The Scalable Question Answering over Linked Data (SQA) Challenge 2018

Giulio Napolitano[1], Ricardo Usbeck[2], and Axel-Cyrille Ngonga Ngomo[2]

Fraunhofer-Institute IAIS, Sankt Augustin, Germany
giulio.napolitano@iais.fraunhofer.de
Data Science Group, University of Paderborn, Germany
ricardo.usbeck|axel.ngonga@uni-paderborn.de

1 Introduction

Question answering (QA) systems, which source answers to natural language questions from Semantic Web data, have recently shifted from the research realm to become commercially viable products. Increasing investments have refined an interaction paradigm that allows end users to profit from the expressive power of Semantic Web standards, while at the same time hiding their complexity behind intuitive and easy-to-use interfaces. Not surprisingly, after the first excitement we did not witness a cooling-down phenomenon: regular interactions with question answering systems have become more and more natural. As consumers' expectations around the capabilities of systems able to answer questions formulated in natural language keep growing, so is the availability of such systems in various settings, devices and languages. Increasing usage in real (non-experimental) settings have boosted the demand for resilient systems, which can cope with high volume demand.

The Scalable Question Answering (SQA) challenge stems from the long-standing Question Answering over Linked Data (QALD)[1] challenge series, aiming at providing an up-to-date benchmark for assessing and comparing state-of-the-art-systems that mediate between a large volume of users, expressing their information needs in natural language, and RDF data. It thus targets all researchers and practitioners working on querying Linked Data, natural language processing for question answering, 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 increasing volume of requests that QA systems have to process, as well as with the large, heterogeneous and distributed nature of Semantic Web data.

The SQA challenge at ESWC2018 was supported by the EU project HOBBIT[2]. The employment of the platform provided by the HOBBIT project guaranteed a robust, controlled setting offering rigorous evaluation protocols. This now popular platform for the benchmarking of linked data has attracted a high number of users, as semantic data and knowledge bases are gaining relevance

[1] http://www.sc.cit-ec.uni-bielefeld.de/qald/
[2] https://project-hobbit.eu/

also for information retrieval and search, both in academia and for industrial players.

2 Task and Dataset Creation

The key difficulty for Scalable Question Answering over Linked Data is in the need to translate a user's information request into such a form that it can be efficiently evaluated using standard Semantic Web query processing and inferencing techniques. Therefore, the main task of the SQA challenge was the following:

> Given an RDF dataset and and a large volume of natural language questions, return the correct answers (or SPARQL queries that retrieves those answers).

Successful approaches to Question Answering are able to scale up to big data volumes, handle a vast amount of questions and accelerate the question answering process (e.g. by parallelization), so that the highest possible number of questions can be answered as accurately as possible in the shortest time. The focus of this task is to withstand the confrontation of the large data volume while returning correct answers for as many questions as possible.

Dataset

Training set. The training set consists of the questions compiled for the award-nominated LC-QuAD dataset [4], which comprises 5000 questions of variable complexity and their corresponding SPARQL queries over DBpedia (2016-10 dump) as the RDF knowledge base[3]. The questions were compiled on the basis of query templates which are instantiated by seed entities from DBpedia into normalised natural question structures. These structures have then been transformed into natural language questions by native speakers. In contrast to the analogous challenge task run at ESWC in 2017 [5], the adoption of this new dataset ensured an increase in the complexity of the questions as the corresponding SPARQL queries are not limited to one triple in the WHERE clause but also include patterns of two or three triples. In the dataset some spelling mistakes and anomalies are also introduced, as a way to simulate a noisy real-world scenario in which questions may be served to the system imperfectly as a result, for instance, of speech recognition failures or typing errors.

Test set. The test set was created by manually paraphrasing 2200 questions from the LC-QuAD dataset. The paraphrased questions where ordered by their Vector Extrema Cosine Similarity score [3] to the original questions and only the first 1830 questions were retained. As in the original source questions, we intentionally left in typos and grammar mistakes in order to reproduce realistic scenarios of imperfect input.

[3] `http://downloads.dbpedia.org/2016-10/core-i18n/`

Data format

The test data for the challenge, without the SPARQL queries, can be found in our project repository `https://hobbitdata.informatik.uni-leipzig.de/SQAOC/`. We used a format similar to the QALD-JSON format[4] and the following sample shows the first two entries of the training set:

```
{
  "dataset": {
    "id": "lcquad-v1"
  },
  "questions": [
    {
      "hybrid": "false",
      "question": [
        {
          "string": "Which comic characters are painted
              by Bill Finger?",
          "language": "en"
        }
      ],
      "onlydbo": true,
      "query": {
        "sparql":
            "SELECT DISTINCT ?uri
            WHERE {?uri <http://dbpedia.org/ontology/
               creator> <http://dbpedia.org/resource/
               Bill_Finger>  .
                ?uri <http://www.w3.org/1999/02/22-rdf-
                   syntax-ns#type> <http://dbpedia.org
                   /ontology/ComicsCharacter>
            }"
      },
      "aggregation": false,
      "_id": "f0a9f1ca14764095ae089b152e0e7f12",
      "id": 0
    },
    {
      "hybrid": "false",
      "question": [
        {
          "string": "Was winston churchill the prime
             minister of Selwyn Lloyd?",
          "language": "en"
        }
```

[4] https://github.com/AKSW/gerbil/wiki/Question-Answering

```
      ],
      "onlydbo": true,
      "query": {
        "sparql":
            "ASK WHERE {<http://dbpedia.org/resource/
               Selwyn_Lloyd> <http://dbpedia.org/
               ontology/primeMinister> <http://dbpedia
               .org/resource/Winston_Churchill>}"
      },
      "aggregation": true,
      "_id": "30b709079ea5421cb33c227c3feb9019",
      "id": 1
    },
...
```


3 Evaluation

The SQA challenge provides an automatic evaluation tool (based on GERBIL QA [6] and integrated into the HOBBIT platform)[5,6] that is open source and available for everyone to re-use. The HOBBIT platform also incorporates a leaderboard feature to facilitate comparable evaluation and result display of systems participating in challenges. The tool is also accessible online, so that participants were able to upload their systems as Docker images and check their (and others') performance via a webservice. The ranking of the systems was based on the usual KPIs (precision, recall and F measure) plus a "response power" measure, which is also taking into account the ability of the systems to cope with high volume demand without failure. The response power is the harmonic mean of three measures: precision, recall and the ratio between processed questions (an empty answer is considered as processed, a missing answer is considered as unprocessed) and total number of questions sent to the system. The final ranking was on

1. response power
2. precision
3. recall
4. F measure

[5] http://gerbil-qa.aksw.org/gerbil/
[6] http://master.project-hobbit.eu/

in that order. For each system q, precision, recall and response power are computed as follows:

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

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

$$\text{response power}(q) = \frac{3}{\frac{1}{\text{precision}(q)} + \frac{1}{\text{recall}(q)} + \frac{\text{processed}}{\text{submitted}}}$$

The benchmark sends to the QA system one question at the start, two more questions after one minute and continues to send *n+1* new questions after n minutes. One minute after the last set of questions is dispatched, the benchmarks closes and the evaluation is generated as explained above. The 1830 questions in the dataset allow the running of the benchmark for one hour but for the SQA challenge we limited to 30 sets of questions.

4 Participating systems

Three teams participated in the SQA challenge. We provide here brief descriptions, please refer to the respective full papers (where they exist) for more detailed explanations.

WDAqua-core1 [1] is built on a rule-based system using a combinatorial approach to generate SPARQL queries from natural language questions. In particular, the system abstracts from the specific syntax of the question and relies on the semantics encoded in the underlying knowledge base. It can answer questions on a number of Knowledge Bases, in different languages, and does not require training.

LAMA [2] was originally developed for QA in French. It was extended for the English language and modified to decompose complex queries, with the aim of improving performance on such queries and reduce response times. The question type (e.g. *Boolean* or *Entity*) is classified by pattern matching and processes by the relevant component to extract entities and properties. Complex questions are decomposed in simple queries by keyword matching.

GQA [7], the Grammatical Question Answering system, is built around a functional programming language with categorial grammar formalism. The question is parsed according to the grammar and the best parse is selected. Finally, this is decomposed into its elements, starting from the innermost, while requests are sent to DBpedia to find the corresponding values and the final answers.

5 Results

The experimental data for the SQA challenge over the test dataset can be found at the following URLs:

– WDAqua: `https://master.project-hobbit.eu/experiments/1527792517766`,
– LAMA: `https://master.project-hobbit.eu/experiments/1528210879939`,
– GQA): `https://master.project-hobbit.eu/experiments/1528283915360`.

By providing human- and machine-readable experimental URIs, we provide deeper insights and repeatable experiment setups.

Note also that the numbers reported here may differ from the publications of the participants, as these figures were not available at the time of participant paper submission.

Test	**WDAqua**	**LAMA**	**GQA**
Response Power	0.472	0.019	0.028
Micro Precision	0.237	0.054	0.216
Micro Recall	0.055	0.001	0.002
Micro F1-measure	0.089	0.001	0.004
Macro Precision	0.367	0.010	0.018
Macro Recall	0.380	0.016	0.019
Macro F1-measure	0.361	0.011	0.019

Table 1. Overview of SQA results.

6 Summary

The Scalable Question Answering over Linked Data challenge introduced a new metric (Response Power) to evaluate the capability of a QA system to perform under increasing stress. For the first time, it also partially employed complex and non-well-formed natural language questions, to make the challenge even closer to real use scenarios. In this challenge, we also kept last year underlying evaluation platform (HOBBIT) based on docker, to account for the need for comparable experiments via webservices. This introduces an entrance threshold for participating teams but ensures a long term comparability of the system performance and a fair and open challenge. Finally, we offered leader boards prior to the actual challenge in order to allow participants to see their performance in comparison with the others. Overall, we are confident that the HOBBIT platform will be able to provide QA challenge support for a long time, making comparable and repeatable question answering research possible.

Acknowledgments. This work was supported by the European Union's H2020 research and innovation action HOBBIT under the Grant Agreement number 688227.

References

1. Dennis Diefenbach, Kamal Singh, and Pierre Maret. On the scalability of the QA system WDAqua-core1. In Diego Reforgiato Recupero and Davide Buscaldi, editors, *Semantic Web Challenges*, Cham, 2018. Springer International Publishing.
2. Nikolay Radoev, Mathieu Tremblay, Amal Zouaq, and Michel Gagnon. LAMA: a language adaptive method for question answering. In Diego Reforgiato Recupero and Davide Buscaldi, editors, *Semantic Web Challenges*, Cham, 2018. Springer International Publishing.
3. Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799, 2017.
4. Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. *LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs*, pages 210–218. Springer International Publishing, Cham, 2017.
5. Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 7th Open Challenge on Question Answering over Linked Data (QALD-7). In Mauro Dragoni, Monika Solanki, and Eva Blomqvist, editors, *Semantic Web Challenges*, pages 59–69, Cham, 2017. Springer International Publishing.
6. 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.
7. Elizaveta Zimina, Jyrki Nummenmaa, Kalervo Jarvelin, Jaakko Peltonen, Kostas Stefanidis, and Heikki Hyyro. GQA: Grammatical question answering for rdf data. In Diego Reforgiato Recupero and Davide Buscaldi, editors, *Semantic Web Challenges*, Cham, 2018. Springer International Publishing.