Querying multilingual DBpedia with QAKiS

Elena Cabrio, Julien Cojan, Fabien Gandon, Amine Hallili

Querying multilingual DBpedia with QAKiS

Elena Cabrio, Julien Cojan, Fabien Gandon, and Amine Hallili

INRIA Sophia Antipolis, France
{firstname.lastname}@inria.fr

Abstract. We present an extension of QAKiS, a system for open domain Question Answering over linked data, that allows to query DBpedia multilingual chapters. Such chapters can contain different information with respect to the English version, e.g. they provide more specificity on certain topics, or fill information gaps. QAKiS exploits the alignment between properties carried out by DBpedia contributors as a mapping from Wikipedia terms to a common ontology, to exploit information coming from DBpedia multilingual chapters, broadening therefore its coverage. For the demo, English, French and German DBpedia chapters are the RDF data sets to be queried using a natural language interface.

1 Question Answering over Linked Data

The Semantic Web community has recently started an effort of internationalization of the DBpedia project (Bizer *et al.* [1]), born to extract structured information from Wikipedia multilingual pages, and to make such structured information accessible on the Web. This demonstration presents an extension of QAKiS, a system for open domain Question Answering over linked data (Cabrio *et al.* [2]), that allows to query DBpedia multilingual chapters. Such chapters, well connected through Wikipedia instance interlinking, can contain different information with respect to the English version. In particular, they can provide *i)* more specificity on certain topics, *ii)* fill information gaps, or *iii)* conceptualize certain relations according to a specific cultural viewpoint. QAKiS exploits the alignment between properties carried out by DBpedia contributors as a mapping from Wikipedia terms to a common ontology, to exploit information coming from DBpedia multilingual chapters. Given the multilingual scenario, attributes are labeled in different natural languages: the common ontology enables to query the multiple DBpedia chapters with the same vocabulary on the mapped data.

The ability to exploit all the amount of multilingual information brings several advantages to QAKiS [3], both considering *i)* the intersection of such resources in different languages to detect contradictions or divergences, and *ii)* the union of such resources, to fill information gap (cross-fertilization among languages) (Rinser *et al.* [4]). Extending QAKiS to query multilingual data sets goes in the direction of enhancing users consumption of semantic data originally produced for a different culture and language, overcoming language barriers[1].

[1] Currently a hot topic, see the Multilingual Question Answering over Linked Data challenge (QALD-3) http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/index.php?x=home

2 Extending QAKiS to query multilingual DBpedia

QAKiS system description. QAKiS (Question Answering wiKiFramework-based System) [2] addresses the task of QA over structured knowledge-bases (e.g. DBpedia), where the relevant information is expressed also in unstructured forms (e.g. Wikipedia pages). It implements a relation-based match for question interpretation, to convert the user question into a query language (e.g. SPARQL). More specifically, it makes use of relational patterns (automatically extracted from Wikipedia and collected in the WikiFramework repository [2]), that capture different ways to express a certain relation in a given language. QAKiS is composed of four main modules (Fig. 1): *i)* the **query generator** takes the user question as input, generates the typed questions, and then generates the SPARQL queries from the retrieved patterns; *ii)* the **pattern matcher** takes as input a typed question, and retrieves the patterns (among those in the repository) matching it with the highest similarity; *iii)* the **sparql package** handles the queries to DBpedia; and *iv)* a **Named Entity (NE) Recognizer**.

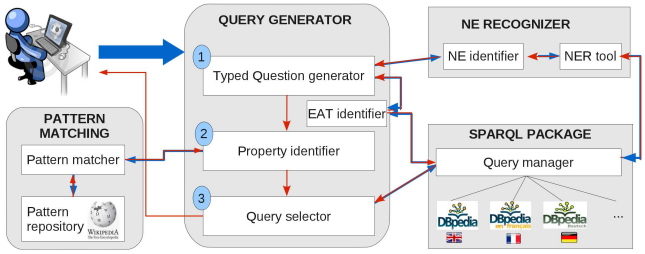


Fig. 1. QAKiS workflow [2]

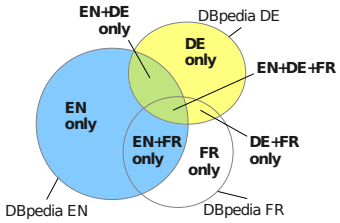
The actual version of QAKiS targets questions containing a Named Entity related to the answer through one property of the ontology, as *Which river does the Brooklyn Bridge cross?*. Such questions match a single pattern (i.e. one relation).

Before running the *pattern matcher* component, the question target is identified combining the output of Stanford NE Recognizer, with a set of strategies that compare it with the instances labels in the DBpedia ontology. Then a *typed question* is generated by replacing the question keywords (e.g. who, where) and the NE by the types and supertypes. A Word Overlap algorithm is then applied to match such typed questions with the patterns for each relation. A similarity score is provided for each match: the highest represents the most likely relation. A set of patterns is retrieved by the pattern matcher component for each typed question, and sorted by decreasing matching score. For each of them, a set of SPARQL queries is generated and then sent to an endpoint for answer retrieval.

Multilingual DBpedia alignment. Multilingual DBpedia chapters[2] have been created following Wikipedia structure: each chapter contains therefore data ex-

[2] `http://wiki.dbpedia.org/Internationalization/Chapters`

tracted from Wikipedia in the corresponding language, and so reflects local specificity. Data from different DBpedia chapters are connected by several alignments: *i) instances* are aligned according to the inter-language links, that are created by Wikipedia editors to relate articles about the same topic in different languages; *ii) properties* that mostly come from template attributes, i.e. structured elements that can be included in Wikipedia pages so as to display structured information. These properties are aligned through mappings manually edited by the DBpedia community. Since in the demo we are focusing on English, French and German DBpedia chapters, Figure 2 shows the additional information coverage provided by the mentioned DBpedia chapters. Areas **FR only**, **DE only** and **DE+FR only** correspond to aligned data made available by French and German DBpedia chapters.



Pairs (subject, property) with values in DBpedia chapters	Number
EN+FR+DE	110 685
EN+DE only	48 804
EN+FR only	408 125
DE+FR only	231 585
EN only	1 790 972
DE only	1 036 953
FR only	411 782

Fig. 2. Relative coverage of English, German and French DBpedia chapters.

QAKiS extension. QAKiS extension to query the ontology properties of multilingual DBpedia is integrated at the *SPARQL package* level. The typed questions generation and the pattern matching steps work as before, but now, instead of sending the query to English DBpedia only, the *Query manager* reformulates the queries and sends them to multiple DBpedia chapters. As only the English chapter contains labels in English, this change has no impact on the NE Recognition. The main difference is in the query selection step. As before, patterns are taken iteratively by decreasing matching score, the generated query is then evaluated and if no results are found the next pattern is considered, and so on. However, as queries are now evaluated on several DBpedia chapters, it is more likely to get results, terminating query selection with a higher matching score. Currently, the results of a SPARQL query are aggregated by the set union. Other strategies could be considered, as using a voting mechanism to select the most frequent answer, or enforcing a priority according to data provenance (e.g. English chapter could be considered as more reliable for questions related to English culture).

3 QAKiS demonstrator

Figure 3 shows QAKiS demo interface.[3] The user can select the DBpedia chapter he wants to query besides English (that must be selected as it is needed for

[3] `http://dbpedia.inria.fr/qakis/`

NER), i.e. French or German DBpedia (box 1). Then the user can either write a question or select among a list of examples, and click on *Get Answers!* (box 2). As output, QAKiS provides: *i)* the user question (the recognized NE is linked to its DBpedia page), *ii)* the generated typed question, *iii)* the pattern matched, *iv)* the SPARQL query sent to the DBpedia SPARQL endpoint, and *v)* the answer (box 3). The demo we will present follows these stages for a variety of queries.

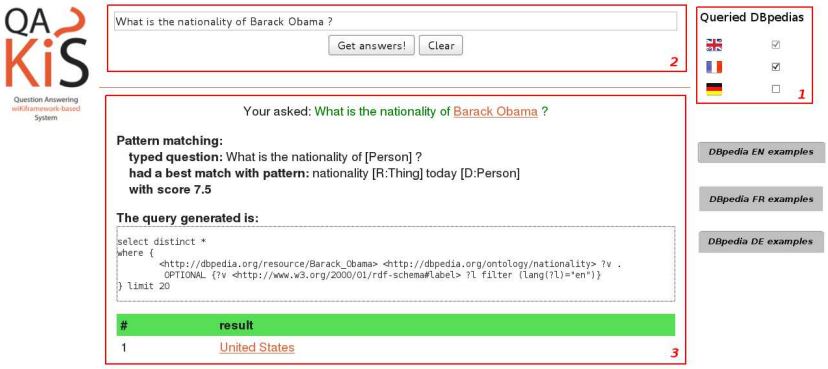


Fig. 3. QAKiS demo interface

3.1 Queries and datasets for demonstration and evaluation

QALD-2. Since QAKiS currently targets only questions containing a NE related to the answer through one property of the ontology (e.g. *In which military conflicts did Lawrence of Arabia participate?*), we extracted from QALD-2 test set[1] the set of questions corresponding to such criterion (i.e. 32 questions). The discarded questions require either some forms of reasoning (e.g. counting or ordering) on data, aggregation (from datasets different from DBpedia), involve n-relations, or they are boolean questions. We run both QAKiS$_{EN}$ (i.e. the system taking part into the challenge) and QAKiS$_{EN+FR}$ and QAKiS$_{EN+DE}$ (the versions enriched with the French and German DBpedia, respectively) on the reduced set of questions. Since the answer to QALD-2 questions can be retrieved in English DBpedia, we do not expect multilingual QAKiS to improve its performances. On the contrary, we want to verify that QAKiS performances do not decrease (due to the choice of the wrong relation triggered by a different pattern that finds an answer in multilingual DBpedia). Even if often extended QAKiS select different patterns with respect to the original system, the selected relation is the same (except than in one case), meaning that generally performances are not worsen by the addition of multilingual DBpedia chapters.

Multilingual DBpedia. Since we are not aware of any reference list of questions whose answers can be found in French or German DBpedia only, we create

our list to evaluate the improvement in QAKiS's coverage as follows: *i)* we take the sample of 32 QALD-2 questions; *ii)* we extract the list of triples present in French and German DBpedia only; in each question we substitute the NE with another entity for which the asked relation can be found respectively in the French or German chapters only. For instance, for the QALD-2 question *How tall is Michael Jordan?*, we substitute the NE *Michael Jordan* with the entity *Margaret Simpson*, for which we know that the relation `height` is missing in English DBpedia, but is present in the French chapter. As a result, we obtain the question *How tall is Margaret Simpson?*, that we submit to QAKiS$_{EN+FR}$. Following the same procedure for German, in *Who developed Skype?* we substituted the NE *Skype* with the entity *IronPython*, obtaining the question *Who developed IronPython?* [4] For some properties (e.g. `Governor`, `Battle`), no additional links are provided by the multilingual chapters, so we discarded the questions asking for those relations. QAKiS precision on the new set of questions over French and German DBpedia is in line with QAKiS$_{EN}$ on English DBpedia ($\sim$ 50%). This evaluation did not have the goal to show improved performances of the extended version of QAKiS with respect to its precision, but to show that the integration of multilingual DBpedia chapters in the system is easily achievable, and that the expected improvements on its coverage are really promising and worth exploring (see Figure 2). To double-check, we run the same set of questions on QAKiS$_{EN}$, and in no cases it was able to detect the correct answer, as expected.

4 Future perspectives

Extensions are planned in several directions, to improve: *i)* the WikiFramework pattern extraction algorithm; *ii)* the question-pattern matching algorithm; *iii)* the system coverage, addressing boolean and n-relation questions. With respect to multilingualism, we plan to *i)* extend QAKiS to query multilingual DBpedia chapters other than the presented ones; *ii)* port QAKiS to multiple languages, i.e. allowing input questions in languages different from English. Moreover, we plan to experiment approaches to automatically extend DBpedia existing alignments.

References

1. C. Bizer et al. DBpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165, Sept. 2009.
2. E. Cabrio, J. Cojan, A. P. Aprosio, B. Magnini, A. Lavelli, and F. Gandon. QAKiS: an open domain qa system based on relational patterns. In *Proceedings of the ISWC 2012 Posters and Demonstrations Track*, Boston, US, November 2012.
3. J. Cojan, E. Cabrio, and F. Gandon. Filling the gaps among DBpedia multilingual chapters for question answering. In *Proceedings of ACM Web Science 2013 (to appear)*, Paris, France, May 2013.
4. D. Rinser, D. Lange, and F. Naumann. Cross-lingual entity matching and infobox alignment in wikipedia. *Information Systems*, 2012.

[4] The obtained set of transformed questions is available online at `http://dbpedia.inria.fr/qakis/`.