Question Answering on the Real Semantic Web

Vanessa López[1], Miriam Fernández[2], Enrico Motta[1], Marta Sabou[1,], Victoria Uren[1]

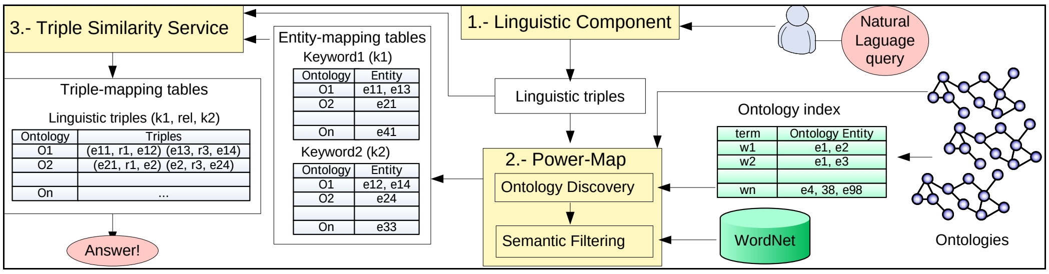
[1] Knowledge Media Institute, The Open University. United Kingdom.
{v.lopez, e.motta, r.m.sabou, v.s.uren}@open.ac.uk
[2] Politécnica Superior, Universidad Autónoma de Madrid. Spain.{miriam.fernandez}@uam.es

Abstract. Restriction to a predefined set of ontologies, and consequently limitation to specific domain environments is a pervading drawback in Semantic Search technologies. In this work we present PowerAqua [2], a multi-ontology-based Question Answering (QA) platform that exploits multiple distributed ontologies and knowledge bases to answer queries in multi-domain environments. The system interprets the user's Natural Language (NL) query using the available semantic information, and translates the user terminology into the ontology terminology (triples), retrieving accurate semantic entity values as response to the user's request.

1 Introduction

The goal of Question Answering (QA) systems is to allow users to ask questions, using their own terminology, and receive a concise answer. A new trend on QA is *ontology based QA* where the power of ontologies as a model of knowledge is and its semantic information is directly exploited for the query analysis and translation (Aqualog [3], ORAKEL [5], GINO [1]). In contrast with traditional NLIDB systems, semantic QA needs very little customization being almost ontology independent. However, they are limited to the knowledge encoded on one, or a set of a priori defined ontologies in the same domain (semantic intranets). As consequence, they are still far away from their successful use as full NL open interfaces to the SW. For instance, neither to involve the user to provide domain specific grammars or vocabulary (ORAKEL), or the use of guided user interfaces (GINO) which generates a dynamic grammar rule for every ontology element, or asking the user every time ambiguity arises (AquaLog) are feasible solutions in the large SW scenario, where portability is not longer enough and openness is required.

In this work we present PowerAqua [2], a platform that evolves from the earlier AquaLog system, designed to take advantage of the vast amount of heterogeneous semantic data offered by the SW in order to interpret a query, without making any assumptions of the relevant ontologies to a particular query a priori.



Triple-mapping tables

Linguistic triples (k1, rel, k2)

Ontology	Triples
O1	(e11, r1, e12) (e13, r3, e14)
O2	(e21, r1, e2) (e2, r3, e24)
On	...

Entity-mapping tables

Keyword1 (k1)

Ontology	Entity
O1	e11, e13
O2	e21
On	e41

Keyword2 (k2)

Ontology	Entity
O1	e12, e14
O2	e24
On	e33

Ontology index

term	Ontology Entity
w1	e1, e2
w2	e1, e3
wn	e4, 38, e98

Figure1: Power Aqua Flow

2 Retrieving answers in open-domain multi-ontology environments

The overall QA processing is illustrated in Figure1. In a first step, the linguistic component [3] analyzes the NL query and translates it into its linguistic triple form. E.g a query “What are the cities of Spain?” has the linguistic triple (<*what-is, cities, Spain*>)

In a second step the Ontology Discovery sub module of PowerMap [2], identifies the set of ontologies likely to provide the information requested by the user. To do so, it searches for approximate syntactic matches within the ontology indexes, using not just the linguistic triple terms, but also lexically related words obtained from WordNet and from the ontologies, used as background knowledge sources. E.g the term *cities* match with the concepts *city, metropolis, etc*. Once the set of possible syntactic mappings have been identified, the PowerMap Semantic Filtering sub module checks its validity using a WordNet-based filtering methodology. This methodology is based on a semantic similarity measure between the set of synsets t obtained for the query term T and the set of synsets c obtained for the matched concept C. To do so, the measure considers the path distance (depth) and the shared information extracted using the WN IS_A hierarchy [4]

$$\text{Semantic Similarity } (t, c) = t \sim c = (2 \text{ x depth } (\text{C.P.I } (t, c))) \;/\; (\text{depth } (t, c) + 2 \text{ x depth}(\text{C.P.I } (t, c)))$$[1]

To elicit the sense of a mapped concept C with respect to a query term T, we intersect (1) $S_{C,T,}$ the set of synsets of C semantically similar to T, with (2) $S^H{}_{C,}$ the set of synsets of C that are semantically similar to any synset of its ontology ancestors. Obviously, if this intersection is empty it means that the sense of the concept in the ontology (2) is different from the sense defined by the query term T (1), and therefore that mapping should be discarded. E.g., the ontological concept *Game* obtained as a synonym of the query term *Sport* should be discarded, as its ontological parent is *Hunted animals.*

$$(1)\ S_{C,T} = \{c \in S_C \mid \exists t \in S_T \text{ such that } t \sim c\} \quad (2) S^H{}_C = \{c \in S_C \mid \forall\ R\ ((R > C) \rightarrow (\exists r \in S_R\ (c \sim r)))\}$$

After this process, PowerMap generates a set of Entity Mapping Tables where each table links a query term with a set of concepts mapped in the different domain ontologies.

In a third step the Triple Similarity Service module takes as input the previously retrieved Entity Mapping Tables and the initial Linguistic triples and extract, by analyzing the ontology relationships, a small set of ontologies that jointly covers the user query. The output of this module is a set of Triple Mapping Tables where each table relates a linguistic triple with all the equivalent ontological triples. Using these triples the information of the Knowledge Bases is analysed to generate the final answer.

To conclude with, PowerAqua balances the heterogeneous and large scale semantic data with giving results in real time across ontologies, to translate user terminology into distributed semantically sound terminology, so that the concepts which are shared by assertions taken from different ontologies have the same sense. The goal is to handle queries which require to be answered not only by consulting a single knowledge source but combining multiple sources, and even domains.

1. Bernstein, A., Kaufmann, E. (2006) GINO - A Guided Input Natural Language Ontology Editor. *In Proc of the International Semantic Web Conference: 144-157*
2. Lopez, V., Sabou, M. and Motta, E. (2006) PowerMap: Mapping the Real Semantic Web on the Fly, International Semantic Web Conference., Georgia, Atlanta.
3. Lopez, V., Motta, E., Uren, V. and Pasin, M. (2007) AquaLog: An ontology-driven Question Answering System for organizational Semantic intranets, *Journal of Web Semantics*, 5, 2, pp. 72-105, Elsevier.
4. Wu Z., and Palmer, M. (2004) Verb Semantics and Lexical Selection. *In Proc of the 32nd Annual Meeting of the Associations for Computational Linguistics.*
5. Cimiano, P., Haase, P., Heizmann, J. (2007) Porting Natural Language Interfaces between Domains -- An Experimental User Study with the ORAKEL System. In *Proc of the Int Conf on Intelligent User Interfaces*.

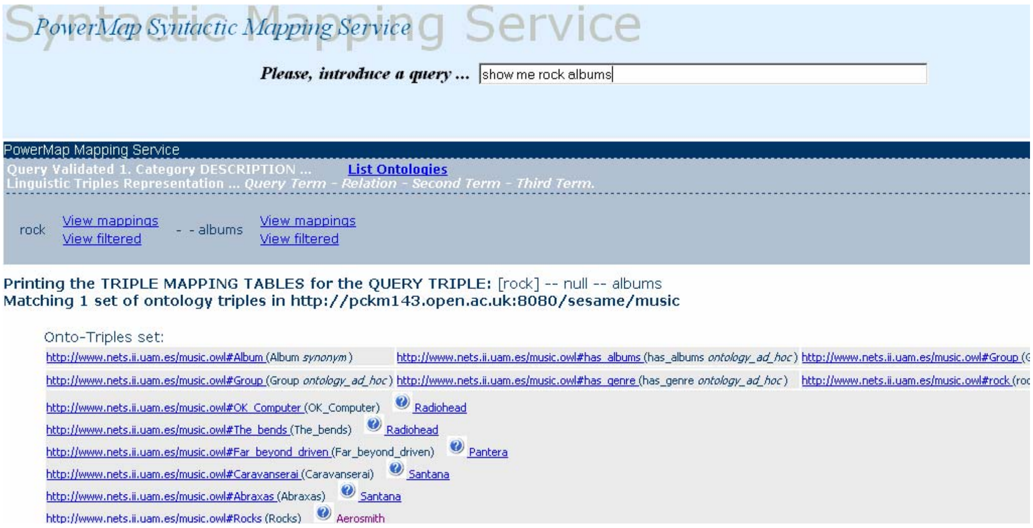
[1] *Uppercase letters denote words and lowercase letters WN synsets. We write depth (CPI (t,c)) to define the depth between the lowest common parent of t and c and the common root in WN*

3 Power Aqua in action: illustrative Example

Consider the simple query "What are the cities of Spain?" (<*what-is, cities, Spain*>), where both the *sweto*[2] and the *agrovoc*[3] ontologies are selected as relevant by PowerMap, as they have candidate matches for both arguments ("Spain" and "cities"), so they potentially cover the linguistic triple. Then, the similarity services are called to try to make sense of each linguistic triple, and its candidate element mappings, by analyzing the ontology taxonomy and relationships. Essentially, the triple similarity services are responsible of mapping each linguistic triple into one or more ontology triples within each relevant ontology, each ontology triple being a complete alternative translation, of the given linguistic triple, if possible.

In this example, both ontologies represent the linguistic relation "cities" as the ontology class "city". Therefore, the RSS generates ontology triples that link the first argument, the class "city", and the second argument together. In the case of *agrovoc*, the second argument is the instance "Spain", therefore, the problem becomes one of finding *ad-hoc* relations which links the *query* term "city" with the instance "Spain" (superclasses and subclasses are considered due to the inheritance of relations through the subsumption hierarchy). In the case that no ontological relations were found, it looks for indirect relations through mediating concepts between both arguments. Then, the answer will be the instances of "city" that has a relationship with "Spain". For the *sweto* ontology, "Spain" corresponds to a literal, therefore it looks for all the instances of "city" that has "Spain" as the value of one of its attributes. The resultant list of instances from both ontologies should be merged to generate a more complete answer.

Furthermore, in order to generate an ontological interpretation of a query, and therefore an answer, nominal compounds terms, like "rock albums", which are translated into two ontology terms, are represented by a new ontology triple that links them together. For instance the resultant ontology triples for "show me rock albums" are: <album, has-albums, group> <group, has genre, rock> in an ontology about music, as seen in Figure 2



Please, introduce a query ... show me rock albums

PowerMap Mapping Service

Query Validated 1. Category DESCRIPTION ... List Ontologies

Linguistic Triples Representation ... Query Term - Relation - Second Term - Third Term.

rock View mappings View filtered - - albums View mappings View filtered

Printing the TRIPLE MAPPING TABLES for the QUERY TRIPLE: [rock] -- null -- albums

Matching 1 set of ontology triples in http://pckm143.open.ac.uk:8080/sesame/music

Onto-Triples set:

http://www.nets.ii.uam.es/music.owl#Album (Album *synonym*) http://www.nets.ii.uam.es/music.owl#has_albums (has_albums *ontology_ad_hoc*) http://www.nets.ii.uam.es/music.owl#Group (G

http://www.nets.ii.uam.es/music.owl#Group (Group *ontology_ad_hoc*) http://www.nets.ii.uam.es/music.owl#has_genre (has_genre *ontology_ad_hoc*) http://www.nets.ii.uam.es/music.owl#rock (roc

http://www.nets.ii.uam.es/music.owl#OK_Computer (OK_Computer) Radiohead

http://www.nets.ii.uam.es/music.owl#The_bends (The_bends) Radiohead

http://www.nets.ii.uam.es/music.owl#Far_beyond_driven (Far_beyond_driven) Pantera

http://www.nets.ii.uam.es/music.owl#Caravanserai (Caravanserai) Santana

http://www.nets.ii.uam.es/music.owl#Abraxas (Abraxas) Santana

http://www.nets.ii.uam.es/music.owl#Rocks (Rocks) Aerosmith

Figure 2. Screenshot of the example "Show me rock albums"

[2] http://lsdis.cs.uga.edu/projects/semdis/sweto/

[3] http://www.few.vu.nl/wrvhage/oaei2006/