Filling the Gaps Among DBpedia Multilingual Chapters for Question Answering

Julien Cojan, Elena Cabrio and Fabien Gandon
INRIA, 2004 Route des Lucioles
06902 Sophia Antipolis, France
{julien.cojan, elena.cabrio, fabien.gandon}@inria.fr

ABSTRACT

To publish information extracted from multilingual pages of Wikipedia in a structured way, the Semantic Web community has started an effort of internationalization of DBpedia. Multilingual chapters of DBpedia can in fact contain different information with respect to the English version, in particular they provide more specificity on certain topics, or fill information gaps. DBpedia multilingual chapters are well connected through instance interlinking, extracted from Wikipedia. An alignment between properties is also carried out by DBpedia contributors as a mapping from the terms used in Wikipedia to a common ontology, enabling the exploitation of information coming from the multilingual chapters of DBpedia. However, the mapping process is currently incomplete, it is time consuming since it is manually performed, and may lead to the introduction of redundant terms in the ontology, as it becomes difficult to navigate through the existing vocabulary. In this paper we propose an approach to automatically extend the existing alignments, and we integrate it in a question answering system over linked data. We report on experiments carried out applying the QAKiS (Question Answering wiKiframework-based) system on the English and French DBpedia chapters, and we show that the use of such approach broadens its coverage.

Author Keywords

Linked Data, DBpedia, Property Alignment, Question Answering

ACM Classification Keywords

I.2.7 Artificial Intelligence: Natural Language Processing

INTRODUCTION

With the goal of extracting structured information from multilingual pages of Wikipedia, and to make such structured information accessible on the Web, the Semantic Web community has started the DBpedia project (Bizer et al. [2]). DBpedia multilingual chapters are well connected through instance interlinking, extracted from Wikipedia. An alignment between properties is also carried out by DBpedia contributors as a mapping from the terms used in Wikipedia to a common ontology, enabling the exploitation of information coming from the multilingual chapters of DBpedia. At the same time, multilingual chapters of DBpedia can contain different information with respect to the English version, in particular they provide more specificity on certain topics, or fill information gaps. For instance, when looking for the nationality of Barack Obama on the English chapter of DBpedia, we can notice that there is no property *nationality* directly linking Obama to the United States. Such information can instead be found in the French version of DBpedia, the second biggest chapter. Moreover, the knowledge of certain instances and the conceptualization of certain relations can be biased according to different cultures, and this is reflected in the structure and content of such collaboratively constructed semantic resources. For instance, no information is provided in English Wikipedia and DBpedia for the French musical group "Les Frères Jacques", or for for French writer Jean-Bernard Pouy.

Being able to exploit all the amount of multilingual information would bring several advantages to systems that harvest information from Wikipedia and DBpedia automatically, 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). Also Rinser et al. [12] highlight the importance of mapping the attributes of the infoboxes across different language versions, to increase the information quality and quantity in Wikipedia. Moreover, in the context of Natural Language (NL) question answering over linked data, a system able to exploit information coming from the multilingual and parallel versions of DBpedia would increase its probability to retrieve a correct answer to provide to the user (i.e. its recall would be improved).

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. Unfortunately, the cross-language mapping process of properties among multilingual DBpedia chapters is currently incomplete, it is time consuming since it is manually performed, and may lead to the introduction of redundant terms in the ontology, as it becomes difficult to navigate through the existing vocabulary. Moreover, several problems arise concerning both the variety and ambiguity of properties extracted from Wikipedia Infoboxes (e.g. attributes names are not always sound, often cryptic or abbreviated), and the fact that they are language sensitive.

WebSci'13, May 2 – May 4, 2013, Paris, France.

In this paper, we tackle the following research question: How to fill the gaps among DBpedia multilingual chapters for Question Answering? Given its complexity, in this work we narrow its scope, answering to the following subquestions: (1) how to benefit from querying multilingual DBpedia data sets in the current mapping progress?, (2) how to safely extend the property alignment?, and (3) how to integrate it into an existing QA system? (in other words, can a QA system benefit from querying multilingual chapters?). In this paper, we do not make use of general ontology alignment techniques, and we do not stand on the related discussions[1]. We rather exploit the existing man made alignments.

As a case study, we propose a comparative analysis of the English and French DBpedia chapters, highlighting the current status of the properties alignment between the two chapters, and providing an in-depth analysis of them. Moreover, we propose an approach to automatically extend the existing alignments taking into consideration the structure of Wikipedia and DBpedia, and we integrate it in a question answering system over linked data. Taking advantage of its modular architecture, we carried out our experiments using the QAKiS (Question Answering wiKiframework-based System) (Cabrio et al. [5]), and we show that the possibility to query multilingual datasets broaden its coverage.

The work we propose is highly interdisciplinary, since the issues we address here have the ultimate goal of bridging Natural Language Processing and the Semantic Web, to enhance interactions between non-expert users and the huge and heterogeneous amount of data available on the Web.

The reminder of the paper is as follows. Section provides an analysis of the current status of properties alignment in multilingual chapters of DBpedia, with a focus on the English and French versions. Section describes the approaches to extend the current mappings that we propose, while Section describes the ongoing work of integration of such approaches into a QA system, and the experimental setting. Section lists the related work in the literature; conclusions end the paper.

DBPEDIA PROPERTIES ALIGNMENT

In the first part of this section, we describe the DBpedia project, and the ongoing cross-language mapping process. In the second part, as a case study we propose a comparative analysis of the English and French DBpedia chapters, highlighting the current status of the properties alignment between the two biggest chapters, and providing the statistics arising from our in-depth analysis.

Existing alignments

As introduced before, DBpedia [2] is a community effort to extract structured data from Wikipedia, and to publish it on the Linked Data. Initially, it only contained data extracted from English Wikipedia, while in the most recent period efforts to integrate data extracted from chapters of languages different from English have arisen (for instance for German, Spanish, French and Italian). However, in the current state of affairs the content is still focused on the English chapter, due to the fact that naming conventions limit the coverage of other chapters, and the fact that English is the biggest chapter.

Multilingual DBpedia chapters[2] have been created following Wikipedia structure (Kontokostas et al. [7]): each chapter contains therefore data extracted from Wikipedia in the corresponding language, and so reflects local specificity. Data are published in RDF, and are structured in triples `<subject, predicate, object>` where the *subject* is an instance corresponding to a Wikipedia page, the *predicate* is a property from the DBpedia ontology or from other vocabularies (foaf, dublin core, georss, ...), and the *object* is either a literal value or another instance.

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. As shown in (Rinser et al. [12]) these correspondences are far from being perfect, but a simple filter applied before data publication in DBpedia significantly improves its quality; *ii) properties* mostly come from template attributes, i.e. structured elements that can be included in Wikipedia pages so as to display structured information, the most common being the infoboxes. The generic template extraction that creates properties URIs from their textual names, has the inconvenient of generating a large variety of properties, as well as ambiguous terms. For instance, both properties `propEn:birthDate`[3] and `propEn:dateOfBirth` appear in English DBpedia with the same meaning. Conversely, the property `propEn:start` is used to indicate both the starting place of a route (e.g. the fist station on a railway line), and the date of the beginning of an event. Moreover, as introduced before, the terms used for properties are language dependent.

To overcome these limitations, a common ontology and mappings from template definitions to the ontology vocabulary are being collaboratively edited by the DBpedia community.[4] For instance, the attributes *date of birth* and *birth date* are mapped to the ontology property **`dbo:birthDate`**[3] in the description of a person, and the attribute *start* is mapped to **`dbo:routeStart`** when describing a road, and to **`dbo:startDate`** when describing an event. This term normalization effort has the goal to improve properties alignment among DBpedia multilingual chapters. It is however an ongoing work, and needs constant maintenance as Wikipedia templates evolve in time. Assistance tools for mapping editions, as well as automated techniques to extend the resulting alignments are becoming therefore important issues.

[1] see the Ontology Alignment Initiative **`http://oaei.ontologymatching.org/`**

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

[3] For simplification, we use here the shorthand `propEn:` for **`http://en.dbpedia.org/property/`** and **`dbo:`** for **`http://dbpedia.org/ontology/`**

[4] On the wiki **`http://mappings.dbpedia.org`**

Comparing English and French chapters

As a case study to analyze the current state of affairs of properties alignment in multilingual chapters of DBpedia, we consider the datasets of English and French DBpedia. While the English chapter is the biggest and the most complete, with about 400 million triples[2] and 345 templates mapped, the French chapter is the second chapter in size (∼130 million triples, and 42 templates mapped).

In our analysis, for each object property **prop** we compare the triples <subject, **prop**, object> from English and French DBpedia on aligned pairs of instances subject and object. That is, triples <$\text{subject}_{\text{fr}}$, **prop**, $\text{object}_{\text{fr}}$> from French DBpedia are transposed into <$\text{subject}_{\text{en}}$, **prop**, $\text{object}_{\text{en}}$>, where $\text{subject}_{\text{en}}$ and $\text{object}_{\text{en}}$ are respectively instances of English DBpedia related to $\text{subject}_{\text{fr}}$ and $\text{object}_{\text{fr}}$ through the relation owl:sameAs. These triples are compared with triples <subject, **prop**, object> from English DBpedia such that subject and object are also related to French instances with relation owl:sameAs.

Figure 1 describes the possible outcomes of such comparison. In case *a)* we have the same property in both the English and the French chapters. For instance, for the subject *Barack Obama* the property birthPlace is present in both the English and the French versions, with the same value. In this case, the French chapter does not bring new information, except a confirmation of values found in English chapter. In *b)* we also have the same property in both English and French chapters, but this time with different values. In *c)* we have values for the property in the English chapter only, in the example of *Barack Obama*, the property residence is present for the English DBpedia chapter (with the value *White House*), while it is missing in the French version. In *d)* we have a value for the property in the French chapter only, again in the example of *Barack Obama*, the property nationality is missing for the English DBpedia chapter, while it is present in the French version (with the value *États-Unis*, i.e. *United States*).

There can be two reasons for the value difference in case *b)* (Figure 1): *i)* there is a disagreement between the two datasets, produced either by an error in one of them, or reflecting a different point of view (in particular for properties of type owl:functionalProperty); or *ii)* the values reported in the two chapters are complementary, often providing a different granularity level (e.g. city vs country for the birth place of *Henry Lawson*). The first case can be interestingly exploited to automatically detect inconsistencies among the data, that can help the Wikipedia community to improve information quality across language versions. The second one brings additional information on the subject, but it could also help to infer relationships between the values (for instance that the city where *Henry Lawson* was born is in his country of birth).

The same comparison is also performed for datatype properties over triples <subject, **prop**, **val**> with aligned instances subject. For every property **prop**, we count: *a)* how many subject have the same values with **prop** in French and English, *b)* how many have at least one different value, and how many have only values either *c)* in the English or *d)* in the French DBpedia. We observed that the ratio between the number of values that are the same in English and French chapters and the number of values that are different is lower for datatype properties than for object properties. This is true in particular for string literals, as most of them are expressed in their respective chapter language (we did not compare neither instance labels nor abstracts). Nevertheless, we kept these properties in our comparison as some of them bring information that can be exploited in a different language, for instance for people's names.

Reflecting the different progression of the mapping task between French and English DBpedia, 217 ontology properties are currently used in French DBpedia, compared to more than 1000 in English DBpedia.

Table 1 shows some statistics resulting from the comparison between English and French DBpedia. In particular, it shows the top 10 properties for which French DBpedia presents the highest number of values not present in the English version, i.e. the properties to which the French chapter can contribute most (the list is ordered with respect to column *d) only FR value*). The sum of the values of each column (for all 1637 properties of the ontology) is given in the bottom of Table 1. These values give the number of pairs (subject, property) *a)* that have a value in common in English and French chapters, *b)* that have different values in the two chapters, *c)* that have only values in English chapter, *d)* that have only values in French chapter. Two intermediate sums are also given for the object properties and for the datatype properties. These sums show that in general, over the aligned data French and English chapters are quite complementary. About 47% of the data from French DBpedia expressed in the common ontology cannot be found in English DBpedia (column *d)* vs. *a)+b)+d)*), and about 80% of the data from English DBpedia expressed in the common ontology cannot be found in French DBpedia (column *c)* vs. *a)+b)+c)*).

The values provided in Table 1 for the columns *d) only FR value* confirm our starting intuition that being able to exploit multilingual chapters of DBpedia provides an additional amount of information both specific to a certain culture (for instance, concerning French habits, food or minor musical groups), or to fill information gaps (for instance, missing links in the English chapters).

A STEP FURTHER: EXTENDING THE EXISTING ALIGNMENT

One of the reasons why templates mapping is so time consuming is that this process needs to be performed for each template, although many of them share the same attributes. For instance, many infoboxes describe people for what concerns their activities (i.e. Music Composers, Athletes, Football players, Philosophers, and so on). Although they contain specific attributes, others appear very frequently, like, for instance, *name*, *birth day* or *nationality* for a person. But unfortunately, these attributes require that the same property mapping is independently edited.

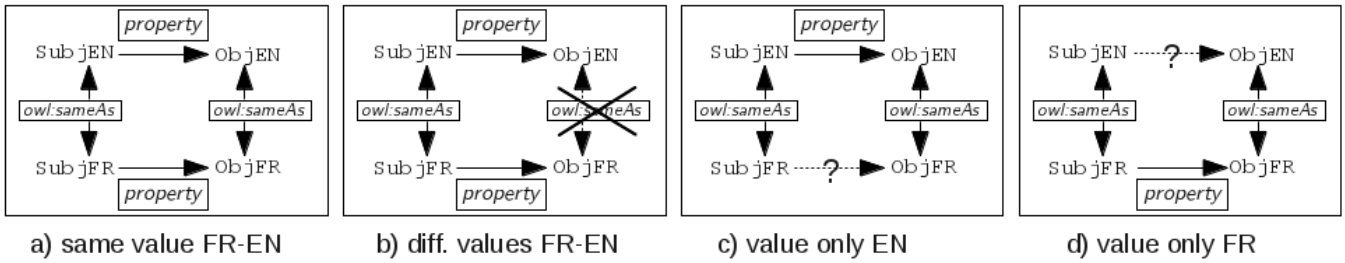

Figure 1. Outcomes of the comparison between EN and FR chapters

ontology property	a) same value FR-EN	b) diff. values FR-EN	c) value only EN	d) value only FR	type
dbo:arrondissement	10551	2421	1000	30110	ObjectProperty
dbo:nationality	1536	437	11825	26074	ObjectProperty
dbo:city	1042	89	2904	22616	ObjectProperty
dbo:birthDate	21360	1380	47104	22303	DatatypeProperty
dbo:birthPlace	14139	1965	49754	15279	ObjectProperty
dbo:title	0	95	4711	13546	DatatypeProperty
dbo:locatedInArea	1092	800	664	12711	ObjectProperty
dbo:region	22178	676	14397	12502	ObjectProperty
dbo:predecessor	1193	99	1973	11925	ObjectProperty
dbo:Person/height	20	1220	6186	9515	DatatypeProperty
total object properties	239321	40232	1046532	305452	-
total datatype properties	104262	134995	976025	155134	-
total	343583	175227	2022557	460586	-

Table 1. Statistics resulting from the comparison of the FR and EN chapters

In our work, we propose an approach to expand the property mappings to all the occurrences of non ambiguous attributes, that is attributes that have always been manually mapped to the same ontology property. This results in the extension of the alignments between the properties textually generated from the attributes, and the ontology properties. And so, it extends the alignment between multilingual datasets.

By non ambiguous attributes, we mean here the terms that have not proven to be ambiguous in the existing mappings. The integration of the extended mappings into the mapping data would require human validation, to check for incorrect alignments. In the following we evaluate the possible gain obtained from the approach we propose. We use a simple heuristic to select mappings that are likely to be correctly propagated: we select only the attributes that have been mapped several times before, and always to the same ontology property.

Extended alignments results

Figure 2 shows the mapping frequency of non ambiguous attributes in French DBpedia to the DBpedia ontology properties. Summing up, there are 47 attributes that are mapped at least twice, 18 attributes mapped at least three times (i.e. *lieu de décès* → **dbo:deathPlace**), and only one mapped at least ten times (i.e. *nom* → `foaf:name`). Table 2 shows the most frequent mappings.

Since we assume that the mapping frequency is a good indicator of the correctness of the mapping, in the rest of the section we will consider only the mappings that were mapped at least twice (i.e. frequency $\geq$2). Moreover, we carry out a manual validation of the 47 mappings appearing more than twice, to

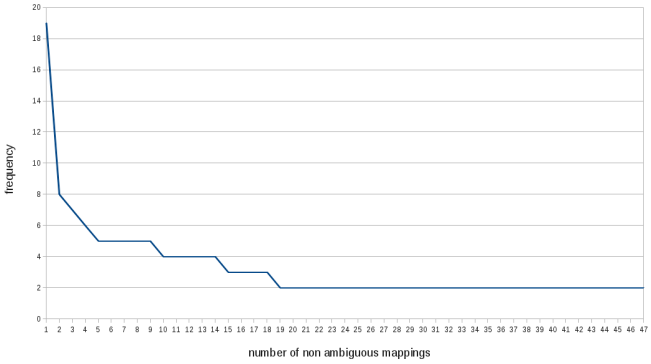

Figure 2. Frequency of non ambiguous mappings in French DBpedia

check if they are correct according to the attribute names. The results of such evaluation confirms that in 83% of the cases (i.e. 35 mappings), the mappings are correct. The validity of the remaining ones can be biased by the context in which they appear, since the attribute terms are either vague, or polysemous (i.e. could have different meanings). For instance, the attribute *division* → **dbo:locatedInArea** seems correct for geographic places but *division* could be used to indicate also a football league or an organization department, and in those cases the mapping is incorrect.

Table 3 provides for each mapping a comparison between the number of instances that have a value for the generic property (build from the attribute occurence), and the number of instances that have a value for the mapped ontology property. For instance, the property `propFr:lieuDeDécès` is present for more than 25,000 instances (column *values for*

generic property (p)	ontology property (po)	values for p	values for p in po range	values for po	values for both	same values
propFr:cp	**dbo:postalCode**	83637	0	57660	57660	0
propFr:lieuDeDécès	**dbo:deathPlace**	25477	14615	17190	13314	7579
propFr:région	**dbo:region**	87917	79853	51713	46077	45993
propFr:nationalité	**dbo:nationality**	44345	10071	46985	34884	8887
propFr:lieuDeNaissance	**dbo:birthPlace**	66262	37326	49430	41716	24569
propFr:altitude	**dbo:elevation**	52928	458	6972	5441	1
total object prop		645719	391044	482444	284201	209692
total datatype prop		680481	111876	517368	259623	59047
total		1326200	502920	999812	543824	268739

Table 3. Values comparison between generic and ontology properties for the extended mappings in French DBpedia. The table contains the number of instances that have values for these properties.

attribute	ontology property	frequency
nom	foaf:name	19
division	**dbo:locatedInArea**	8
nom local	foaf:name	7
cp	**dbo:postalCode**	6
lieu de décès	**dbo:deathPlace**	5

Table 2. Most frequent non ambiguous mappings in French DBpedia.

p, Table 3), and **dbo:deathPlace** for more than 17,000 (column *values for po*). Note that *lieu de décès* is not the only attribute to be mapped to **dbo:deathPlace** (i.e. there is also *lieu décès*, *décès*, and other variants). The column *values for both* indicates how often the mapping *lieu de décès* → **dbo:deathPlace** is actually applied, and it gives the number of instances that have values for both the generic and the ontology property: 13,314. The potential gain for the extension of this mapping is given by the number of instances that have a value for the attribute (ie. for the generic property) but no values for the ontology property, that is 25,477 - 13,314 = 12,163 additional values for **dbo:deathPlace**. Over the 47 mappings that can be extended, the potential gain is 1,326,200 - 999,812 = 326,388, that corresponds to an increase of about 30%.

Column *same values* give the number of instances for which the generic property and the ontology property have the same value. However, the comparison with the number of co-occurrence of the two properties is not fair as the extractor that generates the values for the ontology property is guided by the property signature signature (in the example of **dbo:deathPlace**, the expected value is an instance), whereas the generic property is more sensitive to noise and may generate other output from the same attribute value (for instance a number if the attribute value begins with a street number). So for this comparison, we narrow our scope to the instances for which the generic property values are coherent with the ontology property signature (column *values for p in po range*). Out of the 25,477 instances that have a value for propFr:lieuDeDécès, only 14,615 have an object value. However, every time there is an object value for propFr:lieuDeDécès and a value for **dbo:deathPlace**, these are the same.

In a symmetric way with respect to Figure 2, Figure 3

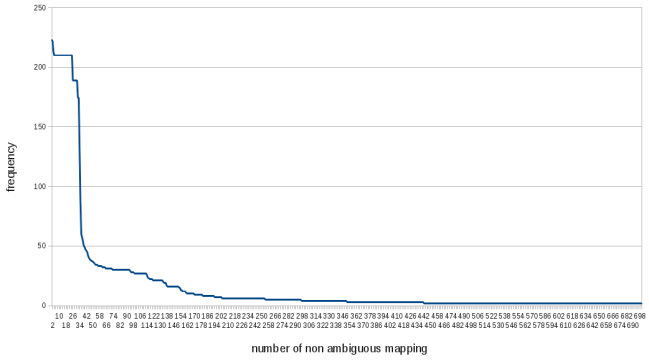

Figure 3. Frequency of non ambiguous mappings in English DBpedia

shows the mapping frequency of non ambiguous attributes in English DBpedia to ontology properties. As expected, many more attributes are mapped more frequently than in the French chapter. More specifically, there are 689 attributes that are mapped at least twice, 296 attributes mapped at least five times, and 160 mapped at least ten times (e.g. *twin* → **dbo:twinCity**). Table 4 shows the most frequent mappings.

attribute	ontology property	frequency
successor	**dbo:successor**	223
president	**dbo:president**	222
twin	**dbo:twinCity**	213
constituencyAm	**dbo:region**	210
majorityFloor	**dbo:majorityFloor**	210
Leader	Leader	210

Table 4. Most frequent non ambiguous mappings in English DBpedia.

Comparing English and French chapters

To evaluate the quality of the data obtained applying the above presented approach to extend the mapping among multilingual versions of DBpedia, we compare the values obtained from the mappings extension for the French chapter, with the values obtained for the English chapter, as previously done in Section for the existing alignments. Table 5 summarizes the results obtained from such comparison. More specifically, it provides the number of values that were added through this process (column *new values wrt. DBpedia En*

generic property	ontology property	new values wrt. DBpedia En and Fr	same values	diff. values
propFr:cp	**dbo:postalCode**	0	0	0
propFr:lieuDeDécès	**dbo:deathPlace**	4393	4016	479
propFr:région	**dbo:region**	16491	18496	3906
propFr:nationalité	**dbo:nationality**	358	870	200
propFr:lieuDeNaissance	**dbo:birthPlace**	6934	7016	862
propFr:altitude	**dbo:elevation**	150	63	97
total object prop		85951	73306	15250
total datatype prop		16155	45177	5001
total		102106	118483	20251

Table 5. Comparison between values obtained with the mappings extension in French DBpedia and English DBpedia

and Fr) with respect to the values already available through ontology properties in English and French DBpedia.

For instance, the mapping extension (*lieu de naissance* → **dbo:birthPlace**) considered earlier generates 6,934 new values. Among the values that were already present in the English chapter, 7,016 are the same and 862 differ (89% identical). We can notice that this is about the same ratio emerged for the comparison between values for the same ontology property in Section , i.e. 14,139 identical values and 1 965 different (87% identical). We can consider it as a positive result, as it suggests that most of the differences in the values are generated by differences between the two chapters of DBpedia, rather than from mappings mistake.

Concerning the 47 mappings described in Section , there are 118,483 identical values (column *same values*, Table 5) with respect to 20,251 different values (column *different values*). If we separately take object properties and datatypes properties, we obtain this time a better correlation between values of English and French chapters for datatype properties (90% instances with same values) than for object properties (82%). This may be explained by the fact that many datatypes are not specified for generic properties, in particular for strings, so we selected the values that fit in the ontology property range, while we have removed values that generated noise in the comparison described in Section .

QA EXPERIMENTAL SETTING

As a case study to experiment the contribution of the proposed approach to automatically extend the existing alignments from English DBpedia to the French chapter in a real application scenario, we integrate it in a QA system over Linked Data, i.e. QAKiS[5] (Section , Cabrio *et al.* [5]). To enhance users interactions with the web of data, query interfaces providing a flexible mapping between natural language expressions, and concepts and relations in structured knowledge bases are becoming particularly relevant. More specifically, QAKiS allows end users to submit a query to an RDF triple store in English and obtain the answer in the same language, hiding the complexity of the non intuitive formal query languages involved in the resolution process. At the same time, the expressiveness of these standards is exploited to scale to the huge amounts of available semantic data. We evaluate the contribution of the approach proposed in the paper with two sets of experiments, described in Section and .

[5] **http://dbpedia.inria.fr/qakis/**

QA system description: QAKiS

QAKiS (Question Answering wiKiFramework-based System) [5] addresses the task of question answering 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 [10]), that capture different ways to express a certain relation in a given language.

QAKiS is composed of four main modules (Fig. 4): *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**.

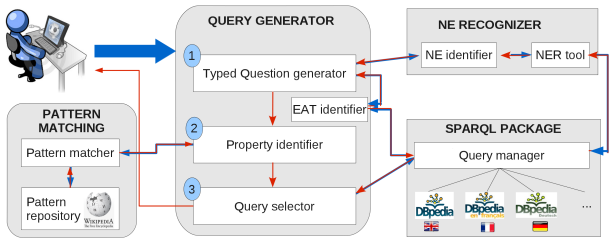

Figure 4. QAKiS workflow

The actual version of QAKiS targets questions containing a NE 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 target of the question is identified using the Stanford Core NLP NE Recognizer, together with a set of strategies based on the comparison with the labels of the instances 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 the SPARQL endpoint for answer retrieval.

QAKiS extension to query French DBpedia

To broaden the system coverage, we extend QAKiS to query the ontology properties of French DBpedia. This new feature is integrated in the query selection step. More specifically, the typed questions are generated as described in the previous section, and the named entity are still recognized basing on the English DBpedia dataset. Also the typed questions-patterns step is not modified. But differently from before, now for each pattern taken with decreasing matching score, the English DBpedia is queried first, then if no result is found, the query is adapted to query French DBpedia terms. If again no results are found, the next pattern is considered. This strategy for querying two DBpedia chapters gives the preference to the pattern matching score, since it represents the confidence of the system with respect to the relevance of the generated query. Then, the preference is assigned to the datasets. Obviously, this preference is arguable. We decided to make the system query the English chapter first as it is the biggest and the most complete chapter, so it is more likely that it contains the answer for a question in English. In the same way, if more than two chapters are to be queried, selecting a priority among them is arbitrary, and can depend by their characteristics and by their completeness. A different strategy we plan to experiment is to query all the multilingual datasets at the same time, and then to aggregate the obtained solution. This can be provided on the base of a voting mechanism, choosing for instance the most frequent answer if a single answer is expected, or combining them if several solutions are expected. An interface to compare the answers given by QAKiS_{EN} that queries only the English DBpedia with the answers of $\text{QAKiS}_{\text{EN+FR}}$ that queries both English and French DBpedia is available at `http://dbpedia.inria.fr/qakis/`.

Evaluation on QALD-2

As a first step of our experiments, we evaluate if the integration of the French DBpedia dataset has an impact on QAKiS performances on the standard benchmark delivered by the QALD-2 challenge organizers[6] (DBpedia track) for comparing different approaches and systems that mediate between a user, expressing his or her information need in natural language, and semantic data. Since in the actual version of the system it 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 the complete benchmark the questions corresponding to such criteria. Out of 100 questions available for testing, the questions containing a NE related to the answer through one property of the ontology amount to 32, that we used in our experiment. 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} (the version enriched with the French DBpedia) on the reduced set of questions.

Since the questions of the challenge have been thought so that the answer is present in the English DBpedia, we do not expect that QAKiS_{EN+FR} will improve its performances. On the contrary, we want to verify that QAKiS performances do not decrease (for instance due to the choice of the wrong relation triggered by a different pattern that finds an answer in French DBpedia).

Out of 32 questions, QAKiS_{EN} correctly answers to 15 questions and partially correctly to 4 questions (e.g. in *Give me all companies in Munich* the list provided by QAKiS using `foundationPlace` as relation and *Munich* as subject, is only partially overlapping with the one proposed by the organizers). Only in one case QAKiS_{EN+FR} negatively influences the system (i.e. for *Give me all movies directed by Francis Ford Coppola.*, where the correct relation `director` is discarded and `foundedBy` is preferred, providing a wrong answer). Even if in a number of cases QAKiS_{EN+FR} selects different patterns with respect to QAKiS_{EN}, the selected relation is the same, meaning that in general performances are not worsen by the addition of French DBpedia.

Evaluation on French DBpedia only

As introduced before, the questions created for QALD-2 challenge are thought to find an answer into the English DBpedia, so they cannot be used to evaluate the contribution resulting from the extension of properties alignments to the French chapter. Since we are not aware of any standard list of questions whose answers can be found in French DBpedia only, we create our reference set to evaluate the extension in QAKiS_{EN+FR}'s coverage performing the following steps:

1. we take the sample of 32 questions from QALD-2;
2. we extract the list of triples present in French DBpedia only (as described in Section);
3. in each question we substitute the named entity with another entity for which the asked relation can be found in the French chapter only.

For instance, for the QALD-2 question *How tall is Michael Jordan?*, we substitute the Named Entity *Michael Jordan* with the entity *Margaret Simpson*, for which we know that the relation `height` is not present in English DBpedia, but it is linked 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 methodology, for the question *List the children of Margaret Thatcher* we substituted the Named Entity *Margaret Thatcher* with the entity *Otto von Bismarck*, obtaining the question *List the children of*

[6] `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/`

Otto von Bismarck. The obtained set of transformed questions is available online.[7]

For some properties (i.e. `Governor`, `Battle`, `FoundationPlace`, `Mission` and `RestingPlace`), no additional links are provided by the French chapter, so we discarded the questions asking for those relations. Out of 27 questions, QAKiS$_{EN+FR}$ correctly answers to 14 questions and partially correctly to 1 questions. To double-check, we run the same set of questions on QAKiS$_{EN}$ (that relies on the English chapter only), and in no cases it was able to detect the correct answer, as expected. This second evaluation did not have the goal to show improved performances of QAKiS$_{EN+FR}$ 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 Table 1).

RELATED WORK

In this paper, we have exploited existing instance and property alignments over DBpedia data to compare and aggregate data from different Wikipedia chapters. These alignments were manually edited –the instance alignment were edited by the Wikipedia community as interlanguage links, and property alignments were edited by the DBpedia community. Questions about alignment techniques, either automated or partially automated are tackled in the broader field of ontology alignment. Rahm and Bernstein [11]; Shvaiko and Euzenat [13] present general surveys on the topic.

Several works tackle the more specific question of data integration from Wikipedia chapters directly from the article content. Rinser et al. [12] provides an overview of instance-based template-attributes matching approaches over multilingual Wikipedia chapters. They also present their own, very thorough approach. First, several criteria are taken into account to improve the instance matching resulting from the inter-language links (i.e. based on this instance alignment, a template alignment is computed according to their use in matched instances). Then, attributes of aligned templates are matched according to the instances and values they relate.

To predict the matching probability of pairs of infobox attribute instances across different language versions, Adar et al. [1] employ self-supervised machine learning with a logistic regression classifier using a broad range of features, such as equality and n-gram/word overlap of attribute keys and values, wiki link overlap, correlation of numerical attributes, and translation-based features. Moreover, Bouma et al. [4] perform a matching of infobox attribute based on instance data. In [3] the same authors describe a system for linking the thesaurus of the Netherlands Institute for Sound and Vision to EnglishWordNet and DBpedia, using EuroWordNet, a multilingual WordNet, and Dutch Wikipedia as intermediaries for the two alignments.

Tacchini et al. [14] provide several strategies for merging data extracted from different chapters of Wikipedia. More specifically, they present a software framework for fusing RDF datasets based on different conflict resolution strategies, and they apply it to fuse infobox data that extracted from multilingual editions of Wikipedia.

Concerning Question Answering, a survey on the QA research field is provided in [9], with a focus on ontology-based QA. Moreover, they examine the potential of the open user friendly interfaces for the SW to support end users in reusing and querying the SW content. State of the art QA systems over Linked Data generally address the issue of question interpretation mapping a natural language question to a triple-based representation. For instance, Freya [6] is an interactive Natural Language Interface for querying ontologies. It uses syntactic parsing in combination with the ontology-based lookup for question interpretation, partly relying on the user's help in selecting the entity that is most appropriate as match for some natural language expression. One of the problem of that approach is that often end-users can be unable to help, in case they are not informed about the modeling and vocabulary of the data. PowerAqua [8] accepts user queries expressed in NL and retrieves answers from multiple semantic sources on the SW. It follows a pipeline architecture, according to which the question is *i)* transformed by the linguistic component into a triple based intermediate format, *ii)* passed to a set of components to identify potentially suitable semantic entities in various ontologies, and then *iii)* the various interpretations produced in different ontologies are merged and ranked for answer retrieval. The major shortcoming of PowerAqua is its limited linguistic coverage.

Pythia [16] relies on a deep linguistic analysis to compositionally construct meaning representations using a vocabulary aligned to the vocabulary of a given ontology. While it can handle linguistically complex questions, Pythia's major drawback is that it requires a lexicon, which up to this moment has to be created manually. It therefore fails to scale to very large data sets. More recently, Unger and colleagues [15] present an approach more similar to the one adopted in QAKiS. Their system (based on Pythia [16]) relies on a linguistic parse of the question to produce a SPARQL template that directly mirrors the internal structure of the question (i.e. a SPARQL template with slots that need to be filled with URIs). This template is then instantiated using statistical entity identification and predicate detection (e.g. applying string similarity as well as natural language patterns extracted from structured data and text documents). However, differently from the other two approaches mentioned before, the last one has not yet been evaluated on the standard data sets of the QALD challenge.

CONCLUSIONS AND FUTURE WORK

The work we have presented here is interdisciplinary with respect to the research fields of Natural Language Processing and the Semantic Web, to enhance interactions between non-expert users and the huge and heterogeneous amount of data available on the Web. More specifically, in this work we have proposed an in-depth comparative analysis of DBpedia multilingual chapters, focusing in particular on the French and the English DBpedia chapters. This showed the fact that their content is complementary: each chapter brings a signifi-

[7] **http://dbpedia.inria.fr/qakis/**

cant amount of data that cannot be found in the other chapter (about half of the data from French DBpedia and 80% of the data from English DBpedia). To perform this comparison, we have first considered the existing alignments and compared the two chapters to highlight their differences. Then, we have proposed an approach to extend the existing properties alignment to all the occurrences of non ambiguous attributes (i.e. attributes that humans have always mapped to the same ontology properties). In this way, we have extended 47 mapping for French DBpedia, with a potential gain of about 30% increase in the number of aligned triples.

Since DBpedia ontology is continuously evolving, maintaining its consistency is a complex task that has to be repeated. Some studies have been carried out to evaluate the quality of DBpedia Ontology, and automatically comparing the values of several chapters as we showed in our work could provide interesting indicators of errors or vandalism in one chapter. Moreover, it could also detect discrepancies among vocabulary used among chapters, or even among topics of the same chapter.

To show the interesting potential for NLP applications resulting from the properties alignment in multilingual DBpedia, we have considered the Question Answering over Linked Data scenario. We have extended the QAKiS system so that it could query the ontology properties of the French DBpedia. We show that this integration extends the system coverage (i.e. the recall), without having a negative impact on its precision.

We plan to extend the presented work in a number of directions. First, we plan to evaluate the contribution of additional multilingual chapters of DBpedia, as for instance calculating the contribution of the French chapters with respect to the sum of the English and German ones, and so on. Moreover, we would like to improve the mapping extension approach by taking into account instance types in order to disambiguate attributes. This should increase the number of mappings that can be extended. We also plan to use alignment tools like Silk[8] to suggest additional property alignments based on the similarity of their use in their respective chapters. For instance, one factor of the similarity measure between two properties could be the number of equivalent pairs (subject, value) that they have in common.

Finally, we plan to take advantage of the analysis provided in this work to implement a mapping assistant that exploits these mappings mining to support the user to edit new mappings. This could also be used to improve consistency among multilingual chapters, and can constitute an help to make property alignments emerge.

REFERENCES

1. Adar, E., Skinner, M., and Weld, D. S. Information arbitrage across multi-lingual wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, ACM (New York, NY, USA, 2009), 94–103.

[8] `http://lod2.eu/Project/Silk.html`

2. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. DBpedia - a crystallization point for the web of data. *Web Semant. 7*, 3 (Sept. 2009), 154–165.
3. Bouma, G. Cross-lingual ontology alignment using eurowordnet and wikipedia. In *LREC*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds., European Language Resources Association (2010).
4. Bouma, G., Duarte, S., and Islam, Z. Cross-lingual alignment and completion of wikipedia templates. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, CLIAWS3 '09, Association for Computational Linguistics (Stroudsburg, PA, USA, 2009), 21–29.
5. Cabrio, E., Cojan, J., Aprosio, A. P., Magnini, B., Lavelli, A., and Gandon, F. Qakis: an open domain qa system based on relational patterns. In *Proceedings of the ISWC 2012 Posters and Demonstrations Track* (Boston, US, November 2012).
6. Damljanovic, D., Agatonovic, M., and Cunningham, H. Freya: an interactive way of querying linked data using natural language. In *Proceedings of the 8th international conference on The Semantic Web*, ESWC'11, Springer-Verlag (Berlin, Heidelberg, 2012), 125–138.
7. Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., and Metakides, G. Internationalization of linked data: The case of the greek dbpedia edition. *Web Semantics: Science, Services and Agents on the World Wide Web 15*, 0 (2012), 51 – 61.
8. Lopez, V., Uren, V. S., Sabou, M., and Motta, E. Cross ontology query answering on the semantic web: an initial evaluation. In *K-CAP* (2009), 17–24.
9. Lopez, V., Uren, V. S., Sabou, M., and Motta, E. Is question answering fit for the semantic web?: A survey. *Semantic Web 2*, 2 (2011), 125–155.
10. Mahendra, R., Wanzare, L., Bernardi, R., Lavelli, A., and Magnini, B. Acquiring relational patterns from wikipedia: A case study. In *Proceedings of the 5th Language and Technology Conference* (Poznan, Poland, 2011).
11. Rahm, E., and Bernstein, P. A. A survey of approaches to automatic schema matching. *The VLDB Journal 10*, 4 (Dec. 2001), 334–350.
12. Rinser, D., Lange, D., and Naumann, F. Cross-lingual entity matching and infobox alignment in wikipedia. *Information Systems*, 0 (2012), –.
13. Shvaiko, P., and Euzenat, J. Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng. 25*, 1 (2013), 158–176.
14. Tacchini, E., Schultz, A., and Bizer, C. Experiments with wikipedia cross-language data fusion. S. Auer, C. Bizer, and G. A. Grimnes, Eds., vol. 449 of *CEUR Workshop Proceedings ISSN 1613-0073* (June 2009).

15. Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.-C., Gerber, D., and Cimiano, P. Template-based question answering over rdf data. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, ACM (New York, NY, USA, 2012), 639–648.

16. Unger, C., and Cimiano, P. Pythia: Compositional meaning construction for ontology-based question answering on the semantic web. In *NLDB* (2011), 153–160.