



REVUE ALGERIENNE DES SCIENCES

Section -A-

Sciences de la Nature et de la Vie

Sciences techniques

<http://univ-eltarf.dz/fr/>



A new approach to switching from natural language to a SPARQL query

Chemam Chaouki ^{a,*}, Zarzour Hafed^b,

^aDépartement d'Informatique – Université Chadli Bendjedid El-Tarf, BP 73, (36000) – Algérie.

^bUniversités Mohammed-Chérif Messaadia, souk ahras, Algérie .

Informations	Résumé
Mots clés : Web, Web of data RDF, SPRQL, Linked data, NLP, http, URI *Correspondance : Che_chaouki@yahoo.fr	With the growing amount of semantic data available on the web, there is a strong need for systems that allow common web users to access this body of knowledge. Question answering systems have received particular attention, as they allow users to express arbitrarily complex information needs in a simple and intuitive manner. The main challenge is to translate users' information needs into a form that can be assessed using standard Semantic Web query processing and inference techniques. Over the past few years, a series of approaches have been developed to address this challenge, showing significant advances towards answering natural language questions in relation to large, heterogeneous sets of structured data. However, only a few systems still deal with the fact that the structured data available today is spread across a large collection of interconnected data sets and that answers to questions can often only be provided if information from multiple sources is combined. . In addition, a lot of information is still only available in textual form, both on the web and in the form of labels and summaries in linked data sources [1]

1. Introduction

A difficult task in answering natural language questions (Q / A for short) on the RDF knowledge graph is how to bridge the gap between unstructured natural language questions (NLQ) and structured RDF data in graph (G). One of the effective tools is the "template," which is often used in many existing RDF Q / A systems. However, few of them study how to automatically generate models. To our knowledge, we are the first to offer a join approach for model generation. Given a workload D of SPARQL queries and a set N of natural language questions, the goal is to find pairs q, n, for $q \in D \wedge n \in N$, where the SPARQL query q is the best match for question n in natural language n. These pairs provide promising advice for automatic model generation. With the ambiguity of natural languages, we model above problems as an uncertain graph join task. We offer several structural and probabilistic pruning techniques to speed up assembly. Extensive experiments on real-world RDF Q / A benchmarks confirm both the effectiveness and efficiency of our approach. [2] Recently, knowledge graphs have gained a lot of attention both in academia and in industry. Since the Resource Description

Framework (RDF) is the de facto standard for a knowledge graph, we focus on the RDF repository in this article. A key issue is how to access knowledge graphs and quickly get the information you want. Although SPARQL is a structural query language on RDF charts, it is not practical for non-professional users to query RDF charts using SPARQL, due to the complexity of the SPARQL syntax and schema. RDF. We illustrate a sample SPARQL query on DBpedia as follows.

```
SELECT ?person
```

```
WHERE { ?person rdf:type Artist .
```

```
?person graduatedFrom Harvard University.}
```

2. Natural Language Processing

Natural language processing (NLP) is one of the most important techniques that change the computer into the human brain. The NLP comes under the field of artificial Intelligence where NLP is used to train the computer to understand the human language [3]. The computer is trained as a human brain for its activities of language processing.

The NLP is majorly aimed for the reading of language, decipher, understanding the language and after this all making sense like the human brain and reacts as per human brain reactions [4]. So doing all these tasks makes the NLP the most important technique for building a system that can reply according to the human brain. This method of language processing is used in many applications.

One of the applications of NLP is seen in the question answering system where the answering for the questions asked by the user is done [5]. This task of answering the question builds the QA system. The NLP helps in this system as it reads the question and tries to answer like a human brain. Whereas the NLP uses machine learning for the evaluation of the proper meaning of the question and then the answering is done.

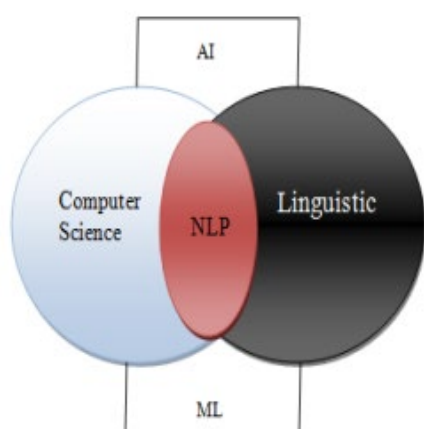


Figure 1 : Natural Language processing

3. Methodology

To reduce the complexity of the SPARQL syntax, RDF question and answer (Q / A) systems provide an easy to use interface for users, which has attracted great attention in both NLP and DB communities (database). A difficult task is to translate natural language questions (NLQ) into structural queries, such as SPARQL, on a large knowledge graph G . one of the effective tools is the model, which is used in many existing RDF Q / A systems. A typical system is EVI (<http://www.evi.com/>, officially True Knowledge), which registered over one million users within four months of its launch in January 2012. EVI was acquired by Amazon in October 2012 and is now part of the Amazon group of companies. EVI aims to directly answer questions posed in plain text in English on the knowledge base. It uses a model-based approach to answering natural language questions, where a model describes how to transform a class of natural language questions into correct

structural queries. However, EVI must manually define the translation models. It is clear that the quality of the models determines the quality of the response. The challenge is how to automatically generate a large number of high quality models. Few of the previous works explore this question, but all have to manually define the models. Obviously, it is expensive to manually define these models, especially for open domain Q / A systems on large-scale RDF knowledge graphs. In this document, we study how to automatically generate models. [1]

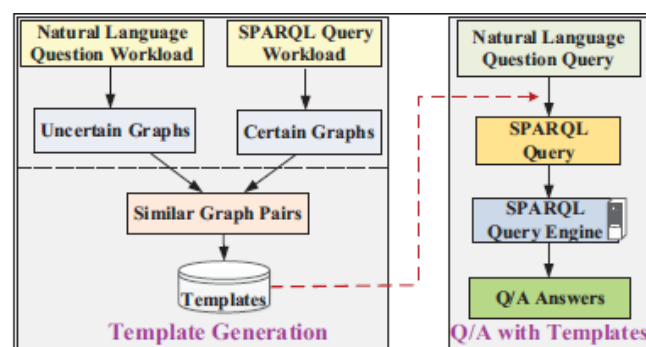


Figure 2: Framework for Model Based RDF Q/A

In figure 2 shows the RDF Q / A framework using templates. It consists of two tasks: how to generate models and how to use these models in RDF Q / A. To solve the first problem (subject of this article), we propose a workload-based approach of queries for automatic generation. of models. Specifically, the inputs for the model building task are two query workloads; one is a D set of SPARQL queries on the RDF repository (such as DBpedia workload1). The other is an N set of natural language questions, which can be collected from some community-based question-and-answer (Q&A) sites (such as Yahoo Answers) or the query engine workload. Search (such as WebQuestions2). **The whole process takes place in three stages as follows.**

Step 1: Uncertain graph generation. According to the method, natural language questions can be interpreted in semantic query graphs. Due to the ambiguity of interpretation, we model the semantic query graph as an uncertain graph, i.e. each vertex / edge has multiple possible labels with different probabilities. Here, the labels are uncertain due to semantic ambiguity. For example, consider the question "which American actor is married to New York City-born Michael Jordan" in Figure 3.

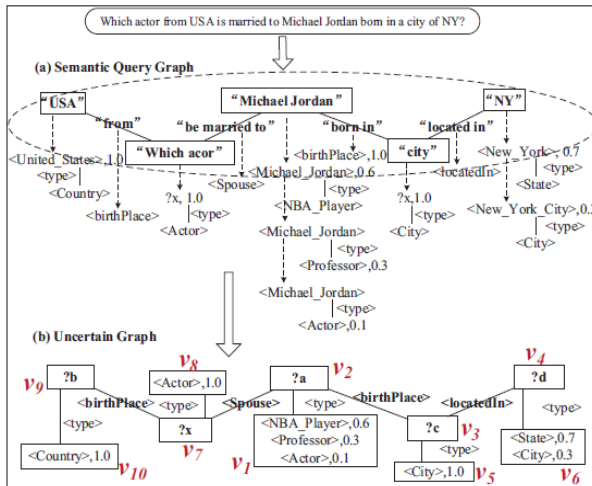


Figure 3: Uncertain graph generation

There are three people named "Michael Jordan" in the DBpedia RDF chart, that is, an NBA star, a professor and an actor, each associated with a probability. "NY" also has two possible connections, namely state and city. Therefore, we can get an uncertain graph containing nodes with multiple labels, each associated with a probability.

Step 2: Find pairs of similar charts. The goal is to find some pairs q, g , for $q \in D \wedge g \in U$, where D corresponds to the SPARQL queries and U are the uncertain graphs derived from the questions in natural language³. Here we use the graph edit distance, which is widely used to measure the similarity of graphs, to calculate the similarity between q and g .

Step 3: Generation of models with pairs of similar charts. Given a returned pair q, g , based on the mapping between q and g (the mapping is found when calculating the graph edit distance in step 2), we can build the models. In order to generate models, the key problem is how to find the pairs of similar graphs efficiently (ie step 2). This is the subject of this article. We propose a new approach, namely, Similarity Join on uncertain charts, denoted SimJ. More exactly, we offer a series of effective terminals to improve performance. Given the models generated, a natural language question is translated into a SPARQL query using the models generated during the first task. Then, the SPARQL query is searched on the knowledge graphs. Existing systems, such as **Jena**, **RDF-3x**, **Virtuoso4**, and **gstore**, can be used to complete the search.). **All the processing takes place in two stages as follows.**

Offline processing: To enable online semantic relation extraction processing, a dictionary of RDF relations and their paraphrases in natural language is automatically constructed in advance. For example, the relationship manager can be expressed as "directed by" or "a movie by". We follow a simple dependency analysis approach based on three points as described in [6]. For each subject-predicate-object triplet (s, p, o) in the data set, these sentences in the s and o confidence coding are retrieved. Then the sentences are analyzed by the dependency analyzer, and the shortest paths between s and o are selected as paraphrases of the relation p . Another challenge is the ambiguity of natural language. A single sentence in the question can often find more than one mapping in the RDF chart. To deal with this, we introduce RDF fragments, which are star-like RDF sub-graphs that discover the neighborhood of each relation, entity, or semantic class. Then we can roughly figure out how a certain semantic element is connected in the graph within a step length threshold. RDF fragments have good filtering and disambiguation power for these inappropriate mappings, while having acceptable space consumptions in many such entities, which share the same fragments. RDF fragments can be indexed by a reverse list.

Online processing: During the online step, the input question q_{NL} is introduced into the following four-step pipeline: First, q_{NL} is parsed in a tNL dependency tree. Second, sentences in q_{NL} that mention any semantic relation are recognized in tNL with the help of the paraphrase dictionary. Sentences for entities and classes are those syntactic arguments in tNL associated with relation sentences. We also unify sentences that refer to the same element by co reference analysis. Third, these sentences are mapped to RDF fragments to find their semantic element matches in RDF graph G . Variables are introduced by question focus, unmatched entity sentences, and RDF fragments of semantic classes that are type constrained variables. Finally, the RDF fragments are attached to the reasonable sub graphic computer of graph G , checking their compatibility against the tNL dependency tree. The results are ranked according to the similarity and semantic consistency score, leading to the target SPARQL.

4. Linked data

The Web data (linked data, in English) is an initiative of W3C (World Wide Web Consortium) to promote the publication of structured data on the Web, not as isolated data silos each other, but by linking them together to form a global information network.

It relies on web standards, such as HTTP and URI - but rather than using those standards just to make it easier for humans to navigate, the web of data extends them to share information evenly between machines. This allows data to be requires automatically, regardless of where it is stored, and without having to duplicate it.

Tim Berners-Lee, director of W3C, coined and defined the term Linked Data or data linked and synonymous Web of data in a book on the future of the semantic web. In France, the term Web of data is used more and more by the community of professionals in the field.

For libraries, it is necessary to participate in the web of data in order to gain visibility and to make this data usable in other contexts. The web of data thus makes it possible to decompartmentalize data in catalogs, so that they are more accessible to users and those in the future. Among other things, it allows data from archives, museums or biographies to be linked to texts in the catalog, thus providing researchers with more complete information.

Tim Berners-Lee has defined four pillars to support the "Web of Data" initiative :

1. use unique URIs to identify things.
2. use HTTP URIs that exist on the Web (URLs, therefore). An HTTP 404 Error simply indicates that the URI used is not explicitly documented.
3. provide usable, human-readable and machine-readable information through the URI, using open formats such as RDF or SparQL. For example, by using the HTTP redirection mechanism (code 302) and the User-Agent variable contained in the headers of HTTP requests, a server can display an XML or RDF page for a machine or an HTML page for the browser. 'one person.
4. Using the initial URI by associating it with external URIs to improve the discovery of other information on the Web.

5. Algorithm for switching from natural language to a SPARQL query :

Algorithm 1. Integrating Adj2ER into existing QA systems

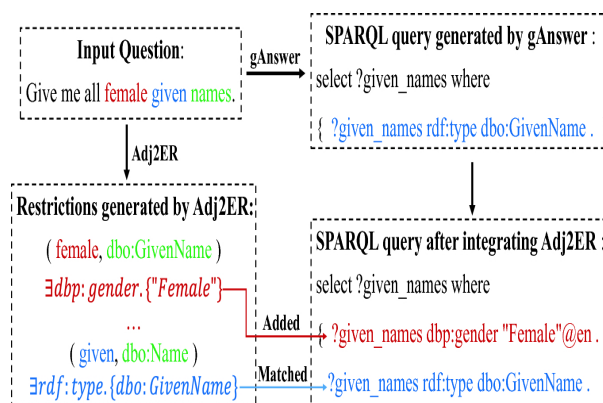
Input: A natural language question Q

Output: A SPARQL query for the input question Q

```

1: procedure GENERATESPARQL( $Q$ )
2:    $S := \text{ExistingQASystem.GENERATESPARQL}(Q)$ ;
3:   for all ( $adj, C$ ) pair in  $Q$  do
4:      $\text{Restrictions} := \text{Adj2ER}(adj, C)$ ;
5:     if  $S$  contains a  $\text{restriction} \in \text{Restrictions}$  then continue;
6:      $S' := S$ ;
7:     Remove all restrictions in  $S'$  which have semantic similarity  $\geq 0.6$  with  $adj$ ;
8:     for all  $\text{restriction} \in \text{Restrictions}$  do
9:        $S^* := \text{Resulting SPARQL for adding } \text{restriction} \text{ to } S'$ ;
10:      if  $S^*$  is a non-empty query then  $S := S^*$ ; break;
11:   return  $S$ ;
```

We have integrated our adjective mapping approach into two state-of-the-art quality assurance systems, namely gAnswer and WDAqua. Who placed first and second respectively in the QALD-9 challenge. The procedure for integration Adj2ER into them is presented in Algorithm 1. An example of integrating Adj2ER into existing quality assurance systems. [7]



	70 QALD questions			50 YA questions			Overall
	P	R	F1	P	R	F1	
gAnswer	30.49%	55.30%	29.75%	16.56%	36.26%	13.97%	23.18%
gAnswer + Adj2ER	44.03%	62.25%	43.02%	37.32%	56.59%	38.59%	41.18%
WDAqua	21.10%	26.64%	17.79%	23.53%	28.10%	22.04%	19.56%
WDAqua + Adj2ER	33.28%	43.86%	32.05%	42.70%	44.88%	40.99%	35.77%

6. Conclusion

In this article, we provide a data-based graphics framework for answering natural language questions about RDF graphics. Different from existing work, we allow sentences and structure ambiguity in the phase of understanding questions. We push the disambiguation in the

evaluation of stage queries. Based on the results of the queries on the RDF charts, we can effectively resolve the ambiguity problem. In other words, we combine disambiguation and query evaluation into a uniform process. Therefore, the data-driven graphing framework not only improves the accuracy but also speeds all the performance of the RDF Q / A system.

References

- [1] Weiguo Zheng, Lei Zou, Xiang Lian and al, How to Build Templates for RDF Question/Answering—An Uncertain Graph Similarity Join Approach.
- [2] Camille Pradel, Ollivier Haemmerlé, Nathalie Hernandez, Passage de la langue naturelle à une requête SPARQL dans le système SWIP.
- [3] H. H. Hsu and N. F. Huang, “Xiao-Shih: The Educational Intelligent Question Answering Bot on Chinese-Based MOOCs,” Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018, pp. 1316–1321, 2019, doi: 10.1109/ICMLA.2018.00213.
- [4] X. Zhang, M. H. Chen, and Y. Qin, “NLP-QA Framework Based on LSTM-RNN,” Proc. - 2nd Int. Conf. Data Sci. Bus. Anal. ICDSBA 2018, pp. 307 –311, 2018, doi: 10.1109/ICDSBA.2018.00065 .
- [5] K. Saengthongpattana, T. Supnithi, and N. Soonthornphisaj, “Quality Classification of ASEAN Wikipedia Articles using Statistical Features,” 2018 Int. Jt. Symp. Artif. Intell. Nat. Lang. Process. iSAI-NLP 2018 - Proc., pp. 1 – 6, 2018, doi: 10.1109/iSAINLP.2018.8692954
- [6] N. Nakashole, G. Weikum, and F. M. Suchanek. Patty: A taxonomy of relational patterns with semantic types. In EMNLP-CoNLL, pages 1135–1145, 2012.
- [7] J Ding et al : The Semantic Web – ISWC 2019: 18th International Semantic Web ..., Partie 1, Page 176, 2019.