PowerAqua: An Ontology Question Answering System for the Semantic Web

Vanessa Lopez and Enrico Motta

Knowledge Media Institute, The Open University.
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom.
{v.lopez, e.motta}@open.ac.uk

Abstract As semantic markup becomes ubiquitous, it will become important to be able to ask queries and obtain answers, using natural language (NL) expressions, rather than the keyword-based retrieval mechanisms used by the current search engines. The Semantic Web (SW) opens the way to novel question answering (QA) systems, which can exploit the availability of distributed semantic markup to provide precise, formally derive answers to questions. On the other hand, the distributed, heterogeneous, large-scale nature of the availability of semantic information raises many difficult challenges. Here, we describe the design of a QA system, PowerAqua, designed to exploit semantic mark-up on the web to provide answers to questions posed in NL. PowerAqua does not assume that the user has any prior information about semantic resources. The system takes an input a NL query, translates it into a set of logical queries, which are then answered by consulting and aggregating information derived from multiple heterogeneous semantic sources..

1 Introduction

The semantic web vision offers a compelling vision in which ontologies play a crucial role on the SW to provide the conceptual infrastructure support for semantic interoperability, addressing data heterogeneity, and opening up opportunities for automated information processing. However, because of the SW's distributes nature, data will inevitable be associated with different ontologies and therefore ontologies themselves will introduce heterogeneity.

Our goal is to design and develop a QA system, able to exploit the availability of distributed, ontology-based semantic markup on the web to provide answers to questions posed in NL. A user must be able to pose queries in NL, without being aware of which information sources exist, the details associated with interacting with each source, or the particular vocabulary used by the sources. We call this system PowerAqua. PowerAqua follow from an earlier system, AquaLog [1], and addresses some of its limitations, as discussed in next section.

2 The AquaLog question answering system

AquaLog is a fully implemented ontology-driven QA system, which takes an ontology and a NL query as an input and returns answers drawn from semantic markup knowledge base compliant with the input ontology. In contrast with much existing work on ontology-driven QA, which tends to focus on the use of ontologies to support query expansion in information retrieval, AquaLog exploits the availability of semantic statements to provide precise answers to complex queries requiring situation-specific knowledge, where multiple pieces of information need to be inferred and combined at run time, rather than retrieving a pre-written paragraph of text. In particular, AquaLog uses generic lexical resources, such as WordNet, as well as reasoning about the ontology structure to make sense of the terms and relations expressed in a query in terms of the concepts familiar to the user, even when they appear not to have any match in the KB/ontology or there are ambiguities.

An important feature of AquaLog is its portability with respect to ontologies. In other words, the time required to configure AquaLog for a particular ontology is negligible. The reason for this

is that the architecture of the system and the reasoning methods are completely domain-independent, relying on an understanding of general-purpose knowledge representation languages, such as OWL, and the use of generic lexical resources, such as WordNet. Finally, AquaLog is interactive, it asks the user for help when is unable to disambiguate terms or relations. AquaLog also includes a learning mechanism, which ensures that, for a given ontology and community of users, its performance improves over time, as the users can easily give feedback and allow AquaLog to learn novel associations between the relations used by users, which are expressed in natural language, and the internal structure of the ontology.

AquaLog present an elegant solution in which different technologies are combined together. AquaLog uses a sequential model, in which NL input is first translated into a set of intermediate representations – these are called *linguistic triples*. The GATE system [2] is used to this purpose. Then these linguistic triples are further processed and interpreted using the available lexical resources and the structure and vocabulary of the ontology to create ontology-compliant triples. AquaLog's intermediate representation is triple based, mainly because representation formalisms for the SW also subscribe to a binary relational model representation.

However, the key limitation in AquaLog for a system targeted at the SW is that AquaLog only makes uses of one ontology at a time. This works well in many scenarios, for instance in company intranets where a shared organizational ontology is used to describe resources. However, if we consider the SW in the large there is a need to compose information from multiple information sources that are autonomously created and maintained. As already pointed out, the SW is heterogeneous in nature and it is not possible to know in advance which semantic data will be relevant to a particular ury The system must be able to automatically locate and aggregate information from the relevant sources, without any pre-formulated assumption about the ontological structure of the relevant information.

3 PowerAqua challenges: an approach to the problem

In this section we shortly examine the specific issues which need to be tackled in order to develop PowerAqua that are not already tackled by AquaLog. For instance, we will not be looking at the problem of translating fro NL into triples: the AquaLog solution, can be simply reused for PowerAqua

Resource discovery and information focusing is not a problem in AquaLog. Given an ontology on the web, which is identified by a URL, it is reasonably simple to retrieve all semantic resources which are based on the ontology in question. In contrast with AquaLog, PowerAqua has to automatically identify semantic markup, which can potentially be relevant to the input query.

Query-driven semantic mapping: User terminology is translated into triples containing heterogeneous terminology across distributed ontologies. In any strategy that focuses on information content, the most critical problem is that of different vocabularies used to describe similar information across domains [8]. Preferably, mappings between ontology and query elements must be determined by analyzing its semantics or meaning codified in the structure of the ontology/KB (concepts, not labels), and not by only a syntactic analysis.

Information correlation: Queries posed by end-users may need to be answered not by a single knowledge source but consulting multiple sources, and therefore, combining the relevant *information* from different repositories. To perform correlation between data from different ontologies we must be able to identify common objects (instances) retrieved from different ontologies, e.g. for intersection, we show only the common objects; and for union, we eliminate the duplicate objects.

1. AquaLog: An Ontology-portable Question Answering System for the Semantic Web. Lopez V., Pasin M. and Motta E. *In proceedings of the 2nd European Semantic Web Conference* (2005)
2. Tablan, V., Maynard, D., Bontcheva, K.: *GATE - A Concise User Guide.* University of Sheffield, UK. http://gate.ac.uk/