The Fifth Open Challenge on
Question Answering Over Linked Data

QALD-5

Document Version: January 30, 2015

QALD-5 is the fifth in a series of evaluation campaigns on question answering over linked data, with a strong emphasis on multilinguality and hybrid approaches using information from both structured and unstructured data. The challenge is part of the Question Answering Lab at CLEF 2015 (http://nlp.uned.es/clef-qa/). It is aimed at all kinds of systems that mediate between a user, expressing his or her information need in natural language, and semantic data. All relevant information for participating in the challenge are given in this document.

1 Introduction

Motivation and Goal

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 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, 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.

Coordinators

- Philipp Cimiano (CITEC, Universität Bielefeld, Germany)
- Vanessa Lopez (IBM Research, Dublin, Ireland)
- Christina Unger (CITEC, Universität Bielefeld, Germany)
- Elena Cabrio (INRIA Sophia-Antipolis Méditerranée, Cedex, France)
- Axel-Cyrille Ngonga Ngomo (Universität Leipzig, Germany)
- Sebastian Walter (CITEC, Universität Bielefeld, Germany)
- Corina Forăscu (Alexandru Ioan Cuza University, Iasi, Romania)

Sponsor

http://www.orange.com/en/home/

2 Relevant information in a nutshell

QALD-5: http://www.sc.cit-ec.uni-bielefeld.de/qald/ > QALD-5

CLEF 2015 QA lab: http://nlp.uned.es/clef-qa/

Registration: http://clef2015-labs-registration.dei.unipd.it

Task

Given an RDF dataset and natural language questions, return the correct answer(s).

Dataset

- DBpedia 3.10
  http://downloads.dbpedia.org/current/en/

For training and test, two kinds of questions are provided: *multilingual questions* in seven languages, and *hybrid questions* that require both structured (RDF) and unstructured (free text) data to be answered.

Training questions:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/5/qald-5_train.xml

Test questions:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/5/qald-5_test_questions.xml

Test questions will be published on April 20.

SPARQL endpoint

During training phase, please use the official DBpedia endpoint:

http://dbpedia.org/sparql

For test phase we will set up a local endpoint that contains a mirror of DBpedia 3.10, in order to ensure availability and consistency of results among systems.

Schedule

Release of training dataset and instructions:	January 30, 2015
Release of test dataset:	April 20, 2015
Submission deadline:	May 8, 2015
Release of evaluation results:	May 15, 2015
Deadline for working notes submission:	June 7, 2015
Workshop:	September, 2015

Submission and evaluation

Submission of results and evaluation is done by means of an online form:
http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=evaltool&q=5

Results for the training phase can be uploaded at any time; results for the test phase can be uploaded from April 20 to May 8.

Prizes

We offer prizes for the top-scoring teams, both in the multilingual and the hybrid track. More details will be announced soon.

Resources

You are free to use any external resources. A list of potentially useful resources and tools is provided in Section 5.

Contact

Updates on the open challenge will be published on the *Interacting with Linked Data* mailing list:

https://lists.techfak.uni-bielefeld.de/cit-ec/mailman/listinfo/ild

In case of question, problems and comments, please contact Christina Unger:
cunger@cit-ec.uni-bielefeld.de

3 Description of the data

3.1 DBpedia 3.10

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 3.10 dataset for English, including links, most importantly to YAGO[1] categories. This dataset comprises all files provided at:

- http://downloads.dbpedia.org/current/en/
- http://downloads.dbpedia.org/current/links/

In order to work with the dataset, you can either load it into your favorite triple store, or access it via a SPARQL endpoint. The official DBpedia SPARQL endpoint can be accessed at http://dbpedia.org/sparql/. We will also provide a local endpoint later, with respect to which evaluation will take place.

Namespaces that are used in the provided training and test queries are the following ones:

- *DBpedia:*

```
dbo:  <http://dbpedia.org/ontology/>
dbp:  <http://dbpedia.org/property/>
res:  <http://dbpedia.org/resource/>
```

- *RDF(S) and XSD:*

```
rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
xsd:  <http://www.w3.org/2001/XMLSchema#>
```

- *Others:*

```
yago:  <http://dbpedia.org/class/yago/>
foaf:  <http://xmlns.com/foaf/0.1/>
```

3.2 The questions

In order to get acquainted with the dataset and possible questions, a set of training questions can be downloaded at the following location:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-5_train.xml

This set comprises:

- 300 *multilingual questions*
- 40 *hybrid questions*

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

Multilingual questions are provided in seven different languages (English, German, Spanish, Italian, French, Dutch, and Romanian) and can be answered with respect to DBpedia 3.10. 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 ti be retrieved from the free text abstracts.

3.2.1 Annotations

Annotations are provided in the following XML format. The overall document is enclosed by a tag that specifies an ID for the dataset indicating the domain and whether it is train or test (i.e. `qald-4_train` and `qald-4_test`).

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

Each of the questions specifies an ID (don't worry if they are not ordered) together with the following attributes:

- `answertype` gives 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 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`. This format is also required when you submit results containing a date as answer.
  - `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 some YAGO category.

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

Most importantly, each question specifies a question string and a corresponding query (both enclosed in `<![CDATA[...]]>` tags) as well as the correct answer(s). They slightly differ for multilingual and hybrid questions.

3.2.2 Multilingual questions

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:

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

<string lang="en">Which book has the most pages?</string>
<string lang="de">Welches Buch hat die meisten Seiten?</string>
<string lang="es">¿Que libro tiene el mayor numero de paginas?</string>
<string lang="it">Quale libro ha il maggior numero di pagine?</string>
<string lang="fr">Quel livre a le plus de pages?</string>
<string lang="nl">Welk boek heeft de meeste pagina's?</string>
<string lang="ro">Ce carte are cele mai multe pagini?</string>

<keywords lang="en">book, the most pages</keywords>
<keywords lang="de">Buch, meisten Seiten</keywords>
<keywords lang="es">libro, mayor numero paginas</keywords>
<keywords lang="it">libro, maggior numero di pagine</keywords>
<keywords lang="fr">livre, le plus de pages</keywords>
<keywords lang="nl">boek, meeste pagina's</keywords>
<keywords lang="ro">carte, cele mai multe pagini</keywords>

<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>

</question>
```

That is, all question strings and keywords have a language attribute (`lang`) with one of the following values: `en` (English), `de` (German), `es` (Spanish), `it` (Italian), `fr` (French), `nl` (Dutch), `ro` (Romanian).

As an additional challenge, a few of the training and test questions are out of scope, i.e. they cannot be answered with respect to DBpedia. The query is specified as `OUT OF SCOPE` and the answer set is empty. Here is an example:

```
<question id="99" answertype="resource"
                  aggregation="false"
                  onlydbo="false"
                  hybrid="false">

<string lang="en">
Give me all animal species that live in the Amazon rainforest.
</string>
...

<query>
OUT OF SCOPE
</query>

```

```
<answers/>

</question>
```

3.2.3 Hybrid questions

The hybrid questions are partly based on the questions used in the INEX Linked Data track[2] 2013. For a description of the track and the corresponding data, see http://www.clef-initiative.eu/documents/71612/2b349f08-de37-41a9-bb62-40c91f1daa0b.

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`, e.g. as follows:

```
res:Australian_Shelduck dbo:abstract
'The Australian Shelduck, Tadorna tadornoides, is a shelduck,
a group of large goose-like birds which are part of the bird
family Anatidae.  The genus name Tadorna comes from Celtic
roots and means "pied waterfowl".  They are protected under
the National Parks and Wildlife Act, 1974.'@en .
```

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 dbo: <http://dbpedia.org/ontology/>
SELECT DISTINCT ?uri
WHERE {
        <http://dbpedia.org/page/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

[2] https://inex.mmci.uni-saarland.de/

```
<http://dbpedia.org/page/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/alternative rock band
Blur [...]
```

All queries are designed in a way that they require both RDF data and free text to be answered. Note that the free text given in the pseudo query corresponds to the natural language question, not the way the relevant information in the abstract is provided. Retrieving the relevant information from the abstracts thus usually requires some kind of textual entailment. Also note that the pseudo queries cannot be evaluated against the SPARQL endpoint.

4 Description of the task

The general task of QALD-5 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.

The training questions described in the previous section are published so you can familiarize yourself with the dataset and the kind of questions that QALD asks. During test phase, from April 20 to May 8, a set of different (but similar) questions without annotations will be provided at the following location:

> http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/5/qald-5_test_questions.xml

You can participate working on the multilingual or the hybrid questions (or on both). Evaluation of performance on multilingual and hybrid questions is done separately, see 4.3 below.

4.1 Registration

In order to participate, please register for the CLEF 2015 QA lab at

http://clef2015-labs-registration.dei.unipd.it

This is not a registration for CLEF and as such is not strictly binding, but we will later use this to identify you during test phase.

4.2 Submission

Results can be submitted from April 20 to May 8 via the same online form used during training phase (note the drop down box that allows you to specify `test` instead of `training`):

> http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=evaltool&q=5

The only difference is that evaluation results are not displayed. You can upload results as often as you like (for example trying different configurations of your system); in this case the file with the best results will count.

All submissions are required to comply with the XML format specified in the previous section. For all questions, the dataset ID and question IDs are obligatory. Also you have to specify the answer(s) your system returned for that question. For questions that are out of the dataset's scope, the file should contain an empty answer set.

You are allowed to change the natural language question or keywords, for example by inserting quotes around named entities, by reformulating expressions your system struggles with, or even by using some controlled language format. If you do so, please document these changes, i.e. replace the provided question string or keywords by the input you used.

Also, it is preferred if your submission leaves out all question strings and keywords except for the ones in the language your system worked on. So if you have a Spanish question answering system, please only provide the Spanish question string and/or keywords in your submission. Otherwise please mark the language in either the system name or configuration slot, when uploading it. This way we can properly honour your multilingual efforts.

4.3 Evaluation measures

Evaluation of performance on multilingual and hybrid questions is done separately.

For both question types, participating systems will be evaluated with respect to precision and recall. Moreover, participants are encouraged to report performance, i.e. the average time their system takes to answer a query.

For each of the questions, your specified answers will be compared to the answers provided by the gold standard XML document. The evaluation tool computes precision, recall and F-measure for every question q:[3]

$$\begin{aligned} \textit{Recall}(q) &= \frac{\text{number of correct system answers}}{\text{number of gold standard answers}} \\ \textit{Precision}(q) &= \frac{\text{number of correct system answers}}{\text{number of system answers}} \\ \textit{F-measure}(q) &= \frac{2 \times \textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}} \end{aligned}$$

The tool then also computes the overall precision and recall taking the average mean of all single precision and recall values, as well as the overall F-measure.

All these results are printed in a simple HTML output; additionally you get a list of all question that your tool failed to capture correctly.

You are allowed to submit results as often as you wish.

4.4 Prizes

We offer prizes for the top-scoring teams, both in the multilingual and the hybrid track. More details will be announced soon.

[3] In the case of out-of-scope questions, an empty answer set counts as precision and recall 1, while a non-empty answer set counts as precision and recall 0.

5 Useful resources and tools

Prominent tools for indexing and searching datasets and text collections are:

- Lucene and Solr
  http://lucene.apache.org/
  http://lucene.apache.org/solr/
- Terrier
  http://terrier.org/

Building question answering systems is a complex task; it thus helps to exploit high-level tools for component integration as well as existing architectures for question answering systems:

- Apache UIMA
  http://uima.apache.org
- Open Advancement of Question Answering Systems (OAQA)
  http://oaqa.github.io
- Open Knowledgebase and Question Answering (OKBQA)
  http://www.okbqa.org
  https://github.com/okbqa
- openQA
  http://aksw.org/Projects/openQA.html

In the remainder of the section we provide a list of resources and tools that can be exploited especially for the linguistic analysis of a question and the matching of natural language expressions with vocabulary elements from a dataset.

Lexical resources

- WordNet
  http://wordnet.princeton.edu/
- Wiktionary
  http://www.wiktionary.org/
  API: https://www.mediawiki.org/wiki/API:Main_page
- FrameNet
  https://framenet.icsi.berkeley.edu/fndrupal/
- English lexicon for DBpedia 3.8 (in *lemon*[4] format)
  http://lemon-model.net/lexica/dbpedia_en/
- PATTY (collection of semantically-typed relational patterns)
  http://www.mpi-inf.mpg.de/yago-naga/patty/

Text processing

- GATE (General Architecture for Text Engineering)
  http://gate.ac.uk/
- NLTK (Natural Language Toolkit)
  http://nltk.org/

[4] http://lemon-model.net

- Stanford NLP
  http://www-nlp.stanford.edu/software/index.shtml
- LingPipe
  http://alias-i.com/lingpipe/index.html

Romanian:

- http://tutankhamon.racai.ro/webservices/TextProcessing.aspx

Dependency parser:

- MALT
  http://www.maltparser.org/
  Languages (pre-trained): English, French, Swedish
- Stanford parser
  http://nlp.stanford.edu/software/lex-parser.shtml
  Languages: English, German, Chinese, and others
- CHAOS
  http://art.uniroma2.it/external/chaosproject/
  Languages: English, Italian

Named Entity Recognition

- DBpedia Spotlight
  http://spotlight.dbpedia.org
- FOX (Federated Knowledge Extraction Framework)
  http://fox.aksw.org
- NERD (Named Entity Recognition and Disambiguation)
  http://nerd.eurecom.fr/
- Stanford Named Entity Recognizer
  http://nlp.stanford.edu/software/CRF-NER.shtml

String similarity and semantic relatedness

- Wikipedia Miner
  http://wikipedia-miner.cms.waikato.ac.nz/
- WS4J (Java API for several semantic relatedness algorithms)
  https://code.google.com/p/ws4j/
- SecondString (string matching)
  http://secondstring.sourceforge.net

Textual Entailment

- DIRT
  Paraphrase Collection: http://aclweb.org/aclwiki/index.php?title=DIRT_Paraphrase_Collection
  Demo: http://demo.patrickpantel.com/demos/lexsem/paraphrase.htm
- PPDB (The Paraphrase Database)
  http://www.cis.upenn.edu/~ccb/ppdb/

Translation systems

- English ↔ {Romanian,German,Spanish}
  http://www.racai.ro/tools/translation/racai-translation-system/

Language-specific resources and tools

Romanian:

- http://nlptools.info.uaic.ro/Resources.jsp

Anything missing?

If you know of a cool resource or tool that we forgot to include (especially for the challenge languages other than English), please drop us a note!