Hackathon on Question Answering based on automatically generated grammars (5-9 July 2021)

home page: https://scdemo.techfak.uni-bielefeld.de/qahackathon/

Extension for Dataset of Italian Cultural Heritage

Gennaro Nolano (gnolano@unior.it)
Mohammad Fazleh Elahi (melahi@techfak.uni-bielefeld.de)

The QA system was extended in order to implement data from the Knowledge Graph ArCo, which aggregates knowledge for Italian Cultural Heritage through the use of several conceptual ontologies and data from the Italian National Catalogue of Cultural properties.

Together with the extension to a new dataset, we also extended the grammar to cover Italian questions in the specific domain of Cultural Heritage: in particular we implemented:

- 3 lexical entries for NounPPFrames (*scheda di, data di creazione di, autore di);*
- 2 lexical entries for TransitiveVerbFrames (*creare* and *rappresentare);*
- 2 lexical entries for InTransitiveVerbFrames (*costruito nel* and *si trova a*).

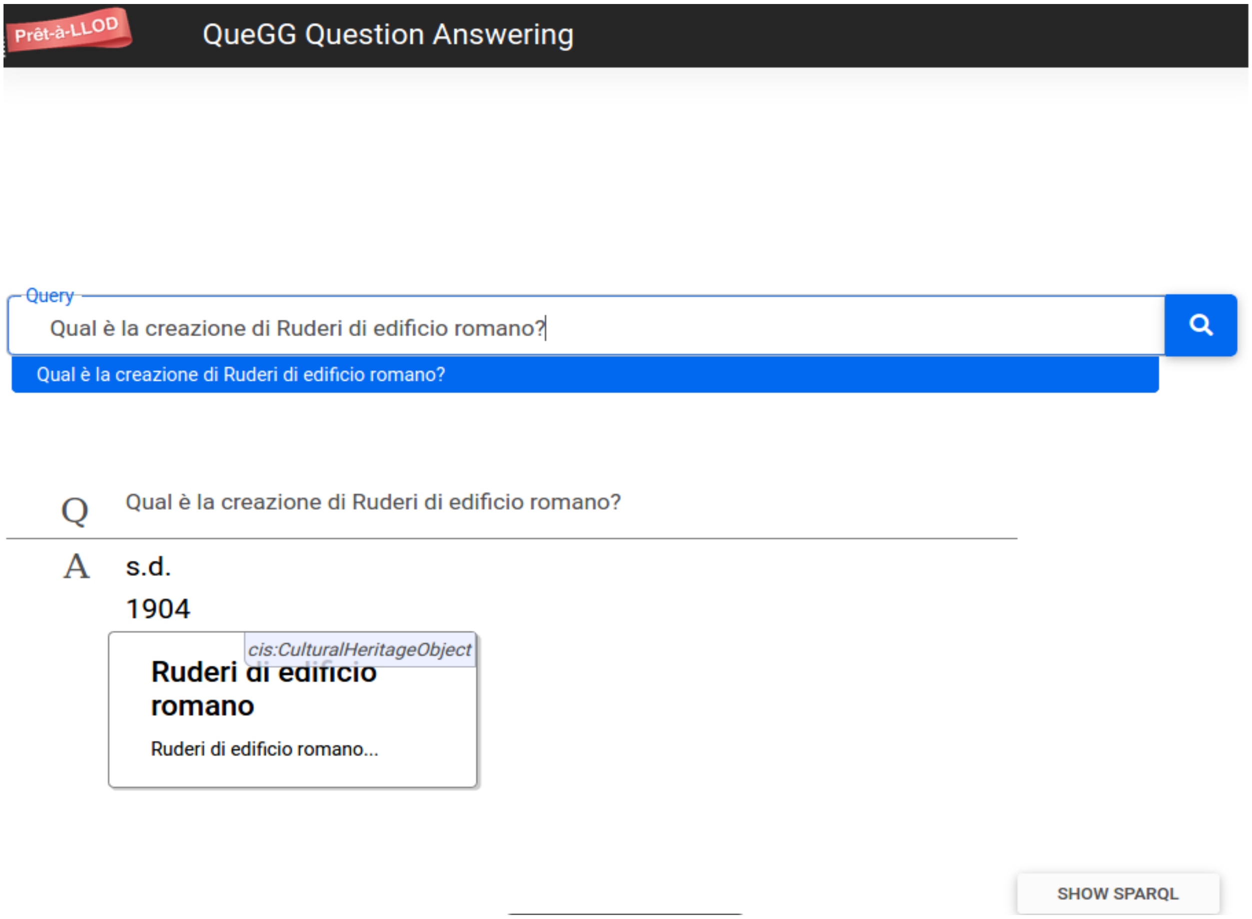
Furthermore, to fully integrate Italian language, we also worked on the creation of a new base lemon file that would correctly understand Italian sentences.

This model, while still exemplary, represents an important step in the future development of a Knowledge Based Question Answering for Italian language in the domain of Cultural Heritage.

The ArCo Knowledge Graph is integrated as one of the datasets in the grammar generation project. There are a couple of issues in the KG such as rdfs:labels are often missing. The sparql with language with filer italian (IT) does not work. Therefore, the generalized sparql creation module of the grammar generation project is extended for the dataset. Along with that we also addressed italian questions for DBpedia and Wikidata.
The extension can be found here
(https://github.com/fazleh2010/question-grammar-generator.git -b italian)

DataSet	Input	output	Number of questions
Arco	lexicalEntries	questions-sparql	600
DBpedia	lexicalEntries	questions-sparql	577
WikiData	lexicalEntries	questions-sparql	352

Table 1. Example of lexical entries and questions for different dataset for italian



Qual è la creazione di Ruderi di edificio romano?

Qual è la creazione di Ruderi di edificio romano?

Q Qual è la creazione di Ruderi di edificio romano?

A s.d.
1904

Ruderi di edificio romano

Ruderi di edificio romano...

Fig 2. A snapshot of example question of Italian over Cultural Heritage dataset (ArCo)

We have developed a question answering system for Italian for three datasets. The grammar coverage can be extended to million questions by adding lexical entries (i.e. adding rows in the XSL sheet). There are some issues for generating questions for TransitiveVerbFrames and InTransitiveVerbFrames for Italian, which can be resolved as future work.

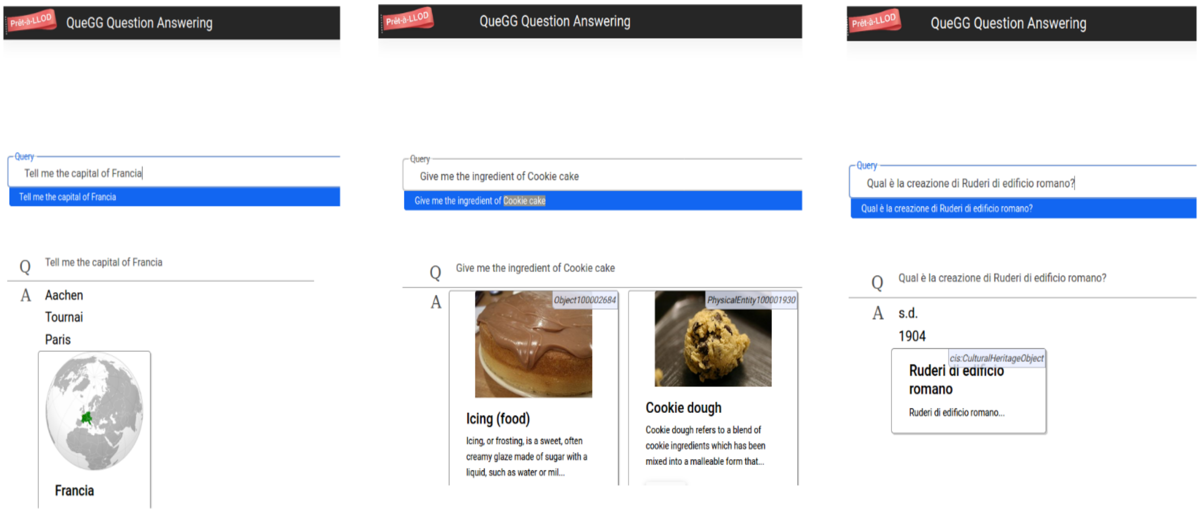
QueGG-web component

Frank Grimm (fgrimm@techfak.uni-bielefeld.de)

The web component for QueGG now supports data uploads targeting various languages and datasets. The configuration format, adapted by extending the grammar generation configuration, contains information on language and the SPARQL endpoint of the dataset. Additionally, resource lookup for improved answer rendering is now fully configurable through search terms and a SPARQL query that dynamically assembles labels, descriptions, and preview images for any ontological resource encountered in a question or answer. The full source code is available at https://github.com/ag-sc/QueGG-web alongside various Docker
and docker-compose configurations to compile and run.
Examples at https://github.com/ag-sc/QueGG-web/tree/main/example now include (images below):

- DBpedia
- Wikidata
- Italian cultural heritage sites (https://dati.beniculturali.it)

Different languages and datasets currently require separate instances, future improvements might include the capability to merge all of them into a single one.



Tell me the capital of Francia

Tell me the capital of Francia

Tell me the capital of Francia

Aachen
Tournai
Paris

Francia

Give me the ingredient of Cookie cake

Give me the ingredient of Cookie cake

Give me the ingredient of Cookie cake

Object100002684

Icing (food)

Icing, or frosting, is a sweet, often creamy glaze made of sugar with a liquid, such as water or mil...

PhysicalEntity100001930

Cookie dough

Cookie dough refers to a blend of cookie ingredients which has been mixed into a malleable form that...

Qual è la creazione di Ruderi di edificio romano?

Qual è la creazione di Ruderi di edificio romano?

Qual è la creazione di Ruderi di edificio romano?

s.d.
1904

cis:CulturalHeritageObject

Ruderi di edificio romano

Ruderi di edificio romano...

Fig. Example of running the QA system for different dataset and language

NLP Script for adding linguistic information

Andreas Röhler (andreas.roehler@online.de)

To extend the grammar coverage in automatic grammar generation, it is necessary to add a lexical entry for syntactic frame (see recipe). When adding a lexical entry manually, it is necessary to provide linguistic information. For example: for the transitive verb *direct*, it is necessary to input different forms of the verb such as *directed* (past tense), *directs* (3rd person singular), etc.

A script is provided which takes a GrammarFrame and fills in the needed conjugated verb forms. The script will reduce manual tasks for lexical entry creation for grammar generation.

At a first attempt calls to Wordnet where checked, which wasn't successful. No way to call for detailed conjugated verb forms was found. Another API accessible from the net couldn't be used due to license related issues. Therefore, we used a resource (https://github.com/clips/pattern) which provides comma-separated english verb forms, suitable for a scriptings input. It is possible to extend it for other languages. Several errors in the source have been fixed, missing entries filled. Corrected forms are committed onto https://github.com/andreas-roehler/pattern/blob/master/pattern/text/en/en-verbs.txt. Also filed a PR against the orginal source: https://github.com/clips/pattern/pull/323

As first proof of concept an Emacs script was run against "DbpediaFrameIntransitive - QALD Train - not solved.csv" and Script "fill-grammar-frame.el" and csv are committed onto the extension-branch of project (https://github.com/fazleh2010/question-grammar-generator.git -b extension) .Script is in Emacs Lisp for now, translating into Python or anything should be feasible. When used against a different frame, the regular expressions addressing the slots might need adoption. A PR for the stuff mentioned: https://github.com/fazleh2010/question-grammar-generator/pull/7

QA system over linked data in industrial context

Name: Anonymous, Ireland

- Attended the presentation and demo
- Analysis of generation of questions over RDF data in the context of QA system and grammar generation
- Finding issues for integrating Dataset (in RDF format) in industrial context.