

QALD-4 Open Challenge

DOCUMENT VERSION: April 1, 2014

QALD-4 is the fourth in a series of evaluation campaigns on multilingual question answering over linked data, with a strong emphasis on interlinked data and hybrid approaches using information from both structured and unstructured data. The challenge is part of the Question Answering Lab at CLEF 2014 (<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	2
2	Relevant information in a nutshell	3
3	Task 1: Multilingual question answering over DBpedia	5
3.1	Dataset: DBpedia 3.9	5
3.2	Training phase	6
3.3	Test phase	8
4	Task 2: Biomedical question answering over interlinked data	10
4.1	Datasets: SIDER, Diseasome, Drugbank	10
4.2	Training phase	10
4.3	Test phase	11
5	Task 3: Hybrid question answering	12
5.1	Dataset: DBpedia 3.9	12
5.2	Training phase	12
5.3	Test phase	13
6	Available resources and tools	15
7	Participant's challenge	17

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, only few systems yet address the fact that the structured data available nowadays is distributed among a large collection of inter-connected datasets, and that answers to questions can often only be provided if information from several sources are combined. In addition, 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 Forascu (Alexandru Ioan Cuza University, Iasi, Romania)

2 Relevant information in a nutshell

Workshop Website: <http://www.sc.cit-ec.uni-bielefeld.de/qald/>

SPARQL endpoint (for all tasks):

<http://vtentacle.techfak.uni-bielefeld.de:443/sparql>

In order to see a list of all graphs loaded in the triple store, send the following query:

```
1 SELECT DISTINCT ?g WHERE {  
2   GRAPH ?g { ?s ?p ?o }  
3 }
```

Task 1: Multilingual question answering over DBpedia

Dataset:

- DBpedia 3.9 (<http://downloads.dbpedia.org/3.9/en/>)

Training questions:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-4_multilingual_train.xml
- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-4_multilingual_train_withanswers.xml

Test questions:

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

Task 2: Biomedical question answering over interlinked data

Datasets:

- SIDER (<http://sideeffects.embl.de>)
- Disasome (<http://wifo5-03.informatik.uni-mannheim.de/diseasome/>)
- Drugbank (<http://www.drugbank.ca>)

Training questions:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-4_biomedical_train.xml
- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-4_biomedical_train_withanswers.xml

Test questions:

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

Task 3: Hybrid question answering

Dataset:

- DBpedia 3.9 (<http://downloads.dbpedia.org/3.9/en/>)

Training questions:

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

Test questions:

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

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=4>

Results for the training phase can be uploaded at any time; results for the test phase can be uploaded from April 1 to May 1, 2014.

Resources

You are free to use all resources. A list of potentially useful resources and tools is provided in Section 6.

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 Task 1: Multilingual question answering over DBpedia

Task:

Given a natural language question or keywords, either retrieve the correct answer(s) from a given RDF repository, or provide a SPARQL query that retrieves these answer(s).

The provided dataset is English DBpedia 3.9. The questions and keywords are provided in seven languages: English, German, Spanish, Italian, French, Dutch¹, and Romanian. 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.

3.1 Dataset: DBpedia 3.9

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 3.9 dataset for English, including links, most importantly to YAGO² categories. This dataset comprises all files provided at:

- <http://downloads.dbpedia.org/3.9/en/>
- <http://downloads.dbpedia.org/3.9/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 also provide an endpoint at the following location, with respect to which evaluation will take place:

<http://vtentacle.techfak.uni-bielefeld.de:443/sparql>

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#>>

¹A special thanks to Gosse Bouma.

²For detailed information on the YAGO class hierarchy, please see <http://www.mpi-inf.mpg.de/yago-naga/yago/>.

```
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 Training phase

In order to get acquainted with the dataset and possible questions, a set of 200 training questions can be downloaded at the following location:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-4_multilingual_train.xml (without answers)
- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-4_multilingual_train_withanswers.xml (with answers)

All training questions are annotated with keywords, corresponding SPARQL queries and, if indicated, answers retrieved from the provided SPARQL endpoint. 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_multilingual_train` and `qald-4_multilingual_test`).

```
1 <dataset id="qald-4_train_multilingual">
2 <question id="1"> ... </question>
3 ...
4 <question id="200"> ... </question>
5 </dataset>
```

Each of the questions specifies an ID for the question (don't worry if they are not ordered) together with a range of other attributes explained below, the natural language string of the question in seven languages (English, German, Spanish, Italian, French, Dutch, and Romanian), keywords in the same seven languages, a corresponding SPARQL query, as well as the answers this query returns. Here is an example:

```
1 <question id="36" answertype="resource"
2           aggregation="false"
3           onlydbo="true">
4
5 <string lang="en">
6 Through which countries does the Yenisei river flow?
7 </string>
8 ...
9 <keywords lang="en">
10 Yenisei river, flow through, country
11 </keywords>
12 ...
13
```

```

14 <query>
15 PREFIX res: <http://dbpedia.org/resource/>
16 PREFIX dbo: <http://dbpedia.org/ontology/>
17 SELECT DISTINCT ?uri
18 WHERE {
19     res:Yenisei_River dbo:country ?uri .
20 }
21 </query>
22
23 <answers>
24 <answer>
25 <uri>http://dbpedia.org/resource/Mongolia</uri>
26 </answer>
27 <answer>
28 <uri>http://dbpedia.org/resource/Russia</uri>
29 </answer>
30 </answers>
31
32 </question>

```

The following attributes are specified for each question along with its ID:

- **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. This format is also required when you submit results containing a date as answer.
 - **boolean**: Either `true` or `false`.

Answer of these types are required to be enclosed by the corresponding tag, i.e. `<number>47</number>`, `<string>Valentina Tereshkova</string>` and `<boolean>>true</boolean>`.

- **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 possibly relies on the DBpedia property namespace (`http://dbpedia.org/property/`), FOAF or some YAGO category.

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 the dataset. The query is specified as `OUT OF SCOPE` and the answer set is empty. Here is an example from the DBpedia training question set:

```
1 <question id="95" answertype="number"
2           aggregation="false"
3           onlydbo="false">
4
5 <string lang="en">
6 How many big fires struck Paris during the Middle Ages?
7 </string>
8 ...
9
10 <query>
11 OUT OF SCOPE
12 </query>
13
14 <answers/>
15
16 </question>
```

For evaluation, your system should in these cases specify `OUT OF SCOPE` as query and/or an empty answer set, just like in the example.

3.3 Test phase

During test phase, from April 1 to May 1, a set of 50 different questions without annotations is provided at the following location:

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

Results can be submitted from April 1 to May 1, 2014, 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=4>

The only difference is that evaluation results are not displayed. You can upload results as often as you like (e.g., 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 above. For all questions, the dataset ID and question IDs are obligatory. Beyond that, you are free to specify either a SPARQL query or the answers (or both), depending on which of them your system returns. You are also allowed to change the natural language question or keywords (insert quotes, reformulate, use some controlled language format, and the like). 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 Dutch question answering system, please only provide the Dutch

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 multilinguality efforts.

Evaluation measures

For each of the questions, your specified answers, or the answers your specified SPARQL query retrieves, 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

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

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

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

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.

³In 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.

4 Task 2: Biomedical question answering over interlinked data

Task:

Given a natural language question or keywords, either retrieve the correct answer(s) from a given RDF repository, or provide a SPARQL query that retrieves these answer(s).

The RDF repository contains three biomedical datasets: SIDER, Diseasome, and Drugbank. The focus of the task is on interlinked data. Thus most of the questions require the integration of information from at least two of those datasets. 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.

4.1 Datasets: SIDER, Diseasome, Drugbank

Also for the life sciences, linked data plays a bigger and bigger role. Already a tenth of the Linked Open Data cloud⁴ consists of biomedical datasets. Especially the pharmaceutical industry is starting to embrace linked data, corroborated by a range of large-scale European projects and task forces such as Linking Open Drug Data⁵ (LODD).

In the context of QALD-4 we provide the following interlinked biomedical datasets:

- SIDER, describing drugs and their side effects
<http://sideeffects.embl.de>
- Diseasome, encompassing description of diseases and genetic disorders
<http://wifo5-03.informatik.uni-mannheim.de/diseasome/>
- Drugbank, describing FDA-approved active compounds of medication
<http://www.drugbank.ca>

You can either download a dump at the specified URL, or access them via the challenge's SPARQL endpoint:

<http://vtentacle.techfak.uni-bielefeld.de:443/sparql>

4.2 Training phase

The training question set comprises 25 questions over the biomedical datasets SIDER, Diseasome and Drugbank. All training questions are provided in an XML format similar to the one explained in Section 3.2. The dataset id is `qald-4.biomedical.train` (and `qald-4.biomedical.test` for the test questions).

⁴<http://lod-cloud.net>

⁵<http://www.w3.org/wiki/HCLSIG/LODD>

Most of the questions require information from at least two of the datasets. Here is an example query (ommitting prefix definitions), representing the question **What are the side effects of drugs used for Tuberculosis?**

```
1 SELECT DISTINCT ?x
2 WHERE {
3     disease:1154 diseasome:possibleDrug ?v2 .
4     ?v2 rdf:type drugbank:drugs .
5     ?v3 owl:sameAs ?v2 .
6     ?v3 sider:sideEffect ?x .
7 }
```

Note that the drugs used for Tuberculosis are retrieve from Diseasome, and their side effects are retrieved from SIDER. The link between the relevant resources in these datasets (bound to ?v2 and ?v3) is established using the OWL property `sameAs`.

4.3 Test phase

During test phase, from April 1 to May 1, a set of 25 similar questions is provided at the following location:

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

Results can be submitted from April 1 to May 1, 2014, 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=4>

All submissions are required to comply with the XML format used for the training questions.

Evaluation measures are the same as for the other tasks, cf. Section 3.3.

5 Task 3: Hybrid question answering

Task:

Given a natural language question or keywords, retrieve the correct answer(s) from a given repository containing both RDF data and free text.

The provided dataset is English DBpedia 3.9. The focus of the task is on hybrid question answering, i.e. the integration of both 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. 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.

5.1 Dataset: DBpedia 3.9

The provided dataset is DBpedia 3.9, cf. Section 3.1 above. It can be accessed via the same SPARQL endpoint:

<http://vtentacle.techfak.uni-bielefeld.de:443/sparql>

However, for this task, not only the RDF triples 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 .
```

5.2 Training phase

In order to get acquainted with the dataset and possible questions, a set of 25 training questions can be downloaded at the following location:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/4/qald-4_train_hybrid.xml (with answers)

The training questions are provided in an XML format that is very similar to the one explained in Section 3.2 above⁶. The dataset id is `qald-4_hybrid_train` (and `qald-4_hybrid_test` for the test questions). All questions are annotated with a pseudo query and the correct answers. The pseudo query is like an RDF

⁶The question attributes `aggregation` and `onlydbo` here refer only to the RDF part of the pseudo query.

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 pseudo query for the question: Which recipients of the Victoria Cross died in the Battle of Arnhem?

```
1 PREFIX dbo: <http://dbpedia.org/ontology/>
2 PREFIX res: <http://dbpedia.org/resource/>
3 SELECT DISTINCT ?uri
4 WHERE {
5   ?uri dbo:award res:Victoria_Cross .
6   ?uri text:"died in" text:"Battle of Arnhem" .
7 }
```

This pseudo query contains two triples, one an RDF triple and the other containing free text as property and object. The way to answer the question is to retrieve all recipients of the Victoria Cross using the triple

```
?uri dbo:award res:Victoria_Cross .
```

and then check the abstract of all returned URIs whether they contain the information that they died in the Battle of Arnhem. For example, the abstract for John Hollington Grayburn contains the following text:

```
he went into action in the Battle of Arnhem [...]
but was killed after standing up in full view of a German tank
```

All queries are designed in a way that they require both RDF data and free text to be answered. In the case of the above question, for instance, not all abstracts of all recipients of the Victoria Cross mention this award, while the battle in which they died is not explicitly given in the available RDF data.

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. Therefore, when submitting results, you are required to provide answers (in the same format as described in Section 3.2).

5.3 Test phase

During test phase, from April 1 to May 1, a set of 10 similar questions is provided at the following location:

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

Results can be submitted from April 1 to May 1, 2014, 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=4>

All submissions are required to comply with the XML format used for the training questions.

Evaluation measures are the same as for the other tasks, cf. Section 3.3.

6 Available resources and tools

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*⁷ 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/>
- 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

⁷<http://lemon-model.net>

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!

7 Participant's challenge

Are there questions that your tool is very good at but that might prove difficult for others? Are there questions that are very interesting but are not among the training questions? Then send in these questions and challenge others!

If there are any questions you would like to contribute, please send an email with the questions and, ideally, corresponding SPARQL queries to Christina Unger: cunger@cit-ec.uni-bielefeld.de. The questions will then be added to the document and published on the ILD mailing list.