Question Answering over Linked Data (QALD-5)

Christina Unger[1], Corina Forascu[2], Vanessa Lopez[3], Axel-Cyrille Ngonga Ngomo[4], Elena Cabrio[5], Philipp Cimiano[1], and Sebastian Walter[1]

[1] CITEC, Bielefeld University, Germany
cunger@cit-ec.uni-bielefeld.de
cimiano@cit-ec.uni-bielefeld.de
swalter@techfak.uni-bielefeld.de
[2] Alexandru Ioan Cuza University of Iasi, Romania
corinfor@info.uaic.ro
[3] IBM Research, Dublin, Ireland
vanlopez@ie.ibm.com
[4] AKSW, University of Leipzig, Germany
ngonga@informatik.uni-leipzig.de
[5] INRIA Sophia-Antipolis Méditerrané, Cedex, France
elena.cabrio@inria.fr

1 Introduction

While more and more structured data is published on the web, the question of how typical web users can access this body of knowledge becomes of crucial importance. Over the past years, there is a growing amount of research on interaction paradigms that allow end users to profit from the expressive power of Semantic Web standards while at the same time hiding their complexity behind an intuitive and easy-to-use interface. Especially natural language interfaces have received wide attention, as they allow users to express arbitrarily complex information needs in an intuitive fashion and, at least in principle, in their own language. Multilingualism has, in fact, become an issue of major interest for the Semantic Web community, as both the number of actors creating and publishing data all 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. The key challenge is to translate the user's information needs into a form such that they can be evaluated using standard Semantic Web query processing and inferencing techniques. Over the past years, a range of approaches have been developed to address this challenge, showing significant advances towards answering natural language questions with respect to large, heterogeneous sets of structured data. However, a lot of information is still available only in textual form, both on the web and in the form of labels and abstracts in linked data sources. Therefore approaches are needed that can not only deal with the specific character of structured data but also with finding information in several sources, processing both structured and unstructured information, and combining such gathered information into one answer.

With the increasing 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. Especially question answering systems have received wide attention, as they allow users to express arbitrarily complex information needs in an easy and intuitive fashion (for an overview see [3]). The key challenge lies in translating the users' information needs into a form such that they can be evaluated using standard Semantic Web query processing and inferencing techniques. Over the past years, a range of approaches have been developed to address this challenge, showing significant advances towards answering natural language questions with respect to large, heterogeneous sets of structured data. However, only a small number of systems yet address the fact that the amount of users speaking native languages other than English is growing substantially. Also, a lot of information is still available only in textual form, both on the web and in the form of labels and abstracts in linked data sources. Therefore approaches are needed that can not only deal with the specific character of structured data but also with finding information in several sources, processing both structured and unstructured information, possibly in different languages, and combining such gathered information into one answer.

The main objective of the open challenge on *question answering over linked data*[6] [2] (QALD) is to provide up-to-date, demanding benchmarks that establish a standard against which question answering systems over structured data can be evaluated and compared. QALD-5 is the fifth instalment of the QALD open challenge and focuses on multilingual and hybrid question answering as part of the Question Answering Lab[7] at CLEF 2015.

2 Task description

QALD aims at all question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data. The general task is the following one:

> Given a natural language question or keywords, retrieve the correct answer(s) from a given repository containing both RDF data and free text, in this case the English DBpedia 2014 dataset[8] with free text abstracts.

To get acquainted with the dataset and possible questions, a set of training questions was provided, comprising 300 multilingual questions as well as 40 hybrid questions. These questions were compiled from the QALD-4 training and test questions, slightly modified in order to account for changes in the DBpedia dataset. In the case of hybrid questions they were also building on the data provided by the INEX Linked Data track[9]. Later, systems were evaluated on 60 different test questions, comprising 50 multilingual ones and 10 hybrid ones. These questions were mainly devised by the challenge organizers.

[6] `http://www.sc.cit-ec.uni-bielefeld.de/qald`
[7] `http://nlp.uned.es/clef-qa/`
[8] `http://dbpedia.org`
[9] `http://inex.mmci.uni-saarland.de/tracks/dc/index.html`

Multilingual questions are provided in seven different languages (English, German, Spanish, Italian, French, Dutch, and Romanian) and can be answered with respect to the provided RDF data. They are annotated with corresponding SPARQL queries and answers retrieved from the provided SPARQL endpoint.

Hybrid questions are provided in English and can be answered only by integrating structured data (RDF) and unstructured data (free text available in the DBpedia abstracts). The questions thus all require information from both RDF and free text. They are annotated with pseudo-queries that show which part is contained in the RDF data and which part has to retrieved from the free text abstracts.

Annotations are provided in an XML format. The overall document is enclosed by a tag that specifies an ID for the dataset indicating whether it belongs to training or test (i.e. `qald-5_train` and `qald-5_test`).

```
<dataset  id="qald-5_train">
<question id="1">   ... </question>
...
<question id="340"> ... </question>
</dataset>
```

For each of the questions, a question string and a corresponding query as well as the correct answer(s) were provided. In addition to a unique ID, questions were also annotated with the following attributes:

- `answertype` specifies the expected answer type, which can be one the following: `resource` (one or many resources, for which the URI is provided), `string` (a string value), `number` (a numerical value such as `47` or `1.8`), `date` (e.g. `1983-11-02`), `boolean` (either `true` or `false`).
- `hybrid` specifies whether the question is a hybrid question, i.e. requires the use of both RDF and free text data
- `aggregation` indicates whether any operations beyond triple pattern matching are required to answer the question (e.g., counting, filters, ordering, etc.).
- `onlydbo` reports whether the query relies solely on concepts from the DBpedia ontology. If the value is `false`, the query might rely on the DBpedia property namespace (`http://dbpedia.org/property/`), FOAF or YAGO.

For hybrid questions, the attributes `aggregation` and `onlydbo` refer to the RDF part of the query only.

For multilingual questions, the question string is provided in seven languages: English, German, Spanish, Italian, French, Dutch, and Romanian, together with keywords in the same seven languages. The corresponding SPARQL query can be executed against the DBpedia endpoint in order to retrieve the specified answers. Here is an example, leaving out string tags and keywords:

```
<question id="272" answertype="resource"
                   aggregation="true"
                   onlydbo="true"
                   hybrid="false">
```

```
Which book has the most pages?
Welches Buch hat die meisten Seiten?
¿Que libro tiene el mayor numero de paginas?
Quale libro ha il maggior numero di pagine?
Quel livre a le plus de pages?
Welk boek heeft de meeste pagina's?
Ce carte are cele mai multe pagini?

<query>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?uri
WHERE {
        ?uri rdf:type dbo:Book .
        ?uri dbo:numberOfPages ?n .
}
ORDER BY DESC(?n)
OFFSET 0 LIMIT 1
</query>

<answers>
<answer>
http://dbpedia.org/resource/The_Tolkien_Reader
</answer>
</answers>

</question>
```

For the hybrid questions, not only the RDF triples comprised by DBpedia are relevant, but also the English abstracts. They are related to a resource by means of the property `abstract`. The questions are annotated with answers as well as a pseudo query that indicates which information from the RDF and which information from the free text abstracts have to be combined in order to find the answer(s). The pseudo query is like an RDF query but can contain free text as subject, property, or object of a triple. This free text is marked as `text:"..."`. Here is an example:

```
<question id="335" answertype="resource"
                   aggregation="false"
                   onlydbo="true"
                   hybrid="true">

<string lang="en">
Who is the front man of the band that wrote Coffee & TV?
</string>

<pseudoquery>
PREFIX res: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?uri
```

```
WHERE {
        res:Coffee_&_TV dbo:musicalArtist ?x .
        ?x dbo:bandMember ?uri .
        ?uri text:"is" text:"frontman" .
}
</pseudoquery>

<answers>
<answer>http://dbpedia.org/resource/Damon_Albarn</answer>
</answers>

</question>
```

The pseudo query contains three triples—two RDF triples and the third containing free text as property and object. The way to answer the question is to first retrieve the band members of the musical artist associated with the song Coffee & TV using the triples

```
res:Coffee_&_TV dbo:musicalArtist ?x .
?x dbo:bandMember ?uri .
```

and then check the abstract of the returned URIs for the information whether they are the frontman of the band. In this case, the abstract of Damon Albarn contains the following sentence:

```
He is best known for being the frontman of the Britpop/alter-
native rock band Blur [...]
```

Overall, of the 350 training questions, 59 questions require aggregation and 102 questions require namespaces other than from the DBpedia ontology (21 of which use the YAGO namespace, 2 require FOAF, and all others rely on the DBpedia property namespace). Of the 60 test questions, 15 questions require aggregation and 12 cannot be answered with the DBpedia ontology only (3 of which use the YAGO namespace, all others rely on the DBpedia property namespace). As an additional challenge, 14 training and 1 test question are out of scope, i.e. they cannot be answered with respect to the dataset. One example is `Give me all animal species that live in the Amazon rainforest`.

3 Evaluation measures

The results submitted by participating systems were automatically compared to the gold standard results and evaluated with respect to precision and recall. 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. In the results reported below, precision, recall and F-measure values refer to the averaged values.

4 Participating systems

Seven teams participated in QALD-5. Two participants submitted results only for the multilingual questions, two participants submitted results only for the hybrid questions, and three participants submitted results for both kinds of questions. Although the overall number of participants is one less than in last year's challenge, the number of participating hybrid question answering systems increased from one to five, which shows an important advancement in the field. However, all systems still processed only the English questions, not yet addressing the issue of multilinguality.

In the following, we give some details on the participating systems.

Xser [7] takes as input a natural language question in English, and retrieves an answer in two steps. First the user query is linguistically analyzed in order to detect predicate argument structures through a semantic parser. Second the query is instantiated with respect to the knowledge base. Besides the DAG dependency parsing it relies on a structured prediction approach implemented using a Collins-style hidden perceptron. The system requires training data but among all participants obtained the highest precision and recall values.

APEQ presents an approach to QA over linked data that is based on graph traversal techniques. The question are first analyzed in terms of phrase structure. A main entity is determined using some heuristics and the RDF graph is explored from that main entity outwards to discover relations to the other entities mentioned in the query, guided by relations in the parse tree. A number of path ranking measures are proposed to rank the different graphs. The best scoring entity according to the path measures is returned.

QAnswer [5] first parses the question with Stanford CoreNLP to generate a directed graph, where the vertices correspond to the tokens of the question annotated with lemma and part-of-speech tags, and the edges correspond to the collapsed dependencies. To detect DBpedia individuals, types and properties in such graph, specific methods are respectively applied (also exploiting expressions extracted from Wikipedia). Among the graphs generated applying such strategies, only the most probable is then selected (relying on a set of scores), and missing entities are inferred, while existing ones are validated using the ontology. The SPARQL query is finally generated as the last step, creating triples and subqueries based on the graph structure and the direction of the properties. In the current implementation, QAnswer targets `onlydbo` questions only.

SemGraphQA [1] is a graph-based approach to transforming natural language questions into SPARQL queries. First, phrases in the question are matched

with elements in the knowledge base (classes, properties, and individuals). In parallel, a syntactic graph is built by dependency parsing the question. Those syntactic graphs are then transformed into possible semantic graphs, the structure of which is guided by both the syntactic structure and the possible mappings of phrases to knowledge base elements. The resulting semantic graphs comprises all possible, coherent interpretations, which are scored and finally converted into SPARQL queries. This approach requires no training data and can easily be ported to new datasets.

YodaQA targets both multilingual and hybrid questions. It first represents the input question as a bag-of-features (e.g. keywords, keyphrases and concept clues that crisply match Wikipedia titles), then generates a set of candidate answers by performing a search in the knowledge bases according to such features (either directly using search results as candidate answers or filtering relevant passages from these and generating candidate answers from the selected passages). Various answer features are then generated based e.g. on the lexical types determination, coercion to question type, distance from clues in passages or text overlap with clues. A machine learning classifier (logistic regression) is finally applied to score the answers by their features.

ISOFT [4] focuses on hybrid queries. It first analyses the natural language question, which includes named entity recognition, determining the expected answer type, and decomposing the question into phrases. The phrases are then searched for in a text database, a processed and annotated version of the text corpus. In case this search fails or if the phrase interpretation requires aggregation operations (e.g. superlatives), the system builds a corresponding SPARQL query to search the RDF database for an answer. Finally, phrase interpretations are combined and the results are filtered according to the expected answer type.

HAWK [6] also focuses on hybrid queries. The framework begins by generating a dependency parse tree of the user query. The resulting parse tree is pruned by using manually crafted rules. The resulting pruned tree is then used to generate potential SPARQL queries. To this end, entities and nouns are recognized by using FOX and AGDISTIS. If no matching resource is found for a given entity then a slot for a text query is created. Each of the edge in the tree is mapped to a basic graph pattern. Valid combinations of basic graph patterns (according to the ontology of the target knowledge base) are kept as potential query candidates. The resulting hybrid queries are finally ranked using a ranking function learned out of the test dataset. The ranked SPARQL queries are issued in order.

5 Results

Tables 1 and 2 report on the results obtained by the participating systems on the multilingual and hybrid questions, respectively. The first column specifies the system name (together with the language it processed in case of multilingual questions), *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-1 measure of 1, *Partial* specifies how many of the questions were answered with an F-1 measure strictly between 0 and 1, *Recall*, *Precision* and *F-1* report the measures with respect to the number of processed questions. *F-1 Global* in addition reports the F-1 measure with respect to the total number of questions.

Table 1. *Results for multilingual question answering. Total number of questions: 50*

	Processed	Right	Partial	Recall	Precision	F-1	F-1 Global
Xser (en)	42	26	7	0.72	0.74	**0.73**	**0.63**
APEQ (en)	26	8	5	0.48	0.40	**0.44**	**0.23**
QAnswer (en)	37	9	4	0.35	0.46	**0.40**	**0.30**
SemGraphQA (en)	31	7	3	0.32	0.31	**0.31**	**0.20**
YodaQA (en)	33	8	2	0.25	0.28	**0.26**	**0.18**

Table 2. *Results for hybrid question answering. Total number of questions: 10*

	Processed	Right	Partial	Recall	Precision	F-1	F-1 Global
ISOFT	3	2	1	1.00	0.78	**0.87**	**0.26**
HAWK	3	1	0	0.33	0.33	**0.33**	**0.10**
YodaQA	10	1	0	0.10	0.10	**0.10**	**0.10**
SemGraphQA	6	0	0	0.00	0.20	**0.00**	**0.00**
Xser	3	0	0	0.0	0.00	**0.00**	**0.00**

The results for multilingual question answering show a slight improvement over last year's challenges, with an average F-measure of 0.43 (compared to an average F-measure of 0.33 last year). This shows that the systems address more of the difficulties contained in the QALD benchmark, while the level of complexity of the questions remains demanding. Similar to earlier challenges, the biggest problem is still the matching of natural language expressions to correct vocabulary elements, especially when the semantic structure of the question and the structure of the query differ. For example, the following questions were not answered by any of the participating systems:
Which animals are critically endangered?

```
SELECT DISTINCT ?uri
WHERE {
        ?uri rdf:type dbpedia-owl:Animal .
        ?uri dbpedia-owl:conservationStatus 'CR' .
}
```

How many scientists graduated from an Ivy League university?

```
SELECT DISTINCT count (?uri)
```

```
WHERE {
        ?uri rdf:type dbpedia-owl:Scientist .
        ?uri dbpedia-owl:almaMater ?university .
        ?university dbpedia-owl:affiliation dbpedia:
            Ivy_League .
}
```

Finally, for the first time in the still young history of QALD, a sponsorship by *Orange*[10] allows us to award prizes for the best systems in both tiers, multilingual and hybrid question answering, in particular Xser, ISOFT and HAWK.

6 Future perspectives

QALD-5, the fifth edition of the QALD challenge, was successful in attracting participants working on hybrid question answering, i.e. answering user questions by fusing information from both RDF data and free text. But although one of the key aspects of the QALD challenge is multilinguality, all participating systems worked on English data only. This shows that the multilingual scenario is still not broadly addressed. There are two measures we plan to try in future challenges: First, to directly reach out to people working on question answering (e.g. in Korean, Vietnamese, and other languages), in order to add those languages to the QALD benchmark. And second, to announce a special award to the first participating system(s) processing questions in another language than English.

In future challenges we want also want to emphasize further aspects of question answering over linked data, such as querying data cubes, in order to continue to provide a state-of-the-art benchmark for systems that offer end users an intuitive and easy-to-use access to the huge amount of data present on the Semantic Web.

References

1. Romain Beaumont, Brigitte Grau, and Anne-Laure Ligozat. SemGraphQA@QALD5: LIMSI participation at QALD5@CLEF. In *CLEF 2015 Working Notes Papers*, 2015.
2. Vanessa Lopez, Christina Unger, Philipp Cimiano, and Enrico Motta. Evaluation question answering over linked data. *Journal of Web Semantics*, in press.
3. Vanessa Lopez, Victoria S. Uren, Marta Sabou, and Enrico Motta. Is question answering fit for the semantic web?: A survey. *Semantic Web*, 2(2):125–155, 2011.
4. Seonyeong Park, Soonchoul Kwon, Byungsoo Kim, and Gary Geunbae Lee. ISOFT at QALD-5: Hybrid question answering system over linked data and text data. In *CLEF 2015 Working Notes Papers*, 2015.
5. Stefan Ruseti, Alexandru Mirea, Traian Rebedea, and Stefan Trausan-Matu. QAnswer - enhanced entity matching for question answering over linked data. In *CLEF 2015 Working Notes Papers*, 2015.

[10] `http://www.orange.com/en/home`

6. Ricardo Usbeck and Axel-Cyrille Ngonga Ngomo. HAWK@QALD5 – trying to answer hybrid questions with various simple ranking techniques. In *CLEF 2015 Working Notes Papers*, 2015.
7. Kun Xu, Yansong Feng, and Dongyan Zhao. Answering natural language questions via phrasal semantic parsing. In *CLEF 2014 Working Notes Papers*, 2014.