sides.”. So, automatic

generation of these questions must reflect comprehension of interconnected concepts in a subject, a task that is complex and demanding.

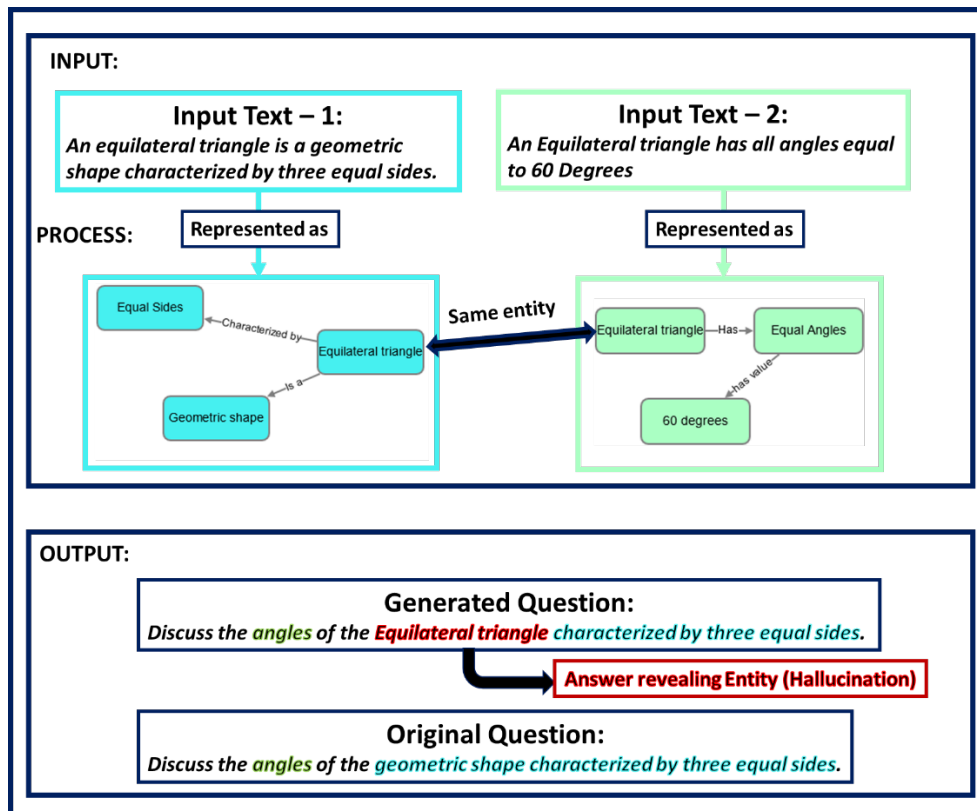


Figure 1 Illustration of answer-revealing Questions

In this regard, Retrieval Augmented Models (RAM) have gained attention in Natural Language Processing (NLP) for enhancing text processing and information retrieval. However, their application in Automatic Question Generation (AQG) remains limited [4,5,6]. This research addresses this gap by exploring RAM's functional workflow, consisting of three stages. Initially, named entities in the source text are expanded to broaden context and coverage, which is essential for improving retrieval accuracy and generating relevant, diverse questions. Next, the enriched text is encoded and indexed to create a contextual representation, ensuring accurate understanding and processing for retrieval tasks. The final stage involves searching the encoded text, ranking the results based on relevance, and decoding them to produce enriched text to improve the question generation process. Here, the first phase—Input Text Enrichment via Named Entity Expansion—is important for the workflow's efficiency, especially in auto-generating exam questions, as it enhances the diversity and volume of questions by providing a comprehensive topic understanding. So approaches to bring out efficacy in this step are studied and application of various Ontology Mapping tools are widely seen [1-3,7-9]. Ontology mapping is a complex process that aligns named entities to different ontologies, given the variability in entity nomenclature and definitions. In this regard, SPARQL is used to query the graph neighborhood of entities and their relationships within the ontology, facilitating precise named entity expansion. This paper examines ontology mapping, focusing on SPARQL querying techniques such as multi-querying, step-back querying, and sub-querying. This paper examines ontology mapping, focusing on SPARQL querying techniques such as multi-querying, step-back querying, and sub-querying. Multi-querying involves

executing multiple related queries to gather comprehensive information on a topic. Step-back querying entails reviewing and querying previous steps or layers of data to refine the search results. Sub-querying involves nesting down to synthesis a complex query from smaller sequential queries to extract specific information. By exploring these techniques, the study aims to enhance the accuracy of named entity expansion, ultimately improving the effectiveness of AQG through RAM.

2. Related Work

The role of ontologies and Ontology Mappings are evergreen in Automated Question Generation (AQG), particularly in text expansion [4-6]. By leveraging ontology models and question templates, innovative methods have been developed to automate question generation across various subjects [5,6,10]. In this line, the Sequence Generation Model based on Domain Ontology for Mathematical Question Tagging utilize domain ontologies to enhance deep learning models' comprehension of textual information, thereby improving question quality and relevance. Ontology-based approaches are also noted to maintain question consistency by deconstructing information into SPARQL queries, which are then converted into questions, achieving accuracy rates up to 90.71% [6-8]. Additionally, automatic ontology enrichment techniques have been applied successfully to extract knowledge from texts and enrich initial ontologies, demonstrating the efficacy of natural language processing and ontology enrichment in automated question generation. Furthermore, ontology mapping tools and methodologies [1-3,5-8] are employed with ontologies and thesauruses as background knowledge to advance educational ontology mapping and the overall efficiency of ontology alignment processes.

Meanwhile, SPARQL querying techniques have also been evolving to address challenges in RDF data processing [7,8]. Techniques such as multi-querying, sub-querying, and step-back querying are noted to enhance query performance and flexibility. Multi-querying allows the submission of a small query set instead of a single query, improving query representation. Sub-querying, involving the use of UNION and OPTIONAL operators [7-9]. Step-back querying optimizes SPARQL queries over decentralized knowledge graphs, addressing issues like cardinality estimation and data fragmentation. Moreover, the use of SPARQL in AQG recently relies on incorporating generative pre-trained language models (PLMs) like T5 and BART are notable. The use of SPARQL in knowledge-based question generation (KBQG) tasks handles complex operations such as aggregation and comparison. Furthermore, translating natural language competency questions [5,8] into SPARQL queries has been proposed to integrate ontologies, enabling efficient exploration of entity-relationships.

However, the application of RAM in AQG remains underexplored, which is highly essential to handle the limitations of pretrained Large Language Models (LLMs) in dealing with the domain-specific knowledge for Multi-hop Question Generation. This research gap emphasises the need for analysis of OM's contributive factors [1,2,8,9]. In this line, an exploration of diverse SPARQL query types such as multi-querying, step-back querying, and sub-querying are done and empirically analysed in the proposed work. Based on these observations, the Research Questions (RQ) and Research Objectives (RO) are given below.

RQ-1: How can RAM be effectively utilized to incorporate domain-specific knowledge for Multi-hop Question Generation using LLMs ?

RQ-2: What are the impacts of diverse SPARQL querying techniques on the accuracy of named entity expansion and efficiency in RAM-based AQG systems?

RO-1: To study the effectiveness of RAM-AQG integration in overcoming the limits of incorporating domain-specific knowledge LLM based Multi-hop Question Generation.

RO-2: To analyze the significance of OM's contributive factors such as multi-querying, step-back querying, and sub-querying on the accuracy of named entity expansion for AQG.

3. Methodology

The proposed methodology employs an Anchor-Align approach, a variant of CLASH approach to link the named entities in the sources text to that in the Ontology to bring out named entity expansion. If all entities are connected by one or more edges, the named entity network becomes a connected graph. Conversely, if an entity remains disconnected due to a lack of association with other entities, it becomes impossible to infer any meaningful information about the entity from the network. Once processed, the Ontology is updated using shared attributes: an entity vector (n-dimensional vector) and the similarity between two entities (cosine similarity between these vectors).

3.1. Step-1: Anchoring

In this phase, the source text is pre-processed using various Natural Language Processing (NLP) techniques such as tokenisation, case folding, entity extraction etc. Entity extraction based categorisation of source text using IsimScore (**Equation 1**) calculated with bigram relevance and semantic relationships intersection for further Ontology Mapping (OM) based entity expansion process. The random subset selection process is based on the assumption that using 63% of records in source data and $\log(F)$ of Feature set can generate feasible number of Random Subsets. In addition, heuristics are applied (**Table 1**) to carry out additive feature extraction for interpretable ensemble model generation in the consecutive stages of the work.

$$ISimScore = \frac{\prod_{POS=NN} Sim(E_{i+1}|E_i, E_{j+1}|E_j)}{\sum_{i=1}^n \sum_{j=1}^m Sim(E_i, E_j)} \quad (1)$$

Where, Sim = Similarity between unigram/bigram entities, E_{i+1}, E_i = Entities extracted from input text and $E_{j+1}|E_j$ = Entities extracted from knowledge graphs.

Table 1 Decision Heuristic for KG Triples

Heuristic ID	Axioms		Decision Heuristic	Inference
	Ontology1	Ontology2		
SSR	AO_1	AO_2	$AO_1 \subset AO_2$	AO_1
ERR	$\exists R_1 AO_1$	$\exists R_2 AO_2$	$(R_1 \subset R_2) \wedge (AO_1 \subset AO_2)$	$\exists R_1 AO_1$
ECR	$\forall R_1 AO_1$	$\forall R_1 AO_2$	$(AO_1 \subset AO_2)$	$\exists R_1 AO_1$

SSR - Superclass Subclass Relation, ERR - Existential Role Relation, ECR – Existential Composite Relation.

3.2. Step-2: Augmenting

The methodology for anchoring links between entities of different ontologies in the ontology mapping process focuses on establishing Basic Mappings as an interim outcome. This process is predicated on the execution of SPARQL queries, primarily employing three query types: sub query, multi query, and step-back query. Sub queries are utilized to extract specific subsets of data from a larger dataset, identifying entities that meet particular criteria within a single ontology. Multi queries perform simultaneous searches across multiple ontologies, facilitating the identification of potential links by comparing entities in parallel. Step-back queries trace relationships by moving backward through the data hierarchy, ensuring the consistency and contextual integrity of mapped entities. In the Basic Mapping process for two conceptual ontologies, sub queries extract relevant properties from the source ontology. Multi queries then identify equivalent properties in the target ontology by retrieving and comparing similar entities. Step-back queries verify and refine these mappings by ensuring contextual consistency. Using the above specified two-step empirical technique the mappings are generated as entity-relationship sets, which is then encoding into vectors for further text processing.

4. Experimental Study

Three dataset under varied domain and size are considered for empirical analysis and comprehensive evaluation of the OM process in the proposed work. The first dataset, The BPMN ontology [12] serves as a structured framework for representing Business Process Model and Notation (BPMN) concepts in a machine-readable format. It encompasses 270 classes, 176 object properties, and 70 data properties, enabling a comprehensive representation of BPMN elements, attributes, and relationships. The 270 classes represent core elements like processes, activities, events, gateways, data objects, and participants, which are being mapped into classes using retinal shapes such as triangles, polygons etc. Object properties detail relationships like control flow (e.g., sequenceFlow), associations (e.g., dataInputAssociation), containment (e.g., contains), and interactions (e.g., participant). The second dataset is the Shape Ontology (SO) [14]. The shape ontology comprises 40 classes, with enduring or continuant entities forming the top-level category and specific shapes as subtypes. It defines 15 data properties, categorizing shapes by characteristics such as dimensionality, symmetry, boundary conditions, and curvature. The ontology includes 30 object properties, detailing relationships like hasShape (between objects and geometric shapes) and approximates (between individual objects and shape property types). The third one is the Geometry Ontology (GO) [13]. The Geometry Ontology encompasses 9 classes, including foundational geometric representations such as Point, LineString, Polygon, and their composite forms like MultiLineString, Triangles, MultiPolygon etc. he ontology includes 10 object properties, including "geometry, symmetry" linking resources to their geometric shapes, and "boundary," grouping properties defining polygonal boundaries.

4.1. Results and Discussion

The study (Table 2) shows that the proposed approach surpasses the baselines through improved triple type driven relation extraction for entity context based AQG.

Table 2 Comparison of Proposed approach with baseline Mappers

Mappers/ Ontologies	SO-BPMN			SO-GO		
	Precision	Recall	F1-score	Precision	Recall	F1-score
LogMap	0.65	0.62	0.65	0.68	0.66	0.66
ODGOMS	0.78	0.73	0.75	0.78	0.73	0.75
Blooms	0.8	0.77	0.78	0.81	0.79	0.8
Proposed Approach	0.83	0.82	0.83	0.89	0.88	0.88

It is noted that a wide margin of 2-3% accuracy improvement can be achieved using the proposed A-A approach in entity selection for AQQ. In particular, the generation of ERR category triples (relationships) are more benefited when used with sub-query and multi-query types, showing an accuracy improvement of around 4%. On the other side, the URR category triples are commonly achieving around 94% upon all three query types, where Basic Mapping (BM) is relatively improved. In this line, it is noted that the proposed A-A approach can also generate improved mappings, only when Basic Mapping (BM) process is required. **Figure 2** shows the sample mapping generated using the three diverse query types. Based on these observations (**Table 3**) the proposed FFRF model is inferred to have improved potential to automatically generate questions, balancing diversity and relevance factors.

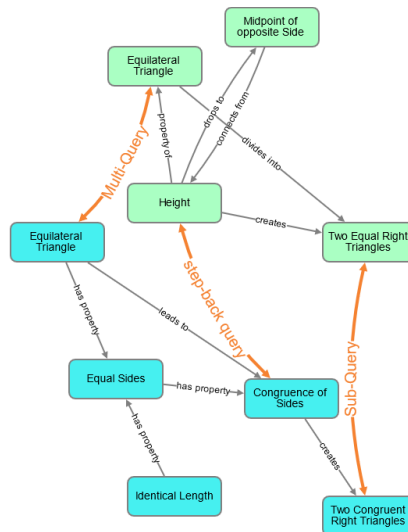
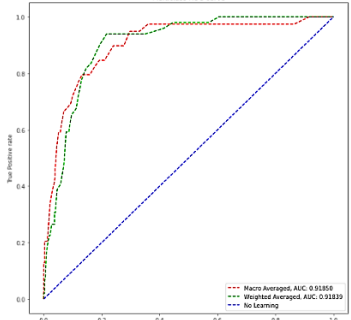
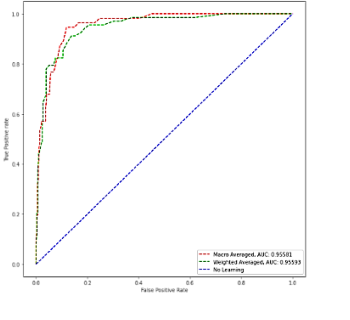
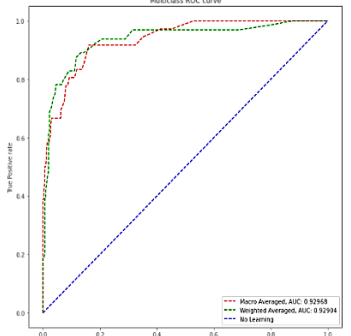
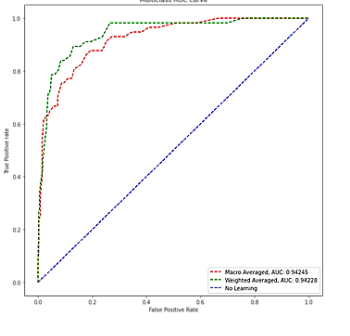


Figure 2 Sample Outcome of Mappings using different query types

Table 3 Proposed Approach Comparison - Triple Type vs Query Type

Triple Type and Generated Question	Multi Query/Sub Query	Stepback Query
<p>Triple Type: SSR</p> <p>Generated Question: List down the properties of an equilateral triangle and illustrate the role of symmetry in geometric applications</p>		

<p>Triple Type: ERR Generated Question: Discuss the geometric characteristics of an equilateral triangle and elaborate on the relationship between its sides and angles.</p>		
<p>Triple Type: ECR Generated Question: Define equilateral triangle and compare its geometric properties with other types.</p>		

Conclusion and Future Work

In conclusion, this research paper explores the utilization of three SPARQL query types to enhance OM and RAM for AQG. The study focused on addressing challenges in generating non-answer revealing multi-hop questions using Retrieval Augmented Models (RAM). By employing SPARQL techniques like multi-querying, step-back querying, and sub-querying, the research aimed to improve the accuracy of named entity set expansion, a critical requirement for AQG tasks. Empirical analysis across diverse datasets shows significant advancements in triple type-driven relation extraction, achieving notable improvements in precision, recall, and F1-score metrics compared to baseline methods. The paper highlights the significance of OM across three common triple types, facilitated by SPARQL to complement Large Language Models (LLMs) with reliance on traditional approaches. Future research could further broaden OM across different domains, and integrate advanced techniques into RAM-based AQG systems.

Acknowledgement

This research was carried out at E-Learning and HCI Lab, Department of Computer Applications, National Institute of Technology, Tiruchirappalli.

References

1. Ivanova, Tatyana. (2022). A Methodology for Mapping Educational Domain Ontologies Using Top Level Ontologies. 1-4. 10.1109/InfoTech55606.2022.9897119.
2. Iglesias-Molina, Ana & Cimmino Arriaga, Andrea & Ruckhaus, Edna & Chaves-Fraga, David & García Castro, Raúl & Corcho, Oscar. (2022). An Ontological Approach for Representing Declarative Mapping Languages. Semantic Web. 15. 191-221. 10.3233/SW-223224.

3. Sadiq, Muhammad & Amin, Muhammad & Bilal, Hafiz & Hussain, Musarrat & Hassan, Anees Ul & Lee, Sungyoung. (2018). LogMap-P: On matching ontologies in parallel. 1-5. 10.1145/3164541.3164589.
4. Huang, Tao & Hu, Shengze & Lin, Keke & Yang, Huali & Zhang, Hao & Song, Houbing & Lyu, Zhihan. (2023). Sequence Generation Model Integrating Domain Ontology for Mathematical question tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 10.1145/3593804.
5. Kusuma, Selvia & Siahaan, Daniel & Fatichah, Chastine. (2022). Automatic question generation with various difficulty levels based on knowledge ontology using a query template. *Knowledge-Based Systems*. 249. 108906. 10.1016/j.knosys.2022.108906.
6. Raboanary, T.H., Wang, S., & Keet, C.M. (2021). Generating Answerable Questions from Ontologies for Educational Exercises. *International Conference on Metadata and Semantics Research*.
7. Xiong, G., Bao, J., Zhao, W., Wu, Y., & He, X. (2022). AutoQGS: Auto-Prompt for Low-Resource Knowledge-based Question Generation from SPARQL. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.
8. Benhocine, Karim & Hansali, Adel & Zemmouchi-Ghomari, Leila & Ghomari, Abdessamed Réda. (2022). Towards an automatic SPARQL query generation from ontology competency questions. *International Journal of Computers and Applications*. 10.1080/1206212X.2022.2031722.
9. Chen, Yanji & Kokar, Mieczyslaw & Moskal, Jakub. (2021). SPARQL Query Generator (SQG). *Journal on Data Semantics*. 10. 1-17. 10.1007/s13740-021-00133-y.
10. Emerson, J., & Chali, Y. (2023). Transformer-Based Multi-Hop Question Generation (Student Abstract). *AAAI Conference on Artificial Intelligence*.
11. Kulshreshtha, Saurabh & Rumshisky, Anna. (2023). Reasoning Circuits: Few-shot Multi-hop Question Generation with Structured Rationales. 59-77. 10.18653/v1/2023.nlrse-1.6.
12. Singer, R. (2019). An ontological analysis of business process modeling and execution. *arXiv preprint arXiv:1905.00499*.
13. Katsumi, M. Geometry Ontology. Enterprise Integration Lab. Retrieved June 16, 2024, from <https://enterpriseintegrationlab.github.io/icity/Geom/doc/index-en.html>
14. Rovetto, R. J. (2011). *The Shape of Shapes: An Ontological Exploration*. Shapes, 1.