Overview of CLEF Question Answering Track 2014

Anselmo Peñas[1], Christina Unger[2], and Axel-Cyrille Ngonga Ngomo[3]

[1] NLP&IR group, UNED, Spain
anselmo@lsi.uned.es
[2] CITEC, Bielefeld University, Germany
cunger@cit-ec.uni-bielefeld.de
[3] University of Leipzig, Germany
ngonga@informatik.uni-leipzig.de

Abstract. This paper describes the CLEF QA Track 2014. In the current general scenario for the CLEF QA Track, the starting point is always a natural language question. However, answering some questions may need to query Linked Data (especially if aggregations or logical inferences are required), some questions may need textual inferences and querying free text, and finally, answering some queries may require both sources of information. The track was divided into three tasks: QALD focused on translating natural language questions into SPARQL queries; BioASQ focused on the biomedical domain, and Entrance Exams focused on answering questions to assess machine reading capabilities.

1 Introduction

In the current general scenario for the CLEF QA Track, the starting point is always a natural language question. However, answering some questions may need to query Linked Data (especially if aggregations or logical inferences are required), some questions may need textual inferences and querying free text, and finally, answering some queries may require both sources of information.

As a matter of example related to CLEF eHealth, consider the use case where patients receive medical reports that they don't understand. Given that report, patients have lots of questions. Some of them will need general definitions as one can find in Wikipedia. Some might need more complex answers about the relations between symptoms, treatments, etc. The final goal, then, is to help users understand the given document by answering their questions.

So, given this general scenario, CLEF QA Track will work on two instances of it: one targeted to (bio)medical experts (the BioASQ task) and a second instance targeted to open domains (the QALD and Entrance Exams tasks). In the first one, medical knowledge bases (KBs), ontologies and articles must be taken into account. In the second one, general resources such as Wikipedia articles and DBpedia are considered.

E. Kanoulas et al. (Eds.): CLEF 2014, LNCS 8685, pp. 300–306, 2014.

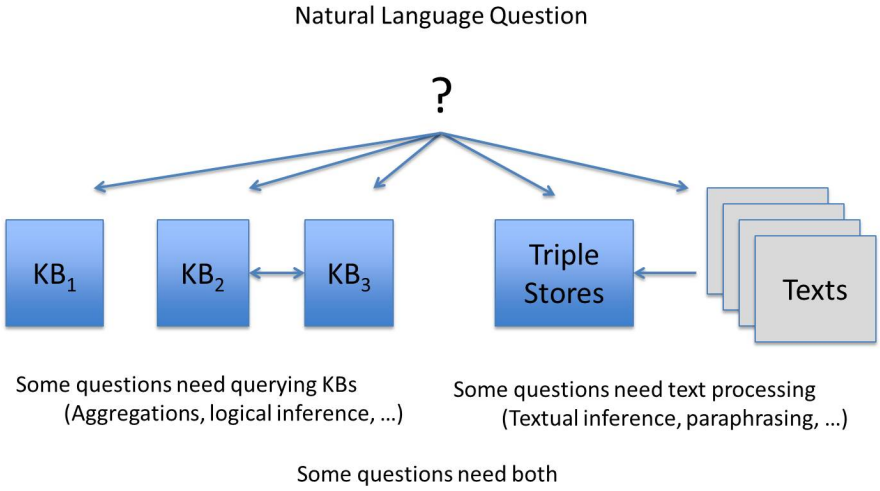

Fig. 1. General scenario of CLEF QA Track

2 Tasks

The CLEF QA Track was divided into the following tasks and subtasks:

2.1 QALD: Question Answering over Linked Data

QALD-4 is the fourth in a series of evaluation campaigns on multilingual question answering over linked data[1] [1], this time with a strong emphasis on interlinked datasets and hybrid approaches using information from both structured and unstructured data.

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 inference techniques.

QALD-4 proposed the following tasks to participants:

2.2 Task QALD-4.1: Multilingual Question Answering

Task QALD-4.1 is the core task of QALD and aims at all question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data. Given the English DBpedia 3.9 dataset and a natural language question or set of keywords in one of seven languages (English, Spanish, German, Italian, French, Dutch, Romanian), the participating systems had to return either the correct answers, or a SPARQL query that retrieves these answers.

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

To get acquainted with the dataset and possible questions, a set of 200 training questions was provided. These questions were compiled from the QALD-3 training and test questions, slightly modified in order to account for changes in the DBpedia dataset. Later, systems were evaluated on 50 different test questions. These questions were mainly devised by the challenge organizers. All training questions were manually annotated with keywords, corresponding SPARQL queries and with answers retrieved from the provided SPARQL endpoint.

2.3 Task QALD-4.2: Biomedical Question Answering over Interlinked Data

Also for the life sciences, linked data plays a bigger and bigger role. Already a tenth of the Linked Open Data cloud consists of biomedical datasets. Especially biomedical data is distributed among a large collection of interconnected datasets, and answers to questions can often only be provided if information from several sources are combined. Task QALD-4.2 therefore focuses on interlinked data.

Given the following three biomedical datasets and a natural language question or set of keywords in English, the participating systems had to return either the correct answers or a SPARQL query that retrieves the answers:

- SIDER[2], describing drugs and their side effects
- Diseasome[3], encompassing description of diseases and genetic disorders
- Drugbank[4], describing FDA-approved active compounds of medication

The training question set comprised 25 questions over those datasets. Later, participating systems were evaluated on 25 similar test questions. Since the focus of the task is on interlinked data, most of the questions require the integration of information from at least two of those datasets.

2.4 Task QALD-4.3: Hybrid Question Answering

A lot of information is still available only in textual form, both on the web and in the form of labels and abstracts in linked data sources. Task QALD-4.3 therefore focuses on the integration of both structured and unstructured information in order to gather answers. Given English DBpedia 3.9, containing both RDF data and free text available in the DBpedia abstracts, and a natural language question or keywords, participating systems had to retrieve the correct answer(s). A set of 25 training questions was provided.

2.5 BioASQ: Biomedical Semantic Indexing and Question Answering

Bio ASQ[5] [2] aims at assessing:

---

[2] http://sideeffects.embl.de
[3] http://wifo5-03.informatik.uni-mannheim.de/diseasome/
[4] http://www.drugbank.ca
[5] http://www.bioasq.org/participate/challenges

- large-scale classification of biomedical documents onto ontology concepts (semantic indexing),
- classification of biomedical questions onto relevant concepts,
- retrieval of relevant document snippets, concepts and knowledge base triples,
- delivery of the retrieved information in a concise and user-understandable form.

The challenge comprised two tasks: (1) a large-scale semantic indexing task and (2) a question answering task.

2.6 Task BioASQ 1: Large-Scale Semantic Indexing

The goal is to classify documents from the PubMed digital library unto concepts of the MeSH2 hierarchy. Here, new PubMed articles that are not yet annotated are collected on a weekly basis.

These articles are used as test sets for the evaluation of the participating systems. As soon as the annotations are available from the PubMed curators, the performance of each system is calculated by using standard information retrieval measures as well as hierarchical ones.

In order to provide an on-line and large-scale scenario, the task was divided into three independent batches. In each batch 5 test sets of biomedical articles were released consecutively. Each of these test sets were released in a weekly basis and the participants had 21 hours to provide their answers.

2.7 Task BioASQ 2: Biomedical Semantic Question Answering

The goal of this task was to provide a large-scale question answering challenge where the systems should be able to cope with all the stages of a question answering task, including the retrieval of relevant concepts and articles, as well as the provision of natural language answers.

It comprised two phases: In phase A, BioASQ released questions in English from benchmark datasets created by a group of biomedical experts. There were four types of questions: yes/no questions, factoid questions, list questions and summary questions. Participants had to respond with relevant concepts (from specific terminologies and ontologies), relevant articles (PubMed and PubMedCentral articles), relevant snippets extracted from the relevant articles and relevant RDF triples (from specific ontologies).

In phase B, the released questions contained the correct answers for the required elements (concepts, articles, snippets and RDF triples) of the first phase. The participants had to answer with exact answers as well as with paragraph-sized summaries in natural language (dubbed ideal answers).

The task was split into five independent batches. The two phases for each batch were run with a time gap of 24 hours. For each phase, the participants had 24 hours to submit their answers. The evaluation in phase B was carried out manually by biomedical experts on the ideal answers provided by the systems.

2.8 Entrance Exams Task

The challenge of Entrance Exams[6] [3] aims at evaluating systems reading capabilities under the same conditions humans are evaluated to enter the University.

Participant systems are asked to read a given document and answer a set of questions. Questions are given in multiple-choice format, with several options from which a single answer must be selected. Systems have to answer questions by referring to "common sense knowledge" that high school students who aim to enter the university are expected to have. The exercise do not intend to restrict question types, and the level of inference required to respond is very high.

Exams were created by the Japanese National Center for University Admissions Tests, and the "Entrance Exams" corpus is provided by NII's Todai Robot Project and NTCIR RITE.

For each examination, one text is given, and five questions on the given text are asked. Each question has four choices. For this year campaign, we reused as development data the 12 examinations from last year's campaign. For testing, we provided 12 new documents where a total of 60 questions and 240 candidate answers had to be validated.

As a novelty this year, data sets for development and testing originally in English were manually translated into Russian, French, Spanish and Italian. They are parallel translations of texts, questions and candidate answers.

In addition to the official data, we collected four more unoffcial translations into French. Despite they preserve original meaning, each translation has its particularities that produce different effects on systems performance: text simplification, lexical variation, different uses of anaphora, overall quality, etc. This data is extremely useful to get insights about systems and their level of inference.

Systems received evaluation scores from two different perspectives: at the question answering level and at the test reading level.

3 Participation

Table 1 shows the distribution of the 30 participants among the exercises proposed by the QA Track.

Table 1. Number of participants in CLEF QA Track 2014

Task	# Registered	Sub-task	# Participants
QALD-4	22	QALD -4.1	6 (English)
		QALD-4,2	3 (English)
		QALD-4.3	(1) (English)
BioASQ	25	BioASQ 1	8 (English)
		BioASQ 2	7 (English)
Entrance Exams	20	Entrance Exams	5 (English) 1 (French)
Total	67	-	30

[6] http://nlp.uned.es/entrance-exams

QALD-4, the fourth edition of the QALD challenge, has attracted a higher number of participants than previous editions, showing that there is a growing interest among researchers to provide end users with an intuitive and easy-to-use access to the huge amount of data present on the Semantic Web. Although one of the aspects of Task QALD-4.1 was multilinguality, all participating systems worked on English data only. This shows that the multilingual scenario is not yet broadly addressed, although it is starting to attract attention. Similarly, research teams start to look at hybrid question answering, although Task QALD-4.3 did not have participating systems yet.

The participation to the second BioASQ challenge signalizes an uptake of the significance of biomedical question answering in the research community, monitoring an increased participation in both Tasks.

With respect to Entrance Exams, 39 systems were presented by the 5 participating teams. This is a similar level of participation than in the previous edition. However, only one team has participated in the two editions. Despite the benchmarks were provided also in Russian, Spanish, Italian and French, all systems run for English and only one for French.

4 Main Conclusions

Readers are referred to the overview papers where a more detailed description of each task is given, together with their evaluation methodology and a general description of the participating systems. Here we draw the main general conclusions derived from 2014 campaign.

Systems performance seems to be improved in all tasks. In the case of BioASQ, the baselines used this year incorporated techniques from last year's winning systems. Best systems outperformed these baselines suggesting an improvement of both large-scale classification systems and question answering.

The results in Entrance Exams were also better than in last edition. At the reading perspective evaluation, we have already three systems (two teams) able to pass at least half of the reading tests.

With respect to earlier challenges of QALD, question answering systems have become more versatile: There is no particular type of questions that systems struggle with, rather most of them can handle all answer types as well as aggregation. The biggest problem, however, remains the matching of natural language questions to correct vocabulary elements.

Something similar was also noticed in Entrance Exams. In this task, there is a big lexical gap between the supporting text, the question and the candidate answer. The level of textual inferences that current systems perform is not enough yet to solve the majority of questions. Therefore, one of the main conclusions of the track is that more resources have to be developed to assess inference in the framework of question answering.

The results show that real question answering is a task far from being solved. However, the CLEF QA Track is providing the benchmarks able to assess real progress in the field along future years.

Acknowledgements. Anselmo Peñas was supported by CHIST-ERA READERS project (MINECO PCIN-2013-002-C02-01). Christina Unger was funded by the EU project PortDial (FP7-296170).

References

1. Unger, C., Forascu, C., Lopez, V., Ngomo, A.-C.N., Cabrio, E., Cimiano, P., Walter, S.: Question Answering over Linked Data (QALD-4). In: CLEF 2014 Working Notes, Sheffield (2014)
2. Balikas, G., Partalas, I., Ngomo, A.-C.N., Krithara, A., Gaussier, E., Paliouras, G.: Results of the BioASQ Track of the Question Answering Lab at CLEF 2014. In: CLEF 2014 Working Notes, Sheffield (2014)
3. Peñas, A., Miyao, Y., Rodrigo, Á., Hovy, E., Kando, N.: Overview of CLEF QA Entrance Exams Task 2014. In: CLEF 2014 Working Notes, Sheffield (2014)