QAnswer: A Question Answering prototype bridging the gap between a considerable part of the LOD cloud and end-users

Dennis Diefenbach, Pedro Migliatti, Omar Qawasmeh, Vincent Lully, Kamal Singh, Pierre Maret

To cite this version:

Dennis Diefenbach, Pedro Migliatti, Omar Qawasmeh, Vincent Lully, Kamal Singh, et al.. QAnswer: A Question Answering prototype bridging the gap between a considerable part of the LOD cloud and end-users. The Web Conference, May 2019, San Francisco, United States. 10.1145/3308558.3314124 . hal-02121400

HAL Id: hal-02121400

https://hal.archives-ouvertes.fr/hal-02121400

Submitted on 6 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
QAnswer: A Question Answering prototype bridging the gap between a considerable part of the LOD cloud and end-users

Dennis Diefenbach
Université de Lyon, CNRS UMR 5516
Laboratoire Hubert Curien
Saint-Etienne, France
dennis.diefenbach@univ-st-etienne.fr

Pedro Henrique Migliatti
Universidade Federal de São Carlos, MaLL
São Carlos, Brasil
619744@comp.ufscar.br

Omar Qawasmeh
Université de Lyon, CNRS UMR 5516
Laboratoire Hubert Curien
Saint-Etienne, France
omar.alqawasmeh@univ-st-etienne.fr

Vincent Lully
Sorbonne Université
Paris, France
vincent.lully@sorbonne-universite.fr

Kamal Singh
Université de Lyon, CNRS UMR 5516
Laboratoire Hubert Curien
Saint-Etienne, France
kamal.singh@univ-st-etienne.fr

Pierre Maret
Université de Lyon, CNRS UMR 5516
Laboratoire Hubert Curien
Saint-Etienne, France
pierre.maret@univ-st-etienne.fr

ABSTRACT

We present QAnswer, a Question Answering system which queries at the same time 3 core datasets of the Semantic Web, that are relevant for end-users. These datasets are Wikidata with Lexemes, LinkedGeodata and Musicbrainz. Additionally, it is possible to query these datasets in English, German, French, Italian, Spanish, Portuguese, Arabic and Chinese. Moreover, QAnswer includes a fallback option to the search engine Qwant when the answer to a question cannot be found in the datasets mentioned above. These features make QAnswer as the first prototype of a Question Answering System over a considerable part of the LOD cloud.

CCS CONCEPTS
• Information systems → Web search engines; Information retrieval; Resource Description Framework (RDF).

KEYWORDS
Question Answering, Semantic Web, Multilingual, Multi-Knowledge-Base

ACM Reference Format:

1 INTRODUCTION

In the last decade, some new datasets adhering to Semantic Web standards were published on the Web. This growth can be seen by looking at the Linked Open Data Cloud\(^1\) (LOD cloud), which collects datasets that have been published using Semantic Web Technologies. Though in 2007, the LOD cloud contained 12 datasets, it now contains 1,231\(^2\) datasets. The LOD cloud does not only contain a lot of datasets, but at the same time some of these datasets are very large. A dump of the LOD cloud called LOD-a-lot\(^3\) contains around 5 TB of structured data in uncompressed N-triples format.

While Semantic Web standards are designed to make data to be machine comprehensible, they do not allow, at the same time, easy accessibility of the data to non experts and in particular to end-users. Question Answering (QA) is seen as the technology that can allow to bridge this gap between Semantic Web data and end-users. While the industry presents QA solutions on their proprietary Knowledge Base (such as Google and Baidu), no system reached a peak which allows querying a substantial part of the LOD cloud.

Table 1 shows a list of QA systems which we are aware of, that are available online and query part of the LOD cloud. We indicate the datasets which they are able to query and the languages which are supported. For a general list and an overview of QA systems over KB, we refer to [4]. QAnswer represents a breakthrough since it allows to query many more datasets, at the same time, in real-time and in many more languages. The main algorithm behind QAnswer is described in [3]. In the following text, we show how the algorithm in [3] was implemented to create a first prototype of a QA System over the Semantic Web and which are the improvements over previous works.

2 DESCRIPTION

The algorithm behind QAnswer is described in [3]. It was shown that it has the following distinctive features:

- **Multilingual**, it supports multiple languages. In the previous work, it was shown that the algorithm works for English, German, French, Italian, Spanish, Portuguese. Moreover, the algorithm can easily be adapted to new languages.
Table 1: List of Question Answering systems that query part of the LOD cloud and that are available online.

<table>
<thead>
<tr>
<th>QA system</th>
<th>Lang</th>
<th>KBs</th>
<th>Url</th>
</tr>
</thead>
<tbody>
<tr>
<td>gAnswer[10]</td>
<td>en</td>
<td>DBpedia</td>
<td><a href="http://ganswer.gstore-pku.com/">http://ganswer.gstore-pku.com/</a></td>
</tr>
<tr>
<td>QAnswer</td>
<td>en, fr, de, es, pt, zh, ar</td>
<td>Wikidata, LinkedGeoData, MusicBrainz, DBpedia, DBLP</td>
<td><a href="http://qanswer.eu/qa">http://qanswer.eu/qa</a></td>
</tr>
</tbody>
</table>

- **Robust**, users ask questions using keywords, natural language questions and even malformed questions, i.e., syntactically wrong questions. The algorithm is robust enough to deal with all these scenarios, but not to spelling mistakes.
- **Real-time**, the algorithm can answer questions in real-time, i.e., on existing benchmarks like QALD and SimpleQuestions an average run-time of 2 seconds per query is realistic.
- **Low hardware footprint**, the algorithm uses specific indexes that guarantee low disk and memory footprint. Our demo will show that QAnswer can query 700Gb of n-triples and can be run on a (standard) laptop having 4 cores and 16 Gb of RAM.
- **Portable**, making question answering over a new dataset can be difficult. Some approaches need a lot of training data, other are not designed to be portable at all. Our system is designed such that any new dataset can be used as a base for a new QA system.
- **Multi-Knowledge-base**, the algorithm allows to query multiple Knowledge-bases at the same time.
- **Precision and Recall**, the algorithm was tested on multiple benchmarks and can compete with most of the existing approaches [3].

We are using the algorithm described in [3] to query what we believe are the three most significant datasets in the LOD cloud for end-users: Wikidata, LinkedGeoData and MusicBrainz. While, DBpedia and Freebase can also be queried, we consider these datasets as outdated since both are not maintained anymore.

We now describe the improvements over previous works brought by QAnswer.

- **Non-european languages**: While it was shown that the algorithm in [3] can be used to answer questions in multiple languages, it was only tested over European languages.

We are using the algorithm described in [3] to query Wikidata. When Wikidata was new, it contained only Q-items, i.e., resources describing a thing or an idea, but not the word describing it. Since 2018, Wikidata...
was extended to also store new information such as words, phrases and sentences, in many languages. This information is stored in new types of entities, called Lexemes (L), Forms (F) and Senses (S). The Lexeme extension can be seen as a structured representation of the Wiktionary\(^4\). There exists an earlier attempt to semantify Wiktionary by DBpedia project\(^5\). This project is unfortunately not maintained anymore and the extraction framework does not work anymore due to changes in the structure of Wiktionary. While the Lexemes in Wikidata still represent a small portion of the information contained in Wiktionary, we believe that the active Wikidata community will be able to fill this gap. Example questions that can be answered using this data are:

- "what is the pronunciation of magic?", i.e. searching for the pronunciation of a word
- "What is the plural of magic?", i.e., searching for forms of a word

A screenshot of an example can be seen in Figure 3.

- **Musicbrainz**: Musicbrainz is one of the largest music databases that is open and available online. The data is natively stored in a relational database (i.e. a PostgreSQL database). In previous work, it was already possible to query MusicBrainz, but it relied on the dump provided by the following RML mappings: https://github.com/LinkedBrainz/MusicBrainz-R2RML. We improved the mappings in different ways. The most radical change was related to how some information is organized. For example in the original LinkedBrainz dump every album appears multiple times (every time for each publication year and country where it was released). The same problem appears for songs. In particular the dump didn’t allow to recognize that the different publications of the albums and songs were in fact referring to the same album and song. This means that when asking for albums or songs, many duplicates appeared, a behaviour that is not expected by users and also not explicable for them. We therefore restructured the export to avoid this. Other changes represent a better coverage to external links like to Wikipedia, Wikidata and Social Media. Moreover, the dump is enriched with the links to the covers of albums from http://coverartarchive.org. Finally, we enriched the artists, albums and songs with links to youtube so that users can effectively hear the pieces they are querying. Example requests where Musicbrainz is used to answer some queries are:

- "albums by green day", i.e., searching for albums of a band
- "record label from turin", i.e. searching for record labels in a city,
- "songs blink-182", i.e., searching for songs of a band

A screenshot of an example can be seen in Figure 4.

\(^4\)https://www.wiktionary.org  
\(^5\)https://wiki.dbpedia.org/wiktionary-rdf-extraction
A demo of the current version can be found under:

http://qanswer.eu/qa

Moreover, by clicking on the above examples you will be redirected to the online demo.

4 CONCLUSION

We presented QAnswer a QA system which queries 3 key Semantic Web datasets at the same time, namely, Wikidata with lexemes, LinkedGeoData and Musicbrainz. These datasets contain a huge amount of information like books, films, persons, music, streets, points of interest and many more. The data can be queried using natural language in 8 different languages, namely English, German, French, Italian, Spanish, Portuguese, Arabic and Chinese. In particular two non-European languages, Arabic and Chinese, are included. All together this represents a first prototype of a QA system querying a considerable part of the Semantic Web.

This represents also a step towards new challenges in this area like: correctly selecting which Knowledge Base to query, new scalability scenarios, making use of links between the datasets to deal with redundant information, studying the impact of the dataset quality on the QA performance, and studying the problem of not answering a question.

In future, we would like to make the algorithm and infrastructure publicly available via web APIs so that new RDF datasets can be indexed and accessed using natural language. We believe that this work will further boost the publication of RDF data and therefore the expansion of the Semantic Web.

REFERENCES


Figure 5: Screen shot of QAnswer for "java integer to string"