Enriching Ontologies with Linguistic Content: an Evaluation Framework

Alessandro Oltramari, Armando Stellato

Laboratory for Applied Ontology (ISTC-CNR), University of Rome, Tor Vergata
Trento, Rome
oltramari@loa-cnr.it, stellato@info.uniroma2.it

Abstract

In this paper, we present a framework for representing and evaluating integrations between ontological and linguistic resources, which originates and improves previous research reported in (Pazienza & Stellato, 2006b; Pazienza, Sguera, & Stellato, 2007) and articulates into two results: first, a set of coordinated RDF vocabularies providing descriptors for representing linguistic resources and their software counterparts, as well as offering metadata for describing the linguistic enrichment of ontologies, both on quantitative and qualitative grounds. The second result is a software library for evaluating the quality of automatic linguistic enrichment tools.

The Linguistic Watermark suite of RDF vocabularies, in the newly presented form, provides to our framework shared vocabularies for addressing the knowledge about heterogeneous linguistic resources, for accessing and managing their content on a common basis through dedicated software components and for representing the integration of this content inside ontologies. This last part constitutes the bridge towards our novel evaluation framework, which produces quality reports based on assessed evaluation metrics taken from the Information Retrieval tradition (Van Rijsbergen, 1975) and adapted to this task. We hope that this framework could provide a stable and reusable tool for evaluating the quality of competing algorithmic solutions for linguistic enrichment of ontologies.

1. Introduction

Semantic Web ontologies represent the shared vocabularies through which machines can read and access content from the Web, or even communicate between them, to exchange information or cooperate for achieving some goal. This definition implicitly assumes that in an heterogeneous scenario like the whole WWW, the same concepts will be represented by the same ontologies and that, therefore, ontological models of data will be consistent; conversely, sensible effort will be put in trying to match these "not-so-shared" vocabularies. If that general assumption may hold true for reduced-size, very specific and data-oriented ontologies (e.g. the WGS84 Geo Positioning RDF vocabulary[1], which contains only a few properties for describing latitude, longitude and point-in-space concepts), for larger domain descriptions, requiring different levels of abstraction and different perspectives depending on local needs, we expect to see several, different ontologies arise from independent organizations, often addressing overlapping domains.

Two issues then urge to be solved: first, facilitating people and automated systems in performing alignments between ontologies where they represent the same concepts and, secondly, make their vocabularies more explicit to humans, so that they can be re-used consistently in different scenarios and by different actors; in this sense, logical consistency may only help in restricting the range of possible interpretations which may be assigned to logical symbols, while common-sense human reasoning using these vocabularies may beneficiate a lot by the presence of clear and exhaustive documentation. Extensive use of Natural Language contents, providing free descriptions, synonymical expressions and translations in different idioms of the intended meaning of a vocabulary, appears thus as the most intuitive kind of documentation for data structures such as ontologies, dealing with representation of domains. Several efforts have been undertaken to cover different aspects of this problem, motivating the adoption of linguistic resources for enriching ontology vocabularies with natural language contents[2] (Pazienza & Stellato, 2006b; Prevot, Borgo, & Oltramari, 2005; Scheffczyk, Baker, & Narayanan 2006; Philpot, Hovy, & Pantel 2005; Huang 2004), showing useful applications exploiting these combined resources (Basili, Vindigni, & Zanzotto, 2003; Peter, Sack, & Beckstein 2006; Cappelli, Giovannetti & Michelassi 2004), providing standards for representing this enrichment/integration, like in SKOS[3] (Simple Knowledge Organization Systems) and in (Buitelaar, et al., 2006), and promoting the development of techniques for automating this task (Pazienza & Stellato, 2006c).

In this paper, we present an ontological and software framework for describing, referring and managing heterogeneous linguistic resources and for using their content to enrich and document ontological objects. This work, which originates ad completes previous research reported in (Pazienza & Stellato, 2006b; Pazienza, Sguera, & Stellato, 2007) articulates into two results: first, a set of coordinated RDF vocabularies providing descriptors for representing linguistic resources (ranging from lexical to frame-based ones) and their software counterparts (data structures, access libraries etc…), as well as offering metadata for describing the linguistic enrichment of ontologies, both on quantitative and qualitative grounds. The second result is a software library for evaluating the quality of automatic linguistic enrichment tools, through comparison of enriched ontologies compiled against the above vocabularies.

2. Related works

The actual practice of enriching ontologies with linguistic content basically depends on the multifariousness of lexical resources and on the explicit linguistic information

[1] http://www.w3.org/2003/01/geo/wgs84_pos

[2] The enrichment of ontologies with linguistic contents fosters the construction of peculiar kinds of semantic resources, which we could refer to as "hybrid" knowledge resources.

[3] http://www.w3.org/TR/swbp-skos-core-guide/

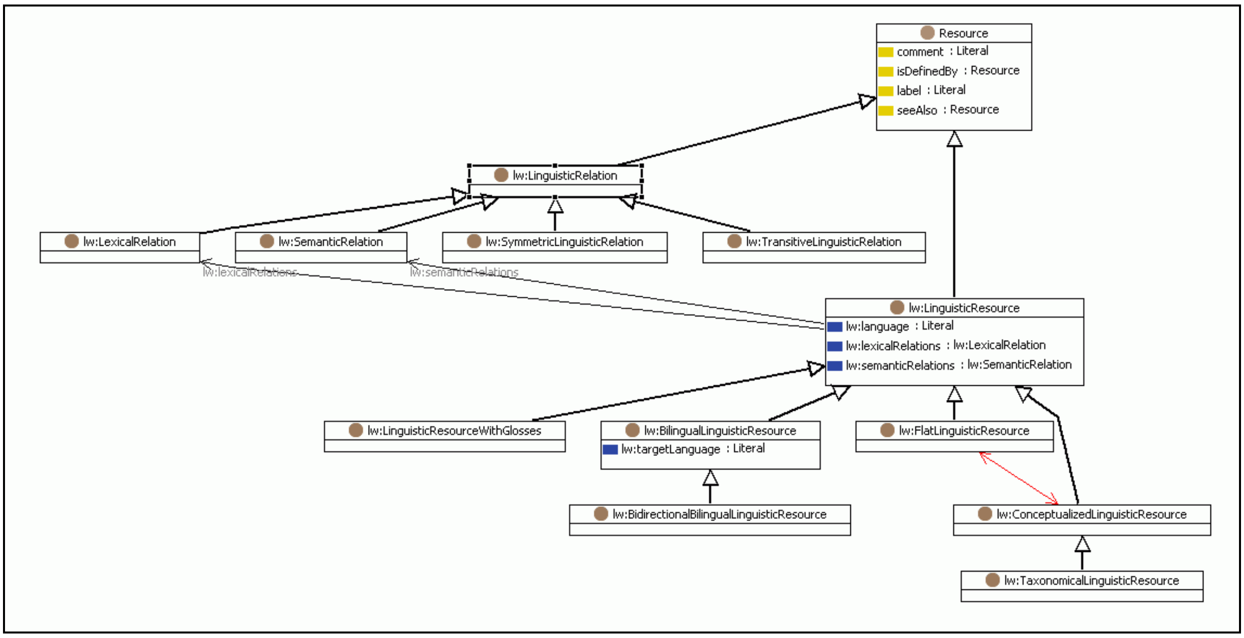
Resource
comment : Literal
isDefinedBy : Resource
label : Literal
seeAlso : Resource
lw:LinguisticRelation
lw:LexicalRelation
lw:SemanticRelation
lw:SymmetricLinguisticRelation
lw:TransitiveLinguisticRelation
lw:lexicalRelations
lw:semanticRelations
lw:LinguisticResource
lw:language : Literal
lw:lexicalRelations : lw:LexicalRelation
lw:semanticRelations : lw:SemanticRelation
lw:LinguisticResourceWithGlosses
lw:BilingualLinguisticResource
lw:targetLanguage : Literal
lw:FlatLinguisticResource
lw:BidirectionalBilingualLinguisticResource
lw:ConceptualizedLinguisticResource
lw:TaxonomicalLinguisticResource

Figure 1: An excerpt (focused on description of Linguistic Resources) from the Linguistic Watermark vocabulary

they expose (Pazienza and Stellato 2006c). Multilingual scenarios also demand for a proper lexicalization of ontological content according to different idioms and languages. From simple vocabularies of terms to wordnet-like structures, distinct lexical models need a solid and comprehensive framework of representation to enable a full-operational integration with ontologies. One example of this research trend is represented by the W3C initiative of translating WordNet to RDF/OWL, whose aim is to enable porting that kind of resource into Semantic Web infrastructure. Moreover, the integration between frame-based lexical databases and ontologies complicates the overall scenario and constitutes another important aspect of the above-mentioned process and a relatively brand-new trend in the scientific community. In a nutshell, the main rationale behind the notion of "frame semantics" (Fillmore 1968) is that meaning is represented by generalizations from stereotyped situations (frames). Berkeley FrameNet Project (Baker, Fillmore & Lowe; 1998) has been designed on the basis of that principle: nouns, verbs, and possibly modifiers (adjectives and adverbs) are clustered according to conceptual structures (e.g., the *commercial transaction* frame) and syntactic combinatory possibilities (valences). Several language-specific framenets have also emerged in the latest years according to Berkeley's model. The value of porting these kind of lexical databases into Semantic Web basically depends on the exploitation of their peculiar semantic structure for the enrichment of ontologies: this task may correspond to supply a formal semantics to frames (i.e. OWL semantics) or, besides re-engineering frame-based resources according to WWW standards, to use suitable pointers to link ontological categories and relations with frames. Similar issues arise from the task of interfacing ontologies with VerbNet (Kipper, Trang Dang, & Palmer, 2000), a project in which PropBank (Palmer, Kingsbury & Gildea, 2005) verb types are mapped to Levin Classes (Levin 1993): here the resource is organized into verb classes and alternations, without considering the role of nouns and modifiers in conceptual structures.

Despite the large interest in this area, standards for representing layered ontological-linguistic knowledge hardly finds a place in the Semantic Web stream of innovation, and while it has been shown that these processes can be handled with different levels of automation, no evaluation framework has been proposed until now.

3. The Linguistic Watermark Suite

The Linguistic Watermark suite of RDF vocabularies is composed of three ontologies:

- The *Linguistic Watermark* (*LW*) vocabulary, describing linguistic resources through their purposes and structure organization
- The *Ontological Linguistic Watermark* (*OLW*) vocabulary: a set of metadata descriptors for characterizing the linguistic expressivity of ontologies
- The *LW Linguistic Interfaces* vocabulary (*LWLI*), providing concepts for describing software libraries which grant access to specific (or ranges of) linguistic resources.

3.1. The Linguistic Watermark (LW) Vocabulary

While the Linguistic Watermark vocabulary partially covers general linguistic concepts like term, word, lexical/semantic relation, frame, agent etc... its main objective is to provide descriptors or characterizing the purpose and structure of linguistic resources: whether they represent translation vocabularies, synonyms collections, lexicons, frame based resources or terminologies, if they are organized around some kind of semantic structure or merely <entry, description> pairs etc..

Though originally conceived to cover any kind of Linguistic Resource, the first version of the Linguistic Watermark (figure 1) was limited to represent only lexical resources: by proper combination of its LW ontological

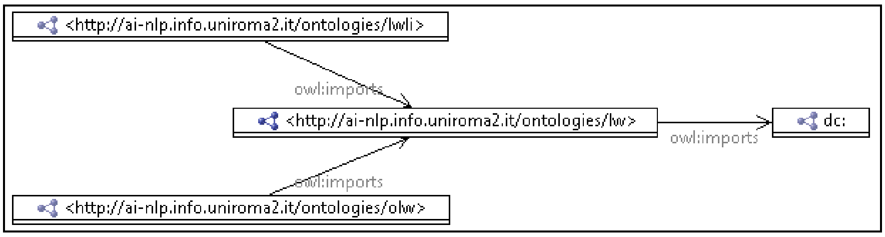
<http://ai-nlp.info.uniroma2.it/ontologies/lwli>
owl:imports
<http://ai-nlp.info.uniroma2.it/ontologies/lw>
owl:imports
dc:
owl:imports
<http://ai-nlp.info.uniroma2.it/ontologies/olw>

Figure 2: owl:imports relationships between ontologies in the Linguistic Watermark suite

descriptors, one could be able to represent very different linguistic resources, from simple synonym dictionaries, to complex resources such as WordNet (Miller et al, 1993). This provided a shared and homogeneous vocabulary upon which multilingual (and multi-resource) applications could be defined.

In this work we have extended le LW vocabulary into two main directions:

- *Instantiation*: now the vocabulary is not only used to describe linguistic resources, but even to predicate over their content (see section 4.2.2 for details)
- *Frames description*: covering frame/class based linguistic resources, such as FrameNet and VerbNet.

FrameNet and VerbNet have been modeled as distinct specializations of the newly introduced class FrameBasedResource, which is a rdfs:subClassOf of ConceptualizedLinguisticResource. This modeling choice mainly depends on the intrinsic nature of so-called "building blocks" of frame-based resources: "frames" are the organizational units of FrameNet corresponding to general schemas of specific situations. They are normally constituted by "Frame Elements", such as *Buyer* and *Seller* (in the *Commercial Transaction* frame), which are to be conceived as conceptual parts of a frame. The notion of "Frame Element" is very close to the basic notion of "Thematic Role", which is more general and domain-independent and actually adopted as the basic unit of VerbNet: some typical examples of thematic roles are *Agent*, *Patient*, *Duration*, *Destination*.

Resources of type FrameBasedResource adopt a specialization of SemanticIndex, namely Frame, which is structured according to variable sets of objects called FrameElement.

Another important issue concern relations holding between frames. Seven types of parent/child relation are used in FrameNet, namely "subframe", "inheritance", "perspective on", "using", "causative of", "inchoative of", "see also" and one type of temporal ordering relation, that is "precedes". Although is not our aim here to focus on the semantics of these relations, clearly they are not lexical ones: they pertain to the conceptual level and are used to structure the set of frames (up to date, around 1000) which compose FrameNet. Nonetheless, they can be mapped through instances of the already existing SemanticRelation class. It is relevant to notice that some frame relations are transitive (as hyponymy in wordnet-like linguistic resources); for instance, the ordering relation "precedes", which establishes a chronological nesting within frames (and frame elements too).

A crucial aspect in making the LW a vocabulary for describing instantiable linguistic resources is the link between SemanticIndex and LexicalUnit class. In general, semantic indexes can be thought as conceptual objects which can, depending on the purpose and semantics of the considered resources, be associated to simple or compound words, which are actually kinds of lexical units. According to this modeling perspective, the relation lexicalUnit has been created, holding between LexicalUnit and SemanticIndex: for instance, the verb "purchase" (simple word) is both the lexical unit of the frame *Commercial Transaction* and of the WordNet's synset <buy, purchase>[4]. This example shows how the LW model is able to capture different uses in different lexical resources of the same linguistic units. The semantics of each instantiation of the lexicalUnit property depend on the considered resource, while the LW library may offer homogeneous API for inspecting different linguistic resources, for showing their content on automatically generated GUIs or enabling its integration inside other representation formalisms, such as ontologies. This generalization thus boosts reuse and integration of several resources in several application contexts.

3.2. The Ontological Linguistic Watermark (OLW)

The characterization given by the OLW is expressed in terms of the linguistic content of the described ontology and with respect to the resources which have been adopted for enriching its concepts. As stated in (Pazienza, Sguera, & Stellato, 2007), where its adoption has been considered in a scenario involving Semantic Coordination of FIPA agents, its metadata assume great significance in all the contexts where ontologies sharing a common domain, but no explicit semantic bridging between their respective vocabularies, need to be automatically aligned or merged. Resource-based algorithms for ontology alignment and semantic coordination agents can in fact inspect the OLW data of the ontologies to be compared and configure at best the resources and facilities to be used for matching their content. This is an aspect which has often been underestimated in literature: setting up the resources to be adopted in a realistic scenario, while being not a trivial task, influences dramatically the outcome and performances of any mediation activity.

The LWLI takes its roots from the first version of the Linguistic Watermark software library[5] – developed by the University of Rome, Tor Vergata – a component providing uniform access to different and heterogeneous linguistic resources, which has been used in several resource-based tools, such as the OntoLing Protégé plug-

[4] Gloss: "obtain by purchase; acquire by means of a financial transaction"; "The family purchased a new car"; "The conglomerate acquired a new company"; "She buys for the big department store".

[5] http://ai-nlp.info.uniroma2.it/software/LinguisticWatermark/

in (Pazienza & Stellato, 2006). The LW presented in that work, was just a class diagram offering several interfaces and abstract classes whose combination could be used to describe the main aspects of a linguistic resource: implementing the proper subset of those (software) interfaces would result in the definition of a linguistic wrapper for accessing a particular linguistic resource. The LW library thus offered a combination of descriptive (with regard to the resources to be wrapped) and operative aspects (delineating the operations which the required wrapper had to implement). Later on, the requirements which brought to developing the OLW, demanded a formal ontological representation, merely focused on resource description, to be extracted from the original class diagram, which led to the LW.

Now, the time has come to close the circle, and with the LWLI we recovered the original intent of the LW library.

3.3. The LW Linguistic Interfaces vocabulary (LWLI)

LWLI contains concepts describing parameters needed by software libraries for setting up access to their target linguistic resources. This third ontology completely migrates the original framework to RDF, thus providing a complete vocabulary at the hand of Semantic Web tools which rely on the use of linguistic resources or are even expressly dedicated to the integration of ontologies with linguistic resources.

The LWLI includes concepts like:

- LinguisticInterface: for describing a specific implementation of a wrapper for a linguistic resource
- LinguisticInterfaceConfiguration: representing instances of basic runtime configurations for a given LinguisticInterface.
- LinguisticInterfaceInstanceConfiguration: each instance of this class provides data for completing a single runtime configuration for accessing a specific linguistic resource, basing on partial configuration from a given LinguisticInterfaceConfiguration

and properties for specifying these configuration settings, among which, we list the following ones:

- configuredInterface: this property tells which LinguisticInterface is being configured through the described configuration
- interfaceableResource: tells which linguistic resources are made accessible through the described Linguistic Interface
- ConfigurationProperty: a property defining configuration parameters for accessing a linguistic resource through a dedicated linguistic interface. This property is never instantiated, though it has a few relevant subproperties for telling whether a given configuration parameter points to the file system, if a property is relevant for configuring a linguistic interface as a whole, or just for accessing specific resources etc..

As for the LW, even this vocabulary provides an upper ontology which, though extensible in principle to match the specification of each represented software library, already contains all the required descriptors for automatically driving different linguistic resources under a shared knowledge model.

To have an example, consider the following use case: we are trying to describe the fictitious YAWW (Yet Another WordNet Wrapper) library. First of all, we declare yaww as a new instance of LinguisticInterface. Then, we should consider all the parameters that the wrapper needs for its configuration, distinguishing those needed to make the interface – as a whole – work, from those which are necessary for granting access to different WordNet versions installed on the host. These parameters should be used to instantiate properties for the two configuration classes LinguisticInterfaceConfiguration (the one related to general interface configuration), and LinguisticInterfaceInstanceConfiguration, for setting up access to specific resources.

We could even add more information at conceptual level, by adding specific subclasses, YAWWInterfaceConfig and YAWWInstanceConfig, respectively, to the two configuration classes above, and binding them, through property restrictions, to ad-hoc configuration properties, like the one which is described next.

Being YAWW a wrapper for WordNet, we would probably need to define a configuration property for specifying the path to the dictionary folders of the various installed wordnets we want to access; by first, we declare the owl:DataTypeProperty wnDictPath, then we state it as being rdfs:subPropertyOf of two available subproperties of lwli:ConfigurationProperty: the first one, lwli:InstanceProperty, tells that the its instantiated value represents a parameter for accessing a given wordnet (the one installed in that path) and not for configuring the whole library (and thus, that it has to be attached to a given YAWWInstanceConfig), while the second one, lwli:FileProperty informs that this property points to a file in the file system, so that applications based on this vocabulary, could in case apply necessary filechecking mechanisms, as well as find appropriate graphical interface widgets – a file chooser dialog, for example – when interacting with the user for filling the value of this parameter.

Though we added specific subclasses and subproperties (thus extending the conceptual part of the ontology), the software *interface*, which is based on the sole LWLI, does not need any changes, and thus the same for any application software based on LWLI, which can now benefit of the new added resource wrapper, without any development effort.

4. An improved Integration Framework

In this section we describe the new libraries and tools which have been developed with the intent of providing a consistent and homogeneous layer for integrating ontologies and linguistic resources, also taking into account the variety of proposed standards and research results which have arisen in these last years

4.1. The new Linguistic Watermark library

Following the recent improvements on the LW suite, we have released a new version of the Linguistic Watermark library, which offers java API for accessing linguistic resources through dedicated Linguistic Interfaces, both entities being defined according to the LW and LWLI vocabularies. In particular, a mapping between the above ontologies and newly added java interfaces allows implemented java wrappers for linguistic resources to

```
<wn20schema:NounSynset rdf:about="wn20instances:synset-entity-noun-1" rdfs:label="entity">
    <wn20schema:synsetId>100001740</wn20schema:synsetId>
</wn20schema:NounSynset>

<rdf:Description rdf:about="wn20schema:Synset">
    <rdfs:subClassOf rdf:resource="lw:SemanticIndex"/>
</rdf:Description>

<someOntology:Noun>
    <olw:semanticDescriptor rdf:resource="wn20instances:synset-entity-noun-1">
</someOntology:Noun>
```

Figure 3: an example of resource wrapping: binding WordNet-RDF synsets to a class concept

declare themselves as new instances of the LinguisticInterface class and accept strongly typed configuration parameters, thus enabling data consistency checks and providing hooks for automatic generation of configuration user interfaces for hosting applications.

4.2. The OLW library and OLW vocabulary improvements

With the specific aim of obtaining a stable range of instruments for enriching ontologies with lexical content, and of formalizing the model and associated format for representing this information, we have developed a dedicated component which, together with the LW library, can be embedded in ontology based tools and applications needing to incorporate linguistic content.

4.2.1. Issues in representing the integrated information

So far, in tools exploiting the Linguistic Watermark framework, like the already cited OntoLing, the association between linguistic content and ontological data has been projected over standard RDFS/OWL predicates. Thus, the rdfs:label property were used for addressing short lexical objects like terms, words (used both to provide synonymical expressions as well as to provide translation for different languages) or even conceptual entities like WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1993) synsets, while rdfs:comment has been commonly associated to wider descriptions like those which could be extracted from word glosses and terminology definitions.

This choice, though guaranteeing a complete adherence to widely accepted standards on the one side, offered poor representation primitives: two major problems concerned the loss of information about the nature of the attached linguistic objects, which became mere strings pointed by the rdfs properties, and difficulty in the integration of artificial entities. As an example, a WordNet synset, being a kind of lw:SemanticIndex, were linked to ontology objects through the rdfs:label property, filling the xml:lang attribute of this predicate with a short namespace for indicating its association to WordNet (and the specific WordNet version), while xml:lang requires codes conforming to the official standard code ISO 3166-1-alpha-2. Clearly, a compromise between popularity, immediateness and completeness of the model needed to be found.

4.2.2. The OLW integration model

In modeling our framework for the integration of ontological and linguistic content, we have taken into consideration the following requisites, which should allow for:

1. Reporting quantitative and qualitative information on the overall process of enriching an ontology with content from a linguistic resource (this was the primary objective of the OLW metadata ontology)
2. Keeping track (at least maintain the possibility to do that) of the source used for enriching the content
3. Being able to properly map different kind of linguistic entities (words, linguistic/semantic relations etc…) with (structures of) ontological objects
4. Giving the user the possibility of adopting resources' specific objects (e.g. FrameNet frames or WordNet synsets) for enriching an ontology
5. Embedding existing models for integration of ontologies and linguistic entities, still respecting the above priorities
6. Assessing reliable links between ontological and linguistic objects as well as taking into account for probabilistic matches produced by automatic enrichment tools (which could also be used for evaluation purposes)

The first requisite has been satisfied by defining a set of meta-descriptors – represented through object properties with domain set to owl:Ontology – for providing an overview of the "linguistic expressiveness" of ontologies. These properties may prove to be helpful for services/agents which, having to map/merge/align/mediate different ontologies, may be willing to invoke the proper linguistic resources for supporting this task. These mediators can thus beneficiate of the overall statistical information provided by the OWL metadata, without inspecting the entire ontologies' content. This part of the OLW has already been described in details in (Pazienza, Sguera & Stellato; 2007).

The second, third and fourth requisites have been accomplished by extending the LW; in its first incarnation, which served solely as a conceptual driver for the software library, the LW was able to express descriptions of linguistic resources, without predicating about their specific content. Now it has been extended to make possible the instantiation of objects from the described resources. The example in Figure 3 shows fragments originating from three different ontologies: the first fragment is a description of WordNet synset 100001740

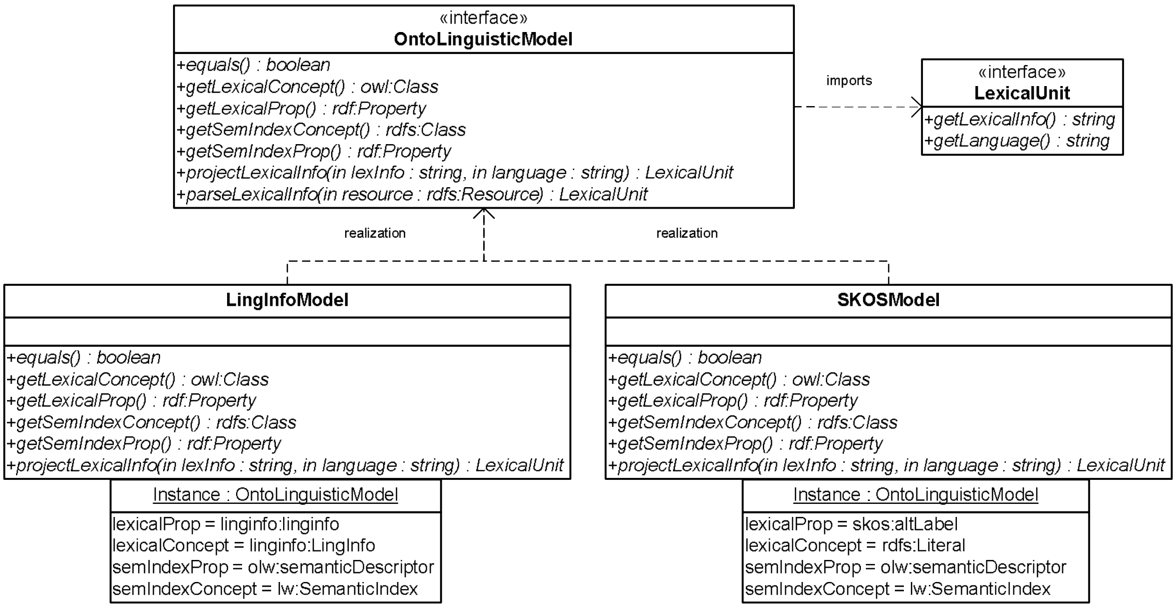
«interface»
OntoLinguisticModel
+equals() : boolean
+getLexicalConcept() : owl:Class
+getLexicalProp() : rdf:Property
+getSemIndexConcept() : rdfs:Class
+getSemIndexProp() : rdf:Property
+projectLexicalInfo(in lexInfo : string, in language : string) : LexicalUnit
+parseLexicalInfo(in resource : rdfs:Resource) : LexicalUnit
imports
«interface»
LexicalUnit
+getLexicalInfo() : string
+getLanguage() : string
realization
realization
LingInfoModel
+equals() : boolean
+getLexicalConcept() : owl:Class
+getLexicalProp() : rdf:Property
+getSemIndexConcept() : rdfs:Class
+getSemIndexProp() : rdf:Property
+projectLexicalInfo(in lexInfo : string, in language : string) : LexicalUnit
Instance : OntoLinguisticModel
lexicalProp = linginfo:linginfo
lexicalConcept = linginfo:LingInfo
semIndexProp = olw:semanticDescriptor
semIndexConcept = lw:SemanticIndex
SKOSModel
+equals() : boolean
+getLexicalConcept() : owl:Class
+getLexicalProp() : rdf:Property
+getSemIndexConcept() : rdfs:Class
+getSemIndexProp() : rdf:Property
+projectLexicalInfo(in lexInfo : string, in language : string) : LexicalUnit
Instance : OntoLinguisticModel
lexicalProp = skos:altLabel
lexicalConcept = rdfs:Literal
semIndexProp = olw:semanticDescriptor
semIndexConcept = lw:SemanticIndex

Figure 4: two examples of OntoLinguisticModel implementation

originating from the WordNet-RDF vocabulary developed by the WordNet task force of the W3C (http://www.w3.org/TR/wordnet-rdf/); the second one is the binding of concept wn20schema:Synset to the lw:SemanticIndex, through a rdfs:subClassOf relationship. Finally, a certain Noun concept coming from a fictitious ontology is enriched with the meaning expressed by the above synset, through the owl:semanticDescriptor property. With this extensible pattern, the LW+OLW offer reusable vocabularies for describing linguistic resources which drive the behavior of software applications serving the same task, while specific extensions (both in terms of ontologies and software components) can be added to describe specific lexical and semantic objects from new resources, without requiring modifications to the core vocabulary nor to the original application.

4.2.3. Compatibility with existing (proposed) models

As previously mentioned, several formats exists or have been proposed for integrating ontological content with linguistic information.

While we did not intend to propose a new one, we tried to obtain cross-compatibility with available standards and proposed models, by gearing our software library with a OntoLinguisticModel interface, consisting of a series of enrichment/retrieval operations defined upon abstract "slots" for representing linguistic information. These slots can be then implemented according to a specific onto-linguistic representation model, by specifying the properties and concepts used to map integrate linguistic information with ontological one.

Obviously, it is impossible to foresee in advance all the characteristics of each model/interface-implementation which could be integrated in the future, thus we provided a specific *project/decode* feature for projecting the linguistic information extracted from linguistic resources according to the LW ontology, towards the (possibly more fine-grained) adopted ontolinguistic model. For evaluative (see next section) and comparative purpose in general, we demand to each specific implementation the specifications of equivalence between the locally defined linguistic objects.

Implementations of OntoLinguisticModel have been developed for the traditionally adopted RDFS annotation properties (rdfs:label and rdfs:comment), for the base SKOS vocabulary (by extending the above with skos:prefLabel and skos:altLabel), for SKOS + SKOS-Mapping[6] vocabularies (thus including skos:broader/skos:narrower and skos:related, to map ontology concepts with instances of lw:SemanticIndex from the LW ontology) and, finally, for the LingInfo model, by wrapping the linginfo:linginfo property and linginfo:LingInfo class.

4.2.4. The OLW integration model

Figure 4 shows (hiding minor details) how two available linguistic models have been mapped to our meta-model and wrapped inside our library. In the reported examples, pointers to lw:SemanticIndex have been implemented by using OLW and LW descriptors, since there were no correspondence for them in the addressed models. Notice how the main mapping completely hides any information associated to more complex specifications of the concepts of the wrapped models. For example, in the LingInfo wrapper, the lexical element associated to an ontology object is bound to the linginfo:term property of the created linginfo:LingInfo object (while it is directly mapped to the value of skos:altLabel in the SKOS case); in the same manner, the language parameter of the projectLexicalInfo() method is associated to the linginfo:lang property for the same object, whereas it is directly mapped to the xml:lang attribute of the skos:altLabel property in the SKOS case.

A similar process will be carried out in the future for frame-based resources, once RDF descriptions and research about mapping of their content to ontologies will reach full maturity and stableness. The above integration model satisfied our fifth requirement, while the resolution

[6] http://www.w3.org/2004/02/skos/mapping/spec/

of the sixth one is part of the discussion presented in the next section.

5. The evaluation framework

The newly developed OLW Library provides a framework for evaluating the quality of algorithms for Linguistic Enrichment of ontologies with respect to previously defined reference standards.

Linguistic Enrichment algorithms can be evaluated by comparing the results of an Enrichment Process (E) to a reference enrichment document, which we call "the Oracle" (O). The usual approach for evaluating the results of process E is to consider them as sets of correspondences and to apply precision and recall originating from Information Retrieval (Van Rijsbergen, 1975) and adapted to the matching task. Precision and recall are thus the ratio of the number of true positive $|O \cap E|$ on that of the retrieved correspondences ($|E|$) and those expected ($|O|$) respectively.

The OLW library can accept pairs of linguistic enrichment documents (that is: ontologies with integrated linguistic content), where one is the Oracle and the other one is the result to be tested, providing that the following extensions are included in the library and properly configured:

- *Enrichment Model* and related software extension (see section 4.2.3)
- *Resource*(s) *description* (and their wrapper implementation) used for enrichment (see sections 3.1 and 4.2.2)
- *Match Specification and Evaluation (MSE)* extension, if different enrichment entries differ from simple links between ontological and linguistic objects

With the ones above, the library is able to seek the enrichment properties (at least, those which need to be considered) in the ontology documents (first extension) and to properly identify the elements used for the enrichment (second extension).

The third one is an extension needed for those cases where an algorithm produces any kind of probabilistic/quantitative result, so that the enrichment links in the tested document cannot be evaluated just in terms of correct/wrong matches versus those in the Oracle.

If this extension is included, an ontological representation for qualifying its results is to be provided (usually, it just requires a property with domain set to the adopted enrichment properties, that is olw:lexicalization olw:semanticDescriptor and range set to the description of the non-conventional link). A proper extension module for the library needs then to be plugged, with a parser for the above description and associated modifiers for adapting the precision/recall measure to the introduced range of values.

Inter-annotator agreement can as well be measured against two reports about the enrichment, compiled by human annotators (with no further requirement apart from the ones above).

6. Conclusions

In this paper we presented the Linguistic Watermark suite, a set of RDF vocabularies used to uniformly represent linguistic knowledge in heterogeneous linguistic resources and to enable shared integration-with and accessibility-from different computational ontologies. In this context the main features of LW library have been also illustrated, a set of JAVA-based software tools and interfaces developed for integrating ontologies and linguistic resources. This library exploits LW vocabularies to establish adequate mappings between linguistic resources and linguistic interfaces, helping knowledge engineers to implement their hybrid semantic systems. We expect that our work may give a contribution to the standardization of models, methodologies and tools for the effective integration of ontologies and linguistic resources; moreover, the possibly adoption by R&D communities of the general framework we presented might inspire, in the next future, new contests for the evaluation of linguistic enrichment of ontologies.

7. References

Baker, C., Fillmore, C., & Lowe, J. (1998). The Berkeley FrameNet project. *COLING-ACL*. Montreal, Canada.

Basili, R., Vindigni, M., & Zanzotto, F. M. (2003). Integrating Ontological and Linguistic Knowledge for Conceptual Information Extraction. *IEEE/WIC International Conference on Web Intelligence.* Washington, DC, USA.

Beth, L. (1993). *English verb classes and alternations: A preliminary investigation* (Vol. XVIII). Chicago: University of Chicago Press.

Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., et al. (2006). LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, hosted by LREC Conference. Genoa, Italy.

Cappelli, A., Giovannetti, E., & Michelassi, P. (2004). Ontological Knowledge and Language in Modelling Classical Architectonic Structures. *Ontology and Lexical Resources – OntoLex 2004)*, hosted by LREC Conference. Lisboa, Portugal.

Euzenat, J. (2004). An API for Ontology Alignment. In S. A. McIlraith, D. Plexousakis, & F. van Harmelen (Ed.), *The Semantic Web - ISWC 2004: Third International Semantic Web Conference. 3298*, pp. 698-712. Hiroshima, Japan: Springer.

Euzenat, J. (2007). Semantic Precision and Recall for Ontology Alignment Evaluation. In M. M. Veloso (Ed.), *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence*, (pp. 348-353). Hyderabad, India, January 6-12.

Fillmore, C. (1968). The Case for Case. In E. Bach, R. T. Harms, & B. a. Harms (Ed.), *Universals in Linguistic Theory* (pp. 1-88). New York: Holt, Rinehart, and Winston.

Huang, C. (2004). Sinica BOW: Integrating bilingual WordNet and SUMO Ontology. *Ontology and Lexical Resources – OntoLex 2004,)*, hosted by LREC Conference. Lisboa, Portugal.

Kipper, K., Trang Dang, H., & Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. *AAAI-2000 Seventeenth National Conference on Artificial Intelligence.* Austin, TX.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1993). *Introduction to WordNet: An On-line Lexical Database.*

Palmer, M., Kingsbury, P., & Gildea, D. (2005). The Proposition Bank: An annotated Corpus of Semantic Roles. *Computational Linguistics , 31* (1), 71-106.

Pazienza, M. T., & Stellato, A. (2006). An open and scalable framework for enriching ontologies with natural language content. *The 19th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE'06), special session on Ontology & Text.* Annecy, France.

Pazienza, M. T., & Stellato, A. (2006b). Exploiting Linguistic Resources for building linguistically motivated ontologies in the Semantic Web. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, hosted by LREC Conference. Genoa, Italy.

Pazienza, M. T., & Stellato, A. (2006c). Linguistic Enrichment of Ontologies: a methodological framework. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006).* Genoa, Italy.

Pazienza, M. T., Sguera, S., & Stellato, A. (2007). Let's talk about our "being": A linguistic-based ontology framework for coordinating agents. (R. Ferrario, & L. Prévot, Eds.) *Applied Ontology, special issue on Formal Ontologies for Communicating Agents , 2* (3-4), 305-332.

Peter, H., Sack, H. Beckstein, C. (2006). SMARTINDEXER – Amalgamating Ontologies and Lexical Resources for Document Indexing. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006)*, hosted by LREC Conference. Genoa, Italy

Philpot, A., Hovy, E., & Pantel, P. (2005). The Omega Ontology. Ontology and Lexical Resources. *OntoLex2005 - Ontologies and Lexical Resources.* Jeju Island, South Korea.

Prevot, L., Borgo, S., & Oltramari, A. (2005). Interfacing Ontologies and Lexical Resources. *Workshop on Ontologies and Lexical resources (OntoLex2005),* hosted by IJCNLP Conference. Jeju Island, South Korea.

Scheffczyk, J., Baker, C. F., & Narayanan, S. (2006). Ontology-based Reasoning about Lexical Resources. *Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006).* Genoa, Italy.

Van Rijsbergen, C. J. (1975). *Information Retrieval.* London, United Kingdom: Butterworths.