Linguistic Enrichment Of Ontologies: a methodological framework

Maria Teresa Pazienza, Armando Stellato

AI Research Group, DISP, University of Rome, Tor Vergata
Via del Politecnico 1 00133 ROMA (ITALY)
{pazienza,stellato}@info.uniroma2.it

Abstract

We introduce here a framework for adding Linguistic Expressivity to conceptual knowledge, as represented in ontologies. 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 in fact the need for a linguistically motivated 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 content mediation between autonomous agents gets far more chances than otherwise.

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.

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.

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.

We introduced the expression "Linguistic Enrichment of Ontologies" to identify a series of different processes sharing the common objective of augmenting the linguistic expressivity of an ontology through the exploitation of existing Linguistic Resources (LRs, from now on). These processes strongly depend on the selected LRs but also on the task the ontology is dedicated to. In the following sections (sections 1.1-1.3) we describe some of the possible scenarios in which different enrichment tasks may be required, we then provide more information about one of these tasks (section 2) and will describe a framework (sections 3 and 4) for automatizing much of the work it requires. Finally (section 5), experimental evidence and quality of the suggested methods will be discussed.

1.1. 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 an agent society with knowledge mediators relying on a common form of knowledge. This common knowledge is represented by so called "upper ontologies", or "upper models" which contain a first stratification of general concepts. In a few cases (Beneventano et al., 2003), instead of an ontology, the semantic structure of an existing (WordNet: Fellbaum, 1998) linguistic resource has been adopted as a interlingua for guaranteeing communication between autonomous distributed agents.

1.2. Explicit Linguistic Enrichment

In case of no committed semantic agreement 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). In-deed, 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 onto-logical data, facilitate reuse of existing knowledge, which is made more comprehensible also for humans.

1.3. Producing Multilingual Ontologies

Though English is commonly agreed to be a "lingua franca" all over the world, much effort must be (and is being) spent to preserve other idioms expressing different

cultures. Multilinguality has been cited as one of the six challenges for the Semantic Web (Benjamins et al., 2004). Exploitation of existing bilingual resources may thus 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.

2. Techniques for Semantic Linguistic Enrichment of Ontologies

In this work, we focus on the first of previously mentioned tasks: semantic enrichment of ontologies. This represents in fact a first necessary step through which all of the other tasks may be accomplished.

We thus designed a semantic enrichment process which can be run either semi-automatically, prompting ontology developers with suggestions to be supervised (approved, rejected or demanded), or executed as a totally automated procedure. These two options represent in fact desirable features for any application intended to support a linguistically motivated ontology development.

For our experimental setup, we adopted the terminology we defined in the Linguistic Watermark[1] (LW, from now on) framework (Pazienza & Stellato, 2006): a collection of interfaces for describing and manipulating linguistic resources. Through instantiation of these interfaces, ontology development applications may provide a uniform framework for accessing linguistic knowledge from different LRs, and use this content to enrich formal ontological data.

2.1. 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 re-sources: OWL and WordNet.

OWL (Dean & Schreiber ed, 2004) 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 (Fellbaum, 1998) 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 exist for many other languages (Vossen, 1998; Stamou et al, 2002) which have thus favored a large diffusion of the model which inspired the original English version. As WordNet model closely matches the LW configuration which best fits for the semantic enrichment task, we will adopt 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.

2.2. The Semantic Enrichment task

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

Depending on their characterizing Watermark, not all LRs are exploitable for semantic enrichment of an ontology; in particular, only those resources whose model is compliant with the `ConceptualizedLR` (see Linguistic Watermark specifications) and at least one of `TaxonomicalLR` and `LRWithGlosses` interfaces, can be considered for this task.

Before detailing the model underlying our enrichment process, we describe a few empirical results we collected during our research. These results took the form of morphosyntactic and semantic evidences observed over several pairs of ontologies and linguistic resources.

All the reported examples refer to semantic enrichment of a DAML ontology[2] about baseball, downloaded from the DAML library of ontologies[3], using WordNet as a source for linguistic knowledge.

2.3. Taxonomy-Alignment evidences

In case the semantic structure of a given LR is organized as a taxonomy of broader/narrower linguistic concepts (the LR is a `TaxonomicalLR`), similarities between this taxonomy and that of the ontology may provide useful evidences for an enrichment task. The IS-A relation of ontologies (under the considered logic or frame based models) has well defined semantics, while taxonomical links of LRs may often bear informal and ambiguous relationships; 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-

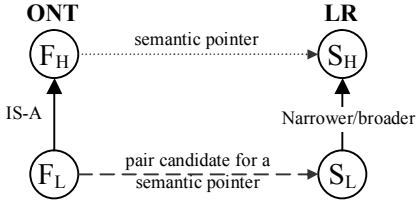

Figure 1: The sense-alignment square

[1] Linguistic Watermark is publicly available at: http://ai-nlp.info.uniroma2.it/software/LinguisticWatermark/

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

[3] http://www.daml.org/ontologies/

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 Figure 1, 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:

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 Word-Net 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 the *Hit*–`noun.124696` and the *Double*–`noun.39042` result as good candidates for being linked through a semantic pointer.

Analogously, a cross-link between a candidate pair and a semantic pointer represent a negative evidence for the candidate pair (Figure 2 below):

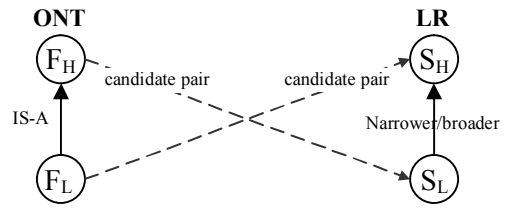

Figure 2: negative evidence for taxonomy-alignment

Though two taxonomical operators may present slight semantic differences, it is very unlikely for a configuration like this to exist so, in most of the cases, the candidate pair must not be connected through a semantic pointer (or the already existing semantic pointer should be verified).

The above examples represent situations involving a semantic pointer and a candidate frame-synset pair, however, in most of the cases, it will happen that there will be no direct cause-effect between an assessed pointer and a candidate pair. It is more frequent to face two (or chains of) candidate pairs each contributing to each other's plausibility: a proper model for representing evidences should take into account these mutual dependencies.

2.4. 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 previous example, we can learn that a "double" is a kind of "base hit" (though the meaning of "hit" is not formally specified), even if the resource lacked of a taxonomical structure, binding the two concepts together in a broader/narrower relation.

A further example is represented by the class *Division*. 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 data 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 labeled 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, 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, contain 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.

3. The Feature Model

To take into account all previous considerations, and to maintain a scalable approach towards new possible strategies and LW 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.

3.1. The Discovery Phase

The linguistic dimension in the two matrices is far broader than the ontological one. An efficient enrichment process should in fact consider a first *discovery* phase in which lexical anchors between the ontology and the LR are thrown. 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 thus 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 in fact a subspace L^A rep-resented by all synsets in L which have been anchored as potential targets for semantic pointers.

3.2. 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 se-mantic 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*; the *Plausibility Matrix*, based on the previously captured evidences, is then evaluated.

There may exist mutual dependencies between contributions of features (which we call dynamic) for different frame-synset pairs (as observed for taxonomy-alignment evidences and for some of the gloss-based evidences). 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:

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

The Plausibility Matrix is thus not a single matrix, but a system which evolves over time, its content being the product of the observed evidences, of the system's history, and (possibly) of human intervention.

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

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

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

4. 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.

4.1. 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) 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 a (number of senses per 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 how the structure of the different features v_i.

4.2. Extracting evidences

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

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

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 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 the average domain ontology, *n* is typically ∞, while its value must be restricted when dealing with very large – and deep – ontologies).

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 proper-ties, 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, proper-ties and individuals has been shown.

4.2.1. Taxonomy-alignment evidences:

These kind of evidences assume the following form:

$$v \doteq \langle frame, synset, \mathrm{sgn} \rangle \tag{6}$$

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:

$$\rho(v_i, t) \doteq \sigma_{TA} \cdot \mathrm{sgn} \cdot p(frame, synset, t-1) \tag{7}$$

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 (as represented in figure 2, where $<F_H,S_L>$ and $<F_L,S_H>$ represent mutual negative influences, so that the plausibility of each pair is decreasing that of the other).

4.2.2. 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 de-scribed 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

```
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    //obj 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.hasRestriction(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 ∪ {c}
end if
return ConceptualSphere
end
```

Figure 3: Algorithm for realizing the conceptual sphere for gloss-based evidences

of raw string matching distances and comparative morphological analysis of the two terms.

4.2.3. 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 2.4)

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
   Frame rc ← find(Ontology, t, MtchLvl),
   if rc≠ null
      Evidences ← Evidences ∪ evd(GG, rc, MtchLvl)
   end if
end for
```

Obviously, if both the previous strategies are applied, the results of the first one must be subtracted from those of the second one, which totally includes them. The second strategy is in fact less effective, on average, than the first one, and is generally used to augment the recall at the cost of a slightly minor precision. The evidences discovered by both strategies must thus be counted only on the first one, which has however a greater impact on the computation of the Plausibility Matrix

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

$$v \doteq \langle MatchingLevel \rangle \tag{8}$$

and their contribution to fse is:

$$\rho(v_i, t) \doteq \sigma_{GR/GG} \cdot MatchingLevel \tag{9}$$

4.2.4. Gloss-overlap between candidate synsets

Humans have the advantage of a wider knowledge about the world with respect to automatic processes. A user performing manual linguistic enrichment knows that the ontology is about baseball and therefore will probably check all the senses whose glosses report this term (see last example in section 2.4).

To reproduce such a behaviour, 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:

$$v \doteq \langle MatchingLevel, object, synset \rangle \tag{10}$$

and ρ assumes is computed this way:

$$\rho(v_i, t) \doteq \sigma_{GO} \cdot MatchingLevel \cdot p(object, synset, t-1) \tag{11}$$

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

4.3. Frame-synset pairs as actors

SA and GO features (and, in general, *dynamic features*) form thus a network of mutual dependencies, where plausibilities of different candidate pairs depend on other pairs' plausibilities. Like in Conway's "Game of Life" (Berlekamp et al., 1982), correlated candidate pairs may associate into sort of "corporations" which tend to augment the *strength* (plausibility, in our case) of each of their members, thus lessening the chances of other candidates which, being cut away from these trust, are deemed to lose their run. At the same time, rare but not unusual "black sheeps" (represented by strong candidate pairs acting as bad evidences for others), may condemn whole sets of potential candidates to lose terrain in favour of others.

4.4. Reliability of gloss-based evidences

It emerges the risk, for gloss-based evidences, that they may be based on a mislead correlation of terms from glosses and labels for concepts, with the former indicating different meanings of those expressed by the related ontological concepts. Though sporadic occurrences of this phenomenon are indeed a possibility for each considered evidence type, their effects are generally cancelled out over large numbers of evidences, which, on average, present right correlations. In some cases the co-occurring terms bear in fact no polysemy at all, moreover, as a general consideration, several studies (Madhu and Lytle, 1965; Resnik, 1997) seem to support the hypothesis of a semantically conservative behavior of words wrt their use in a given specific context, so that even ambiguous

expressions tend to assume the same meaning if considered inside the same ontological framework.

5. Automatic Linguistic Enrichment: experimental results and final remarks

Fine tuning of evidence-typed σ-parameters has been performed over a collection of several small ontologies and/or portions of them. We then ran two experiments on two public domain ontologies, reporting performance in terms of standard precision & recall metrics.

We stress the fact that our framework foresees human effort both as a verification of automatic suggestions and as possible intervention on the enrichment process: a very few human decisions can in fact greatly affect the outcome of the automatic enrichment process, as they represent strong evidences (human choices are considered assessed semantic pointers, and have thus plausibility equal to 1) for correlation-based dynamic evidences.

Nonetheless, our experiments aim at evaluating the enrichment process also as a completely automatic procedure.

Recall has been measured towards the number of concepts which can be enriched with the considered LR. The linguistic resource thus determines the whole search space, and each evaluation of a linguistic enrichment process has only sense if considered wrt a specific LR. Regarding Precision, the "suggest and wait for confirm" threshold-based approach, which is well suited for a human centered process, has been given out for an immediate outcome of the highest ranking synset, chosen among all the candidate ones for every concept.

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), was ~ 9,16. The oracle has been manually produced by two annotators which realized two documents reporting the most evocative synset for each concept. These documents have been compared and a final decision has been taken where discrepancies were found. The observed inter-annotator agreement has been however of 98.76% (one re-discussed decision out of the whole oracle).

The second experiment has been run on an ontology related to the university academic domain[4], 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[5] 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 domain.

We report in the following table evaluation of the algorithm for both the experiments.

Ontology	Precision	Recall
Baseball Ont	80%	39,5%
Moses Italian	81,48%	42,72%

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

Detailed analysis of the test data on the first experiment revealed that, though only 40% of the original corpus (ontology) has been correctly enriched, another 50% contains the right choice as the second or third ranked suggestion. A similar observation holds for precision, where the remaining 20% wrong suggestions gave only some percentage 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.

6. Acknowledgements

This work has been partially funded under the GALILEO Cuspis Project (GJU/05/2412/CTR/CUSPIS) started inside the GALILEO Joint Undertaking User Segment, Call 1A: User Community/GNSS for Special Users Community, under the 6th Framework Programme of European Commission.

7. References

Beneventano D., Bergamaschi S., Guerra, F., Vincini, M: Building an integrated Ontology within SEWASIE system. In proceedings of the First In-ternational Workshop on Semantic Web and Data-bases (SWDB), Co-located with VLDB 2003 Ber-lin, Germany, September 7-8, 2003

Benjamins, V. R., Contreras, J., Corcho, O., and Gómez-Pérez, A.: Six Challenges for the Se-mantic Web. SIGSEMIS Bulletin, April 2004.

Berlekamp, E., Conway, J., and Gut, R.: The game of Life, *Winning Ways for your Mathematical Plays*, vol. 2, Academic Press, 1982, pp. 817-849

Dean, M. and Schreiber, G. editors: OWL Web On-tology Language Guide. 2004. W3C Recom-mendation (10 February 2004).

Fellbaum, C: WordNet - An electronic lexical da-tabase. MIT Press, (1998).

Madhu, Swaminathan and Dean Lytle, 1965. A figure of merit technique for the resolution of non-grammatical ambiguity. *Mechanical Translation*, 8(2):9–13

Pazienza, M.T. and Stellato, A.: Linguistic Enrichment of Ontologies: a methodological framework. *Second*

[4] http://www.mondeca.com/owl/moses/ita.owl

[5] www.cs.umd.edu/projects/plus/DAML/onts/univ1.0.daml

Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006), held jointly with LREC2006, Magazzini del Cotone Conference Center, Genoa, Italy, 24-26 May 2006

Resnik, P., 1997. Selectional preference and sense disambiguation. *In Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How*?, Washington, April 4-5, 1997

Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufiş D., Koeva S., Totkov G., Dutoit D., Grigoriadou M. (2002). BALKANET: A Mul-tilingual Semantic Network for the Balkan Languages. Proceedings of the International Wordnet Conference, January 21-25, Mysore, In-dia, 12-14.

Vossen. P: EuroWordNet: A Multilingual Data-base with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht, 1998