9th Challenge on Question Answering over Linked Data (QALD-9)

Ricardo Usbeck, Ria Hari Gusmita, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo

Data Science Group, Paderborn University, Germany
ricardo.usbeck@uni-paderborn.de

1 Introduction

Recent years have seen a growing amount of research on question answering (QA) over Semantic Web data, shaping an interaction paradigm that allows end users to profit from the expressive power of Semantic Web standards. At the same time, QA systems hide their complexity behind an intuitive and easy-to-use interface. However, the growing amount of data available on the Semantic Web has led to a heterogeneous data landscape where QA systems struggle to keep up with the volume, variety and veracity of the underlying knowledge.

The Question Answering over Linked Data (QALD) challenges aim to provide up-to-date benchmarks for assessing and comparing state-of-the-art systems that mediate between a user, expressing his or her information need in natural language, and RDF data. In the past few years, more than *40 research groups and their systems have taken part in the last nine QALD challenges*. The QALD challenge targets all researchers and practitioners working on querying Linked Data, natural language processing for question answering, multilingual information retrieval and related topics. The main goal is to gain insights into the strengths and shortcomings of different approaches and into possible solutions for coping with the large, heterogeneous and distributed nature of Semantic Web data.

QALD[1] has a 8-year history. The challenge began in 2011 and is developing benchmarks that are increasingly being used as a standard evaluation venue for question answering over Linked Data. Overviews of past instantiations of the challenge are available from the CLEF Working Notes, CEUR workshop notes as well as ESWC proceedings, see Table 1.

This article will give a technical overview of the task and results of the 9th Question Answering over Linked Data challenge.

2 Dataset and Task

The key challenge for QA over Linked Data is to translate a user's natural language query into such a form that it can be evaluated using standard Seman-

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

QALD	URI to proceedings
7	CEURproceedingsofNLIWOD4workshop
7	CEURproceedingsofNLIWOD4workshop
7	https://link.springer.com/chapter/10.1007/978-3-319-69146-6_6
6	https://www.springer.com/us/book/9783319465647
5	https://ceur-ws.org/Vol-1391/173-CR.pdf
4	https://ceur-ws.org/Vol-1180/CLEF2014wn-QA-UngerEt2014.pdf
3	https://pub.uni-bielefeld.de/download/2685575/2698020
2	https://ceur-ws.org/Vol-913/
1	https://qald.sebastianwalter.org/1/documents/qald-1-proceedings.pdf

Table 1. QALD proceedings and their URIs.

tic Web query processing and inferencing techniques. The main task of QALD therefore is the following:

> Given one or several RDF dataset(s) as well as additional knowledge sources and natural language questions or keywords, return the correct answers or a SPARQL query that retrieves these answers.

Data format

All the data for the tasks can be found in our project repository:

- https://github.com/ag-sc/QALD/tree/master/9/data.

We encouraged the use of QALD-JSON format[2] as communication format between the systems and the GERBIL QA platform [3]:

```
{ "id":"3",
  "answertype":"resource",
  "aggregation":false,
  "onlydbo":true,
  "hybrid":false,
  "question":[
    {
      "language":"en",
      "string":"Who was the wife of U.S. president
         Lincoln?",
      "keywords":"U.S. president, Lincoln, wife"
    },
    {
      "language":"nl",
```

[2] https://github.com/AKSW/gerbil/wiki/Question-Answering and the results are formatted according to https://www.w3.org/TR/sparql11-results-json/

```
      "string":"Wie was de vrouw van de Amerikaanse
         president Lincoln?",
      "keywords":"vrouw, president van America, Lincoln
         "
    }
  ],
  "query":{
    "sparql":"PREFIX dbo:<http://dbpedia.org/ontology/>
       PREFIX res:<http://dbpedia.org/resource/>
       SELECT DISTINCT ?uri
       WHERE {res:Abraham_Lincoln dbo:spouse ?uri.}"
  },
  "answers":[
    {
      "head":{
        "vars":[
          "uri"
        ]
      },
      "results":{
        "bindings":[
          {
            "uri":{
              "type":"uri",
              "value":"http://dbpedia.org/resource/
                 Mary_Todd_Lincoln"
            }
          }
        ]
     }}]}
```


Task: Multilingual question answering over DBpedia

Given the diversity of languages used on the web, there is an increasing need to facilitate multilingual access to semantic data. The core task of QALD is thus to retrieve answers from an RDF data repository given an information need expressed in a variety of natural languages.

Training data. The underlying RDF dataset was DBpedia 2016-10. There was no newer dataset of DBpedia available at the point of this challenge. This year, we used a novel approach to updating the dataset. We employed QUANT [2], a Question Answering curation interface especially developed to speed up the update processes of QA benchmarks against new QA pairs or updated knowledge bases.

Thus, we were able to create the largest QALD dataset so far. The training data consisted of **408 questions** compiled and curated from previous challenges. The questions were available in 11 different languages (e.g., English, Spanish, German, Italian, French, Dutch, Romanian or Farsi). Those questions were general, open-domain factual questions, for example:

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

In the current version of the benchmark, the questions vary with respect to their complexity, including questions with counts (e.g., *How many children does Eddie Murphy have?...*), superlatives (e.g., *Which museum in New York has the most visitors?*), comparatives (e.g., *Is Lake Baikal bigger than the Great Bear Lake?*), and temporal aggregators (e.g., *How many companies were founded in the same year as Google?*). Each question is annotated with a manually specified SPARQL query and answers. In the case above, the SPARQL query is as follows:

```
SELECT DISTINCT ?uri
WHERE {
    ?uri a <http://dbpedia.org/ontology/Book> .
    ?uri <http://dbpedia.org/ontology/numberOfPages> ?n .
}
ORDER BY DESC(?n)
OFFSET 0 LIMIT 1
```

And the answer is `<http://dbpedia.org/resource/The_Tolkien_Reader>`.

Test Data. The test dataset consists of 150 similar manually created questions. This year, we removed question types including questions according to RDF types (e.g., *What is backgammon?...*). All questions are compiled from existing, real-world question and query logs as well as past challenges to provide unbiased questions expressing real-world information needs.

Evaluation Metric. The 9th QALD challenge relies on an automatic evaluation tool, namely GERBIL QA [3][3], which is open source and available for everyone to re-use. The GERBIL QA platform is accessible online, so that participants can simply upload the answers produced by their system or even check their system via a webservice. Each experiment has a citable, time-stable and archivable URI that is both human- and machine-readable. However, participating systems have to provide a webservice to participate in the final challenge.[4]

[3] http://gerbil-qa.aksw.org/gerbil/

[4] https://github.com/dice-group/gerbil/wiki/Question-Answering#web-service-interface

The QA systems were evaluated with respect to precision and recall:

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

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

The metrics use the following additional semantic information:

- If the golden answerset is empty and the system does respond with an empty answer, we set precision, recall and F-measure to 1.
- If the golden answerset is empty but the system responds with any answerset, we set precision, recall and F-measure to 0.
- If there is a golden answer but the the QA system responds with an empty answerset, we assume the system could not respond. Thus we set the precision to 0 and the recall and F-measure to 0.
- In any other case, we calculate the standard precision, recall and F-measure per question.

The evaluation also computed the macro and micro F-measure of a system over all test questions. For micro F-measure, we summed up all true and false positives and negatives and calculated the precision, recall and F-measure at the end. For the macro measures, we calculated precision, recall and F-measure per question and averaged the values at the end. **For the final evaluation, we focused only on the Macro F1 QALD metric.** That is, we decided to have a metric more comparable to older QALD challenges and also to follow community requests.[5] This metric uses the previously mentioned additional semantic information with the following exception:

- If the golden answerset is not empty but the QA system responds with an empty answerset, it is assumed that the system determined that it cannot answer the question. Here we set the precision to 1 and the recall and F-measure to 0.

3 Participating systems

After six registrations, **5 teams** were able to join the final evaluation. This indicates, that Web services rather than file-based submissions are finally accepted in the community as means to evaluate scientific systems.

WDAqua [1] is a rule-based system using a combinatorial approach to generate SPARQL queries from natural language questions, leveraging the semantics encoded in the underlying knowledge base. It can answer questions on both DBpedia (supporting English) and Wikidata (supporting English, French, German

[5] `https://github.com/dice-group/gerbil/issues/211`

and Italian). The system, which does not require training, participated in Tasks 1 and 4 of the challenge.

ganswer2 [4] has participated outside the actual challenge this year as a system without a paper submission in Task 1. Zou et al. use a graph-based approach to generate a semantic query graph, which reduced the transformation of natural language to SPARQL to a subgraph matching problem.

TeBaQA by Peter Nancke et al. from Leipzig University in Germany is an unpublished system which is based on learning SPARQL templates from past benchmark challenges and filling them subsequently. The system is available at `http://139.18.2.39:8187/`.

Elon by Szabó Bence et al. from Paderborn University in Germany stems from a student project and is available at `http://qald-beta.cs.upb.de:443/`. It is based on an own dictionary and not yet published.

QASystem by Lukas Blübaum and Nick Düsterhus is also a student project from Paderborn University Germany and available at `http://qald-beta.cs.upb.de:80/`. Their system is able to cope with comparatives and superlatives in questions via hand-crafted rules.

4 Results

All QA systems were run on the QALD-9 train and test dataset in English and GERBIL QA version 0.2.3 and you can find FAIR experiment data for training and test dataset at `http://w3id.org/gerbil/qa/experiment?id=201810080002` and `http://w3id.org/gerbil/qa/experiment?id=201810060001`.

QALD-8 test introduced some **curve balls** which we tried to eliminate in QALD-9 via QUANT [2] to focus on answering the questions rather than cleaning the input. In 2018, the **gAnswer system is the winner** of the QALD-9 challenge, see Table 2.

Annotator	Macro Precision	Macro Recall	Macro F1	Error Count	Average Time/Doc ms	Macro F1 QALD
Elon (WS)	0.049	0.053	0.050	2	219	0.100
QASystem (WS)	0.097	0.116	0.098	0	1014	0.200
TeBaQA (WS)	0.129	0.134	0.130	0	2668	0.222
wdaqua-core1 (DBpedia)	0.261	0.267	0.250	0	661	0.289
gAnswer (WS)	0.293	0.327	0.298	1	3076	0.430

Table 2. QALD-9 final results.

5 Summary

The QALD-9 challenge focused on the successful and long running multilingual QA task. In particular, we wanted to create an updated dataset which is larger

than ever before and has a higher quality due to the employment of QUANT's semi-automatic curation measure. [2]. The teams were required to provide web-services of their systems to participate in the challenge, which will in turn support comparable research in the future in contrast to former XML/JSON file submissions. This increased the entrance requirements for participating teams but ensures long term comparability of the system performance and a fair and open challenge.

In the future, we will further simplify the participation process and offer leaderboards prior to the actual challenge to allow participants to see their performance beforehand. After feedback from the authors, we will likely add new key performance indicators for the capability of a system to know which questions it cannot answer and take confidence scores for answers into account.

Acknowledgements This work has been supported by the H2020 project HOBBIT (GA no. 688227) as well as by the BMVI projects LIMBO (project no. 19F2029C) and OPAL (project no. 19F20284) as well as by the German Federal Ministry of Education and Research (BMBF) within 'KMU-innovativ: Forschung für die zivile Sicherheit' in particular 'Forschung für die zivile Sicherheit' and the project SOLIDE (no. 13N14456). The authors gratefully acknowledge financial support from the German Federal Ministry of Education and Research within Eurostars, a joint programme of EUREKA and the European Community under the project E! 9367 DIESEL and E! 9725 QAMEL.

References

1. Dennis Diefenbach, Kamal Deep Singh, and Pierre Maret. Wdaqua-core0: A question answering component for the research community. In Mauro Dragoni, Monika Solanki, and Eva Blomqvist, editors, *Semantic Web Challenges - 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28 - June 1, 2017, Revised Selected Papers*, volume 769 of *Communications in Computer and Information Science*, pages 84–89. Springer, 2017.
2. Ria Hari Gusmita, Rricha Jalota, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. QUANT - Question Answering Benchmark Curator. In *to be published*, 2018.
3. Ricardo Usbeck, Michael Röder, Michael Hoffmann, Felix Conrad, Jonathan Huthmann, Axel-Cyrille Ngonga-Ngomo, Christian Demmler, and Christina Unger. Benchmarking question answering systems. *Semantic Web Journal*, 2018.
4. Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He, and Dongyan Zhao. Natural language question answering over rdf: a graph data driven approach. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 313–324. ACM, 2014.