QA4LOV: A Natural Language Interface to Linked Open Vocabulary

Ghislain Auguste Atemezing[1], Pierre-Yves Vandenbussche[2]

[1] MONDECA, 35 Boulevard de Strasbourg, Paris, France.
[2] Fujitsu, Galway, Ireland.
ghislain.atemezing@mondeca.com
pierre-yves.vandenbussche@ie.fujitsu.com

Abstract. There is an increasing presence of structured data due to the adoption of Linked data principles on the web. At the same time, web users have different skills and want to be able to interact with Linked datasets in various manner, such as asking questions in natural language. This paper proposed a first implementation of Query Answering system (QA) applied to the Linked Open Vocabularies (LOV) catalogue, mainly focused on metadata information retrieval. The goal is to provide to end users yet another access to metadata information available in LOV using natural language questions.

Keywords: Question Answering, Vocabulary Catalogue, data usage, user experience

1 Introduction

The recent years have seen the adoption of semantic web in many domains, generating a mass of structured data available in RDF. Linked data has contributed to interlink different datasets across domains, where publishers use best practices for producing interoperable datasets by reusing ontologies and making alignments.

However, users need to have a minimal knowledge of SPARQL language and RDF skills to query RDF datasets. It is one barrier that is overcome by Question Answering (QA) systems, which directly take as input questions in natural language. The need for more advanced tools and QA systems that operate over large repositories of Linked Data has also been the motivation for the QALD Question Answering over Linked Data (QALD) series of workshops [1].

Vocabulary catalogues are special datasets where classes and properties are used to model and generate datasets available in the Linked Data space. Most vocabulary catalogues provide terms search and APIs to access their datasets. LOV provides five types of search methods: metadata search, ontology search, APIs access, dump file in RDF and SPARQL endpoint access [3].

This paper presents a prototype for a vocabulary backed question answering system that can transform natural language questions into SPARQL queries,

thus giving the end users access to the information stored in vocabulary repositories. The paper is structured as follows: Section 2 describes the system, followed by the set of questions in Section 3. An evaluation is presented in Section 4 and a short conclusion and future work in Section 5.

2 System Description

The system receives as input a question formulated in English and outputs the query that will retrieve the answer to the question from the LOV catalogue. The architecture of the system is illustrated in Fig. 1 and a screenshot of the live demo in Fig. 2. The system is available at `http://lov.okfn.org/dataset/lov/qa`.

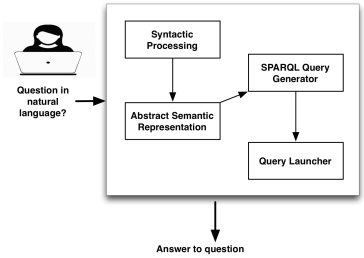

Fig. 1: System architecture

Fig. 2: A screenshot showing the system answering to the contributors of adms

The implementation uses the Quepy tool from Machinalis [3]. The POS tagset used by Quepy is the Penn Tagset [2]. First, regular expressions are defined to match the natural language questions and transform them into an abstract semantic representation. Then, specific templates are defined for the system to handle users' questions. To handle regular expressions, Quepy uses the `refo` library[4] which work with regular expressions as objects.

[3] `https://github.com/machinalis/quepy/`
[4] `https://github.com/machinalis/refo`

A vocabulary is defined by a fixed relation `voaf:Vocabulary` and a POS associated to a `vann:preferredNamespacePrefix`. LOV uses a unique prefix to identify namespaces, which is a string from length 2 to length 17, although the recommendation for publishers is to use a prefix with less than 10 characters.

The syntactic processor is based on regular expressions using POS terms. As a vocabulary is identified by its prefix, we use the following syntactic patterns: NN, NNS, FW, DT, JJ and VBN. Each question from Q1 to Q14 is associated to a unique template. After, when a prefix is recognized, the semantic interpreter uses fixed relations with the English tag, which are properties in RDF triple pattern. The code below show a regex for the contributors of a vocabulary

```
regex1 = Pos("WP") + Lemma("be") + Pos("DT") + Lemma("contributor") + \
        Pos("IN") + Vocabulary()
```

3 Types of Questions

LOV is a highly curated observatory of the semantic vocabularies ecosystem, with the aim of promoting the reuse of well-documented vocabularies in the Linked Data space. Each version of a vocabulary in LOV contains relevant metadata information which can be discovered by agents.

Additionally, users could be interested in other interesting facts, such as the number of versions, the number of datasets using the vocabulary, the number of external vocabularies reusing the vocabulary and the category to which belong the vocabulary. A first set of 14 questions in natural language can be handle by the prototype, covering different type of metadata information available in a vocabulary. Table 1 shows the list of questions, where in the column "Template", [be] can be either the present or the past form of the verb, and [vocab] is the preferred prefix of the vocabulary. The regex uses lemmas to combine different types of the questions accordingly.

4 System Evaluation

The system allows users to interact with the LOV catalogue through the answers generated. Depending on the types of the results (e.g., agents, versions, categories), the system allows users to further explore the dataset with more interactions. All the questions which generated SPARQL query gives satisfactory results. The most challenging issue is to determine the most suitable POS that cover all the vocabulary prefixes. For example, out of 528[5] vocabularies in LOV, 13 of them contain an hyphen and the system can not generate query (e.g., `elseweb-modelling`). Moreover, 21 prefixes contain a number (e.g., `g50k`) and 3 special cases (`homeActivity`, `LiMo`, and `juso.kr`)and are not currently covered by the system. Currently, the system handles 92,99% of the prefixes in LOV.

[5] This number corresponds to the total number of vocabularies inserted in LOV as of January, 8th 2016.

Table 1: Questions in natural language for retrieving metadata information in a vocabulary catalog.

ID	Template	Sample Question
Q1	*What [be] [vocab]?*	*What is prov?*
Q2	*Where [be] [vocab] from?*	*Where is foaf from?*
Q3	*How old [be] [vocab]?*	*How old is prov?*
Q4	*When [be] [vocab] release?*	*When was voaf release?*
Q5	*Who [be] the contributors of [vocab]?*	*Who are the contributors of adms?*
Q6	*When [be] [vocab] last update?*	*When was schema last update?*
Q7	*What [be] the versions of [vocab]?*	*What are the versions of adms?*
Q8	*What [be] the languages of [vocab]?*	*What are the languages of dcat?*
Q9	*Where to find [vocab] documentation?*	*Where to find foaf documentation?*
Q10	*How many vocabularies reuse [vocab]?*	*How many vocabularies reuse adms?*
Q11	*How many datasets use [vocab]?*	*How many datasets use adms?*
Q12	*What [be] the namespace of [vocab]?*	*What is the namespace of dcterms?*
Q13	*What [be] the title of [vocab]?*	*What is the title of foaf?*
Q14	*What [be] the category of [vocab]?*	*What is the category of dcterms?*

5 Conclusion and Future Work

In this paper, we have presented a prototype system for answering a set of questions in natural language backed by a vocabulary catalogue. Accessing LOV dataset by this system will hugely help lay users without SPARQL skills to interact more with the catalogue, and improve also the quality of the metadata by ontology publishers. The implementation uses the LOV dataset in RDF and the Quepy tool. The first results show that the system handles 92,99% of the metadata of vocabularies in the LOV catalogue. We plan to extend the types of queries to more complex ones. Moreover, we can use various semantic relationships in LOV to do query expansion by using for instance sub-properties.

References

1. V. Lopez, C. Unger, P. Cimiano, and E. Motta. Evaluating question answering over linked data. *Web Semantics Science Services And Agents On The World Wide Web*, 21:3–13, 2013.
2. M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
3. P.-Y. Vandenbussche, G. A. Atemezing, M. Poveda-Villalónc, and B. Vatant. LOV: a gateway to reusable semantic vocabularies on the Web. *Semantic Web Journal*, 2015. `http://www.semantic-web-journal.net/system/files/swj974.pdf`.