A Multilingual SPARQL-Based Retrieval Interface for Cultural Heritage Objects

Mariana Damova[1] and Dana Dannélls[2] and Ramona Enache[3]

[1] Mozaika, Bulgaria
mariana.damova@mozajka.co
[2] Språkbanken, University of Gothenburg
dana.dannells@svenska.gu.se
[3] Department of Computer Science and Engineering, University of Gothenburg
ramona.enache@cse.gu.se

1 Introduction

In this paper we present a multilingual SPARQL-based [1] retrieval interface for querying cultural heritage data in natural language (NL). The presented system offers an elegant grammar-based approach which is based on Grammatical Framework (GF) [2], a grammar formalism supporting multilingual applications. Using GF, we are able to present a cross-language SPARQL grammar covering 15 languages and a cross-language retrieval interface that uses this grammar for interacting with the Semantic Web[4]. To our knowledge, this is the first implementation of SPARQL generation and parsing via GF that is published as a knowledge representation infrastructure-based prototype.

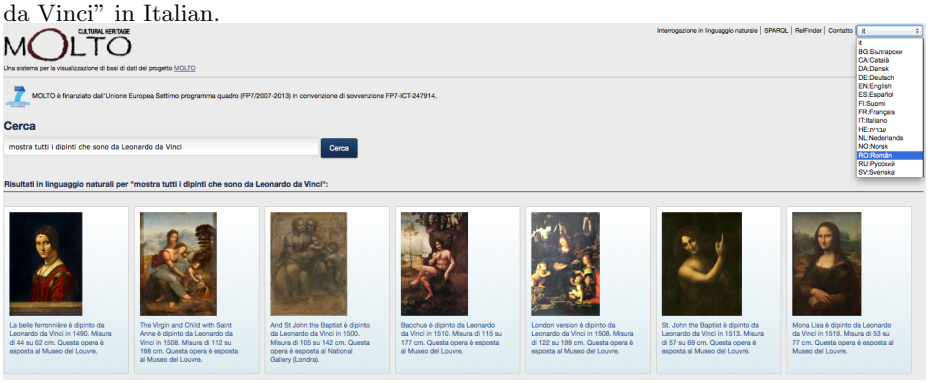
Querying the Semantic Web in natural language, more specifically, using English to formulate SPARQL queries with the help of controlled natural language (CNL) syntax has been developed before [3,4]. Such approaches, based on verbalization methods are adequate for English, but in a multilingual setting where major challenges such as lexical and structural gaps become prominent [5], grammar-based approaches are preferable. The work presented here complements the method proposed by Lopez et al. [6] in that it faces the challenges in realizing NL in real world systems, not only in English, but also in multiple languages.

2 An Interface for Multilingual Queries

Our system follows the approach of the Museum Reason-able View (MRV) of Linked Open Data (LOD) [7]. It provides a unified access to the cultural heritage sources including LOD from DBpedia,[5] among other sources.

[4] A demo is available from the following page: http://museum.ontotext.com/
[5] http://dbpedia.org

Fig. 1. Demo of the natural language query "show all paintings that are by Leonardo da Vinci" in Italian.



The query grammar for this data covers the nine central classes: title, painter, type, colour, size, year, material, museum, place and the major properties describing the relationship between them: hasCreationDate, fromTimePeriodValue, toTimePeriodValue, hasMaterial, hasTitle, hasDimension, hasCurrentLocation, hasColour. The set of SPARQL queries we cover include the famous five WH questions: *who, where, when, how, what*. Table 1 shows some NL queries and their mappings to query variables in SPARQL.

NL Query	SPARQL
Where is Mona Lisa located?	:hasCurrentLocation ?location
What are the colours of Mona Lisa?	:hasColour ?colour
Who painted Mona Lisa?	:createdBy ?painter
When was Mona Lisa painted?	:hasCreationDate ?crdat
How many paintings were painted by Leonardo da Vinci?	?(count(distinct ?painting) as ?count)

Table 1. Queries and query variables

The NL to SPARQL mapping is implemented as a transformation table, which could be extended to cover larger syntactic question variations.

The grammar has a modular structure with three main components: (1) lexicon modules covering ontology classes and properties; (2) data module covering ontology instances; and (3) query module covering NL questions and SPARQL query patterns. It supports NL queries in 15 languages, including: Bulgarian, Finnish, Norwegian, Catalan, French, Romanian, Danish, Hebrew, Russian, Dutch, Italian, Spanish, English, German and Swedish. The system relies on GF grammars, treating SPARQL as yet another language. In the same manner as NL generation, SPARQL patterns are encoded as grammar rules. Because of this compact representation within the same grammar, we can achieve parallel translations between any pair of the 15 languages and SPARQL.

The grammar-based interface provides a mechanism to formulate a query in any of the 15 languages, translate it to SPARQL and view the answers in any of those languages. The answers can be displayed as natural language descriptions or as triples. The latter can then be navigated as linked data. The browsing of the triples can be carried on continuously; by clicking on one of the triples listed in the answers, a new SPARQL query is launched and the results are generated as natural language text via the same grammar-based interface or as triples.

Fig. 2. Example of the query "who painted Guernica?" in 15 languages and in SPARQL.

Bul: к о й н а р и с у в а Guernica **Cat**: per qui és Guernica **Dan**: hvem malede Guernica **Dut**: wie schilderde Guernica **Eng**: who painted Guernica **Fin**: kenen maalaama on Guernica **Fre**: par qui est Guernica **Ger**: wer malte Guernica **Heb**: מי צייר את גרניקה **Ita**: da chi è dipinto Guernica **Nor**: hvem malte Guernica **Ron**: cine a pictat Guernica **Rus**: к т о н а р и с о в а л Guernica **Spa**: quién pintó Guernica **Swe**: vem målade Guernica	**SPARQL**: PREFIX painting: <http://spraakbanken.gu.se/rdf/owl/painting.owl#> PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> SELECT distinct ?author WHERE { ?painting rdf:type painting:Painting . ?painting painting:createdBy ?painter . ?painter rdfs:label ?author . ?painting rdfs:label ?title FILTER (str(?title)= "Guernica") . }

3 Evaluation

Following previous question answering over linked data (QALD) evaluation challenges [5], we divided the evaluation into three parts, each focusing on a specific aspect: (1) user satisfaction, i.e. how many queries were answered; (2) correctness; and (3) coverage, how the system scales up.

For the first parts of the evaluation, we considered a number of random queries in 7 languages and counted the number of corrections that 1-2 native informants would make to the original queries. The results of the evaluation showed that the amount of suggested corrections is relatively low for the majority of the evaluated languages. The overall correctness of the generated queries seem to be representative and acceptable, at least among the users who participated in the evaluation.

Regarding coverage, the grammar allows for paraphrasing most of the question patterns, which sums up, on average to 3 paraphrases per construction in the English grammar. The number of alternatives varies across languages, but

the average across languages ranges between 2 and 3 paraphrases per construction. In addition, the 112 basic query patterns from the query grammar can be combined with logical operators, in order to obtain more complex queries, which sums up to 1159 query patterns that the grammar covers, including WH and Yes/No questions. The additions needed to build the query grammar in order for it to scale up are small, given that the other resources are in place. Also for building the query grammar for a given language, no more than 150 lines of code are needed. This process can be done semi-automatically.

4 Conclusions

We introduce a novel approach to multilingual interaction with the Semantic Web content via GF grammars. The method has been successfully demonstrated for the cultural heritage domain and could subsequently be implemented for other domains or scaled up in terms of languages or content coverage. The main contribution with respect to current state-of-the-art approaches is SPARQL support and question answering in 15 languages.

Acknowledgment

This work was supported by MOLTO European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement FP7-ICT-247914. The authors would like to acknowledge the Centre for Language Technology in Gothenburg.

References

1. Garlik, S.H., Andy, S.: SPARQL 1.1 Query Language. (March 2013) W3C Recommendation.
2. Ranta, A.: Grammatical Framework: Programming with Multilingual Grammars. CSLI Studies in Computational Linguistics. CSLI, Stanford (2011)
3. Ferré, S.: SQUALL: A controlled natural language for querying and updating RDF graphs. In: CNL. (2012) 11–25
4. Ngonga Ngomo, A.C., Bühmann, L., Unger, C., Lehmann, J., Gerber., D.: Sorry, I don't speak SPARQL — translating SPARQL queries into natural language. In: Proceedings of WWW. (2013)
5. Walter, S., Unger, C., Cimiano, P., Bär, D.: Evaluation of a Layered Approach to Question Answering over Linked Data. In: International Semantic Web Conference (2). (2012) 362–374
6. Lopez, V., Fernández, M., Motta, E., Stieler, N.: Poweraqua: Supporting users in querying and exploring the semantic web. Semantic Web **3**(3) (2012) 249–265
7. Damova, M., Dannélls, D.: Reason-able View of Linked Data for cultural heritage. In: Proceedings of the third International Conference on Software, Services and Semantic Technologies (S3T). (2011)