Classifying Inconsistencies in DBpedia Language Specific Chapters

Elena Cabrio, Serena Villata, Fabien Gandon
INRIA
2004 Route des Lucioles BP93, 06902 Sophia Antipolis, France
firstname.lastname@inria.fr

Abstract

This paper proposes a methodology to identify and classify the semantic relations holding among the possible different answers obtained for a certain query on DBpedia language specific chapters. The goal is to reconcile information provided by language specific DBpedia chapters to obtain a consistent results set. Starting from the identified semantic relations between two pieces of information, we further classify them as positive or negative, and we exploit bipolar abstract argumentation to represent the result set as a unique graph, where using argumentation semantics we are able to detect the (possible multiple) consistent sets of elements of the query result. We experimented with the proposed methodology over a sample of triples extracted from 10 DBpedia ontology properties. We define the LingRel ontology to represent how the extracted information from different chapters is related to each other, and we map the properties of the LingRel ontology to the properties of the SIOC-Argumentation ontology to built argumentation graphs. The result is a pilot resource that can be profitably used both to train and to evaluate NLP applications querying linked data in detecting the semantic relations among the extracted values, in order to output consistent information sets.

Keywords: DBpedia language specific chapters, inconsistencies detection, bipolar argumentation

1. Introduction

The Web is an information space which is evolving from sharing textual documents to publishing structured data. In particular, the Linked Data initiative aims at fostering the publication and interlinking of data on the Web, giving birth to the so called *Web of Data*, a dataspace where structured data are interlinked with each other. The best known example of Linked Data is DBpedia (Bizer et al., 2009), a structured information dataset extracted from Wikipedia. The advantage brought by the availability of such information in a structured format (usually expressed in RDF[1]) is that in this way sophisticated queries can be raised against Wikipedia data to extract further information (e.g., using the SPARQL query language[2]), and new links to different datasets on the Web may be exploited. DBpedia presents several language specific chapters[3] extracted from the language specific Wikipedia chapters, where information is presented in a structured format. DBpedia chapters, well connected through Wikipedia instance interlinking, can contain different information from one language to another. In particular, language specific chapters can provide *i)* more details about certain topics, *ii)* fill information gaps (provide information missed in the other DBpedia chapters), or iii) conceptualize certain relations according to a specific cultural viewpoint (Rinser et al., 2012). If on the one side relying on information contained in language specific DBpedia chapters in Natural Language Processing (NLP) interfaces and applications (e.g., NL question answering systems over Linked Data) would bring promising improvements on systems coverage, on the other side they face the following problems: *i)* different results can be obtained for the same query (e.g., English and French DBpedia chapters provide a different date of birth of Louis IX of France), and *ii)* the combination of these query results may lead to inconsistent information sets about the same topic.

To better understand the problem, let us consider for instance a question answering system that takes as input NL questions and queries two (or more) language specific chapters of DBpedia for answer retrieval (e.g. QAKiS[4] (Cabrio et al., 2013a)). Ideally, we would expect that each DBpedia endpoint returns its own answer and all these answers are identical. However, this is not always the case, and three possible situations arise, illustrated in Figure 1 (Cojan et al., 2013): *i)* the same result is provided from both chapters (i.e., both datasets contain the same value for the same relation), *ii)* two different results are provided (i.e., the datasets contain different values for the same relation), *iii)* the result is provided from one chapter only (only one dataset provides the value for the queried relation). Considering the two DBpedia biggest chapters, i.e. English and French ones, on the total of all the properties used in the two chapters, 11.4% of the properties have the same value in the two chapters (∼343,000), 5.8% have different values in the two chapters (∼175,000), 67.3% have a value only in the English chapter (∼2,022,500), and 15.5% has a value only in the French chapter (∼460,500). If the results are the same, the answer to be returned to the user is precisely this result set, but if the results are different then a specific management of the query result has to be done before returning the answer to the user.

While several initiatives are proposed to push DBpedia contributors into extending the property mapping to DBpedia common ontology, in this paper we focus on case *(ii)*, and we address the following research question:

- How to detect the inconsistencies arising from merging information from DBpedia language specific chapters?

In particular, the research question breaks down into the

[1] http://www.w3.org/RDF/
[2] http://www.w3.org/TR/rdf-sparql-query/
[3] http://wiki.dbpedia.org/Internationalization/Chapters
[4] http://qakis.org/qakis2/

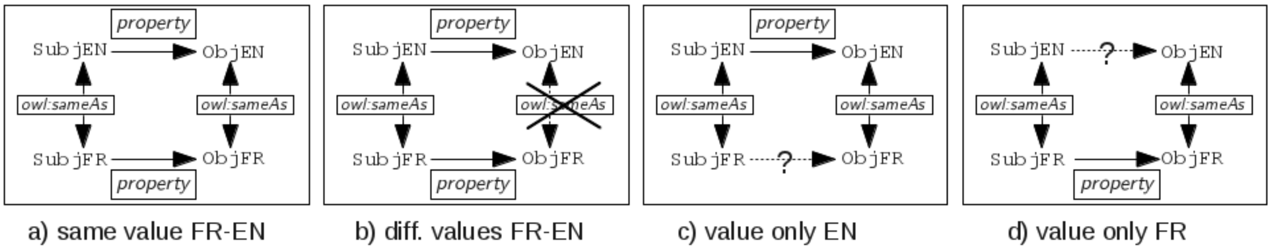

Figure 1: Possible outcomes of the values comparison among DBpedia language specific chapters.

following sub-questions:

1. Which kind of semantic relations can be established to reconcile information provided by language specific DBpedia chapters to obtain a consistent results set?

2. How to compute in an automated way consistent results sets starting from the identified relations?

To answer the research questions, we propose a data-driven approach to identify and classify the semantic relations holding among the possible different answers obtained for a certain query on DBpedia language specific chapters. We consider 17 DBpedia ontology properties, and we collect the data from the different multilingual chapters of DBpedia concerning such properties. We further identify which kind of relations hold between the query results items. In order to provide a machine-readable version of our dataset, we introduce the LingRel ontology, a lightweight vocabulary composed of properties only, where the relations we identified are formally represented.
Starting from the identified semantic relations between two pieces of information, we further classify them as *positive* or *negative*, and we exploit bipolar abstract argumentation (Dung, 1995; Cayrol and Lagasquie-Schiex, 2011) to represent the result set as a unique graph, such that argumentation semantics can be adopted to detect the (possible multiple) consistent sets of elements of the query result. Only this consistent set of information items has to be returned to the user. Again, in order to provide such argumentation graphs in a machine-readable format, we adopt the extended SIOC Argumentation vocabulary[5] proposed by Cabrio et al. (2013c) where two kinds of relations among the nodes of the graph hold, i.e., a *support* relation and an *attack* relation.
The remainder of the paper is as follows. We first present our classification of the relations holding between the values we retrieved from the results set over different language specific DBpedia chapters (Section 2.), and second we adopt argumentation theory to detect the inconsistencies arising in the results set of the query (Section 3.). The feasibility study of our data-driven approach is detailed in Section 4. Section 5. compares the proposed approach with other approaches adopting argumentation theory for ontologies alignment and communities management. Finally, some conclusions are drawn.

[5] `http://bit.ly/SIOC-Argumentation`

2. Inconsistencies classification

The first step of our approach consists in identifying which kind of semantic relation holds among the different values obtained by querying a set of language specific DBpedia chapters with a certain query. More precisely, we want to detect the linguistic phenomena holding among the inconsistent values obtained querying two DBpedia language specific chapters at a time, given a certain subject and a certain ontological property.
In the following, we propose our classification of the linguistic phenomena we detected, defined referring to widely accepted linguistic categories in the literature. We identify the following positive relations between values (mainly discourse and lexical semantics relations):

a) *identity*, i.e. same value but in different languages (missing `SameAs link` in DBpedia)
E.g., `Dairy product` vs `Produits laitiers` (property: `product`); `Yevgeny Kaspersky` vs `Eugene Kaspersky` (property: `owner`)

b) *acronym*, i.e. initial components in a phrase or a word
E.g., `PSDB` vs `Partito della Social Democrazia Brasiliana` (property: `leaderParty`); `RFID` vs `Radio-frequency identification` (property: `product`)

c) *disambiguated entity*, i.e. one value contains in the name the class to which the entity belongs
E.g., `Michael Lewis (Author)` vs `Michael Lewis` (property: `author`); `David Williams (oil baron)` vs `David Williams` (property: `fundedBy`)

d) *identity:stage name*, i.e. pen/stage names pointing to the same entity
E.g., `Lemony Snicket` vs `Daniel Handler` (property: `author`); `The Wachowskis` vs `Andy et Larry Wachowski` (property: `director`)

e) *coreference*, i.e. an expression referring to another expression describing the same thing (in particular, non normalized expressions)
E.g., `William Burroughs` vs `William S. Burroughs` (property: `author`); `Arup` vs `Arup Associates` (property: `architect`)

f) *renaming*, i.e. reformulation of the same entity name in time
E.g., `Charleville (Ardennes)`, old name of

Charleville-Mézières (property: birthPlace); Edo, old name of Tokyo (property: birthPlace)

g) *metonymy*, i.e. a thing/concept not called by its own name, but by the name of something intimately associated with that thing/concept
E.g., Joseph Hanna vs Hanna-Barbera (property: author); Hamburg vs FC St. Pauli (property: owner)

h) *geo-specification*, i.e. ontological geographical knowledge
E.g., Castelnau-Barbarens vs France (property: birthPlace); Queensland vs Australia (property: spokenIn)

i) *meronymy*, i.e. a constituent part of, or a member of something
E.g., Justicialist Party is a part of Front for Victory (property: leaderParty); Bob Marley was a part of The Wailers (property: fundedBy)

j) *hyponymy*, i.e. relation between a specific word and a general word when the former is included within the latter
E.g., alluminio vs metal (property: product); sport car vs autovettura (property: product)

Moreover, we identify the following negative relations (i.e., values mismatches):

k) *text mismatch*, i.e. unrelated entity/thing
E.g., Palermo vs Modene (property: owner); Yves Allegret vs Marc Allegret (property: director)

l) *date mismatch* , i.e. different date for the same event
E.g., 1215-04-25 vs 1214-04-25 (property: birthDate); 1397-08-10 vs 1397-08-16 (property: birthDate)

m) *numerical mismatch*, i.e. different numberical values
E.g., 1.91 vs 1.8 (property: height); 1000 vs 1200 (property: numberOfStudents)

3. Mapping Inconsistencies to Bipolar Argumentation Graphs

To address the problem of inconsistencies detection and reconciliation of information coming from different language specific DBpedia chapters, the second step of our approach consists in the transformation of the detected linguistic phenomena into argumentation-based relations. In particular, we map positive relations (categories **a** to **j**) and negative relations (**k** to **m**) respectively with the *support* and *conflict relations* in bipolar argumentation theory.
An abstract argumentation framework (AAF) (Dung, 1995) aims at representing conflicts among elements called *arguments*. An argumentation framework is based on a binary *attack* relation among arguments, whose role is determined only by their relation with the other arguments.

Definition 1 *An abstract argumentation framework (AAF) is a tuple $\langle A, \rightarrow \rangle$ where A is a finite set of arguments and $\rightarrow$ is a binary attack relation on A.*

An AF encodes, through the conflict (i.e., *attack*) relation, the existing conflicts within a set of arguments. It is then interesting to identify the conflict outcomes, which, roughly speaking, means determining which arguments should be accepted ("they survive the conflict") and which arguments should be rejected, according to some reasonable criterion. (Dung, 1995) presents several acceptability semantics that produce zero, one, or several *consistent* sets of accepted arguments. The set of accepted arguments of an argumentation framework consists of a set of arguments that does not contain an argument conflicting with another argument in the set. This set is called a *conflict free* set of arguments. Following from this notion, an *admissible* set of arguments is required to be both internally coherent (*conflict-free*) and able to defend its elements. In our framework, we adopt admissibility based semantics. Roughly, an argument is accepted if all the arguments attacking it are rejected, and it is rejected if there is at least an argument attacking it which is accepted. The (possibly multiple) sets of accepted arguments computed using one of the acceptability semantics are called *extensions*, and the addition of another argument from outside the set will make the set *inconsistent*. For more details about acceptability semantics, see (Dung, 1995). An example of abstract argumentation framework is visualized in Figure 2.a.
Since we represent in the argumentation graph also positive relations among the arguments, we adopt bipolar argumentation theory (Cayrol and Lagasquie-Schiex, 2011) where in addition to the conflict relation, we have a binary support relation among arguments.

Definition 2 *A bipolar argumentation framework is a tuple $\langle A, \rightarrow, \dashrightarrow \rangle$ where A is a finite set of arguments, $\rightarrow \subseteq A \times A$, and $\dashrightarrow$ is a binary relation called support defined on $A \times A$.*

An example of bipolar argumentation framework is visualized in Figure 2.b where the dotted edge represents the support relation.

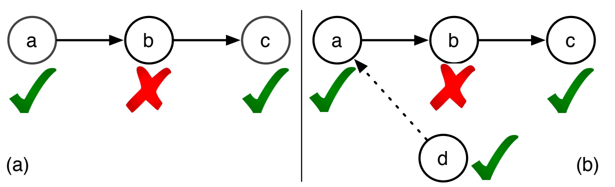

Figure 2: Example of (a) an abstract argumentation framework, and (b) a bipolar argumentation framework.

Bipolar argumentation has been exploited by (Cabrio et al., 2013b) to detect inconsistencies among the answers provided by a QA system over linked data. However, in this work only the subsumption and the identity relations were considered as positive relations among the arguments, and no distinction was provided w.r.t. the different kinds of conflicts.
Among the positive relations, some of them (e.g., identity, coreference) are bidirectional which means that they are translated in the AAF as a bidirectional support relation. Some others are directional (e.g., meronymy, hyponymy)

ontology properties	*# annotated triples*	*# annotated positive relations*										*# annotated negative rel.*		
		a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**
`architect`	21	10	-	1	-	4	-	1	-	-	-	5	-	-
`author`	25	8	-	5	5	2	-	2	-	1	-	2	-	-
`birthDate`	25	-	-	-	-	-	-	-	-	-	-	-	25	-
`birthPlace`	25	2	-	-	-	1	2	-	19	-	-	1	-	-
`capital`	25	6	-	7	-	1	-	-	9	-	-	2	-	-
`deathDate`	25	-	-	-	-	-	-	-	-	-	-	-	25	-
`director`	25	9	-	7	1	4	-	-	-	1	-	3	-	-
`foundedBy`	25	4	-	6	-	2	-	-	-	4	1	8	-	-
`height`	25	3	-	-	-	-	-	-	-	-	-	-	-	22
`leaderParty`	25	5	8	2	-	-	2	-	-	1	-	7	-	-
`nationality`	16	1	-	-	-	12	1	-	1	1	-	-	-	-
`numberOfStudents`	25	-	-	-	-	-	-	-	-	-	-	-	-	25
`owner`	25	4	1	-	-	1	-	6	5	5	-	3	-	-
`product`	25	4	1	-	-	-	-	1	-	1	15	3	-	-
`recordLabel`	25	-	-	-	-	1	-	1	-	10	-	13	-	-
`sourceCountry`	25	3	1	-	-	-	-	-	20	-	-	1	-	-
`spokenIn`	13	2	-	-	-	-	1	-	9	-	-	1	-	-
total	400	61	11	28	6	28	6	11	63	24	16	49	50	47

Table 1: Results of the pilot annotation phase

and we map them as a directional support relation in the argumentation graph. On the contrary, all mismatches are always bidirectional conflicts among the arguments. Figure 3 shows how the examples of *date mismatch* and *meronymy* are represented as bipolar argumentation graphs. In the first case, given that the two arguments attack each other, they are evaluated as *undecided* (e.g., using preferred semantics). The idea is that either we accept one of them rejecting the other, or we evaluate them both as undecided as they contradict each other. In the second case, the directional support relation from the meronymy relation is from the more precise argument `Justicialist Party` to the more general one `Front for Victory`. Both the arguments are accepted, as there are no conflicts between them.

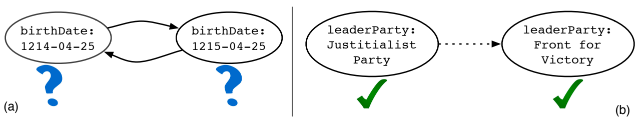

Figure 3: Example of (a) *date mismatch*, and (b) *meronymy*.

The reader may argue that argumentation theory is too much powerful (i.e., too much complicated) for the task of detecting an inconsistency, as in the case of *date mismatch*. For instance, just looking at the kind of properties involved, i.e., functional properties like `birthDate`, would lead to the identification of an inconsistency. However, it must be noticed that when the query is raised against 10 or more DBpedia chapters, then the result is a bipolar argumentation graph where there is a node for each result item provided by each chapter. The idea of using argumentation is that in this way we are able to *solve* the inconsistency and know what is the right result. Consider for instance, a case where we query 5 different DBpedia multilingual chapters: two of them return the values for `birthDate` in Figure 3, and the remaining three return the value `1215-04-25`. These values will support each others, leading to the final choice of the result `1215-04-25`, as the only consistent result to be proposed to the user.

4. Feasibility study

In the following, we describe the ongoing work of creation of a dataset of values extracted from DBpedia language specific chapters, that we annotate with the semantic relations holding among pairs of values, so that to highlight possible inconsistencies. More specifically, in Section 4.1. we detail how we extracted the DBpedia triples to be annotated with the linguistic categories listed in Section 2. In Section 4.2., we present the data translation and the LingRel vocabulary we designed to describe the data. In Section 4.3. we build the argumentation graphs by adopting the SIOC-Argumentation extended ontology.

4.1. Dataset annotation

To assess the feasibility of the proposed approach to build a dataset of possibly inconsistent information, we randomly selected a sample of 17 DBpedia ontology properties (Table 2., column *ontology properties*). We then query English, French and Italian DBpedia chapters in such a way to extract a set of RDF triples (subject, property, object) where the subject and the property are the same, but with a different object or datatype value. More specifically, for each property of the DBpedia ontology we extract 25 triples providing possibly inconsistent values (13 for the property `spokenIn`, 16 for property `nationality`, and 21 for the property `architect`) in at least two DBpedia chapters (e.g. English *vs* French chapter, or English *vs* Italian chapter). We narrow the scope of our search to the instances (subjects) that have only one value for the property under observation.[6] An annotator with skills in linguistics annotated then the relations among the couples of extracted values, applying the category labels proposed in Section 2. Examples 1 and 2 show the XML annotation format of the pairs in the dataset. The information conveyed by Example 1 is that there exists a relation of *metonymy* between the entities *Joseph Barbera* and *Hanna-Barbera*, indicated as `authors` of the show The Ruff & Reddy Show, respectively from the English and the Italian DBpedia chapters. The information conveyed by Example 2 is that there exists

[6] The underlying reason is that in this work we focus on possibly inconsistent information, and not on incomplete lists of values.

a relation of *geo-specification* between the entities *Londra* (i.e. London) and *Camden Town*, indicated as `bithPlace` of the English actor Anthony Head, respectively from the English and the Italian DBpedia chapters.

For each pair in the dataset (each containing therefore a triple with a certain subject and a certain property, and for which the object assumes different values in the language specific DBpedia chapters), the subject of the triple is expressed as the value of the attribute *dbpediasubj*, the property as the value of the attribute *dpediaontologyproperty*. The different values of the object are contained into the elements *value1* and *value2*, respectively. For each value, the DBpedia chapter of provenance is reported (attribute *dbpediachapter*). The relation holding between the two values is reported as attribute of the element *pair* (i.e. *relation*).

Example 1

```
<pair id="9" relation="METONYMY"
dpediaontologyproperty="author"
dbpediasubj="The_Ruff_&_Reddy_Show">
    <value1 dbpediachapter="EN">
    Joseph_Barbera
    </value1>
    <value2 dbpediachapter="IT">
    Hanna-Barbera
    </value2>
</pair>
```

Example 2

```
<pair id="55" relation="GEO-SPECIFICATION"
dpediaontologyproperty="birthPlace"
dbpediasubj="Anthony_Head">
    <value1 dbpediachapter="IT">
    Londra
    </value1>
    <value2 dbpediachapter="EN">
    Camden_Town
    </value2>
</pair>
```

To assess the validity of the annotation task, we calculate the inter-annotator agreement. Another annotator with skills in linguistics has therefore independently annotated a sample of 30 pairs of the data set. While the percentage of agreement between the two annotators is 72%, Fixed-marginal kappa is 0.69. Overall, we consider this value high enough to demonstrate the stability of the categories of the proposed set of semantic relations, thus validating their classification model. Most of the disagreements between annotators were due to misunderstandings in the definition of the categories *identity* and *coreference*, whose distinction can be not very straightforward at first sight.

Table 2. reports on the results of this pilot annotation, as obtained after the reconciliation phase. The resource is currently composed of 400 annotated pairs of values, where the actual distribution of linguistic phenomena holding among possibly inconsistent values in DBpedia chapters emerges. A number of remarks can be made on the data presented in Table 2. First of all, a distinction exists between datatype properties (properties for which the value is a data literal) and object properties (properties for which the value is an individual). For the former properties (e.g. `birthDate`, `numberOfStudents`), a different value in the range identifies always an inconsistent result sets. Possible rules could then be chosen to reconcile such information when possible, for instance defining a tolerance value (applicable to properties such as *numberOfStudents*, but not applicable to functional properties like `birthDate`).

For object properties, we can notice recurrent types of relations occurring between possibly inconsistent values, depending on the class of the range of the property. For instance, for properties as `birthDate` and `sourceCountry`, whose range is of class *place* (or of semantically similar classes as *country* or *PopulatedPlace*), the most recurrent cause of values mismatch is the fact that one DBpedia chapter contains a more fine-grained location with respect to the other chapter (that we call *geo-specification*), as showed for instance in Example 2. The classes of such possibly inconsistent values should therefore be taken into consideration when possible, so that to find an adequate knowledge base supporting a system in reconciling such information (e.g. GeoNames[7] for geographical entities). For more "standard" lexical semantics relations (as *hyponymy*, or *meronymy*) lexical databases such as WordNet (and its multilingual variants EuroWordNet or MultiWordNet) can contribute to the fulfillment of this purpose. The number of pairs labeled with the relation *identity* highlights the fact that the SameAs links between the same instance in different DBpedia chapters are often missing, and that an effort is required by the DBpedia community to fill this gap. Finally, as expected, relations as `renaming` and `identity:stage name` are less frequent than the others in the data, but we consider it important to include them in our classification, since they capture two relations among possibly inconsistent values that were not captured by the other relations. Note that the list of semantic relations we propose in this work is not exhaustive, and reflects the types of relations we detected in the sample of DBpedia triples we analyzed.

4.2. LingRel vocabulary

In order to describe the data using the semantic properties we highlighted in Section 2., we designed a new lightweight vocabulary called LingRel[8]. In particular, LingRel models the ten positive semantic relations we identified on the data. We need to design our own ontology since we could not find another ontology actually modeling the properties we required. However, in the light of the Web of Data philosophy, the properties of LingRel have been aligned with other vocabularies like SKOS, Ontology Patterns, when possible. The properties and their alignment are presented below (XML/RDF format).

```
<owl:ObjectProperty rdf:ID="identity">
<rdfs:label xml:lang="en">identity</rdfs:label>
<rdfs:comment xml:lang="en">The property defines an
identity relation between two elements.</rdfs:comment>
<skos:exactMatch rdf:resource="http://www.w3.org/2002/07/
owl#sameAs"/>
<skos:exactMatch rdf:resource="http://www.w3.org/2004/02/
skos/core#exactMatch"/>
</owl:ObjectProperty>
```

[7] http://www.geonames.org/

[8] http://ns.inria.fr/lingRel/v1

```
<owl:ObjectProperty rdf:ID="acronym">
<rdfs:label xml:lang="en">acronym</rdfs:label>
<rdfs:comment xml:lang="en">The property links an element
to another element representing its abbreviation, possibly
formed from the initial letters of the first element.
</rdfs:comment>
<skos:related rdf:resource="http://purl.org/cerif/frapo/
hasAcronym"/>
<skos:related rdf:resource="http://rdfs.org/scot/
ns#acronym"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="disambiguated_identity">
<rdfs:label xml:lang="en">disambiguated identity
</rdfs:label>
<rdfs:comment xml:lang="en">The property defines an iden-
tity relation between two elements but specifying the
distinctive features.</rdfs:comment>
<skos:related rdf:resource="http://www.w3.org/2002/07/
owl#sameAs"/>
<skos:related rdf:resource="http://www.w3.org/2004/02/
skos/core#exactMatch"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="stage_name">
<rdfs:label xml:lang="en">stage name</rdfs:label>
<rdfs:comment xml:lang="en">The property links an element
to another element representing its stage name.
</rdfs:comment>
<rdfs:subPropertyOf rdf:resource="http://ns.inria.fr/
lingRel/v1#identity"/>
<skos:related rdf:resource="http://www.w3.org/2002/07/
owl#sameAs"/>
<skos:related rdf:resource="http://www.w3.org/2004/02/
skos/core#exactMatch"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="coreference">
<rdfs:label xml:lang="en">coreference</rdfs:label>
<rdfs:comment xml:lang="en">The property links two ele-
ments having the same reference.</rdfs:comment>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="renaming">
<rdfs:label xml:lang="en">renaming</rdfs:label>
<rdfs:comment xml:lang="en">The property links an element
to the another element that is the same element renamed.
</rdfs:comment>
<skos:related rdf:resource="http://www.w3.org/2002/07/
owl#sameAs"/>
<skos:related rdf:resource="http://www.w3.org/2004/02/
skos/core#exactMatch"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="metonymy">
<rdfs:label xml:lang="en">metonymy</rdfs:label>
<rdfs:comment xml:lang="en">The property defines the
substitution of the name of an element for that of the
element meant.</rdfs:comment>
<skos:related rdf:resource="http://www.ontologydesign
patterns.org/ont/lmm/LMM_L1.owl#hasInterpretant"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="geospecification">
<rdfs:label xml:lang="en">geo-specification</rdfs:label>
<rdfs:comment xml:lang="en">The property defines the
geographical specification of an element by another
element.</rdfs:comment>
<rdfs:subPropertyOf rdf:resource="http://ns.inria.fr/
lingRel/v1#meronymy"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="meronymy">
<rdfs:label xml:lang="en">meronymy</rdfs:label>
<rdfs:comment xml:lang="en">The property defines an ele-
ment that denotes part of another element but which is
used to refer  to the whole of it.</rdfs:comment>
<skos:exactMatch rdf:resource="http://purl.org/lingui-
stics/gold/meronym"/>
<rdfs:subPropertyOf rdf:resource="http://www.w3.org/2004/
02/skos/core#narrower"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="hyponymy">
<rdfs:label xml:lang="en">hyponymy</rdfs:label>
<rdfs:comment xml:lang="en">The property links an element
of more specific meaning to a general or superordinate
element applicable to it.
</rdfs:comment>
<skos:exactMatch rdf:resource="rdfs:subClassOf"/>
<rdfs:subPropertyOf rdf:resource="http://www.w3.org/2004/
02/skos/core#narrower"/>
</owl:ObjectProperty>
```

Negative relations, i.e., *text mismatch*, *date mismatch*, and *numerical mismatch*, are more subtle and cannot be represented in the same way in the vocabulary. These relations have been detected in the data, and they characterize the possible kinds of inconsistencies. These inconsistencies arise due to functional properties that present more than one value, i.e., for birth dates, height, etc. In order to include also such data in the machine-readable dataset we create using LingRel, we introduce in the vocabulary also a “negative” relation called `mismatch`, that we use to annotated the data when one of the possible mismatches we identified arises.

```
<owl:ObjectProperty rdf:ID="mismatch">
<rdfs:label xml:lang="en">mismatch</rdfs:label>
<rdfs:comment xml:lang="en">The property links two ele-
ments resulting from a functional property but whose
values are different.</rdfs:comment>
<skos:related rdf:resource="owl:FunctionalProperty"/>
<rdfs:isDefinedBy rdf:resource="http://ns.inria.fr/
lingRel/v1#"/>
</owl:ObjectProperty>
```

We create a machine-readable dataset using the LingRel vocabulary. The dataset is available at `http://www-sop.inria.fr/members/Elena.Cabrio/resources.html`. An example of *stage name* relation annotated using the LingRel vocabulary is visualized below.

```
<http://dbpedia.org/resource/madeleine_wickham>
                                   lingRel:stage_name
<http://it.dbpedia.org/resource/sophie_kinsella> .
```

4.3. SIOC-Argumentation extended vocabulary

We then build the argumentation graphs by adopting the SIOC-Argumentation extended ontology[9] provided in (Cabrio et al., 2013d). The main classes of the vocabulary are highlighted below

```
<rdf:Description rdf:about="http://rdfs.org/sioc/argu-
ment#challengesArg">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/
    owl#ObjectProperty"/>
    <rdfs:domain rdf:resource="http://rdfs.org/sioc/ar-
    gument#Argument"/>
    <rdfs:range rdf:resource="http://rdfs.org/sioc/argu-
    ment#Argument"/>
    <dc:description>Expresses that an argument challenges
    another argument</dc:description>
  </rdf:Description>

  <rdf:Description rdf:about="http://rdfs.org/sioc/argu-
  ment#supportsArg">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/
    owl#ObjectProperty"/>
    <rdfs:domain rdf:resource="http://rdfs.org/sioc/ar-
    gument#Argument"/>
    <rdfs:range rdf:resource="http://rdfs.org/sioc/argu-
    ment#Argument"/>
    <dc:description>Expresses that an argument supports
    another argument</dc:description>
  </rdf:Description>
```

We generated a dataset[10] in the RDF format using such extension of the SIOC Argumentation vocabulary such that

[9] `http://bit.ly/SIOC-Argumentation`

[10] Available in both XML and RDF formats at `http://www-sop.inria.fr/members/Elena.Cabrio/resources.html`

the information items we retrieved from the different chapters are linked by a support or a conflict relation (expressed through the two properties defined in the ontology). The ten positive relations of LingRel are mapped with the support relation in SIOC-Argumentation, and the mismatch relation with the challenges (i.e., attack) relation in SIOC-Argumentation. The final aim of such pilot study is that such dataset can be queried using SPARQL to discover further insightful information. An example of data extracted from the RDF dataset annotated with the extended SIOC Argumentation vocabulary is provided below, where the semantic relation *stage name* is translated into a support relation in the argumentation perspective.

```
<http://dbpedia.org/resource/madeleine_wickham>
                                    a sioc_arg:Argument ;
<http://it.dbpedia.org/resource/sophie_kinsella>
                                    a sioc_arg:Argument ;
<http://dbpedia.org/resource/madeleine_wickham>
                                    sioc_arg:supportsArg
<http://it.dbpedia.org/resource/sophie_kinsella> .
```

5. Related Work

Several works address alignment agreement based on argumentation theory. More precisely, Laera et al. (Laera et al., 2007) address alignment agreement relying on argumentation to deal with the arguments which attack or support the candidate correspondences among ontologies. Doran et al. (Doran et al., 2009) propose a methodology to identify subparts of ontologies which are evaluated as sufficient for reaching an agreement, before the argumentation step takes place, and dos Santos and Euzenat (dos Santos and Euzenat, 2010) present a model for detecting inconsistencies in the selected sets of correspondences relating ontologies. In particular, the model detects logical and argumentation inconsistency to avoid inconsistencies in the agreed alignment. The framework we propose has common points with this line of works, i.e., the use of argumentation theory to select a consistent set of information items, but the scenario in which the two approaches are exploited is different and this leads to a different addressed issues and proposed solutions. Another framework combining argumentation theory and NLP in the Semantic Web has been proposed by Cabrio et al. (Cabrio et al., 2013c). The authors propose an automatic framework to support the management of argumentative discussions in wiki-like platforms. In particular, a natural language module is proposed to detect the arguments in natural language returning the relations among them, and argumentation theory provides the overall view of the argumentative discussion highlighting the accepted arguments. The arguments and their relations are then translated in RDF using the SIOC Argumentation vocabulary to be queried using SPARQL in order to retrieve additional information. In this paper, the arguments are represented by the results of a query over the multilingual chapters of DBpedia, and we use the semantic relations among the information items to provide the relations among the arguments, and then compute the accepted ones.

6. Conclusions and perspectives

The pilot study we have described is a first step toward the creation of a large-scale annotated dataset of DBpedia language specific chapters inconsistencies. We have extracted 400 pairs of values for the 17 DBpedia ontology relations we selected, and we identified ten positive relations holding between such values extracted from the multilingual chapters. Finally, we mapped the relations with the support and attack relations in argumentation, such that argumentation graphs are built to further support inconsistencies detection and discovering information thanks to the machine-readable argumentation graphs. Such resource can be exploited both to train and to evaluate NLP applications querying linked data in detecting the semantic relations among the extracted values, to output consistent information sets (lexical semantics resources can for instance be integrated into such systems to turn the inconsistencies belonging to categories **a** to **j** into support relations). Moreover, pointing out mismatches (categories **k** to **m**) in the queried data improves the systems reliability.
As for future work, we plan to exploit the outcomes of the study presented here to improve our automatic framework to detect the inconsistencies which may arise into a set of answers provided by a QA system over linked data. In our first attempt (Cabrio et al., 2013b), only the subsumption and the identity relations were considered as positive relations among the arguments, and no distinction was provided w.r.t. the different kinds of conflicts. We plan therefore to investigate how to deal with more fine-grained relations among possibly inconsistent answers, and to propose adhoc strategies to reconcile such information.

7. Acknowledgements

We thank Julien Cojan for providing us the statistics about DBpedia language specific chapters. The work of Elena Cabrio was funded by the French Government (National Research Agency, ANR) through the "Investments for the Future" Program (reference # ANR-11-LABX-0031-01).

8. References

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165, September.

Cabrio, E., Cojan, J., Gandon, F., , and Hallili, A. (2013a). Querying multilingual DBpedia with QAKiS. In *Procs of ESWC 2013. Demo papers.*

Cabrio, E., Cojan, J., Villata, S., and Gandon, F. (2013b). Argumentation-based inconsistencies detection for question-answering over dbpedia. In *Proceedings of the ISWC 2013 workshop NLP & DBpedia.*

Cabrio, E., Villata, S., and Gandon, F. (2013c). A support framework for argumentative discussions management in the web. In *Proceedings of the 10th Extended Semantic Web Conference (ESWC 2013).*

Cabrio, E., Villata, S., and Gandon, F. (2013d). A support framework for argumentative discussions management in the web. In Cimiano, P., Corcho, Ó., Presutti, V., Hollink, L., and Rudolph, S., editors, *ESWC*, volume 7882 of *Lecture Notes in Computer Science*, pages 412–426. Springer.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2011). Bipolarity in argumentation graphs: Towards a better understanding. In *Procs of SUM 2011*, LNCS, v. 6929, pages 137–148.

Cojan, J., Cabrio, E., and Gandon, F. (2013). Filling the gaps among DBpedia multilingual chapters for question answering. In *Proceedings of ACM Web Science 2013*.

Doran, P., Tamma, V., Palmisano, I., and Payne, T. (2009). Efficient argumentation over ontology correspondences. In *Procs of the 8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 1241–1242.

dos Santos, C. T. and Euzenat, J. (2010). Consistency-driven argumentation for alignment agreement. In *Proceedings of the 5th International Workshop on Ontology Matching (OM 2010), CEUR Workshop Proceedings 689*.

Dung, P. (1995). On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358.

Laera, L., Blacoe, I., Tamma, V., Payne, T., Euzenat, J., and Bench-Capon, T. (2007). Argumentation over ontology correspondences in MAS. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007)*, pages 1–8.

Rinser, D., Lange, D., and Naumann, F. (2012). Cross-lingual entity matching and infobox alignment in wikipedia. *Information Systems*.