Semantic Web 1 (2016) 1–5
IOS Press

Benchmarking Question Answering Systems

Editor(s): Ruben Verborgh, Ghent University – iMinds, Belgium
Solicited review(s): Philipp Cimiano, University of Bielefeld, Germany; Lei Zou, Peking University, China; one anonymous reviewer
Open review(s): Seifeddine Mechti, University of Tunis

Ricardo Usbeck[c,*], Michael Röder[c], Michael Hoffmann[a], Felix Conrads[c], Jonathan Huthmann[a], Axel-Cyrille Ngonga-Ngomo[c], Christian Demmler[a], and Christina Unger[b]

[a] *AKSW Group, University of Leipzig, Germany*
[b] *CITEC, University of Bielefeld, Germany*
[c] *DICE - Data Science Group, Paderborn University, Germany*

Abstract. The necessity of making the Semantic Web more accessible for lay users, alongside the uptake of interactive systems and smart assistants for the Web, have spawned a new generation of RDF-based question answering systems. However, fair evaluation of these systems remains a challenge due to the different type of answers that they provide. Hence, repeating current published experiments or even benchmarking on the same datasets remains a complex and time-consuming task.

We present a novel online benchmarking platform for question answering (QA) that relies on the FAIR principles to support the fine-grained evaluation of question answering systems. We detail how the platform addresses the fair benchmarking platform of question answering systems through the rewriting of URIs and URLs. In addition, we implement different evaluation metrics, measures, datasets and pre-implemented systems as well as methods to work with novel formats for interactive and non-interactive benchmarking of question answering systems. Our analysis of current frameworks shows that most of the current frameworks are tailored towards particular datasets and challenges but do not provide generic models. In addition, while most frameworks perform well in the annotation of entities and properties, the generation of SPARQL queries from annotated text remains a challenge.

Keywords: Factoid Question Answering, Benchmarking, Repeatable Open Research

1. Introduction

The Web of Data has grown to contain billions of facts pertaining to a large variety of domains. While this wealth of data can be easily accessed by experts, it remains difficult to use for non-experts [7,41]. This need has led to the development of a large number of question answering (QA) and keyword search tools for the Web of Data [37,38,39,40,44]. As benchmarking has been credited with the more rapid advancement of research, many campaigns and challenges (e.g., Question Answering on Linked Data [38, 39,40], BioASQ [37]) have evolved around the QA research field (see Section 2) since the evolution of the first question answering system [17]. A significant improvement in F-measure for question answering frameworks has been achieved in recent years, an increase which is partly due to the existence of such campaigns [20]. However, evaluation datasets, measures and QA system processes are hardly documented. In addition, the few existing testbeds are commonly tailored toward a particular challenge and cannot be used universally. Hence, there is no overview of the performance of frameworks outside of the challenges, making the evaluation of (1) the state of the art and (2) the weaknesses and strengths of existing systems tedious if not impossible.

Motivated by the more than 17,000 experiments that have already been run on GERBIL [43] and the improvement of named entity recognition (NER) and entity linking (EL) systems by over 12% F-measure since the deployment of GERBIL, we address the drawbacks

[*] ricardo.usbeck@uni-paderborn.de

aforementioned by presenting a novel benchmarking platform for question answering systems dubbed **GERBIL QA**. Our platform relies on the foundations provided by the community-approved GERBIL framework for benchmarking Named Entity Recognition and Entity Linking systems [43] (see Figure 1). While we reused the mechanisms provided by GERBIL to store experiments and generate corresponding URIs, we replaced the core components of semantic annotation systems, datasets, metrics and matching procedures since they are not usable for the QA benchmarking task. In particular, we addressed the crucial problem of benchmarking systems which return equivalent URIs, URLs and strings, in a fair manner. In doing so, we provide the QA community with the means to perform citable, comparable and extensible in-depth benchmarking of QA systems.

GERBIL QA follows the FAIR principles [49]:

- **F**indable: All experimental (meta)data is available in persistent RDF as JSON-LD and in a SPARQL endpoint[1] using the rich DataID [6] and DataCube [9] vocabularies.
- **A**ccesible: Experiments can be linked via W3ID[2] URIs using the HTTP protocol for a human- or machine-readable version.
- **I**nteroperatable: All (meta)data and its respective identifiers uses RDF as formal, accessible, shared, and broadly applicable language for knowledge representation.
- **R**e-Usable: Every captured evaluation metric is described via an RDF model and released without any license restrictions. The metadata further describes the provenance of the used system and dataset.

Our approach differs from the state of the art (including GERBIL) and addresses the following drawbacks of existing challenges and systems for evaluating question answering:

- *Datasets*: Current evaluation campaigns and challenges offer a dataset (mostly by mere reference) and a set of questions without any extensibility. We address this drawback by allowing for the user-driven addition of datasets.
- *Reference implementations*: The development of QA systems driven by the objective assessment of the weaknesses of one's own system in comparison to existing solutions was quasi impossible. With GERBIL QA, users can continuously benchmark their systems against the solutions included in the platform.
- *Evaluation*: The fair evaluation of knowledge base-based QA systems across different URLs and URIs (e.g. Wikipedia vs. Freebase) used to refer to the same real-world object, which has not been investigated before this project but is now an integral part of GERBIL QA.

To address these challenges, GERBIL QA provides the following novel contributions:

- We offer 8 metrics for benchmarking QA systems as well as 6 novel QA (sub-)experiment types to (1) allow for a fine-grained evaluation of QA systems and (2) improve the diagnostic process.
- While we reuse the existing GERBIL core, we provide novel matching and metric calculations for QA since existing evaluation platforms do not offer this functionality.
- We integrate 6 existing QA systems into the platform and provide an unprecedented bundle of 22 question answering datasets (QALD-1 to QALD-6 and NLQ) to evaluate these systems. We hence present the first integral comparison of QA systems for Linked Data across challenges.
- Our framework supports both online systems and file-based evaluation campaigns over a large variety of datasets. That is, we allow for the upload of system results as well as datasets on the fly as well as webservices for systems.
- In addition, we support three widely used formats for the interactive communication of QA systems via webservices.

Note that GERBIL QA reuses the mechanisms provided by GERBIL to offer citable, stable experiment URIs and descriptions, which are both human and machine-readable. To this end, GERBIL QA uses the recently proposed DataID [6] ontology which is based on a combination of VoID [2] and DCAT [25] metadata with Prov-O [23] provenance information and ODRL [27] licenses to describe datasets.

A demo of the system is available at `http://gerbil-qa.aksw.org/gerbil/`. Furthermore, we made the datasets, utilities and the source code openly available and extensible.[3] A general overview

[1] `http://gerbil-qa.aksw.org/sparql`

[2] `https://w3id.org/`

[3] `https://github.com/AKSW/NLIWOD` and `https://github.com/AKSW/gerbil/tree/QuestionAnswering`

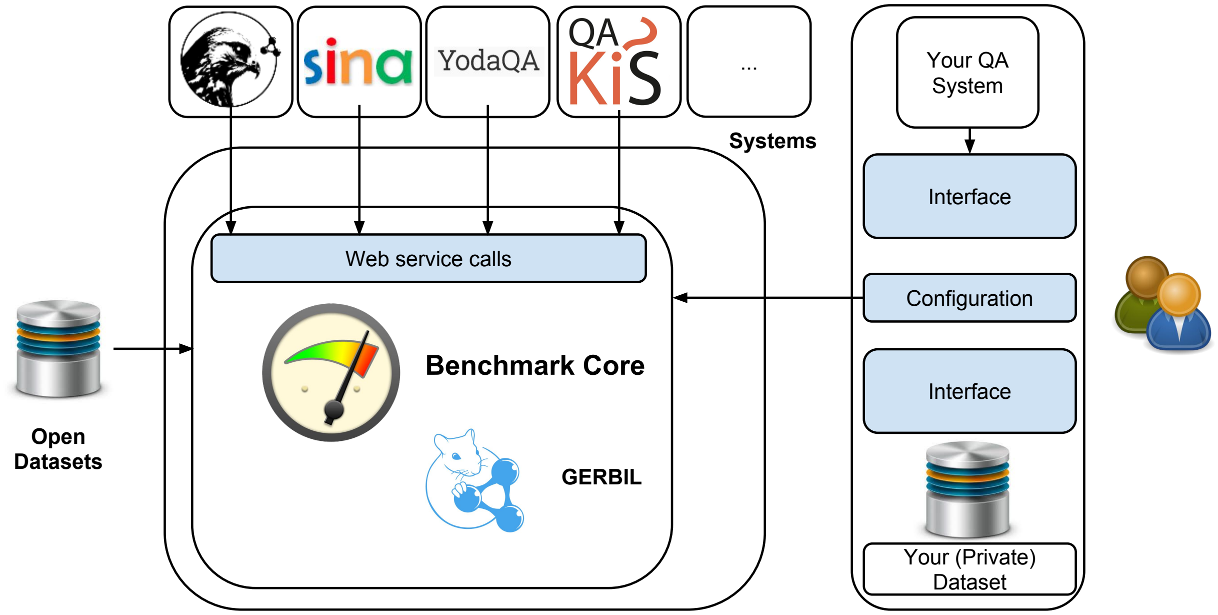

Fig. 1. Overview of the Question Answering Benchmarking platform based on the GERBIL core.

of the GERBIL framework can be found at the project website.[4] Note, while our platform focuses on RDF-based systems, i.e., question answering systems based on Linked Data and other resources providing RDF resources or literals as answers, it can be easily extended to non-RDF systems.

2. Question Answering Benchmarking Campaigns

Like in other disciplines, QA researchers and practitioners require reliable test environments and comparison methods to step-up their development speed and lower entrance barriers. There has thus been a number of challenges and campaigns attracting researchers as well as industry practitioners to QA. Since 1998, the TREC conference, especially the QA track [47], aims to provide domain-independent evaluations over large, unstructured corpora. This seminal campaign pushed research projects forward over the course of its more than ten implementations. The latest TREC-QA tackles the field of live QA[5] where systems answer real-life, real-time questions of users submitted to popular community-based Question and Answer sites. The CLEF campaigns on information retrieval have a more than 10 year tradition in evaluating IR systems [1]. However, here we focus on benchmarking QA systems that are able to return a concise set of answers rather than snippets from documents to a particular keyword query.

Next to that, the BioASQ series [37] challenges semantic indexing as well as QA systems on biomedical data and is currently at its fifth installment. Here, systems have to work on RDF as well as textual data to present matching triples as well as snippets of text. Moreover, the OKBQA[6] is primarily an open QA platform powered by several Korean research institutes such as KAIST. The KAIST institute also released the NLQ datasets within their 3rd hackathon.[7] This dataset is answerable purely by Wikipedia or its machine-readable version – DBpedia – using SPARQL. The well-known QALD (Question Answering over Linked Data) [40] campaign, currently running in its 6th instantiation, is a diverse evaluation series including 1) RDF-based, 2) hybrid, i.e., RDF and textual data, 3) statistical 4) multi-knowledge base and 5) music-domain-based benchmarks.

In the following, we will use the datasets and formats (QALD-XML, QALD-JSON) as a base for our benchmarking suite, since they have been adopted by

[4] https://gerbil.aksw.org

[5] https://sites.google.com/site/trecliveqa2016/

[6] http://www.okbqa.org

[7] http://2015.okbqa.org/nlq

more than 20 QA systems since 2011 (see [20,11] and Table 2). So far, yearly QALD events enable participants to upload XML or JSON-based system answers to previously uploaded files on the QALD website.

In contrast to existing challenges and campaigns, our platform

1. allows the use of curated, updated benchmark datasets (e.g., via Github) instead of once-uploaded-static files and
2. allows to refer to specific experiments using a specific version of datasets by providing a time and date when the experiment was executed. This is a major issue when aiming to run benchmarks developed on previous versions of a dataset whose SPARQL endpoint has been updated over the years (e.g., running QALD-3 on the 2016 DBpedia endpoint) as the results achieved differ completely from those specified in the benchmark with some queries being not possible to execute.
3. In addition, GERBIL QA allows the implementation of wrappers for QA systems respectively using REST interfaces in an interactive manner to benchmark QA systems online and in real-time, (see Section 4).

We refer the interested reader to our dataset project homepage[8] to read more or add novel datasets.

3. Datasets

In its current version, our framework supports 21 QALD campaign datasets[10] as well as the OKBQA NLQ shared task 1 dataset[11] listed in Table 1. The versions of the datasets used here are curated versions of the original datasets with respect to correctness of answers, quality of questions and completeness of metadata. It is important to note that no evaluation campaign, especially QALD and OKBQA, offers endpoints for all knowledge bases, i.e., developers and end users have to set up their own knowledge base (KB) endpoint for the respective version. In Table 1, the Knowledge Base version is the dataset which served as a background for the provided answersets for questions. That is, a certain benchmark dataset was created to work on a certain version of the KB and thus the answers could look different with another version of the KB. Curating these datasets to the most current KB is an open, future task. However, our platform already checks basic assertions to these datasets, such as the existence of answers in the gold standard or syntactical correctness of gold standard SPARQL queries.

In contrast to the existing benchmarking campaigns, GERBIL QA allows supplementary datasets to be added. Users can (1) add them to the project repository and write a dataset wrapper in Java or (2) upload a dataset as a file via our Web-interface for only one particular experiment. The first option enables other users to benchmark with this dataset and can thus spark the generation of new datasets. The second option allows the benchmarking of not yet ready or non-disclosed datasets. In addition to supporting JSON and XML files in the QALD format, GERBIL QA supports the extension dubbed eQALD-JSON, which we developed to address some of the drawbacks of the QALD format.[12]

Existing formats lack the possibility to measure a system's ability to recognize entities, classes or properties. Moreover, the QALD XML and JSON do not allow the benchmarking of systems with respect to their confidence in the computed answer. Thus, the main advantage of eQALD-JSON is that it represents the answers of a QA system, supporting the full set of benchmark types provided by GERBIL QA by explicating annotations, underlying SPARQL queries and more, (see Section 5). In particular, it includes 1) a knowledge base version, 2) questions in multiple languages and equivalent keyword queries, 3) annotations of the question w.r.t. RDF resources and properties, 4) meta-information like answer-type and answer-item-type, 5) a schemaless query[13] and a SPARQL query, and 6) answers from the KB formatted in a manner compliant with the W3C JSON-RDF standard,[14] as well as confidence scores for further evaluations. This format is currently being standardized for the evaluation of natural language interfaces (see Section 7) and is depicted below:

[8] https://github.com/AKSW/NLIWOD/tree/master/qa.datasets

[10] http://qald.sebastianwalter.org/

[11] http://3.okbqa.org/nlq

[12] See also the wiki pages for updates about the formats https://github.com/dice-group/gerbil/wiki/Question-Answering

[13] https://sites.google.com/site/eswcsaq2015/documents

[14] https://www.w3.org/TR/sparql11-results-json/

Table 1
Built-in datasets and their features.

Dataset	#Questions	Knowledge Base
NLQ shared task 1	39	DBpedia 2015-04
QALD1_Test_dbpedia	50	DBpedia 3.6
QALD1_Train_dbpedia	50	DBpedia 3.6
QALD1_Test_musicbrainz	50	MusicBrainz[9] (dump 2011)
QALD1_Train_musicbrainz	50	MusicBrainz (dump 2011)
QALD2_Test_dbpedia	99	DBpedia 3.7
QALD2_Train_dbpedia	100	DBpedia 3.7
QALD3_Test_dbpedia	99	DBpedia 3.8
QALD3_Train_dbpedia	100	DBpedia 3.8
QALD3_Test_esdbpedia	50	DBpedia 3.8 es
QALD3_Train_esdbpedia	50	DBpedia 3.8 es
QALD4_Test_Hybrid	10	DBpedia 3.9 + long abstracts
QALD4_Train_Hybrid	25	DBpedia 3.9 + long abstracts
QALD4_Test_Multilingual	50	DBpedia 3.9
QALD4_Train_Multilingual	200	DBpedia 3.9
QALD5_Test_Hybrid	10	DBpedia 2014 + long abstracts
QALD5_Train_Hybrid	40	DBpedia 2014 + long abstracts
QALD5_Test_Multilingual	49	DBpedia 2014
QALD5_Train_Multilingual	300	DBpedia 2014
QALD6_Train_Hybrid	49	DBpedia 2015-10 + long abstracts
QALD6_Train_Multilingual	333	DBpedia 2015-10
Total	**1431**	

```
{
"dataset": {
"id": "the dataset id",
"metadata": "some metadata..."
},
"questions": [{
"id": "the question id",
"metadata": {
"answertype": "Date|Number|String
|ListOfResource ",
"hybrid": "TRUE|FALSE",
"aggregation": "TRUE|FALSE",
"answeritemtype": [
"e.g., dbo:Person"
]
},
"question": [{
"language": "e.g. en or de",
"string": "The question in that
particular language...",
"keywords": "question as keywords",
"annotations": [{
"char_begin": "5...",
"char_end": "11...",
"URI": "e.g. dbr:Berlin...",
"type": "CLASS|PROPERTY|ENTITY"
}]
}],
"query": {
"SPARQL": "Question as SPARQL",
"schemaless ": "Schema-less SPARQL"
},
"answers": {
"bindings": [{
"result": {
"type": "...",
"value": "..."
}
}],
"confidence": "e.g. 0.9..."
}
}]
}
```

4. Systems

Table 2 shows that many systems of previous challenges and campaigns do not offer webservices, hence increasing the difficulty to benchmark them with novel datasets. Some offer webservice interfaces but they are either not open or demand human input. Other systems do not provide comprehensive answerset representations, e.g., showing whole paragraphs containing the answer to a question instead of a Linked Data URI. Another kind of system participating in past challenges did not leave any trace of fine granular quality assessment as they missed publications and webservices. Thus, the first release of GERBIL QA contains only 6 implemented system webservice clients. These are capable of answering hybrid, multilingual questions or keyword queries. These systems are:

1. HAWK [42], the first hybrid source QA system which processes RDF as well as textual information to answer one input query. HAWK is based on a mix of computational linguistics and semantic annotations to build SPARQL queries.
2. SINA [34], a keyword and natural language query search engine which exploits the structure of RDF graphs to implement an explorative search approach. The system is based on Hidden Markov Models for choosing the correct dataset to query based on a SPARQL generation process.
3. YodaQA [3], a modular, open-source, hybrid approach built on top of the Apache UIMA framework.[15] YodaQA allows easy parallelization and leverages pre-existing NLP UIMA components by representing each artifact (question, search result, passage, candidate answer) as standalone module.
4. QAKIS [7], a language-agnostic QA system grounded in ontology-relation matches. Here, the relation matches are based on surface forms extracted from Wikipedia to enforce a wide variety of context matches. QAKiS matches only one relation per query and moreover relies on basic heuristics which do not account for the variety of natural language in general.
5. QANARY [5] follows the desire to reuse the most components possible to enable a best-of-breed QA system following a new methodology for combining preexisting modules. Thus, QANARY itself is a rapid development environment for new QA systems and a QA system itself.
6. OKBQA [22] was recently introduced by Kim et al. to likewise facilitate a strong collaboration among experts. The Open Knowledge Base Question Answering system thus supports the development of a new QA system reusing collaborative and intuitive ways.

Currently, GERBIL QA supports the addition of 3 types of systems: (1) services implemented as Java-

Table 2
Systems that participated in past QALD challenges. Note, having a publication is optional with QALD. U means unreliable webservice, N not yet implemented due to non-open API, M human interaction needed.

Engine	Reference	Webservice?	Reason for Exclusion
	QALD-1		
FREyA	[10]	—	—
PowerAqua	[24]	✓	U
SWIP	[8]	—	—
	QALD-2		
SemSeK	—	—	—
Alexandria	[48]	✓	N
MHE	—	—	—
QAKiS	[7]	✓	—
	QALD-3		
squal2sparql	[15]	—	—
CASIA	[18]	—	—
Scalewelis	[21]	—	—
RTV	[16]	—	—
Intui2	[13]	—	—
SWIP	[29]	—	—
	QALD-4		
Xser	[50]	—	—
gAnswer	[51]	✓	N
CASIA	[19]	—	—
Intui3	[14]	—	—
ISOFT	[28]	—	—
RO_FII	—	—	—
	QALD-5		
Xser	[50]	—	—
APEQ	—	—	—
QAnswer	[32]	—	—
SemGraphQA	[4]	—	—
YodaQA	[3]	✓	—
ISOFT	[28]	—	—
HAWK	[41]	✓	—
	QALD-6		
CANaLI	[26]	✓	M
PersianQA	—	—	—
UTQA	[46]	—	—
KWGAnswer	—	—	—
NbFramework	—	—	—
SemGraphQA	[4]	—	—
UIQA	—	—	—

[15] https://uima.apache.org/

based wrapper (see above), (2) services configured via the Web-interface as webservice or (3) file uploads. Option (2) demands responses as either QALD-JSON or eQALD-JSON while (3) supports QALD-XML files as well. For option (1), we implemented the 4 systems which were available as webservice and returned Linked Data. We tested option (2) using the recent QANARY [35] framework. Option (3) was tested with the QALD-6 data and will be used for the 7th instantiation of the QALD challenge. This option enables developers to benchmark their system without setting up a webservice endpoint under a public address. Within the main GERBIL platform, experiments and log files remain private until published, i.e., companies and interested parties can test their systems online without fearing premature publication.

5. Experiment Evaluation

In this section, we will explain the different experiment types to evaluate a QA system, as well as how the evaluation metrics are computed and the system answers are compared. Throughout this section, we will use the question *"Who are the children of Ann Dunham?"* as running example.

5.1. Experiment Types

GERBIL QA allows the performance of common components of QA systems (named entity recognition, entity linking, etc.) to be measured, in addition to the benchmarking of whole QA systems. We use the term *sub-experiments* to denote experiments for benchmarking such sub-components. We designed and implemented 5 sub-experiments inspired by past evaluation campaigns. Moreover, we follow the motivation to also measure sub-experiments suggested by Both et al. [5]. For the sub-experiments P2KB and RE2KB we argue also, that in most QA systems, two different components are responsible for linking resources and properties and thus these features must be evaluated independently according to a recent study [33]. The goal is to provide system designers, researchers and decision makers with the opportunity to spot particular flaws in a QA pipeline and gain in-depth insights about the performance on different aspects of systems on diverse datasets.

The data necessary to carry out all five of these sub-experiments can be provided via eQALD-JSON. For four of the five new sub-experiments, the needed data can also be derived from the SPARQL query that might be returned by the QA system via QALD-XML or QALD-JSON.

Table 3
Availability of sub-experiments if the data has the QALD format without a SPARQL query, including a SPARQL query (i.S.q.) and for the eQALD-JSON format.

	QALD	QALD i.S.q.	eQALD-JSON
QA	✓	✓	✓
C2KB		✓	✓
P2KB		✓	✓
RE2KB		✓	✓
AT			✓
AIT2KB		✓	✓

Question Answering (QA). The first experiment is the classic experiment as described by evaluation campaigns like OKBQA and QALD. It aims to measure the capability of a system to answer questions correctly. A system's answer and the corresponding gold standard answer are regarded as the set of URIs and literals. For our running example, GERBIL QA expects a set of URIs containing `dbr:Maya_Soetoro-Ng` and `dbr:Barack_Obama`.[16]
Note, that if a different set is returned, we refer to our matching algorithm to try to match the answers, (see Section 5.3).

Resource to Knowledge Base (C2KB). This sub-experiment aims to identify all relevant resources for the given question. It is known from GERBIL [43] as *Concept to Knowledge Base*. The evaluation calculates the measure's precision, recall and F-measure based on a comparison of the expected resource URIs and the URIs returned by the QA system. Instead of a simple string comparison, we make use of an advanced meaning matching implementation offered by GERBIL, which is explained in the technical report for GERBIL version 1.2.2 [31]. With respect to our running example, GERBIL QA would expect a system to annotate `dbr:Ann_Dunham`.

Properties to Knowledge Base (P2KB). For this experiment, the system must identify all properties that are relevant for the given question. The experiment is evaluated in a manner akin to that of the C2KB experiment. In our case, the correct answer would be `dbo:children`.

[16] The prefix `dbr` is used for `http://dbpedia.org/resource/` while `dbo` is used for `http://dbpedia.org/ontology/` throughout the paper.

Relation to Knowledge Base (RE2KB). This sub-experiment focuses on the triples that have to be extracted from the question and are needed to generate the SPARQL query that would retrieve the correct answers. These triples can contain resources, variables and literals. The evaluation of this sub-experiment calculates precision, recall and F-measure based on the comparison of the expected triples from the gold standard SPARQL query and the triples or the SPARQL query returned by the QA system. For achieving a true positive, a returned triple has to match an expected triple. Two triples are counted as matching if they contain the same resources at the same positions. If they contain variables, the positions of the variables must be the same but the variable names are ignored. If they contain a literal, the value of the literal must be the same. Regarding the running example, we would expect a system to build the triple `dbr:Ann_Dunham dbo:children ?uri.`

Answer Type (AT). The identification of the answer type is an important part of a QA system. We distinguish 5 different answer types extracted from the QALD benchmarking campaign [40], i.e., `date`, `number`, `string`, `boolean` and `resource`, where resource can be a single URI or a set of URIs. A single answer type is expected for each question. This is the type for which the F-measure is calculated. Note that this sub-experiment can only generate meaningful results if the eQALD-JSON is used. For the case of our running example, we expect `resource` as answer type.

Answer Item Type to Knowledge Base (AIT2KB). The answer item types are the `rdf:type` information of the returned resources. Precision, recall and F-measure are calculated based on the set of expected types. If the expected answerset of a question does not contain resources then the set of answer item types is expected to be empty. Here, we would expect to see `dbo:Person` as answer item type because both answers are persons.

5.2. Metrics

GERBIL QA implements 9 evaluation metrics, i.e., micro- as well as macro precision, recall and F-measure, a QALD-specific Macro F1 metric as well as the runtime and number of errors of webservice-based QA systems [43]. As a reminder, the F1-Score is defined as

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad , \qquad (1)$$

with

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{and} \qquad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad , \qquad (3)$$

with respect to a set of provided answers.

For the micro precision, recall and F-measure, we first collect all true and false positives and negatives and only in the end average them. Thus, this measure actually gives more weight on questions which have many answers.

For the macro metric, we calculate the precision, recall and F-measure per question and average these metrics individually at the end. Thus, this measure assigns more meaning to the question of whether a system can answer all questions correctly. Note that it is possible that the F-measure is not between the precision and recall for all macro F-measures.

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.

For the Macro F1 QALD metric, we decided to have a more comparable metric to older QALD challenges and also to follow community requests.[17] 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.

However, GERBIL QA offers the implementation of additional metrics [31]. Thus, it would be possible to use a hierarchical F-measure, e.g., for the AIT2KB sub-experiment [30].

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

5.3. Answer Matching

A general problem of benchmarking current QA systems is the different answerset formats. The example question might be answered with the resources listed above or the names *"Maya Soetoro-Ng"* and *"Barack Obama"*.

Our approach chooses a matching strategy based on the type of response that is expected by the benchmark.

In this case, the gold standard answerset asks for a list of resources, like in the running example above, our approach can handle two types of answersets. First, if the QA system returns an RDF resource, GERBIL QA relies on the transitive closure of resource URIs that are connected by `owl:sameAs` links [31]. One set is generated for the gold standard answer and the other set for the returned resource. If both sets intersect, the answer of the system is correct, i.e., counted as a true positive for the normal precision and recall calculation. This approach enables the benchmarking of QA systems with datasets even if both are not based on the same KB as long as we can find `owl:sameAs` relations between the KBs.

Second, if the answer type demanding a resource is a plain string, GERBIL QA tries to use it as label for the resource. However, a returned label like *"Barack Obama"* might be shared by several resources. In this case, all resources are retrieved and used as input for the resource-based strategy described previously. Note that this might decrease the precision of the system since not all retrieved resources sharing the given label match the expected answers.

Another problem is that in the course of time the solution for a question can differ and QA systems should provide the most recent answers. If a question refers to the current president of the USA, it would generate different results today compared with 2015. To challenge this problem, GERBIL QA provides a way to update answers for older QA datasets. If present, the SPARQL query in the QALD file can be used to ask a specific configurable KB and receive the latest answers. The main drawback of this feature is that the answerset is not manually curated and contains unchecked results.

Strings, Dates and Numbers are currently matched by exact string matches. In the future, we will extend this by more sophisticated matching strategies such as lexical mapping, e.g. towards XSD datatypes.

Besides these result-focused metrics, our method measures the performance of live systems in two ways. First, it computes the average time a system needs to generate a response. Second, it counts the number of errors returned by the system, or that occur during communication with the system.

5.4. Diagnostics

The implemented sub-experiments lead to detailed insights about a system's performance. We created an example experiment[18] with four QA systems, three pre-implemented as well as an uploaded QALD XML answer file, on two datasets, namely QALD-5-train multilingual and hybrid. The uploaded HAWK file suggests an improvement over the pre-implemented HAWK system. The pre-implemented HAWK system however performs better on hybrid questions than on plain English questions. Systems like YODA, which only provide answers without a SPARQL query, cannot be analysed sufficiently. However, systems that also provide a SPARQL query can be analysed on their performance in the sub-experiments.

6. Sustainability Plan and Community

To foster an open community of QA researchers, we need a reliable platform for managing experimental data in a citable and comparable way, both readable for humans and machines. Thus, we published the GERBIL QA platform under the permanent ID `http://w3id.org/gerbil/qa`, which has been registered with W3ID.[19]

We presented this platform as a prototype for the W3C community group for Natural Language Interfaces for the Web of Data[20], which will build recommendations for benchmarks based on it. All experiment data and source code is open source, in particular, underlies a dual-LGPL license or is without any licence restrictions.[21] The project itself is hosted by the AKSW research group, who already maintain more than 50 projects.[22] Furthermore, the research and development unit of the University Leipzig Computation Center keeps daily backups to ensure long-term quotability. GERBIL is open-source software which can be maintained and hosted by anybody.

[18] `http://w3id.org/gerbil/qa/experiment?id=201605010001`

[19] `https://w3id.org/`

[20] `https://www.w3.org/community/nli/`

[21] `https://github.com/AKSW/gerbil/blob/master/LICENSE`

[22] `http://aksw.org/Projects.html`

We are seeing tremendous interest in the platform even though it is not yet published in any conference or journal. GERBIL QA has already been used for 85 experiments including more than 940 sub-experiment executions. For example, the developers of HAWK use the system to measure the performance of different configurations through uploads in this experiment `http://gerbil-qa.aksw.org/gerbil/experiment?id=201610230001`. Although the HAWK optimal configuration is overall better than the HAWK feature configuration, the feature configuration is more able to detect the correct Answer Item Type. Such insights enable researchers and developers to steer the development process more precisely.

7. Conclusion & Future work

We present the first online benchmarking system for question answering approaches over factoid questions. Our platform strives to speed up the development process by offering diverse datasets, systems and interfaces to generate repeatable and citable experiments with in-depth analytics of a system's performance. A known limitation is our focus on RDF-based systems (RDF resource matching, requiring the SPARQL query for sub-experiments), which we seek to circumvent in the future by using a standard to let interfaces communicate the needed information by demanding a SPARQL query within the result set.

In near-future developments, we will add additional metrics such as hierarchical f-measure, novel datasets (e.g. LCQUAD [36] as currently the largest QA benchmark dataset or the Wikidata-based dataset [12] presented at the 3rd NLIWOD workshop) and more systems. Moreover, we will unify the method of matching system answers with gold standard answers, thus pushing a fast-paced, open science movement. We will look into evaluation campaigns such as TREC LiveQA and CLEF to broaden our scope and also include non-factoid QA. Therefore, we need to look into enabling hybrid crowd-based evaluations within the workflow of the existing automatic evaluation. Furthermore, we will add this benchmarking platform to the H2020 HOBBIT project[23] to broadcast our activities. Also, we will bring this development to the W3C community group of Natural Language Interfaces for the Web of Data to standardize system interfaces and allow for an even easier and more concise benchmarking. Finally, GERBIL QA will be used as underlying system for the 7th instantiation of the QALD challenge [45].

Acknowledgements This work has been supported by Eurostars projects DIESEL (E!9367) and QAMEL (E!9725) as well as the European Union's H2020 research and innovation action HOBBIT (GA 688227). We thank the QANARY team for inspiring discussions. Furthermore, we want to thank Jin-Dong Kim for his thoughts on the novel QA format. We also want to acknowledge that this project has been supported 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).

References

[1] M. Agosti, G. M. D. Nunzio, M. Dussin, and N. Ferro. 10 years of CLEF data in DIRECT: where we are and where we can go. In T. Sakai, M. Sanderson, and W. Webber, editors, *Proceedings of the 3rd International Workshop on Evaluating Information Access, EVIA 2010, National Center of Sciences, Tokyo, Japan, June 15, 2010*, pages 16–24. National Institute of Informatics (NII), 2010. ISBN 978-4-86049-054-6. URL `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/EVIA/04-EVIA2010-AgostiM.pdf`.

[2] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets with the void vocabulary, 2011. http://www.w3.org/TR/void/.

[3] P. Baudiš and J. Šedivý. *CLEF'15*, chapter Modeling of the Question Answering Task in the YodaQA System, pages 222–228. Springer International Publishing, 2015.

[4] R. Beaumont, B. Grau, and A.-L. Ligozat. Semgraphqa at qald-5: Limsi participation at qald-5 at clef. In *CLEF (Working Notes)*, 2015.

[5] A. Both, D. Diefenbach, K. Singh, S. Shekarpour, D. Cherix, and C. Lange. Qanary — a methodology for vocabulary-driven open question answering systems. In *Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains - Volume 9678*, pages 625–641, New York, NY, USA, 2016. Springer-Verlag New York, Inc. ISBN 978-3-319-34128-6. . URL `http://dx.doi.org/10.1007/978-3-319-34129-3_38`.

[6] M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. DataID: Towards semantically rich metadata for complex datasets. In *10th International Conference on Semantic Systems 2014*, 2014.

[7] E. Cabrio, J. Cojan, F. Gandon, and A. Hallili. Querying Multilingual DBpedia with QAKiS. In *ESWC*, pages 194–198, 2013.

[8] C. Comparot, O. Haemmerlé, and N. Hernandez. An easy way of expressing conceptual graph queries from keywords

[23] `http://project-hobbit.eu/`

and query patterns. In *International Conference on Conceptual Structures*, pages 84–96. Springer, 2010.

[9] R. Cyganiak, D. Reynolds, and J. Tennison. The RDF Data Cube Vocabulary, 2014. http://www.w3.org/TR/vocab-data-cube/.

[10] D. Damljanovic, M. Agatonovic, and H. Cunningham. Natural language interfaces to ontologies: Combining syntactic analysis and ontology-based lookup through the user interaction. In *Extended Semantic Web Conference*, pages 106–120. Springer, 2010.

[11] D. Diefenbach, V. Lopez, K. Singh, and P. Maret. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information Systems*, Sep 2017. ISSN 0219-3116. . URL `https://doi.org/10.1007/s10115-017-1100-y`.

[12] D. Diefenbach, T. P. Tanon, K. Singh, and P. Maret. Question Answering Benchmarks for Wikidata. In *ISWC 2017*, Vienne, Austria, Oct. 2017. URL `https://hal.archives-ouvertes.fr/hal-01637141`.

[13] C. Dima. Intui2: A prototype system for question answering over linked data. In *CLEF (Working Notes)*, 2013.

[14] C. Dima. Answering natural language questions with intui3. In *CLEF (Working Notes)*, pages 1201–1211, 2014.

[15] S. Ferré. squall2sparql: a translator from controlled english to full sparql 1.1. In *Work. Multilingual Question Answering over Linked Data (QALD-3)*, 2013.

[16] C. Giannone, V. Bellomaria, and R. Basili. A hmm-based approach to question answering against linked data. In *CLEF (Working Notes)*. Citeseer, 2013.

[17] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224. ACM, 1961.

[18] S. He, S. Liu, Y. Chen, G. Zhou, K. Liu, and J. Zhao. Casia@ qald-3: A question answering system over linked data. In *CLEF (Working Notes)*, 2013.

[19] S. He, Y. Zhang, K. Liu, and J. Zhao. Casia@ v2: A mln-based question answering system over linked data. In *CLEF (Working Notes)*, pages 1249–1259, 2014.

[20] K. Höffner, S. Walter, E. Marx, J. Lehmann, A.-C. Ngonga Ngomo, and R. Usbeck. Overcoming Challenges of Semantic Question Answering in the Semantic Web. *Submitted to Semantic Web Journal*, 2016.

[21] G. Joris and S. Ferré. Scalewelis: a scalable query-based faceted search system on top of sparql endpoints. In *Work. Multilingual Question Answering over Linked Data (QALD-3)*, 2013.

[22] J. Kim, G. Choi, J. Kim, E. Kim, and K. Choi. The open framework for developing knowledge base and question answering system. In H. Watanabe, editor, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, December 11-16, 2016, Osaka, Japan*, pages 161–165. ACL, 2016. ISBN 978-4-87974-703-7. URL `http://aclweb.org/anthology/C/C16/C16-2034.pdf`.

[23] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology, 2013. http://www.w3.org/TR/prov-o/.

[24] V. Lopez, E. Motta, and V. Uren. Poweraqua: Fishing the semantic web. In *European Semantic Web Conference*, pages 393–410. Springer, 2006.

[25] F. Maali, J. Erickson, and P. Archer. Data Catalog Vocabulary (DCAT), 2014. http://www.w3.org/TR/vocab-dcat/.

[26] G. M. Mazzeo and C. Zaniolo. Canali: A system for answering controlled natural language questions on rdf knowledge bases, 2016.

[27] M. McRoberts and V. Rodríguez-Doncel. Open Digital Rights Language (ODRL) Ontology, 2014. http://www.w3.org/ns/odrl/2/.

[28] S. Park, S. Kwon, B. Kim, and G. G. Lee. Isoft at qald-5: Hybrid question answering system over linked data and text data. In *CLEF (Working Notes)*, 2015.

[29] C. Pradel, G. Peyet, O. Haemmerlé, and N. Hernandez. Swip at qald-3: results, criticisms and lesson learned. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. PROMISE Network of Excellence, 2013.

[30] M. Röder, R. Usbeck, and A.-C. Ngonga Ngomo. Developing a Sustainable Platform for Entity Annotation Benchmarks. In *ESWC Developers Workshop 2015*, 2015. URL `http://svn.aksw.org/papers/2015/ESWC_GERBIL_semdev/public.pdf`.

[31] M. Röder, R. Usbeck, and A.-C. Ngonga Ngomo. Gerbil's new stunts: Semantic annotation benchmarking improved. Technical report, Leipzig University, 2016. URL `http://svn.aksw.org/papers/2016/ISWC_Gerbil_Update/public.pdf`.

[32] S. Ruseti, A. Mirea, T. Rebedea, and S. Trausan-Matu. Qanswer-enhanced entity matching for question answering over linked data. In *CLEF (Working Notes)*, 2015.

[33] M. Saleem, S. N. Dastjerdi, R. Usbeck, and A. N. Ngomo. Question answering over linked data: What is difficult to answer? what affects the F scores? In R. Usbeck, A. N. Ngomo, J. Kim, K. Choi, P. Cimiano, I. Fundulaki, and A. Krithara, editors, *Joint Proceedings of BLINK2017: 2nd International Workshop on Benchmarking Linked Data and NLIWoD3: Natural Language Interfaces for the Web of Data co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21st - to - 22nd, 2017.*, volume 1932 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017. URL `http://ceur-ws.org/Vol-1932/paper-02.pdf`.

[34] S. Shekarpour, E. Marx, A.-C. N. Ngomo, and S. Auer. Sina: Semantic interpretation of user queries for question answering on interlinked data. *Journal of Web Semantics*, 2014.

[35] K. Singh, A. Both, D. Diefenbach, S. Shekarpour, D. Cherix, and C. Lange15. Qanary–the fast track to creating a question answering system with linked data technology. In *ESWC*, 2016.

[36] P. Trivedi, G. Maheshwari, M. Dubey, and J. Lehmann. Lc-quad: A corpus for complex question answering over knowledge graphs. In C. d'Amato, M. Fernandez, V. Tamma, F. Lecue, P. Cudré-Mauroux, J. Sequeda, C. Lange, and J. Heflin, editors, *The Semantic Web – ISWC 2017*, pages 210–218, Cham, 2017. Springer International Publishing. ISBN 978-3-319-68204-4.

[37] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artières, A. Ngonga, N. Heino, É. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, and G. Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:

138, 2015. . URL http://dx.doi.org/10.1186/s12859-015-0564-6.

[38] C. Unger, C. Forascu, V. Lopez, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. Question answering over linked data (QALD-4). In *CLEF*, pages 1172–1180, 2014.

[39] C. Unger, C. Forascu, V. Lopez, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. Question answering over linked data (QALD-5). In *CLEF*, 2015. URL http://ceur-ws.org/Vol-1391/173-CR.pdf.

[40] C. Unger, A.-C. N. Ngomo, and E. Cabrio. *6th Open Challenge on Question Answering over Linked Data (QALD-6)*, pages 171–177. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46565-4.

[41] R. Usbeck and A.-C. Ngonga Ngomo. HAWK@QALD5 – Trying to answer hybrid questions with various simple ranking techniques. In *CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org/Vol-1391)*, 2015. URL http://svn.aksw.org/papers/2015/CLEF_HAWK/public.pdf.

[42] R. Usbeck, A. N. Ngomo, L. Bühmann, and C. tina Unger. HAWK – Hybrid Question Answering Using Linked Data. In *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings*, pages 353–368, 2015.

[43] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – General Entity Annotation Benchmark Framework. In *24th WWW conference*, 2015.

[44] R. Usbeck, M. Röder, P. Haase, A. Kozlov, M. Saleem, and A.-C. N. Ngomo. Requirements to modern semantic search engines. In *KESW*, 2016.

[45] R. Usbeck, A.-C. N. Ngomo, B. Haarmann, A. Krithara, M. Röder, and G. Napolitano. 7th open challenge on question answering over linked data (qald-7). In *Semantic Web Evaluation Challenge*, pages 59–69. Springer, Cham, 2017.

[46] A. P. B. Veyseh. Cross-lingual question answering using common semantic space. In *Proceedings of TextGraphs@NAACL-HLT 2016: the 10th Workshop on Graph-based Methods for Natural Language Processing, June 17, 2016, San Diego, California, USA*, pages 15–19, 2016. URL http://aclweb.org/anthology/W/W16/W16-1403.pdf.

[47] E. M. Voorhees et al. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82, 1999.

[48] M. Wendt, M. Gerlach, and H. Düwiger. Linguistic modeling of linked open data for question answering. In *Extended Semantic Web Conference*, pages 102–116. Springer, 2012.

[49] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

[50] K. Xu, Y. Feng, and D. Zhao. *Xser@ QALD-4: Answering Natural Language Questions via Phrasal Semantic Parsing*. QALD-4, 2014.

[51] L. Zou, R. Huang, H. Wang, J. X. Yu, W. He, and D. 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.