QALD-3: Multilingual Question Answering over Linked Data

Elena Cabrio[1], Philipp Cimiano[2], Vanessa Lopez[3], Axel-Cyrille Ngonga Ngomo[4], Christina Unger[2], and Sebastian Walter[2]

[1] INRIA Sophia-Antipolis, France
elena.cabrio@inria.fr
[2] CITEC, Universität Bielefeld, Germany
{cimiano,cunger}@cit-ec.uni-bielefeld.de;
swalter@techfak.uni-bielefeld.de
[3] IBM Research, Dublin, Ireland
vanlopez@ie.ibm.com
[4] Universität Leipzig, Germany
ngonga@informatik.uni-leipzig.de

Abstract. The third edition of the open challenge on Question Answering over Linked Data (QALD-3) has put a strong emphasis on multilinguality. This paper provides an overview of the first task, focusing on multilingual question answering, which attracted six teams to submit results.

1 Introduction

Recently there has been much progress towards the goal to provide web users with natural language access to structured data. In particular, natural language interfaces to the Web of Data have the advantage that they can exploit the expressivity of semantic data models to answer complex user queries, while at the same time hiding their complexity from the user. In this context, multilinguality has become more and more important, as both the number of actors creating and publishing data in languages other than English, as well as the amount of users that access this data and speak native languages other than English is growing substantially. In order to achieve the goal that users from all countries have access to the same information, there is an impending need for systems that can help in overcoming language barriers by facilitating multilingual access to semantic data originally produced for a different culture and language.

Since the main objective of the open challenges on *question answering over linked data*[1] (QALD) is to provide an up-to-date and challenging dataset that establishes a standard benchmark against which question answering systems over structured data can be evaluated and compared, we considered it now time to enrich the challenge by aspects related to multilinguality.

[1] http://www.sc.cit-ec.uni-bielefeld.de/qald

2 The datasets

In order to evaluate and compare question answering systems on the task of extracting correct answers for natural language questions or corresponding keywords from given RDF repositories, we provided three datasets:

- English DBpedia 3.8 (`http://dbpedia.org`)
- Spanish DBpedia (`http://es.dbpedia.org`)
- MusicBrainz (`http://musicbrainz.org`)

MusicBrainz is a collaborative effort to create an open content music database. The dataset provided for the challenge is an RDF export containing all classes (artists, albums and tracks) and the most important properties of the MusicBrainz database, building on the Music Ontology[2].

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available as RDF data. The RDF dataset provided for the challenge is the official DBpedia 3.8 dataset for English, including multilingual labels and links, in particular to to YAGO[3] categories and MusicBrainz. Since 2011, information from Wikipedia is extracted also in 15 non-English languages, including Spanish. So far, the English DBpedia contains 400 million RDF triples and the Spanish DBpedia contains almost 100 million RDF triples.

In addition to the datasets, we released 100 English training questions for MusicBrainz and 100 training questions for DBpedia in six different languages: English, Spanish, German, Italian, French and Dutch, as well as slightly adapted 50 training questions over Spanish DBpedia. The questions are of different complexity levels and are annotated with manually specified SPARQL queries and answers, as well as information on the answer type and whether the question requires aggregation operations beyond simple triple matching (e.g. counting and filters) in order to be answered. An example of a question from the DBpedia training set is given in Figure 1, while an example of a question from the Spanish DBpedia training set is shown in Figure 2. Along with a unique ID, the following attributes are specified for each question:

- `answertype` gives the answer type, which can be one the following:
  - `resource`: One or many resources, for which the URI is provided.
  - `string`: A string value such as `Valentina Tereshkova`.
  - `number`: A numerical value such as `47` or `1.8`.
  - `date`: A date provided in the format `YYYY-MM-DD`, e.g. `1983-11-02`.
  - `boolean`: Either `true` or `false`.
- `aggregation` indicates whether any operations beyond triple pattern matching are required to answer the question (e.g., counting, filters, ordering, etc.).

[2] `http://musicontology.com`

[3] For detailed information on the YAGO class hierarchy, please see `http://www.mpi-inf.mpg.de/yago-naga/yago/`.

- onlydbo is given only for questions on English DBpedia and reports whether the query relies exclusively on concepts from the DBpedia ontology; similarly for onlyesdbp for questions on Spanish DBpedia.

During the test phase, participating systems were then evaluated with respect to precision and recall on similarly annotated test questions in the same languages (99 for English DBpedia, 50 for Spanish DBpedia and MusicBrainz each).

```
<question id="40" answertype="resource"
                  aggregation="true"
                  onlydbo="true">
<string lang="en">
What is the highest mountain in Australia?
</string>
<string lang="de">
Was ist der höchste Berg in Australien?
</string>
<string lang="es">
¿Cuál es la montaña más alta de Australia?
</string>
<string lang="it">
Qual è la montagna più alta d'Australia?
</string>
<string lang="fr">
Quelle est la plus haute montagne d'Australie?
</string>
<string lang="nl">
Wat is de hoogste berg van Australië?
</string>
<keywords ... />
<query>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX res: <http://dbpedia.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?uri
WHERE {
   ?uri rdf:type dbo:Mountain .
   ?uri dbo:locatedInArea res:Australia .
   ?uri dbo:elevation ?elevation .
}
ORDER BY DESC(?elevation) LIMIT 1
</query>
</question>
```

Fig. 1. Example question from the English DBpedia training question set, provided in six different languages

```
<question id="37" answertype="resource"
                  aggregation="false"
                  onlyesdbp="true">
<string lang="en">
Give me all films produced by Pedro Almodóvar.
</string>
<string lang="es">
Dame todas las películas producidas por Pedro Almodóvar.
</string>
<query>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX esdbp: <http://es.dbpedia.org/property/>
PREFIX esres: <http://es.dbpedia.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?uri
WHERE {
  ?uri rdf:type dbo:Film .
  ?uri esdbp:producción esres:Pedro_Almodóvar .
}
</query>
</question>
```

Fig. 2. Example question from the Spanish DBpedia training question set

The availability of the DBpedia dataset in two languages and gold standard questions in six languages gives rise to different levels of difficulty of the task, ranging from question answering over English DBpedia in English and question answering over Spanish DBpedia in Spanish, to question answering over English DBpedia in German, French, Italian and Dutch. Including MusicBrainz as an additional dataset furthermore keeps the complexity that the QALD challenges already reached with respect to the major challenges involved in querying linked data, independent of multilinguality.

As an additional challenge, a few of the training and test questions were out of scope, i.e. they could not be answered with respect to the dataset, in order to test systems on the ability to distinguish whether an empty answer is due to a failure of the system or due to the fact that no answer is contained in the data. Further, seven questions were provided that require both the DBpedia and MusicBrainz dataset in order to be answered.

3 Evaluation measures

The results submitted online by participating systems were automatically compared to the gold standard results. Participating systems were evaluated with respect to precision and recall. Moreover, participants were encouraged to report performance, i.e. the average time their system takes to answer a query. For each question q, precision, recall and F-measure were computed as follows:

$$Recall(q) = \frac{\text{number of correct system answers for } q}{\text{number of gold standard answers for } q}$$

$$Precision(q) = \frac{\text{number of correct system answers for } q}{\text{number of system answers for } q}$$

$$F\text{-}Measure(q) = \frac{2 * Precision(q) \times Recall(q)}{Precision(q) + Recall(q)}$$

On the basis of these measures, overall precision and recall values as well as an overall F-measure value were computed as the average mean of the precision, recall and F-measure values for all questions. Below, precision, recall and F-measure values refer to the averaged values.

In order to access the datasets, they could either be downloaded or queried by means of the provided SPARQL endpoints. For the Spanish DBpedia the evaluation took place with respect to the official Spanish DBpedia endpoint. For the English DBpedia, the evaluation took place with respect to the provided SPARQL endpoint (not the official one), in order to ensure invariable and comparable results. Submissions of results by participating systems were required in the provided XML format to facilitate the automatic comparison of the answers provided by the system with the ones provided by the gold standard XML document.

4 Results

Six participating systems submitted results, all of them on the English DBpedia question set and one also on the MusicBrainz question set. In the following, we give some details on the participating systems and the results.

4.1 Overview of the evaluated systems

The six participating systems follow different approaches to interpret and answer questions over linked data. Some approaches rely on linguistic techniques and syntactic patterns, while others implement a statistical approach. Among the external sources used by participating systems, Wikipedia and WordNet are the most commonly exploited for semantic knowledge extraction (e.g., to discover similarities across words). External tools for text processing are used for preprocessing and analysing the queries, in particular Stanford CoreNLP[4], MaltParser[5] and Chaos[6], while Information Retrieval tools such as Lucene[7] are used to either create indexes (e..g, of Wikipedia version) or to obtain similarity scores. In

[4] http://nlp.stanford.edu/software/corenlp.shtml

[5] http://www.maltparser.org/

[6]

[7] http://lucene.apache.org/core/

what follows we provide some basic information about each of the participating systems.

SWIP, submitted by the IRIT group from the University of Toulouse in France, is based on query patterns to address the task of interpreting natural language queries. The query interpretation process consists of two main steps. First, the natural language question is translated into a pivot query, capturing the query focus and the dependencies and relations between substrings of the natural language question. Second, the pivot query is mapped to predefined query patterns, obtaining a list of potential interpretations of the user question. The interpretations are then ranked according to their estimated relevance and proposed to the user in form of reformulated natural language questions.

CASIA, submitted by the National Laboratory of Pattern Recognition and the institute of Automation from the Chinese Academy of Sciences in Beijing, implements a pipeline consisting of question analysis, resource mapping and SPARQL generation. More specifically, the system first transforms natural language questions into a set of *query triples* of the form <subject,predicate,object>, based on a shallow and deep linguistic analysis. Second, it instantiates these query triples with corresponding resources from DBpedia, resulting in *ontology triples*. Third, based on the ontology triples and question type, SPARQL queries are constructed. Finally, the candidate queries are validated and ranked, and the best query is selected.

Squall2sparql, by IRISA in the University of Rennes, France, is a translator from SQUALL, a controlled natural language for English, to SPARQL. Given a SQUALL sentence, the system first translates it into an intermediate logical representation based on Montague grammar. This intermediate representation is then translated into SPARQL by mapping logical constructs to combinations of SPARQL constructs.

Scalewelis, also by IRISA in the University of Rennes, is a faceted search system that guides the user through the search for an answer. Starting from an initial SPARQL query, facets are created for the first 1,000 results retrieved by that query, consisting of the classes the results belong to as well as properties that relate the results to other entities in the dataset. The user's selection of a facet is then used to refine the query until the answer is found.

The RTV system, by the Enterprise Engineering department in the University of Rome Tor Vergata, integrates lexical semantic modelling and statistical inferences within an architecture that decomposes the natural language interpretation task into a cascade of three different stages: i) the selection of salient information from the question (i.e. predicate, arguments and properties), ii) the location of the salient information in the ontology through joint disambiguation of all candidates, and iii) the compilation of the final query against RDF triples. This architecture exploits a Hidden Markov Model (HMM) to select the proper ontological triples according to the graph nature of RDF. In particular, for each query, an HMM model is produced whose Viterbi solution is the comprehensive joint disambiguation across the sentence elements.

Intui2, by the University of Tubingen in Germany, analyses the questions in terms of the syntactic constituents, so-called *synfragments*, they are composed of. Syntactically, a *synfragment* corresponds to a subtree of the syntactic parse tree of the question. Semantically, it is a minimal span of text that can be interpreted as a concept URI, an RDF triple or a complex RDF query. These *synfragments* are then compositionally combined to an interpretation of the whole input question.

4.2 Evaluation results

Tables 1 and 2 show the results obtained by the participating systems over DBpedia and MusicBrainz datasets, respectively. The column *processed* states for how many of the questions the system provided an answer, *right* specifies how many of these questions were answered with an F-measure of 1, and *partially* specifies how many of the questions were answered with an F-measure strictly between 0 and 1. On the DBpedia dataset, the best F-measure was 0.9 and the lowest was 0.17, the average being 0.4. These results are comparable to the results achieved in earlier challenges, showing that the level of complexity of the questions is still very demanding.

System	*Total*	*Processed*	*Right*	*Partially*	*Recall*	*Precision*	*F-measure*
squall2sparql	99	99	80	13	0.88	0.93	0.90
CASIA	99	52	29	8	0.36	0.35	0.36
Scalewelis	99	70	32	1	0.33	0.33	0.33
RTV	99	55	30	4	0.34	0.32	0.33
Intui2	99	99	28	4	0.32	0.32	0.32
SWIP	99	21	15	2	0.16	0.17	0.17

Table 1. Results for DBpedia test set

System	*Total*	*Processed*	*Right*	*Partially*	*Recall*	*Precision*	*F-measure*
SWIP	50	33	24	2	0.51	0.51	0.51

Table 2. Results for MusicBrainz test set

The following questions on DBpedia were answered by all systems:

ID	*Question*
21	What is the capital of Canada?
22	Who is the governor of Wyoming?
30	What is the birth name of Angela Merkel?
68	How many employees does Google have?

And the following questions on DBpedia were answered by no systems:

ID	*Question*
14	Give me all members of Prodigy.
16	Does the new Battlestar Galactica series have more episodes than the old one?
92	Show me all songs from Bruce Springsteen released between 1980 and 1990.
96	Give me all B-sides of the Ramones.

While question answering systems over structured data can greatly benefit from exploiting ontological relationships in order to understand and disambiguate a query, inheriting relationships and linking word meanings across datasets, a great challenge for this type of systems lies in being able to deal with the heterogeneity and noise intrinsic in the large amount of interlinked data. The openness of the domain and the datasets used in this evaluation are large enough to raise both scalability and heterogeneity issues. An important challenge lies in being able to map implicit relations in the input, indicated by light verbs (e.g. by *to be* and *to have*) or light prepositions like *of* and *with* to explicit relations in the corresponding SPARQL query. For example, a linguistically simple question such as question 39 (Give me all companies in Munich) can be translated into any of three RDF properties `dbp:location`, `dbo:headquartered` and `dbo:locationCity`. A question answering system should be aware of all this possible translations to have a good recall.

Of the questions in the test set, 45 queries require to search the answer using other namespaces than the DBpedia ontology (attribute `onlydbo=false`), such as YAGO or FOAF, and 19 queries require aggregation operations (attribute `aggregation=true`), such as comparisons, like in 16 above, superlatives, like in question 15 (What is the longest river?), or filtering, like in 92 above. It is especially on these complex queries that the systems perform poorly on.

Further, the challenge to identify out-of-scope questions was addressed only by one system, squall2sparql. The reason for the excellent performance of this system is due to the fact that the questions have been first manually translated into the SQUALL controlled languages and terms have been mapped to URIs, thus removing many ambiguities already.

The QALD website contains detailed information about the precision and recall per question for all systems, thus providing the basis for a detailed comparison of the different systems.

5 Summary

The goal of the QALD challenge is to provide a non-trivial benchmark that allows to systematically compare different systems under the same conditions. The evaluation results indicate that the challenge is far from easy, with systems being still quite far away from answers all questions correctly. Nevertheless, most systems achieved decent F-measures between 32% and 36%, showing that the task is in principle feasible. We are optimistic that in the future the results on the challenge will steadily increase as the systems are developed further and become more mature.

Unfortunately, none of the systems worked on the Spanish DBpedia dataset and none of the systems used natural language questions other than the English ones. This clearly shows that the state-of-the-art is not yet that advanced and that research is still struggling to provide answers in one language. Nevertheless, the dataset is out for public use and in the next years we will surely see results on other languages published.

Overall, we feel that we have provided a solid basis for future research on question answering over linked data by providing a challenging and exciting benchmark dataset to work on, allowing to systematically compare different systems to each other under the same conditions.