Overview of the CLEF Question Answering Track 2015

Anselmo Peñas[1(✉)], Christina Unger[2], Georgios Paliouras[3], and Ioannis Kakadiaris[4]

[1] NLP&IR Group, UNED, Madrid, Spain
anselmo@lsi.uned.es
[2] CITEC, Bielefeld University, Bielefeld, Germany
cunger@techfak.uni-bielefeld.de
[3] IIT, NCSR Demokritos, Athens, Greece
paliourg@iit.demokritos.gr
[4] CBL, Deptartment of Computer Science, University of Houston, Houston, TX, USA
ioannisk@uh.edu

Abstract. This paper describes the CLEF QA Track 2015. Following the scenario stated last year for the CLEF QA Track, the starting point for accessing information 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. In this edition, the Track was divided into four tasks: (i) *QALD:* focused on translating natural language questions into SPARQL; (ii) *Entrance Exams:* focused on answering questions to assess machine reading capabilities; (iii) BioASQ1 focused on large-scale semantic indexing and (iv) BioASQ2 for Question Answering in the biomedical domain.

1 Introduction

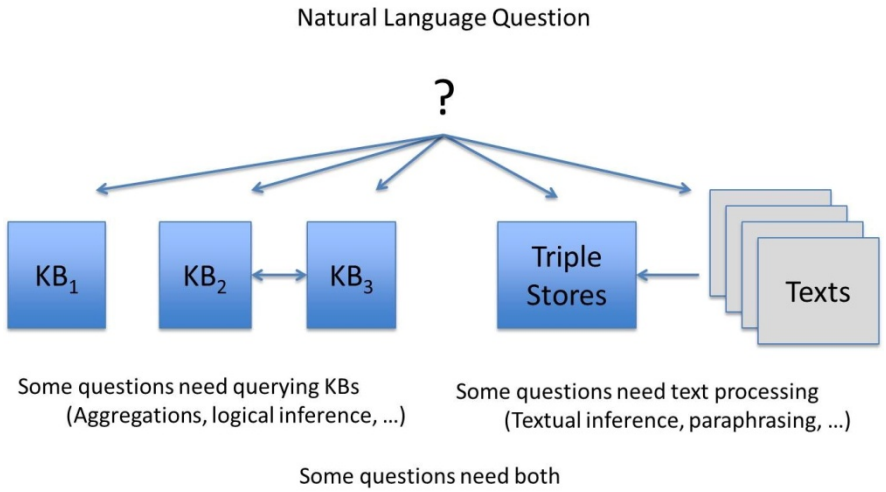
Following last edition of the CLEF QA Track, the starting point is always a Natural Language question that has to be answered against Linked Data, Natural Language or both. Answering some questions may need to query Linked Data (especially if aggregations or logical inferences are required), some questions may need textual inferences and querying free-text, and finally, answering some queries may require both sources of information. The final goal is to help users understand the document by answering their questions.

Thus, given this general scenario, CLEF QA Track will work on two of its instances: one targeted to (bio)medical experts (BioASQ Tasks) and one targeted to Open Domains (QALD and Entrance Exams Tasks). In the former, medical knowledge bases, ontologies and articles must be considered. In the latter, textual documents and general resources such as Wikipedia articles and DBpedia are considered.

J. Mothe et al. (Eds.): CLEF 2015, LNCS 9283, pp. 539–544, 2015.
DOI: 10.1007/978-3-319-24027-5_50

2 Tasks

The CLEF QA Track 2015 was divided into the following tasks:



2.1 QALD: Question Answering Over Linked Data

QALD-5[1] [1] is the fifth in a series of evaluation campaigns on multilingual question answering over linked data, with a strong emphasis on multilingual question answering and hybrid approaches using information from both structured and unstructured data.

The challenge aims at all question answering systems that mediate between a user, expressing his or her information need in natural language, and semantic data. The general task is the following one: Given a natural language question or keywords, retrieve the correct answer(s) from a repository containing both RDF data and free text, in this case the English DBpedia 2014 dataset with free text abstracts.

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-5 provides a benchmark comprising two kinds of questions:

1. *Multilingual questions* are provided in seven different languages (English, German, Spanish, Italian, French, Dutch, and Romanian) and can be answered using the provided RDF data. They are annotated with corresponding SPARQL queries and answers retrieved from the provided SPARQL endpoint.
2. *Hybrid questions* are provided in English and can be answered only by integrating structured data (RDF) and unstructured data (free text available in the DBpedia abstracts). The questions thus all require information from both RDF and free text. They are annotated with pseudo-queries that show which part is contained in the RDF data and which part must be retrieved from the free text abstracts.

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

To get acquainted with the dataset and possible questions, a set of training questions was provided, comprising of 300 multilingual questions and 40 hybrid questions. Later, systems were evaluated on 60 different test questions, comprising of 50 multilingual ones and 10 hybrid ones. Overall, of the 350 training questions, 59 questions require aggregation (e.g., counting, filtering, ordering) and 102 questions require namespaces other than from the DBpedia ontology (21 of which use the YAGO namespace, 2 require FOAF, and all others rely on the DBpedia property namespace). Similarly, of the 60 test questions, 15 questions require aggregation and 12 cannot be answered with the DBpedia ontology only (3 of which use the YAGO namespace, all others rely on the DBpedia property namespace). As an additional challenge, 14 training and 1 test question are out of scope, i.e. they cannot be answered using the dataset.

The results submitted by participating systems were automatically compared to the gold standard results and evaluated using precision and recall metrics.

2.2 Entrance Exams Task

The challenge of Entrance Exams[2] [3] aims at evaluating systems reading capabilities under the same conditions humans are evaluated to enter the University.

Participant systems are asked to ingest a given document and answer a set of questions. Questions are provided in multiple-choice format, with several options from which a single answer must be selected. Systems must answer questions by referring to "common sense knowledge" that high school students who aim to enter the university are expected to have. The exercise does 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 some (between 4 and 8) questions on the given text are asked. Each question has four choices. For this year's campaign, we reused as development data the 24 examinations from the last two years' campaigns. For testing, we provided 19 new documents where 89 questions and 356 candidate answers had to be validated.

Data sets for testing originally in English were manually translated into Russian, French, Spanish, German and Italian. They are parallel translations of texts, questions and candidate answers that offer a benchmark for evaluating systems in different languages.

In addition to the official data, we collected unofficial translations for each language. Although they preserve the original meaning, each translation has its particularities that produce different effects on systems performance: text simplification, lexical variation, different uses of anaphora, and overall quality. This data is useful to obtain insights about systems and their level of inference.

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

Systems were evaluated from two different perspectives: question answering, where the relevant number is the overall number of questions being answered correctly; and reading comprehension, where results were grouped by test (document plus questionnaire) and we measure if machines were able to pass each test.

2.3 BioASQ: Biomedical Semantic Indexing and Question Answering

BioASQ [2] aims at assessing:

- 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: (i) a large-scale semantic indexing task and (ii) a question answering task.

2.3.1 Task BioASQ 1: Large-Scale Semantic Indexing

The goal was to classify documents from the MEDLINE digital library unto concepts of the MeSH2015 hierarchy. New MEDLINE articles not yet annotated are collected weekly. These articles are used as test sets for evaluating the participating systems. As soon as the annotations are available from the MEDLINE curators, the performance of each system is computed using standard information retrieval measures and hierarchical ones.

To provide an on-line and large-scale scenario, the task was divided into three independent batches. In each batch five 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.

2.3.2 Task BioASQ 2: Biomedical Semantic QA

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, and the provision of natural-language answers. This process involves a variety of technologies and methods, ranging from information retrieval from text and knowledge bases to information extraction and summarization.

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 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 partici-

pants had to answer with exact answers and 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 during two consecutive days. 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. Each answer was evaluated along four dimensions: readability, recall, precision and repetition, using a scale from 1 to 5.

3 Participation

Table 1 shows the distribution of the participating teams among the exercises proposed by the CLEF QA Track 2015.

Table 1. Number of participants in CLEF QA Track 2015

Task	# Registered	Sub-task	# Participants
QALD	26	QALD	7 (English)
Entrance Exams	28	Entrance Exams	5 (English) 1 (French)
BioASQ	19	BioASQ 1	18 (English)
	23	BioASQ 2	12 (English)
Total	96	-	43

QALD-5, the fifth edition of the QALD challenge, has attracted seven participants. Two participants submitted results only for the multilingual questions, two participants submitted results only for the hybrid questions, and three participants submitted results for both kinds of questions. Although the overall number of participants is one less than in last year's challenge, the number of participating hybrid question answering systems increased from one to five, which is an important step towards advancement in the field. However, all systems still processed only the English questions, not yet addressing multilingualism.

Continuing the trend that appeared in the second edition of BioASQ, the number of participating teams increased further in the third BioASQ challenge. Particularly encouraging is the increase of participation in the hard QA task (BioASQ2), where by now a corpus of over 1,300 questions has been formed, including associated material (documents, snippets, concepts, triples) and correct answers produced by biomedical experts.

Concerning Entrance Exams, 18 systems were presented by the five participating teams. This represents a lower amount of runs than in the previous edition despite the

fact that the number of participants was the same. Moreover, only one team has participated in the three editions of the task, while there has been two teams taking part also in the last two editions. Although the benchmarks were provided in Russian, Spanish, Italian, German and French, all systems run for English and only one for French.

4 Conclusions

Top systems performance appears to have improved in all tasks.

The average result in Entrance Exams was similar to the last edition, and only the best team from the last edition improved its score in English, obtaining similar results in French. From the reading perspective evaluation we had two systems (from the same team) able to pass at least half of the reading tests.

Concerning earlier challenges of QALD, question answering systems have made an important step towards hybrid question answering, querying not only RDF data but also including information in plain text sources. One of the biggest challenges remains the matching of natural language questions to correct vocabulary elements.

Something similar was also observed 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 adequate yet to solve the majority of questions.

In BioASQ the best systems increased their performance over last year and outperformed clearly all baselines, e.g. the difference between the best system in the semantic indexing task (by University of Fudan, China) and the MTI baseline was 5-6 percentage points throughout the challenge.

Acknowledgements. Anselmo Peñas was supported by CHIST-ERA READERS project (MINECO PCIN-2013-002-C02-01) and the Voxpopuli project (TIN2013-47090-C3-1-P). BioASQ started as an FP7 project, supported by the EC (contract number 318652). The third edition of BioASQ is supported by a conference grant from the NIH/NLM (number 1R13LM012214-01) and sponsored by Viseo.

References

1. Unger, C., Forascu, C., Lopez, L., Ngonga Ngomo, A., Cabrio, E., Cimiano, P., Walter, S.: Question answering over linked data (QALD-5). In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073 (2015)
2. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An Overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. BMC Bioinformatics **16**, 138 (2015)
3. Peñas, A., Miyao, Y., Rodrigo, Á., Hovy, E., Kando, N.: Overview of CLEF QA entrance exams task 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds) CLEF 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073 (2015)