An Environment for Semi-automatic Annotation of Ontological Knowledge with Linguistic Content

Maria Teresa Pazienza and Armando Stellato

AI Research Group, Dept. of Computer Science, Systems and Production
University of Rome, Tor Vergata
Via del Politecnico 1, 00133 Rome, Italy
{pazienza, stellato}@info.uniroma2.it

Abstract. Both the multilingual aspects which characterize the (Semantic) Web and the demand for more easy-to-share forms of knowledge representation, being equally accessible by humans and machines, push the need for a more "linguistically aware" approach to ontology development. Ontologies should thus express knowledge by associating formal content with explicative linguistic expressions, possibly in different languages. By adopting such an approach, the intended meaning of concepts and roles becomes more clearly expressed for humans, thus facilitating (among others) reuse of existing knowledge, while automatic content mediation between autonomous information sources gets far more chances than otherwise. In past work we introduced OntoLing [7], a Protégé plug-in offering a modular and scalable framework for performing manual annotation of ontological data with information from different, heterogeneous linguistic resources. We present now an improved version of OntoLing, which supports the user with automatic suggestions for enriching ontologies with linguistic content. Different specific linguistic enrichment problems are discussed and we show how they have been tackled considering both algorithmic aspects and profiling of user interaction inside the OntoLing framework.

1 Introduction

The multilingual aspects which characterize the (Semantic) Web and the demand for more easy-to-share forms of knowledge representation, being equally accessible by humans and machines, depict a scenario where formal semantics must coexist side-by-side with natural language, all together contributing to the shareability of the content they describe. The role of different cultures and languages is fundamental in a real World *aWare* Web and, though English is widely accepted as a "lingua franca" all over the world, much effort must be spent to preserve other idioms as they express different cultures. As a consequence, multilinguality has been cited as one of the six challenges for the Semantic Web [2].

These premises suggest that semantic web ontologies, delegated to express machine-readable information on the Web, should be enriched to both cover formally expressed conceptual knowledge and expose its content in a linguistically motivated fashion.

Y. Sure and J. Domingue (Eds.): ESWC 2006, LNCS 4011, pp. 442 – 456, 2006.

Even more could be done: revisiting ontology development process under this perspective, would in fact guarantee this scenario to become a suitable framework upon which even machine oriented task, like mediation and discovery, would benefit of this greater expressivity.

Following this intent, in [7,8] we defined OntoLing, a Protégé [5,6] plug-in offering a modular and scalable framework for supporting manual annotation of ontological data with information from different, heterogeneous linguistic resources.

We present now an improved version of OntoLing, which prompts the user with automatic suggestions for enriching ontologies with linguistic content. We explain how and why different kinds of linguistic enrichment processes should be performed and focus our attention on one of these tasks, showing how its automatization has been obtained, considering both algorithmic aspects and profiling of user interaction in the context of OntoLing framework.

2 Linguistic Enrichment of Ontologies: Different Possible Tasks

We introduced the expression "Linguistic Enrichment of Ontologies" to identify a series of different processes sharing the common objective of improving the linguistic expressivity of an ontology, through the exploitation of existing Linguistic Resources (LRs, from now on). The nature of this "linguistic expressivity" strongly depends on the LRs used for linguistic enrichment and on the specific goals the enrichment process will achieve. In the following sections we describe three different enrichment tasks, together with possible scenarios in which these tasks may be applied.

2.1 Using a LR's Semantic Structure as a Controlled Vocabulary: Semantic Enrichment of Ontologies

In this class of Linguistic Enrichment tasks, the semantic structure of a given LR (provided it has one), is used as a controlled vocabulary for the ontology and related application. What is required is just identification of *pointers* from ontological data to semantic elements of the linguistic resource. Access to pure linguistic information is then guaranteed by the links between the semantic and linguistic structure of the LR.

As a first example, consider an NLP ontology-based application, dedicated to whatsoever kind of text analysis task (e.g. Information Extraction), and which is strongly coupled with a semantic lexicon for extracting linguistic information from the text. The semantic pointers are needed to easily move from extracted, neutral, "linguistic information", which is processed in terms of lexical concepts, to "events" which are represented by the ontology.

As a further example, consider a scenario where distributed information sources must be aligned by mediators relying on a common form of knowledge. This committed knowledge is represented by so called "upper ontologies", or "upper models" which contain a first stratification of general concepts. Examples in literature [1] report of adoption, instead of an ontology, of the semantic structure of an existing linguistic resource [4] as a interlingua for guaranteeing communication between autonomous distributed agents.

2.2 Explicit Linguistic Enrichment

When no semantic commitment has been established between autonomously developed information sources, no further solution exists for reaching semantic interoperability than relying on the very last form of *shared* knowledge representation: natural language. It is the form used by humans to pass from their own conceptualization of the world, to any form of shareable communication, being it spoken, written, or even related to formal representations of knowledge (also a good programming style ask for variables and functions being expressed through *evocative* labels). Indeed, stating direct links between ontological content (which is often scarcely modeled, upon a linguistic point of view) and linguistic expressions, may represent the only viable solution to increase the shareability of the represented knowledge.

Moreover, the improved range of expressions for denoting a concept and the (possible) presence of natural language descriptions for ontological data, facilitate reuse of existing knowledge, which is made more comprehensible also for humans.

2.3 Producing Multilingual Ontologies

Exploitation of existing bilingual resources may help in the development of multilingual ontologies, in which different multilingual expressions coexist and share the same ontological knowledge. The multilingual enrichment process, mainly if considered upon already enriched ontologies, may beneficiate of a greater linguistic expressivity of the source data and thus exploit different techniques for obtaining proper translations for ontology concepts and roles.

3 Techniques for Semantic/Linguistic Enrichment of Ontologies

While OntoLing's underlying model for accessing LRs is thought for supporting all of the above tasks, in this work we focus on techniques and solutions for automatizing the first task which has been presented: semantic enrichment of ontologies. This represents in fact a first necessary step through which all of the other tasks may be accomplished.

3.1 The Linguistic Enrichment Environment: Adopted Terminology

For sake of clarity, we will adopt from now on a terminology inherited from two well known standards for ontological and linguistic resources: OWL and WordNet model.

OWL [3] has recently been accepted as a W3C recommendation for the representation of ontologies on the Web, so we have adopted its ontological model for our framework and will use its nomenclature for distinguishing ontological objects into *classes*, *properties* (*object properties* and *datatype properties*) and *individuals*. Frame based models for knowledge representation can equally be considered inside this framework, with *slots* taking the role of properties and *instances* acting as individuals of the OWL model. We adopt in fact the term *frame* to address any ontological object whose type needs not to be specified.

WordNet [4] is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs,

adjectives and adverbs are organized into synonym sets (synsets), each representing one underlying lexical concept. Several wordnets have been developed for other languages [11, 9], which have thus favored a large diffusion of the model which inspired the original English version. As only those LRs which expose (cfr. [7, 8]) a semantic structure (like WordNet) may be elected for the semantic enrichment task, we decided to adopt notation borrowed from the WordNet Model to address linguistic elements from LRs. We thus use terms like *synset* (or *lexical concept*, or *semantic element*), *sense* and *synonym*, under the meaning they assume in WordNet-like lexical databases.

We prefer in general to avoid use of term *concept* in any formal statement, as it is adopted in different communities with different meanings: a synset is a *lexical concept* in WordNet, while an OWL class implements a *concept* in Description Logics theory, furthermore, other ontology traditions use "concept" to mean every generic ontology construct, thus including properties and instances other than classes.

3.2 The Semantic Enrichment Task

Objective of semantic enrichment task is to identify *pointers* from ontological objects (*frames*) to semantic entities (e.g. *synsets*, for WordNet) of a linguistic resource.

Before detailing our semantic enrichment process, we describe a few empirical results we collected during our research. These results took the form of morphosyntactic and semantic evidences recognized over several pairs of ontologies and linguistic resources, which could be used to guide the enrichment process.

All the reported examples refer to semantic enrichment of a DAML ontology[1] about baseball, downloaded from the DAML library of ontologies[2], using WordNet as a source for linguistic knowledge.

3.3 Taxonomy-Alignment Evidences

In case the semantic structure of a given LR has been organized as a taxonomy of broader/narrower linguistic concepts, similarities between this taxonomy and that of the ontology may provide useful evidences for an enrichment task. The IS-A relation of ontologies has however well defined semantics, while taxonomical links of LRs may often confuse different informal and/or ambiguous relationships (specialization, part-of, relatedness etc…); nonetheless, an analysis of these similarities typically leads to interesting and reliable results. The intuition behind this strategy is that *if* a semantic pointer links a frame-synset pair $<F,S>$, *then* other frame-synset pairs (where the frame is more specific/more generic that F and the synset is narrower/broader than S), have a good probability of being linked through a semantic pointer. We call this phenomenon the "sense-alignment square".

In Fig. 1 below, the semantic pointer between F_H and S_H already exists and represents an evidence for assessing a new semantic pointer over the pair $<F_L, S_L>$.

An example of this configuration is represented by the class labeled as *Hit* in the baseball ontology: this class has been eligible for 14 potential senses in WordNet. Of these 14 senses one is represented by the synset `noun.124696`, whose gloss states:

[1] http://www.daml.org/2001/08/baseball/baseball-ont for the original DAML version.
[2] http://www.daml.org/ontologies/

a successful stroke in an athletic contest (especially in baseball); "he came all the way around on Williams"hit"

This synset is more general than another WordNet synset, `noun.39042`, which is described by the following gloss:

a base hit on which the batter stops safely at second base; "he hit a double to deep centerfield"

and which has among its synonyms the word "double". Finally, closing the alignment-square, *Double* is another class of the ontology, which is a subclass of *Hit*. Thanks to this evidence, both *Hit*-`noun.124696` and *Double*-`noun.39042` pairs result as good candidates for being linked through a semantic pointer.

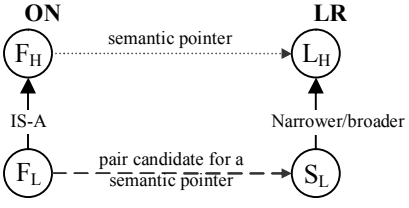

Fig. 1. The sense-alignment square

3.4 Evidences Resulting upon Analysis of Glosses from the Linguistic Resource

Glosses offer natural language descriptions of concepts. Though their content is generally intended as an easy reference for human readability, it represents indeed a useful mean for discovering relations which have no explicit semantic counterpart in the resource they come from.

From the glosses reported in the previous example, we could learn that a "double" is a kind of "base hit" (though the meaning of "hit" is not formally specified by the gloss), even if the resource lacked of a taxonomical structure, thus binding the two lexical concepts together in a broader/narrower relation.

A further example is represented by the class *Division* (again in the DAML baseball ontology). WordNet offers 12 different senses for the term "division". The gloss of the correct synset, `noun.7741947`, states:

a league ranked by quality; "he played baseball in class D for two years"; "Princeton is in the NCAA Division 1-AA".

Again, we could learn that a "division" is a "league", and *League* is one of the classes of the ontology. This case is however different from the previous one: in fact in the ontology tree, *Division* has not been conceived as a type of *League*. Nonetheless, a further analysis of ontological context reveals that *Division* appears in the restricted range of a property of class *League*. The co-occurrence of these two terms in the gloss, together with the presence of the range restriction binding the two classes labelled by the terms, suggests `noun.7741947` as a potential candidate for *Division*.

There are however cases where a supposed interesting relation is not formally expressed in the ontology. An example is given by the class *Out*: we report here the gloss of its correct matching synset:

(baseball) a failure by a batter or runner to reach a base safely in baseball; "you only get 3 outs per inning".

we observe that "base" is a term appearing in the above gloss and that, at the same time, *Base* is a class in the ontology. Unfortunately, *Base* is not bound by any ontological relation to *Out*. Should this combination be discarded as a mere fortuity? May be not: the baseball ontology for example, with its 104 frames (considering classes and properties), may in fact be considered as a very domain-specific representation, where the sole presence of few concepts is enough to consider them semantically related in some way.

A final consideration: it may happen that glosses describing synsets which are candidate for enrichment of different ontology frames, contain common references to concepts of which no trace is present in the ontology. Oddly enough, the ontology about baseball which we used for our examples, contains no specific lexical nor conceptual reference to "baseball" itself! On the other hand, many WordNet definitions contain the word baseball in their glosses, so that, in those cases, it is quite easy for a human to immediately choose the right sense from the given set of candidates, just after a glimpse at the list of glosses. An automatic process should be able to discover even these "hidden" correlations and weight their effectiveness appropriately.

4 The Feature Model

To take into account all previous considerations, and to maintain a scalable approach towards new possible strategies and LR configurations, we adopted a probabilistic model based on a feature space which is produced upon the observed evidences.

We have thus defined a *Plausibility Matrix* M_P as a two-dimensional matrix on a $O \times L$ space, where O is the cardinality of the ontological objects and L is the cardinality of the semantic data in the linguistic resource. Each element $M_P(i,j)$ of the matrix represents the plausibility that the ontological object i be matched with the lexical concept j. Analogously, an *Evidence Matrix* M_E contains in each element $M_E(i,j)$ the set of evidences which contribute to the computation of element $M_P(i,j)$ in the Plausibility Matrix.

The Discovery Phase. The linguistic dimension in the two matrices is far broader than the ontological one. An efficient enrichment process should thus consider a first discovery phase in which lexical anchors between the ontology and the LR are thrown to define possible candidates for linguistic enrichment. Each anchor represents a potential pointer from the ontology to the LR, and is discovered thanks to lexical similarity measures (use of string matching distances, possibly made smarter through knowledge of morphosyntactic properties of the natural language under analysis). In this phase it is important to drop as many anchors as possible, as they will represent the whole search space which is screened during the linguistic enrichment process. The trade-off is therefore lightly biased towards recall rather than precision, as the latter, in this case, is only important for reducing the computational cost of the process. The result of the discovery phase is thus a subspace L^A represented by all synsets in L which have been anchored as potential targets for semantic pointers.

The Semantic Enrichment Function. Once an L^A space has been extracted, we can then define the linguistic enrichment function f^{se} :

$$f^{se} : O \times L^A \mapsto [0..1] \tag{1}$$

This function maps pairs of elements from the ontology and the (restricted) linguistic resources into a confidence interval [0..1] representing the plausibility for assessing the presence of a semantic pointer between them.

The whole function f^{se} is realized through two main phases: by first the analysis of the linguistic and semantic similarities of the ontology and of the LR will lead the production of the *Evidence Matrix* M_E; the *Plausibility Matrix* M_P, based on the previously captured evidences, is then evaluated upon M_E.

There may exist mutual dependencies between contributions of features for different frame-synset pairs. For this reason, f^{se} is actually an iterative process $f^{se} = f^{se}(t)$; in particular, computation of the plausibility matrix takes this general form:

$$M_P(t) = f\left(M_E, M_P(t-1), M_P(0)\right) \tag{2}$$

To adopt a smarter notation for addressing plausibilities of single frame-synset pairs, we define:

$$p(F,S,t) \stackrel{def}{=} M_P(F,S) \quad \text{with } M_P = M_P(t) \tag{3}$$

Finally, we define a *candidate pair* <*F*,*S*> as a pair of elements $F \in O$ and $S \in L^C$, where $p(F,S,0) \neq 0$.

5 Instantiating f^{se}

The formulas in equations (1,2) are declarative forms representing classes of functions for realizing a semantic enrichment process, which are compatible with our model. In this section we present our realization of the semantic enrichment function, according to the two defined phases.

5.1 Computing Plausibilities

In our experiments, we specified this function according to the following desiderata:

1. *prizing* candidate pairs characterized by positive evidences
2. *punishing* candidate pairs characterized by negative evidences
3. evaluate quantitative factors associated to different kind of evidences (representing the *strength*, or *presence*, of the evidence)
4. take into account inherent polysemy of every label associated to ontology concepts

The following equation has thus been conceived for computing elements of the Plausibility Matrix:

$$p(t) = \frac{p_0 + \left(1 - \prod_{i=1}^{n}\left(1 - \rho(v_i, t)\right)\right) \cdot (1 - p_0)}{1 + \left(1 - \prod_{i=1}^{m}\left(1 - \rho(v_i, t)\right)\right) \cdot \left(\frac{1}{p_0} - 1\right)} \tag{4}$$

$p(t)$ is actually a smarter notation (to avoid abuse of indices in the formula) for $p(F, S, t)$, while $p_0 = p(0)$. p_0 value depends on τ_{high} and τ_{low}, two parameters representing the threshold over (resp. under) which a frame-synset pair must automatically be accepted (rejected), and on the ambiguity (number of senses for word) of the term denoting F, according to the following formula:

$$p_0 \doteq \frac{\tau_{high} - \tau_{low}}{a} + \tau_{low} \tag{5}$$

For each evidence v_i, a weighted feature is then computed through the function $\rho(v_i,t)$, whose value depends on the type of evidence v_i and on the instantiation of its associated parameters. In the following section details are provided about the structure of the different features v_i.

5.2 Extracting Evidences

Following the experiences we summarized in section 3, we formalized methods for extracting interesting evidences and for mapping their content into features for our f^{ee} function.

First of all, we define the search space over ontological relations which is investigated for every class of evidences:

Def. A *conceptual sphere* of a frame F over a set of relations R is a collection of frames linked to F through a relation $r \in R$. If r is a transitive relation, its closure may be limited to n allowed *hops*, depending on ontology's size; n is called the *range* of the sphere wrt the r dimension.

The conceptual sphere (sometimes called *context* in literature) for the Taxonomy-Alignment evidences has obviously been defined over the sole IS-A relationship, and its allowed range depends on the dimension of the ontology.

For gloss-based evidences we restricted the IS-A relation to cover only super concepts of the frame to be enriched; moreover, we considered both domain and range specifications of properties, and range restrictions of properties for specific classes. Computation of the sphere also depends on the nature of the ontological object under analysis. In figure 3 the algorithm for computing the conceptual sphere for classes, properties and individuals has been shown.

Taxonomy-alignment evidences: These kind of evidences assume the following form:

$$v \doteq \langle frame, synset, \mathrm{sgn} \rangle$$

where frame-synset is a *candidate pair* whose alignment influences the plausibility of the candidate pair which is being evaluated. The associated weighted features are computed through this formula:

$$\rho(v_i,t) \doteq \sigma_{TA} \cdot \mathrm{sgn} \cdot p(frame, synset, t-1)$$

where σ_{SA} is a coefficient related to this type of evidences and p(*frame*, *synset*, *t*-1) is the plausibility of the <*frame*, *synset*> pair at time *t*-1. sgn is 1 if *v* is a positive evidence, -1 if it is a negative one. Negative features for this kind of evidence are represented by configurations like that in fig. 2 below:

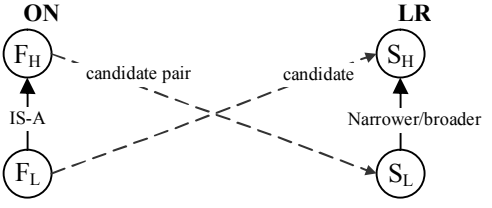

Fig. 2. negative evidence for sense-alignment

Here, <F_H,S_L> and <F_L,S_H> represent mutual negative influences, so that the plausibility of each pair is decreasing that of the other.

Gloss-mentioned Related Concepts: The strategy for extracting these evidences is based on the intuition that the glosses of the candidate synsets which best define a given frame *F*, may contain linguistic references to other concepts contained in the *conceptual sphere* of *F*.

The extraction of this kind of evidences is described by the following algorithm:

```
for each Frame rc ∈ ConceptualSphere do
    MtchLvl ← match(rc, gloss),
    if MtchLvl ≠ 0
    Evidences ← Evidences ∪ evd(GR, rc, MtchLvl)
    end if
end for
```

where *Evidences* is the set of evidences related to a given <*F*,*S*> pair, *ConceptualSphere* is the conceptual sphere built around *F* and *gloss* is the gloss of *S*. `GR` is a tag denoting membership of the extracted evidences to this class of features. *MtchLvl* is a degree of lexical similarity between the term from the gloss and the label of the matching concept: this value is obtained on the basis of raw string matching distances and comparative morphological analysis of the two terms.

Gloss-mentioned Generic concepts: Sometimes glosses of a candidate synset may disclose useful correlations between ontology concepts, which are unfortunately not captured by existing ontological relationships. In most cases nothing could be done and this phenomenon should simply be treated as a lack of information: the concepts can be recognized, upon human common sense, as potentially related (and they actually represent an evidence for a correct semantic pointer!), but they are not connected by any sort of relationship in the ontology (see related example in section 3.4)

Should the ontology be of modest size, offering a specification of a conceptualization of a very limited domain, it is nonetheless possible to consider each concept as somewhat related to the others. Under this hypothesis, given a <*F*, *S*> pair and a gloss *gloss* for synset *S*, this strategy considers as an evidence every occurrence of a term inside *gloss* which is also a label for a frame, even if no apparent relation with *F* exists.

```
for each term t ∈ gloss do
    F rc ← find(Ontology, t, MtchLvl),
    if rc≠ null
        Evidences ← Evidences ∪ evd(GG, rc, MtchLvl)
    end if
end for
```

Both these two gloss-based features are defined by the following expression:

$$v \doteq \langle MatchingLevel \rangle$$

and their contribution to f^{se} is:

$$\rho(v_i, t) \doteq \sigma_{GR/GG} \cdot MatchingLevel$$

```
computeConceptualSphere(Frame frm, int DepthRange) SET OF Frame
input frm: the class, property or individual which has been selected for linguistic enrichment
       DepthRange: the number of allowed hops along the IS-A relation for retrieving super concepts of frm
output     ConceptualSphere: the conceptual sphere surrounding frm
begin
FrameType type ← getOntoType(frm)
SET OF Frame ConceptualSphere ← {}
if (type = class  or type = property)
    ConceptualSphere ← ConceptualSphere ∪ getSuperConcepts(frm, DepthRange)
else  //frm is an instance
    Classes ← getClasses(frm)
    for each class ∈ Classes do
            ConceptualSphere ← ConceptualSphere ∪ {class} ∪ getSuperConcepts(class, DepthRange)
    end for
end if
if (type = class)
    for each property p, class c | frm.hasRestriction(p,c) or c.harRestriction(p,frm)
            ConceptualSphere ← ConceptualSphere ∪ { c } ∪ { p }
if (type = instance)
    for each property p ∈ ( frm.getOwnRelationalProperties() ) do
            ConceptualSphere ← ConceptualSphere ∪ { p } ∪ frm.getOwnPropertyValues(p)
end if
if (type = property)
    for each class c ∈ ( domain(frm) ∪ range(frm) ) do
            ConceptualSphere ← ConceptualSphere ∪ {class}
end if
return ConceptualSphere
end
```

Fig. 3. Algorithm for realizing the conceptual sphere for gloss-based evidences

Gloss-overlap between candidate synsets: A user manually doing linguistic enrichment knows the domain covered by the ontology and therefore would prefer senses whose glosses report domain related terms (see last example in section 3.4).

Analogously, this strategy checks for possible term overlaps between glosses of synsets which appear as candidates for enriching concepts appearing each in the conceptual sphere of the other. Of course, overlapping terms must be properly filtered, to remove co-occurrences of articles, particles and very common words.

Instead of adopting large stop-lists, which may reveal to be incomplete, we exploit the whole set of glosses of the same resource which is used for linguistic enrichment, as a large corpus for statistically determining the distribution of terms. Thresholds may then be established for filtering very common terms which bear no informative evidence. Formally:

```
for each Frame rf_i ∈ ConceptualSphere do
   for each synset s_ij ∈ candidateSynsets(rf_i) do
      let rfgloss[i,j] ← s_j.getGloss()
   end for
   for each term t, t ∈ gloss and t ∈ rfgloss[i,j]
      let freq = LR.getGlossFrequency(t)
      if !filter(freq)
      Evidences ← Evidences ∪ evd(GO, rf_i, s_i, freq)
      end if
   end for
end for
```

As for taxonomy-alignment, even this third gloss-based strategy produces mutual influences among features: the collected evidences are in fact dependent upon the plausibility of candidate <*rc*, s_i> pairs. Their structure is in fact:

$$v \doteq \langle MatchingLevel, frame, synset \rangle$$

and ρ assumes is computed this way:

$$\rho(v_i, t) \doteq \sigma_{GO} \cdot MatchingLevel \cdot p(frame, synset, t-1)$$

MatchingLevel is in this case also dependant on the frequency of the observed overlapping term.

6 Supporting Linguistic Enrichment of Ontologies in OntoLing

In line with OntoLing's highly modular architecture, we defined abstract layers for supporting automatic linguistic enrichment of ontologies at different levels. The schema in figure 4 extends OntoLing architecture [7] with new interfaces for:

- accessing a generic module for linguistic enrichment
- invoking standard methods for storing/caching information necessary for the enrichment task, from both the ontology the linguistic resource

We have provided a first realization of the enrichment interface through the implementation of the previously discussed techniques for semantic enrichment of ontologies. The storage and caching API have been realized according to diverse technologies and solutions, each of them thought for matching specific requirements. Mainly, these solutions can be split into two main categories:

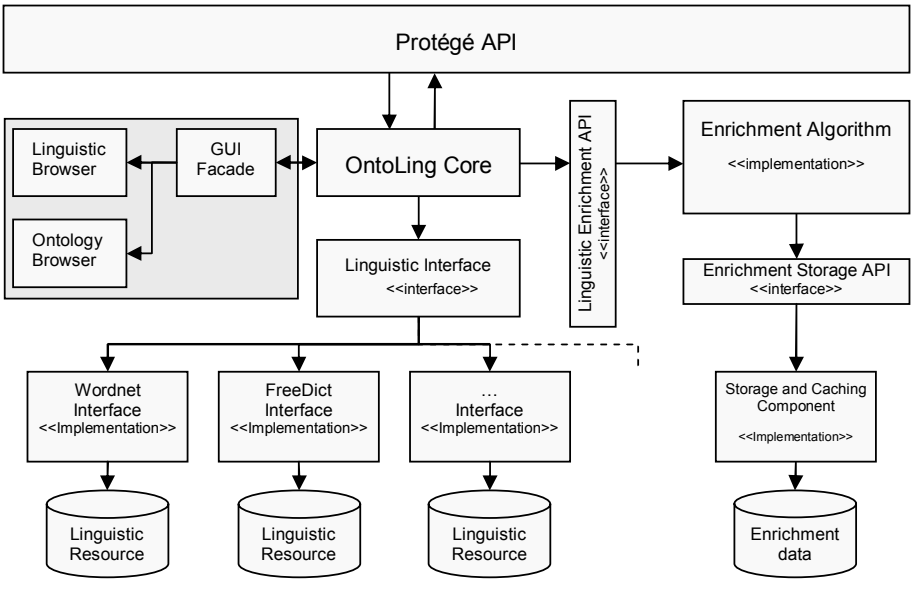

Fig. 4. OntoLing Architecture

- disk storage/caching of data
- in-memory storage

The first class has been thought to provide a scalable environment where even thousands of ontological objects, linguistic elements and relations between them can be easily handled. We provided one implementation for this category, based on use of database technologies. Different drivers may be loaded at run time for accessing available and preferred DBMSs. A dedicated driver for a popular java embedded DBMS [12] has been bundled into the application, to make OntoLing immediately operative without need of any external technology.

Aim of the second class of solutions is to maximize performances by storing data directly in memory, thus providing fast access to ontological and linguistic information during the enrichment process. This approach is ideal whenever size of the ontology and complexity of the linguistic resource do not require massive memory usage. Currently, two implementations are available to realize this solution:

- A prolog DB, which represents linguistic and ontological objects (and the relations between them) into sets of prolog facts
- An specific driver for an in-memory DBMS, sharing the same SQL implementation of already described DBMS solution.

Finally, a new interface has been produced for interacting with the user, which can initially choose between three different modalities:

1. manual enrichment (classic OntoLing behavior and interface)
2. completely automatic enrichment (which can be lately verified by the user and corrected wherever necessary)
3. step-by-step verification of prompted suggestions

In both modalities 2 and 3 the user can in any moment choose to stop the process and cycle through classes, properties and instances to verify their enrichment status. Fig. 5 provides an example of a step-by-step supervised process of semantic enrichment, by showing a dialog window which lets the user choose between different (WordNet) senses of the word "hit". The user can cycle through ontological data by selecting elements from the list on the left. Different colors in the central table indicate whenever a sense has been suggested by OntoLing, inspected, selected or confirmed by the user. Supplementary interfaces and interaction modalities will be developed in next releases of OntoLing also for other kinds of linguistic enrichment tasks. At present time, it is however possible to automatically pass from senses chosen during the semantic enrichment process, to their related linguistic information (synonyms and/or glosses) and use it for directly enriching ontological objects.

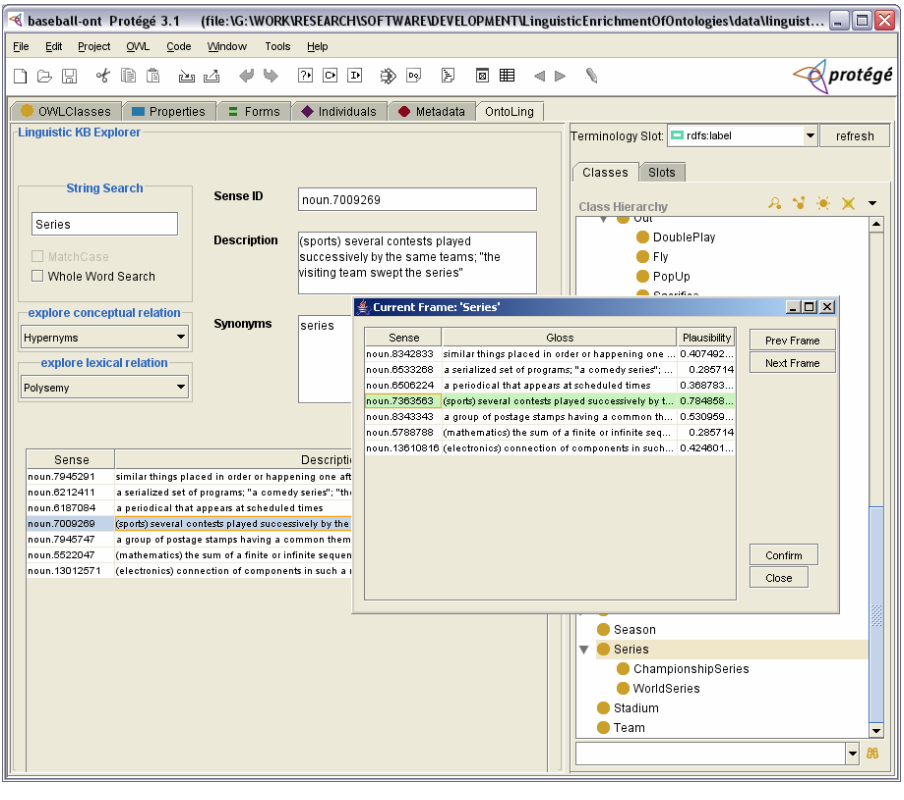

Fig. 5. Prompting the user with suggested WordNet senses for the word "hit"

7 Automatic Semantic Enrichment: Experimental Results and Final Remarks

To evaluate our enrichment process, we ran two experiments on enriching two public domain ontologies with synsets from WordNet. In reporting performances, standard Precision & Recall metrics have been adopted, instead of simple Hit Percentages, because for any given Frame, the system may propose a suggestion (right or wrong) or not. We also reported F-measure [10] which combines recall and precision in a single efficiency measure (it is the harmonic mean of precision and recall):

$$F = 2*(recall * precision)/(recall + precision)$$

The first experiment has been performed on the baseball ontology chosen for our examples. The ontology, is composed of 78 classes, 26 properties and 13 individuals. Of these objects, 60 classes and 21 properties were considered for semantic enrichment (we performed the experiment limiting to the ontology schema, so we provide statistics only for classes and properties) during the discovery phase. The number of non ambiguous concepts (including both classes and properties) is 20 (~ 24,7% of the whole concept set) while the average ambiguity, (measured as the average polysemy of considered terms, wrt WordNet synset structure), is ~ 9,16. Two annotators were initially hired to realize two documents (one per annotator) reporting the most evocative synset for each concept. The documents have then been compared and a final decision has been taken where discrepancies were found, to produce the oracle used in the experiments. The observed inter-annotator agreement on the two original documents has been however of 98.76% (one re-discussed decision out of the whole set).

Recall has been measured towards the number of concepts which can be enriched with the considered LR. The terms offered by any linguistic resource represent in fact the whole search space, and each evaluation of a linguistic enrichment process has only sense if considered wrt a specific LR. Fine tuning of evidence-typed σ-parameters has been performed over a collection of several small ontologies and/or portions of them, before running the experiment, whose results are reported in table 1.

The second experiment has been run on an ontology related to the university academic domain[3], developed in the context of the EU funded project MOSES (IST-2001-37244). This ontology has been built, in OWL language, over a preexisting DAML ontology[4] from the official DAML repository and finalized for representing the Italian university domain. As a consequence, while the original language in which concepts were expressed was English, many of the concepts added for describing the Italian academic institutions had only Italian labels. Though we plan for the future to define a two step enrichment process which is able to rely on multiple linguistic resources (for different languages) even for dealing with this kind of situations, we evaluated our algorithms over those parts of the ontology which were eligible for monolingual enrichment. More than half of the classes (100 out of 192) emerged during the discovery phase, while only a very small part of the properties (9 out of 100) have been discovered: this is probably due to the large amount of properties added during the customization to the Italian

[3] http://www.mondeca.com/owl/moses/ita.owl
[4] http://www.cs.umd.edu/projects/plus/DAML/onts/univ1.0.daml

Table 1. Evaluation of linguistic enrichment over two publicly available ontologies

Ontology	Precision	Recall	F-Measure
Baseball Ontology	80%	39,5%	52,89%
Moses Italian	81,48%	42,72%	56,05%

domain. We report in table 1 evaluation of the algorithm for both the experiments. Detailed analysis of the test data on the first experiment revealed that, though only 40% of the original corpus (ontology) has been correctly annotated with WordNet synsets, another 50% contains the right choice in a high ranked position (second or third suggestion, or even first but under the established plausibility threshold).

A similar observation holds for precision, where the 20% wrong hits gave only few plausibility points over the correct ones. This reveals to be in line with the intended nature of the task, which is to be seen as part of a computer-aided, linguistically motivated approach to ontology development, more than a mere disambiguation problem.

References

1. Beneventano D., Bergamaschi S., Guerra, F., Vincini, M: Building an integrated Ontology within SEWASIE system. In proceedings of the First International Workshop on Semantic Web and Databases (SWDB), Co-located with VLDB 2003 Berlin, Germany, September 7-8, 2003
2. V. R. Benjamins, J. Contreras, O. Corcho and A. Gómez-Pérez. Six Challenges for the Semantic Web. *SIGSEMIS Bulletin*, April 2004.
3. M. Dean and G. Schreiber, editors: OWL Web Ontology Language Guide. 2004. W3C Recommendation (10 February 2004).
4. C. Fellbaum: WordNet - An electronic lexical database. MIT Press, (1998).
5. J. Gennari, M. Musen, R. Fergerson, W. Grosso, M. Crubézy, H. Eriksson, N. Noy, and S. Tu. The evolution of Protégé-2000: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1):89–123, 2003.
6. H. Knublauch, R. W. Fergerson, N. F. Noy, M. A. Musen. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications *Third International Semantic Web Conference - ISWC 2004*, Hiroshima, Japan. 2004
7. M. T. Pazienza, A. Stellato: The Protégé OntoLing Plugin: Linguistic Enrichment of Ontologies in the Semantic Web. In *Poster Proceedings of the 4th International Semantic Web Conference (ISWC-2005)* Galway, Ireland, November, 2005
8. M.T. Pazienza, A. Stellato: Linguistically motivated Ontology Mapping for the Semantic Web. *Semantic Web Applications and Perspectives 2nd Italian Semantic Web Workshop (SWAP 2005)*, December 2005
9. S. Stamou, K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufiş, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou (2002). *BALKANET: A Multilingual Semantic Network for the Balkan Languages*. Proceedings of the International Wordnet Conference, January 21-25, Mysore, India, 12-14.
10. C. J. Van Rijsbergen, *Information Retrieval*. 2nd edition, London, Butterworths, 1979
11. P. Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht, 1998
12. http://www.daffodildb.com/