Multilingual Question Answering Over Linked Data

QALD-3 Open Challenge

Document Version: February 2, 2013

The QALD-3 open challenge is the third instalment of the *Question Answering over Linked Data* benchmark and is organized as a half-day lab at CLEF 2013. QALD-3 offers two tasks: *multilingual question answering*, aimed at all kinds of question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data, and *ontology lexicalization*, aimed at all methods that (semi-)automatically create lexicalizations for ontology concepts. All relevant information for participating in the challenge are given in this document.

1 Introduction

Motivation and Goal

While more and more semantic 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. 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. To this end, systems have to deal with a heterogeneous, distributed and very large set of highly interconnected data. The availability of such an amount of open and structured data has no precedents in computer science and approaches that can deal with the specific character of linked data are urgently needed.

In addition, multilinguality has 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. 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.

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)

2 Relevant information in a nutshell

Workshop Website: http://www.sc.cit-ec.uni-bielefeld.de/ild/

Task 1: Multilingual question answering

Datasets:

- DBpedia 3.8
  http://downloads.dbpedia.org/3.8/en/
- Spanish DBpedia
  http://es.dbpedia.org/DBpediaESdata/
- MusicBrainz
  http://greententacle.techfak.uni-bielefeld.de/~cunger/qald2/musicbrainz.tar.gz (226.8 MB)

SPARQL endpoint (for all three datasets):
http://vtentacle.techfak.uni-bielefeld.de:443/sparql

Training questions:
http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/

- dbpedia-train.xml and dbpedia-train-answers.xml
- esdbpedia-train.xml and esdbpedia-train-answers.xml
- musicbrainz-train.xml and musicbrainz-train-answers.xml

Test questions will be made available on May 1, 2013.

Participant's challenge:
http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/participants-challenge.xml

Task 2: Ontology lexicalization

Training concepts (10 classes and 30 properties) from the DBpedia ontology:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/dbpedia_train_classes_properties.txt

Corresponding *lemon* lexicon containing lexicalizations of those concepts:

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/dbpedia_train_lexicon_en.ttl

Test concepts will be made available on May 1, 2013.

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=3

Results for training the training phase can be uploaded at any time; results for the test phase can be uploaded from May 1 to May 31, 2013.

Resources

You are free to use all resources (e.g., WordNet, GeoNames, dictionary tools, and so on).

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

This task is aimed at all kinds of question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data.

3.1 Datasets

In order to evaluate and compare question answering systems, we provide three RDF datasets: English DBpedia 3.8 (http://dbpedia.org), Spanish DBpedia (http://es.dbpedia.org), and MusicBrainz (musicbrainz.org). In order to work with the datasets, you can either download them or use the provided SPARQL endpoint.

3.1.1 English DBpedia 3.8

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 links, most importantly to YAGO[1] categories and MusicBrainz. This dataset comprises all files provided at:

- http://downloads.dbpedia.org/3.8/en/
- http://downloads.dbpedia.org/3.8/links/

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

```
 dbo: <http://dbpedia.org/ontology/>
 dbp: <http://dbpedia.org/property/>
 res: <http://dbpedia.org/resource/>
yago: <http://dbpedia.org/class/yago/>
foaf: <http://xmlns.com/foaf/0.1/>
 xsd: <http://www.w3.org/2001/XMLSchema#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
 rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

3.1.2 Spanish DBpedia

Since 2011, information from Wikipedia is extracted also in 15 non-English languages, including Spanish. So far, the Spanish DBpedia contains almost 100 million RDF triples. The data can be downloaded at http://es.dbpedia.org/DBpediaESdata/.

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

Relevant namespaces that are used in the provided training and test queries are the following (in addition to the usual `rdf`, `rdfs` and `xsd`):

```
esdbo: <http://es.dbpedia.org/ontology/>
esdbp: <http://es.dbpedia.org/property/>
esres: <http://es.dbpedia.org/resource/>
```

3.1.3 MusicBrainz

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. A package containing all RDF data[2] can be downloaded from the following location:
http://greententacle.techfak.uni-bielefeld.de/~cunger/qald2/musicbrainz.tar.gz (226.8 MB)

The following namespaces are used in the provided training and test queries:

```
   mo: <http://purl.org/ontology/mo/>
  bio: <http://purl.org/vocab/bio/0.1/>
  rel: <http://purl.org/vocab/relationship/>
event: <http://purl.org/NET/c4dm/event.owl#>
   tl: <http://purl.org/NET/c4dm/timeline.owl#>
 foaf: <http://xmlns.com/foaf/0.1/>
   dc: <http://purl.org/dc/elements/1.1/>
  xsd: <http://www.w3.org/2001/XMLSchema#>
  rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

The RDF export builds on the Music Ontology, a specification of which can be found at http://musicontology.com. Examples for how to model data w.r.t. this specification are given in the Music Ontology Wiki: http://wiki.musicontology.com/index.php/Examples. In the following, we briefly describe the most important classes and relations relevant for the challenge.

There are three major classes:

- `mo:MusicArtist` and its subtypes `mo:MusicGroup` for bands and orchestra, and `mo:SoloMusicArtist` for persons (independent of whether they are solo artists or members of a group)
- `mo:Record`
- `mo:Track`

Artists have a birth and death date modelled by means of the BIO vocabulary[3]. For example, the following SPARQL query extracts the birth and death date of John Lennon:

[2] In fact it contains only a subset of all track information, due to performance problems.
[3] http://vocab.org/bio/0.1/

```
SELECT ?birthdate deathdate WHERE {
   ?artist foaf:name 'John Lennon'.
   ?artist bio:event ?event1 .
   ?event1 rdf:type bio:Birth .
   ?event1 bio:date ?birthdate .
   ?artist bio:event ?event2 .
   ?event2 rdf:type bio:Death .
   ?event2 bio:date ?deathdate .
}
```

In exactly the same way, the corresponding dates for music groups are formulated (where birth date can be read as founding date and death date as the date the group broke up).

Artists are related among each other through relations like `rel:spouseOf`, `rel:parentOf`, `siblingOf` and `rel:collaboratesWith` from the RELATIONSHIP vocabulary[4].

Membership in a group is expressed in two ways: by means of the simple relation `mo:member_of`, indicating that someone is or was member of a group, and by means of the Event and Timeline Ontology[5]. Using the former, the following triple expresses that John Lennon is or was a member of The Beatles:

```
<http://zitgist.com/music/artist/4d5447d7-c61c-4120-ba1b-d7f471d385b9>
mo:member_of
<http://zitgist.com/music/artist/b10bbbfc-cf9e-42e0-be17-e2c3e1d2600d>
```

Or, in a more human-readable way:

```
?artist foaf:name 'John Lennon' .
?artist mo:member_of ?band .
?band foaf:name 'The Beatles' .
```

In order to also express time information, the more complex Event and Timeline Ontology representation has to be used. For example, the following triples express that Pete Best was a member of The Beatles from August 12, 1960 until August 16, 1962.

```
?artist foaf:name 'Pete Best' .
?event rdf:type mo:membership .
?event event:agent ?artist .
?event mo:group ?band .
?band foaf:name 'The Beatles' .
?event event:time ?time .
?time tl:start '1960-08-12'^^xsd:date .
?time tl:end '1962-08-16'^^xsd:date .
```

Records are related to their creator through the property `foaf:maker`, and through `mo:releaseType` to the type of record (`mo:album`, `mo:single`, `mo:ep`,

[4] http://vocab.org/relationship/

[5] http://purl.org/NET/c4dm/event.owl and http://purl.org/NET/c4dm/timeline.owl.

mo:soundtrack, mo:live, mo:compilation, mo:remix, mo:interview, and mo:audiobook). For example, the following SPARQL query extracts all live albums by Slayer:

```
SELECT ?album WHERE {
   ?album mo:release_type mo:live .
   ?album foaf:maker ?artist .
   ?artist foaf:name 'Slayer'.
}
```

The dataset also contains relations between records and artists, specifying their role during the record creation, for example mo:performer, mo:singer, mo:composer, mo:producer, and mo:lyricist.

Tracks are also related to their creator through the property foaf:maker, to their duration through tl:duration, and through mo:trackNum to their position in the track list of a record. For example, the following SPARQL query extracts the title of the first track of Abbey Road:

```
SELECT ?title WHERE {
   ?album dc:title 'Abbey Road' .
   ?album mo:track ?track .
   ?track mo:trackNum '1' .
   ?track dc:title ?title .
}
```

3.1.4 SPARQL endpoint

DBpedia provides official SPARQL endpoints:

- *English DBpedia:* http://dbpedia.org/sparql/
- *Spanish DBpedia:* http://es.dbpedia.org/sparql

We also provide a SPARQL endpoint for both DBpedia datasets as well as the MusicBrainz dataset at the following location:

http://greententacle.techfak.uni-bielefeld.de:5171/sparql

Evaluation will take place with respect to this SPARQL endpoint (and not the official DBpedia endpoints, for example), in order to ensure invariable and therefore comparable results.

3.2 Training and test phase

The task is to extract correct answers for natural language questions or corresponding keywords from one of the given RDF repositories. 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.

3.2.1 Training questions

In order to get acquainted with the datasets and possible questions, a set of 100 training questions for each dataset can be downloaded at the following locations: `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/`

- *English DBpedia*

  `dbpedia-train.xml` (without answers)
  `dbpedia-train-answers.xml` (with answers)

- *Spanish DBpedia*

  `esdbpedia-train.xml` (without answers)
  `esdbpedia-train-answers.xml` (with answers)

- *MusicBrainz*

  `musicbrainz-train.xml` (without answers)
  `musicbrainz-train-answers.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. `dbpedia-train`, `dbpedia-test`, `esdbpedia-train`, `esdbpedia-test`, `musicbrainz-train`, `musicbrainz-test`).

```
<dataset id="dbpedia-train">
  <question id="1">...</question>
  ...
  <question id="100">...</question>
</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 six languages (English, German, Spanish, Italian, French, and Dutch), keywords in the same six languages, a corresponding SPARQL query, as well as the answers this query returns. Here is an example:

```
<question id="36" answertype="resource" aggregation="false" onlydbo="false">
   <string lang="en">Through which countries does the Yenisei river flow?</string>
   <string lang="de">Durch welche Länder fließt der Yenisei?</string>
   <string lang="es">Por qué países fluye el río Yenisei?</string>
   <string lang="it">Attraverso quali stati scorre il fiume Yenisei?</string>
   <string lang="fr">...</string>
```

```
  <string lang="nl">Door welke landen stroomt de Jenisej?</string>
  <keywords lang=en>Yenisei river, flow through, country</keywords>
  ...
  <query>
   PREFIX res:  <http://dbpedia.org/resource/>
   PREFIX dbp:  <http://dbpedia.org/property/>
   SELECT DISTINCT ?uri WHERE {
      res:Yenisei_River dbp:country ?uri .
      OPTIONAL { ?uri rdfs:label ?string . FILTER (lang(?string) = "en") }
   }
   </query>
  <answers>
    <answer>
     <uri>http://dbpedia.org/resource/Mongolia</uri>
    </answer>
    <answer>
     <uri>http://dbpedia.org/resource/Russia</uri>
    </answer>
  </answers>
</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>` (except for resources, for which a URI and/or a string should be provided, see the cave example above).
- `aggregation` indicates whether any operations beyond triple pattern matching are required to answer the question (e.g., counting, filters, ordering, etc.).
- `onlydbo` is given only for DBpedia questions and reports whether the query relies solely on concepts from the DBpedia ontology.

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:

```
<question id="94" answertype="number" aggregation="false" onlydbo="false">
   <string lang="en">Which budget did Zdenek Sverak's first movie have?</string>
   <query>
   OUT OF SCOPE
   </query>
   <answers />
</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.

3.2.2 Submitting results during test phase

During test phase, a set of different questions for each dataset without annotations are provided at the following location:

`http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/`

- `dbpedia-test-questions.xml`
- `musicbrainz-test-questions.xml`

Note that there will be only one question set for DBpedia, applying to both the English and the Spanish DBpedia. On which one you worked has to be indicated in your submission in the `dataset id` tag.

Results can be submitted during the whole of May, 2013, via the same online form used during training phase (note the drop down box that will then allow you to specify *test* instead of *training*):

`http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=evaltool&q=3`

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.

3.2.3 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:[6]

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

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

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

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.

3.3 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!

In order to make a start, we provide a few questions that cannot be answered over DBpedia or MusicBrainz alone, but require the combination of both datasets. You can access them at the following location:
http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/participants-challenge.xml

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.

[6] 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.

4 Task 2: Ontology lexicalization

Multilingual information access can be facilitated by the availability of lexica in different languages, for example allowing for an easy mapping of Spanish, German, and French natural language expressions to English ontology labels. The task consists in finding English lexicalizations of a set of classes and properties from the DBpedia ontology, for example in a Wikipedia corpus.

4.1 Training data

The training data consists of a set of 10 classes and 30 properties from the DBpedia ontology, as well as a *lemon* lexicon containing lexicalizations of those classes and properties.

- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/dbpedia_train_classes_properties.txt
- http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/3/dbpedia_train_lexicon_en.ttl

Detailed information on *lemon* can be found on http://lemon-model.net, including a cookbook which contains all you need to know in order to create *lemon* lexica: http://lemon-model.net/lemon-cookbook.pdf

A suitable corpus for finding English lexicalizations is Wikipedia. You can either download a Wikipedia data dump at http://en.wikipedia.org/wiki/Wikipedia:Database_download, or directly download an already cleaned up part of English Wikipedia: https://www.dropbox.com/s/chncp9zjy81r42n/wikipediaSentences.tar.bz2 (1.54 GB).

4.2 Submitting results during test phase

Submitted lexicalizations are expected to be in *lemon* format. The training lexicon uses LexInfo (lexinfo.net/ontology/2.0/lexinfo.owl) as linguistic ontology, mainly for reasons of readability, but you may as well use ISOcat (http://www.isocat.org), for example.

Lexica can be submitted during the whole of May, 2013, via the same online form used for the question answering task:

http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=evaltool&q=3

Evaluation results will be displayed during training phase but not during test phase. 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.

4.3 Evaluation measures

For each class and property, the submitted lexical entries will be compared to the gold standard lexical entries along two dimensions: i) lexical precision, lexical recall and lexical F-measure, and ii) lexical accuracy. In the first dimension, we evaluate how many of the gold standard entries for a property are in the submitted lexicon (*recall*), and how many of the submitted entries are among the gold standard entries (*precision*), where two entries count as the same lexicalization if their lemma, part of speech and sense coincide. Thus lexical precision P_{lex} and recall R_{lex} for a class or property *uri* are defined as follows:

$$\begin{aligned} P_{lex}(uri) &= \frac{|entries_{submitted}(uri) \cap entries_{gold}(uri)|}{|entries_{submitted}(uri)|} \\ R_{lex}(uri) &= \frac{|entries_{submitted}(uri) \cap entries_{gold}(uri)|}{|entries_{gold}(uri)|} \end{aligned}$$

Where $entries_{submitted}(uri)$ is the set of entries for the class or property *uri* in the submitted lexicon, while $entries_{gold}(uri)$ is the set of entries for *uri* in the gold standard lexicon. The F-measure $F_{lex}(uri)$ is then computed as the harmonic mean of $P_{lex}(uri)$ and $R_{lex}(uri)$, as usual.

The second dimension, lexical accuracy, serves to evaluate whether the specified subcategorization frame and its arguments are correct, and whether these syntactic arguments have been mapped correctly to the semantic arguments (domain and range) of the property in question. The accuracy of a submitted lexical entry l_{auto} for a class or property *uri* w.r.t. the corresponding gold standard entry l_{gold} is therefore defined as:

$$\begin{aligned} A_{uri}(l_{submitted}) = & frameEq(l_{submitted}, l_{gold}) + \frac{|args(l_{submitted}) \cap args(l_{gold})|}{|args(l_{gold})|} \\ & + \frac{\sum_{a \in args(l_{submitted})} map(a)}{|args(l_{submitted})|} \end{aligned}$$

where $frameEq(l_1, l_2)$ is 1 if the subcategorization frame of l_1 is the same as the subcategorization frame of l_2, and 0 otherwise, where $args(l)$ returns the syntactic arguments of l's frame, and where

$$map(a) = \begin{cases} 1, & \textit{if } a \textit{ has been mapped to the correct semantic argument of } p \\ 0, & \textit{otherwise} \end{cases}$$

When comparing the argument mapping of a submitted entry with that of a gold standard entry, we only consider the class of the argument, simply being *subject* or *object*. This abstracts from the specific type of subject (e.g. copulative subject) and object (e.g. indirect object, prepositional object, etc.) and therefore evaluates the argument mappings independently of the correctness of the frame and frame arguments.

The lexical accuracy $A_{lex}(uri)$ for a class or property *uri* is then computed as the average mean of the accuracy values of each generated lexicalization.

All measures are computed for each concept (class and property) and then averaged for all concepts.

5 Contact and trouble shooting

If you have any questions or comments, including worries about the training and test questions, trouble with the datasets, the SPARQL end point, or the online submission and evaluation form, please contact Christina Unger: `cunger@cit-ec.uni-bielefeld.de`.