QALM: a Benchmark for Question Answering over Linked Merchant Websites Data

Amine Hallili[1], Elena Cabrio[2,3], and Catherine Faron Zucker[1]

[1] Univ. Nice Sophia Antipolis, CNRS, I3S, UMR 7271, Sophia Antipolis, France
amine.hallili@inria.fr; faron@unice.fr
[2] INRIA Sophia Antipolis Méditerranée, Sophia Antipolis, France
elena.cabrio@inria.fr
[3] EURECOM, Sophia Antipolis, France

Abstract. This paper presents a benchmark for training and evaluating Question Answering Systems aiming at mediating between a user, expressing his or her information needs in natural language, and semantic data in the commercial domain of the mobile phones industry. We first describe the RDF dataset we extracted through the APIs of merchant websites, and the schemas on which it relies. We then present the methodology we applied to create a set of natural language questions expressing possible user needs in the above mentioned domain. Such question set has then been further annotated both with the corresponding SPARQL queries, and with the correct answers retrieved from the dataset.

1 Introduction

The evolution of the e-commerce domain, especially the Business To Client (B2C), has encouraged the implementation and the use of dedicated applications (e.g. Question Answering Systems) trying to provide end-users with a better experience. At the same time, the user's needs are getting more and more complex and specific, especially when it comes to commercial products whose questions concern more often their technical aspects (e.g. price, color, seller, etc.). Several systems are proposing solutions to answer to these needs, but many challenges have not been overcome yet, leaving room for improvement. For instance, federating several commercial knowledge bases in one knowledge base has not been accomplished yet. Also, understanding and interpreting complex natural language questions also known as n-relation questions seems to be one of the ambitious topics that systems are currently trying to figure out.

In this paper we present a benchmark for training and evaluating Question Answering (QA) Systems aiming at mediating between a user, expressing his or her information need in natural language, and semantic data in the commercial domain of the mobile phone industry. We first describe the RDF dataset that we have extracted through the APIs of merchant sites, and the schemas on which it relies. We then present the methodology we applied to create a set of natural language questions expressing possible user needs in the above mentioned domain.

Such question set has then be further annotated both with the corresponding SPARQL queries, and with the correct answers retrieved from the dataset.

2 A Merchant Sites Dataset for the Mobile Phones Industry

This section describes the QALM (Question Answering over Linked Merchant websites) ontology (Section 2.1), and the RDF dataset (Section 2.2) we built by extracting a sample of data from a set of commercial websites.

2.1 QALM Ontology

The QALM RDF dataset relies on two ontologies: the Merchant Site Ontology (MSO) and the Phone Ontology (PO). Together they build up the QALM Ontology.[4] MSO models general concepts of merchant websites, and it is aligned to the commercial part of the *Schema.org* ontology. MSO is composed of 5 classes: `mso:Product`, `mso:Seller`, `mso:Organization`, `mso:Store`, `mso:ParcelDelivery`, and of 29 properties (e.g. `mso:price`, `mso:url`, `mso:location`, `mso:seller`) declared as subclasses and subproperties of Schema.org classes and properties. We added to them multilingual labels (both in English and in French), that can be exploited by QA systems in particular for property identification in the question interpretation step. We relied on WordNet synonyms [2] to extract as much labels as possible. For example, the property `mso:price` has the following English labels: "price", "cost", "value", "tariff", "amount", and the following French labels: "prix", "coût", "coûter", "valoir", "tarif", "s'élever".

PO is a domain ontology modeling concepts specific to the phone industry. It is composed of 7 classes (e.g. `po:Phone`, `po:Accessory`) which are declared as subclasses of `mso:Product`, and of 35 properties (e.g. `po:handsetType`, `po:operatingSystem`, `po:phoneStyle`).

2.2 QALM RDF Dataset

Our final goal is to build a unified RDF dataset integrating commercial product descriptions from various e-commerce websites. In order to achieve this goal, we analyze the web services of the e-commerce websites regardless of their type (either SOAP or REST). To feed our dataset, we create a mapping between the remote calls to the web services and the ontology properties, that we store in a separate file for reuse. In particular, we built the QALM RDF dataset by extracting data from eBay[5] and BestBuy[6] commercial websites through BestBuy Web service and eBay API. The extracted raw data is transformed into RDF triples by applying the above described mapping between the QALM ontology

[4] Available at `www.i3s.unice.fr/qalm/ontology`

[5] `http://www.ebay.com/`

[6] `http://www.bestbuy.com/`

and the API/web service. For instance, the method `getPrice()` in the eBay API is mapped to the property `mso:price` in the QALM ontology. Currently, the QALM dataset comprises 500000 product descriptions and up to 15 millions triples extracted from eBay and BestBuy.[7]

3 QALM Question Set

In order to train and to evaluate a QA system mediating between a user and semantic data in the QALM dataset, a set of questions representing users requests in the phone industry domain is required. Up to our knowledge, the only available standard sets of questions to evaluate QA systems over linked data are the ones released by the organizers of the QALD (Question Answering over Linked Data) challenges.[8] However such questions are over the English DBpedia dataset[9], and therefore cover several topics. For this reason, we created a set of natural language questions for the specific commercial domain of the phone industry, following the guidelines described by the QALD organizers for the creation of their question sets [1]. More specifically, these questions were created by 12 external people (students and researchers in other groups) with no background in question answering, in order to avoid a bias towards a particular approach. To accomplish the task of question creation, each person was given *i)* the list of the product types present in the QALM dataset (mainly composed of IT products as phones and accessories); *ii)* the list of the properties of the QALM ontology presented as product features in which they could be interested in; and they were asked to produce *i)* both 1-relation and 2-relation questions, and *ii)* at least 5 questions each. The questions were designed to present potential user questions and to include a wide range of challenges such as lexical ambiguities and complex syntactical structures. Such questions were then annotated with the corresponding SPARQL queries, and the correct answers retrieved from the dataset, in order to consider them as a reliable goldstandard for our benchmark.

The final question set comprises 70 questions; it is divided into a training set[10] and a test set of respectively 40 and 30 questions. Annotations are provided in XML format, and according to QALD guidelines, the following attributes are specified for each question along with its ID: *aggregation* (indicates whether any operation beyond triple pattern matching is required to answer the question, e.g., counting, filtering, ordering), *answertype* (gives the answer type: resource, string, boolean, double, date). We also added the attribute *relations*, to indicate whether the question is connected to its answer through one or more properties of the ontology (values: 1, n). Finally, for each question the corresponding SPARQL query is provided, as well as the answers this query returns. Examples 1 and 2 show some questions from the collected question set, connected to their answers through 1 property or more than 1 property of the ontology, respectively. In

[7] Available at `www.i3s.unice.fr/QALM/qalm.rdf`

[8] `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/`

[9] `http://dbpedia.org`

[10] Available at `www.i3s.unice.fr/QALM/training_questions.xml`

particular, questions 14 and 50 from Example 2 require also to carry out some reasoning on the results, in order to rank them and to produce the correct answer.

Example 1. 1-relation questions.
id=36. *Give me the manufacturers who supply on-ear headphones.*
id=52. *What colors are available for the Samsung Galaxy 5 ?*
id=61. *Which products of Alcatel are available online?*

Example 2. n-relations questions.
id=14. *Which cell phone case (any manufacturer) has the most ratings?*
id=50. *What is the highest camera resolution of phones manufactured by Motorola?*
id=58. *I would like to know in which stores I can buy Apple phones.*

4 Conclusions and Ongoing Work

This paper presented a benchmark to train and test QA systems, composed of *i)* the QALM ontologies; *ii)* the QALM RDF dataset of product descriptions extracted from eBay and BestBuy; and *iii)* the QALM Question Set, containing 70 natural language questions in the commercial domain of phones and accessories.

As for future work, we will consider aligning the QALM ontology to the *GoodRelations* ontology to fully cover the commercial domain, and to benefit from the semantics captured in this ontology. We also consider improving the QALM RDF dataset by *i)* extracting RDF data from additional commercial websites that provide web services or APIs; and *ii)* directly extracting RDF data in the Schema.org ontology from commercial websites whose pages are automatically generated with Schema.org markup (e.g. Magento, OSCommerce, Genesis2.0, Prestashop), to extend the number of addressed commercial websites.

In parallel, we are currently developing the SynchroBot QA system [3], an ontology-based chatbot for the e-commerce domain. We will evaluate it by using the proposed QALM benchmark.

Acknowledgements

We thank Amazon, eBay and BestBuy for contributing to this work by sharing with us public data about their commercial products. The work of E. Cabrio was funded by the French Government through the ANR-11-LABX-0031-01 program.

References

1. Cimiano, P., Lopez, V., Unger, C., Cabrio, E., Ngomo, A.C.N., Walter, S.: Multilingual question answering over linked data (qald-3): Lab overview. In: CLEF. pp. 321–332 (2013)
2. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
3. Hallili, A.: Toward an ontology-based chatbot endowed with natural language processing and generation. In: Proc. of ESSLLI 2014 - Student Session, Poster paper (2014)