Medical Semantic Question Answering Framework on RDF Data Cubes

Usman Akhtar, Sungyoung Lee

Department of Computer Science and Engineering, Kyung Hee Univerity,
{usman,sylee}@oslab.khu.ac.kr

Abstract. In this paper, we have proposed a framework to support the semantic question answering over the RDF data cube that is published according to the Linked Open Data (LOD) principles. As statistical data published all over the Internet there is a need to empowers the non-experts to query in the form of the natural language. But, the existing question answering system unable to support query on the statistical data in the form of the RDF cube. The current research is motivated by the need of the clinical organizations, who wish to develop a platform for analyzing the clinical data across multiple clinical sites. Linked open data (LOD) provides a support to published statistical data in the form of the RDF cube. Our proposed framework will provide a support to interact in the form of the natural language question answering that will produce the SPARQL query to extract the answer from the RDF data cube. In future, we will develop the benchmark to calculate the accuracy of the answer.

Keywords: Semantic Question Answering (SQA) · RDF Data Cutes · Question Answering Framework ·

1 Introduction

An increasing amount of the statistical data is published on the Linked Open Data (LOD) cloud. Getting insights from the data in more intuitive ways are becoming important. Systems for the Semantic Questions Answering (SQA) plays a vital role to connect with linked open data and provides an intuitive interface by translating natural languages queries into SPARQL syntax. Statistical data need more advanced querying methods to empowers non-experts users to draw their own conclusions. Semantic question answering is extremely important in the following application involving Linked Data to access public data sources.

— **Healthcare and Life Sciences (HCLS):** Statistical data in the form of the RDF data cubes influences decisions in a domain such as health care and life sciences. Many clinical datasets are often composed of the numerical observations as well as statistical information such as clinical trial data which is often composed of patient attribute [3].

— **Biomedical Question Answering:** In biomedical, workers want to express their information needs in natural language. BIOASQ [5] encourages the

participant to adopt semantic indexing as a mean to combine the information from the multiple sources of different types such as biomedical articles and ontologies. But, this system typically lacks supports for the RDF data Cubes, where clinical data represented in the form of multi-dimensional data.

We have motivated the need of the semantic question answering, where statistical data in the form of the RDF data cubes. However, there exist some important challenges that need to be tackled by the Semantic questioning answering system, including the following:

— **Lack of processing RDF cubes by SQA systems:** One of the major limitations of SQA system is a lack of processing of the statistical data in the form of the RDF data cute. Statistical data is different than other data and can not be queried by the existing linked open data querying approaches. However, the current SQA system provide translating natural language into SPARQL, which is a native language to query the RDF knowledge bases [2, 3]

— **Enabling Access Over Statistical Data:** Current query federation approaches enables the integration of the multiple data sources but they do not consider the methods to access the statistical data while maintaining the good performance [5].

Although, there are a number of benefits to publishing data in multi-dimensional, such as statistics in Linked Open Data (LOD) cloud using LOD publishing principles. First, the data become web addressable and allow a consumer to annotate and link the data. Secondly, data can be flexible and combined with the other data using Linked data technology. Finally, data can be reusable and access by using the SPARQL, one of the example is linkedspending[1], which contains government spendings from all over the world as linked data.

In this paper, we have proposed a framework to handle the statistical data natural language query that works on the RDF data cube using SPARQL query template. In section 2, we have discussed the preliminaries related to the RDF data cube and what operations are allowed in the RDF data cube. In section 3, We have explicitly mentioned the problem of the existing SAQ system while dealing with the RDF data cube. In section 4, we have proposed the framework to handle statistical queries. Finally, we have discussed the initial results in the discussion and give directions to the future work.

2 Preliminaries

Statistical data can be expressed in the form of the RDF data cute[2], also called as an OLAP cube or hypercube, which usually consist of the multi-dimensional datasets. Data cube represent the multidimensional numerical data and consist of the array of cells. And each cell is identified by the associated dimension and mostly sparse.

[1] http://linkedspending.aksw.org/
[2] https://dvcs.w3.org/hg/gld/raw-file/29a3dd6dc12c/data-cube/index.html

Definition 1. *RDF data cube format: In linked open data, RDF data cube vocabulary*[3] *allows expressing the data cute. Each RDF data cube consists of model and observations. RDF data cube supports three main operations dicing, slicing and rolling which create a subcube of the main data cube.*

Definition 2. *There is a difference between the Question Answering (QA) and the Semantic Question Answering (SAQ). In QA, users ask questions in natural language using their own terminology. In SQA, the natural language question is transformed into the formal query using SPARQL.*

3 Problem

In summary, from the prior work, we can represent the medical statistical data as RDF data cube. However, how to execute the SPARQL queries over the data cubes is still the challenging issue. We have not seen any work looking at putting all aspect of RDF Cubes together. As argued in the introduction we can not use existing Question Answering (QA). Since they do not provide the query template to match the RDF cube.

3.1 Solution Strategies

Question answering over the linked data is an active research area but there are only a few solutions exists in the semantic question answering over the RDF data cube. From the literature, we have shortlisted the two strategies that can be implemented on the current problem, but each strategy has its own strength and weakness.

Template Based Question Answering: In this approach, domain independent and domain dependent lexicon and it use SQA pipeline. First, users supplies a natural language question and then the tagger identifies the part of the speech such as nouns and lexicon is created which parse the question. Then the construction of the semantic representation and transformed into the SPARQL query. For indexing, Apache Solr is used which perform faster on statistical data. Finally, the answer is presented as a list to improve the SPARQL query using the AutoSPARQL [4] algorithm. The work is limited in support the RDF data cube, the generated SPARQL query do not provide the accurate results. We decide to use this approach in order to reduce the complexity. But, in our algorithm, we have used the modified version.

Extracting RDF from Text: A bootstrapping strategy that extracts the RDF from the text, the main focus is to extract the natural language patterns from the unstructured data. A similar type of approach such as BOA [1] the goal of this approach is to extract the structured data as RDF from unstructured data. BOA approach uses the Linked data web for knowledge. First, this approach will extract the natural language pattern from the Linked Data Web making the use

[3] https://www.w3.org/TR/vocab-data-cube/

of the data web as a background knowledge. Secondly, BOA generates RDF and that can be queried easily via SPARQL. On the statistical data, this approach can solve the problem but the current SPARQL syntax generated by BOA will not work on RDF cube.

4 Our Proposed Approach

We propose a framework by combining different approaches to Semantic Question Answering (SQA) over RDF data cute, where natural language questions convert into SPARQL query and work on the statistical data. To design, we take into the account of the existing SAQ architecture over the linked data. In the following, we will explain our proposed framework in detail. But one of the biggest challenges that we face in realization how to combine the different approaches in order to work on RDF data cube. The framework has a main three main states (Interpretation, Matching, Execution). As shown in the Fig. 1, a user posted the natural language question over the Linked open data cube that is stored in the triple store. The framework generates the equivalent SPARQL query and generates the answer.

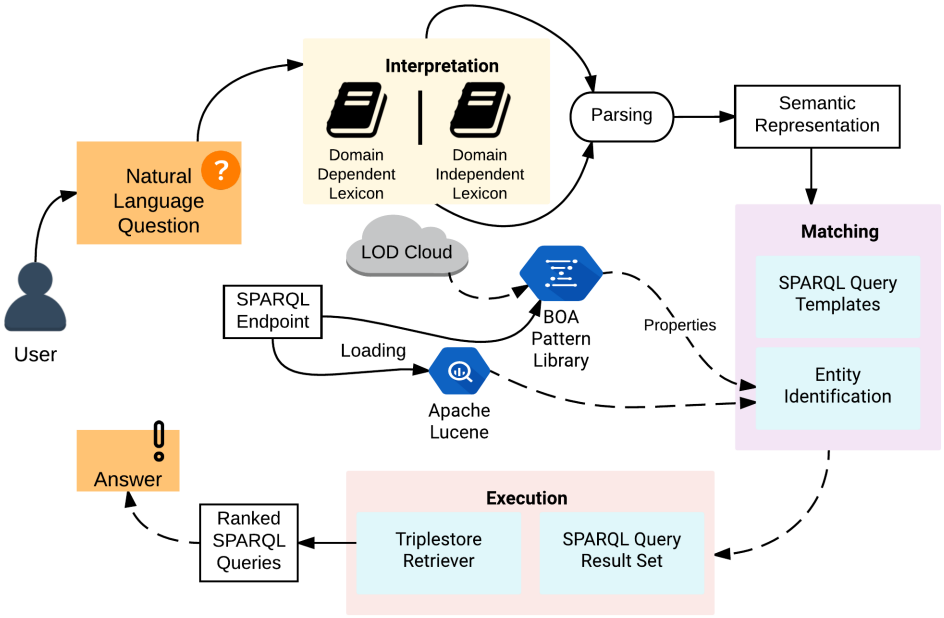

Fig. 1: Our proposed framework handle the naive user question and produce equivalent SPARQL query to run over the RDF data cube.

4.1 Interpretation Phase

The first step is a crucial step of the core module when the natural language question comes then this phase referring to the RDF cube operations are detected.

Normally, these operations are detected by certain key phrases and are detected using the regular expressions. In this phase, it is determined that how the input questions will be identified by rest of the framework. In the pre-processing step dataset index that is initialized by using Apache Lucene[4] and then the terms can be found by the index. At the end, of this step, a syntactic parse tree is generated for the question.

4.2 Matching Phase

After the interpretation of the given question is generated, the question is then split into the phrases and mapped to the RDF data cube. The matching phase extracts the answers from the sources according to the delivered interpretation content. The output of the interpretation can be SPARQL query which can be handled by the RDF data cube store. In the statistical data, the answer type can be a countable, uncountable or temporal example such as integer is countable and double is uncountable and a year is temporal.

4.3 Execution Phase

After the answer is extracted from the RDF data cube then it can be ambiguous and redundant. In the simplest example, question explicitly mentions in a dataset description such a "located hospitals in 2017", for this case it can be found by matching. Finally, the highest ranking query will be executed first and the final ranking is based on the returned answer.

5 Discussion

As more and more Question Answering system is emerged every year to support the statistical data access of the RDF data cube, there is a need of the query builder, that select and combine the SPARQL features to access the linked open data sets that are published over the cloud.

5.1 Research Findings

We believe that the RDF data cube will be a strong baseline in the new research sub-field. Accessing statistical medial data is still challenging. Over the past years, a lot of research focusing on the interaction paradigms that allow the end user to easily access the interface and use natural language in a more intuitive way. Although, the key challenges are in translating the user's information using the standard web query processing technique and evaluate the accuracy. On the other hand, multilingual questions have become an issue for the semantic web community, there is a need of the system that can help in overcoming the language barriers by facilitating the multilingual access to semantic data.

[4] https://lucene.apache.org/core/

6 Conclusion and Future Work

We have introduced the framework to support the question answering over the RDF data cube. This work is motivated by the need of the clinical organizations who wish to develop a platform for analyzing the clinical dataset in the form of the natural language questions. Our framework will support the question answering over the Linked Data. One of the strengths of the proposed approach is to generate the SPARQL query template that can extract the answer from the RDF data cube. In the future work, we plan to continue contributing the development the full system and open source along with the benchmarks. We will investigate, how to integrate the Semantic question answering technique to the existing question answering system.

Acknowledgments. This work was supported by the Industrial Core Technology Development Program (10049079), Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) and This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning(2011-0030079).

References

1. Daniel Gerber and Axel-Cyrille Ngonga Ngomo. Extracting multilingual natural-language patterns for rdf predicates. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 87–96. Springer, 2012.
2. Konrad Höffner and Jens Lehmann. Towards question answering on statistical linked data. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 61–64. ACM, 2014.
3. Konrad Hoffner, Jens Lehmann, and Ricardo Usbeck. CubeQA-question answering on RDF data cubes. In *International Semantic Web Conference*, pages 325–340. Springer, 2016.
4. Jens Lehmann and Lorenz Bühmann. Autosparql: Let users query your knowledge base. *The Semantic Web: Research and Applications*, pages 63–79, 2011.
5. George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*, 2012.