Linguistically motivated Ontology Mapping for the Semantic Web

Maria Teresa Pazienza, 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. Knowledge Sharing is a crucial issue in the Semantic Web: SW services expose and share knowledge content (expressed through ontologies and related knowledge bases) arising from distinct languages, locales, and personal perspectives; in this scenario, semantic alignment approaches play a pivotal role, providing viable solutions for integrating heterogeneous resources, still maintaining their local independence. We focus here on a 3-step approach to ontology mapping, which is strongly based on the exploitation of (monolingual and multilingual) linguistic resources for content publishing and discovery, and on a human intervention for supervising the process and assessing semantic links between mapped resources. Our methodology is also being supported by the development of dedicated tools for accompanying knowledge engineers and users across the different steps of creating and integrating ontology resources.

1 Introduction

Ontology Mapping is a task requiring considerable human involvement, which can only in part be mitigated by the support of specific ad-hoc tools.

In fact, ontologies structure their content according to the formal semantics of their underlying model as well as to the domain conceptualization which developers are acquainted with. As this conceptualization may change from person to person, the inner meaning of their formal content is inherently bound to the specific applications which rely on them. Mapping the content of two ontologies is thus like committing to a shared interpretation of the two semantic structures, which is a model for both and which is compliant with the applications which insist on them.

If the inner meaning of independently developed ontological knowledge is thus a shadowed and ambiguously identifiable concept, the same should not hold for its surface expression, which could at least provide strong evidences for recognizing identities and similarities across different information sources. This is indeed the task which can mostly be supported, being it related to the discovery of objective and measurable clues and facts, and thus being easily managed in an automatic approach.

Ontologies, as they are organized nowadays, offer instead a completely different view: concepts expressed by hardly recognizable labels, lexical ambiguity represented by phenomena like synonymy and polysemy and use of different idioms which derive from diverse cultures, all together make a strong opposition towards the readability of their content, thus turning away the dream of a scenario where automatic semantic interoperability is made possible.

We thus put forward a 3-step methodology for knowledge integration, which consists in:

1. expressing ontological content in a linguistically motivated fashion, as a necessary part of the development of ontologies
2. automatically discovering linguistic and semantic evidences to suggest conceptual similarities during automatic ontology alignment.
3. supporting users in the process of producing assessed ontology mapping documents, offering reliable knowledge for providing semantic links across different information sources

To this end, exploitation of existing linguistic resources is an ideal feature for any tool wishing to support users during the first two steps. Providing a clear and uniform interface for Linguistic Resources (from now on, LRs) is thus strongly expected, as it would represent an abstraction guaranteeing independence from the intended task, models and implementations, and allowing for easy scalability towards newly available resources.

2 "Lexicalizing" The Semantic Web: A 3-Step Approach To Knowledge Integration

In this paper, we give an overview of our 3-step methodology, presenting our framework for a linguistically motivated approach to ontology development and discussing how this approach could give raise to new scenarios for ontology mapping. In this section, a brief introduction is given for each of the three presented steps, while more details are provided in the next sections.

2.1 Linguistically motivated Ontology Development

As a first important step of our methodology, we should reconsider the process of Ontology Development to include the enrichment of ontological content with proper lexical expressions in natural language. Knowledge Integration would benefit of such a new perspective over ontological data, as it provides, once for all, a whole series of evidences which can then be exploited by ontology mediators to align different conceptualizations.

Ontology Development tools should reflect this need, supporting users with dedicated interfaces for browsing linguistic resources: these are to be integrated with classic views over knowledge data such as class trees, slot and instance lists, offering a set of functionalities for linguistically enriching concepts and, possibly, for building new ontological knowledge starting from linguistic one.

Unfortunately, while ontologies have undergone a process of standardization which culminated, in 2004, with the promotion of OWL [4] as the official ontology language for the semantic web, linguistic resources still maintain heterogeneous formats and follow different models, which make tricky the development of such an interface. We will address this problem in section 3 and discuss our approach in defining a uniform model for accessing linguistic resources; then in section 4 we will introduce Ontoling [10], a Plugin for the Protégé [6] ontology development environment dedicated to linguistic enrichment of ontologies. Section 5 provides instead a methodology for automatically performing the enrichment process.

2.2 Automatic Ontology Alignment

Exploitation of linguistic resources is also a characterizing feature of our ontology mapping approach. Once ontologies have been linguistically enriched, they offer a wide range of evidences for guessing proper semantic connections with other ontologies. Natural languages descriptions of concepts, presence of synonyms and possible translations into different idioms, all of them properly added during the linguistic enrichment phase, offer reliable lexical anchors for supporting an ontology mapping process. This phase may be considered inside two different scenarios:

1. it can either be considered as part of a machine-supported task which aims at discovering possible mappings between ontologies, which need then to be accepted through human intervention (see next phase), or seen as
2. a completely automatic process, as for a dialogue between two agents (or Semantic Web Services), where each of them is trying to properly match concepts expressed by its interlocutor, against its internal representation of the domain. In this case the process may not always bring to correct matches; also, it is really difficult to automatically discover complex relationships which go far beyond simple 1-to-1 mappings. On the other side, the idea of a web populated by independent agents exchanging information and presenting the collected results to the user ([12, 13]) is indeed a fascinating one.

In section 6 we discuss our approach to automatic alignment of ontologies and depict a scenario for semantic coordination of distributed, heterogeneous, information sources, which is strongly driven by exploitation of linguistic resources.

2.3 Assessing semantic alignments into ontology mapping documents

When automatic mapping discovery is seen as part of a more complex alignment task (scenario 1 of phase 2, described above), that phase has to be followed by another one where a final and reliable mapping document is produced. In this last phase the screening of all the discovered semantic links produces suggestions for a user-centered process in which a complete and sound mapping document is finalized. With complete and sound we mean a set of mappings where every concept from the considered ontologies has been covered (completeness) and where all the reported mappings commit to a shared interpretation which is a model for both ontologies (soundness).

In this scenario, ontology mapping representations should account for the needs of both phase 2 and 3, so that, ideally, an ontology mapping tool could first allow an automatic matcher to produce a draft mapping document, which needs then only small changes (or no modification at all) where the suggested mappings are accepted by the user. In [11] we presented XeOML, an extensible mapping language, with the characteristic of separating structural aspects of mappings from any kind of additional information connected with their assertion (motivations, perspectives etc…). Extensions to the language could thus be developed to realize the specific views we need in our framework.

3 Linguistic Enrichment of Ontologies

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.

The nature of these processes strongly depends on the LR being exploited for enriching the ontology and on the specific goals which the enrichment needs to achieve.

In the discussed framework, these goals are represented by the necessity of making ontologies easy shareable in a distributed environment, even when no agreement is established in advance about how their content must be exposed. Boosting conceptual knowledge with information coming from LRs may thus provide a first "common sense" layer upon which heterogeneous information sources may be compared and hypotheses for their alignment can be established; successive semantic analysis can then verify these hypotheses and let mappings be assessed between the information sources.

3.1 A classification of Linguistic Enrichment tasks

Though this intent is clear, still heterogeneities in the way ontologies are shared and distributed, can influence adoption of different techniques for ontology enrichment.

Here we describe some of the possible enrichment tasks we contemplate in our framework.

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 representing ontological knowledge. What is required is just identification of pointers from ontological data to semantic elements of the linguistic resource. Access to pure linguistic information (if needed) is then guaranteed by the links between the semantic and linguistic structure of the LR.

A possible scenario is represented by an agent society with knowledge mediators relying on a common form of knowledge. This common knowledge is traditionally represented by so called "upper ontologies", or "upper models" which contain a first stratification of general concepts. However in a few cases [1], instead of an ontology, WordNet [5] synsets have been adopted as a interlingua for guaranteeing communication between autonomous distributed agents. When a LR is used in this way, there are no formal links between distributed ontologies (i.e. ontologies do not need to "OWL-import" any reference ontology, so there is no need to maintain consistency of a unique, big ontology constituted of the global reference ontology and of all connected local ontologies); instead loose mappings (in the spirit of [9]) between their contents and the semantic structure of a LR provide a common reference vocabulary for enabling semantic coordination between different peers.

The supposition behind this scenario is that distributed peers (agents/services and their related ontologies) bear some form of semantic commitment towards one or more LRs, which are elected as interlingua for communication.

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). 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.

Due to inherent ambiguity of natural language, this kind of linguistic enrichment provide less reliable evidences for ontology mediators. Nonetheless, redundancy of expressions for denoting the same concept may cancel out ambiguity issues and provide instead more clues for identifying similarities. Moreover, a language-aware approach to ontology mapping able to recognize potentially ambiguous information, is less prone to those semantic mismatches which represent typical pitfalls for pure string-matching or term-matching based approaches.

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. As a consequence, Multilinguality has been cited as one of the six challenges for the Semantic Web [2].

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.

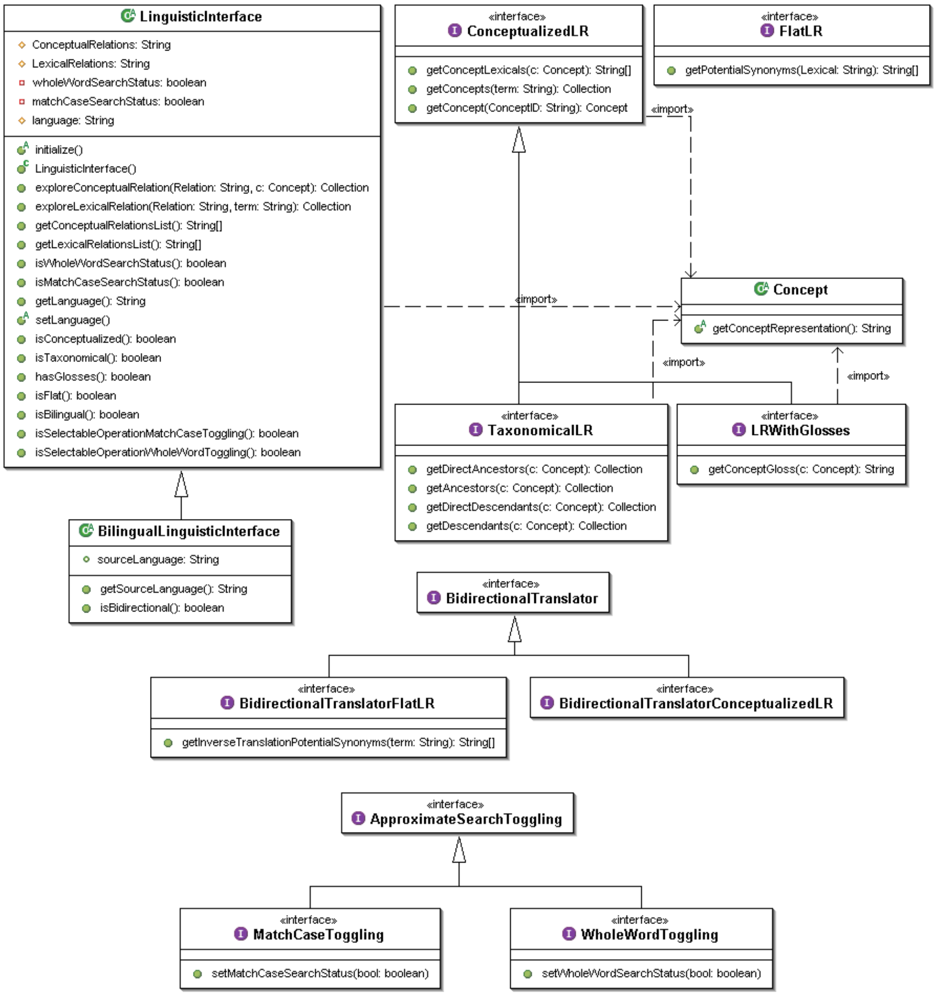

LinguisticInterface
ConceptualRelations: String
LexicalRelations: String
wholeWordSearchStatus: boolean
matchCaseSearchStatus: boolean
language: String
initialize()
LinguisticInterface()
exploreConceptualRelation(Relation: String, c: Concept): Collection
exploreLexicalRelation(Relation: String, term: String): Collection
getConceptualRelationsList(): String[]
getLexicalRelationsList(): String[]
isWholeWordSearchStatus(): boolean
isMatchCaseSearchStatus(): boolean
getLanguage(): String
setLanguage()
isConceptualized(): boolean
isTaxonomical(): boolean
hasGlosses(): boolean
isFlat(): boolean
isBilingual(): boolean
isSelectableOperationMatchCaseToggling(): boolean
isSelectableOperationWholeWordToggling(): boolean
«interface»
ConceptualizedLR
getConceptLexicals(c: Concept): String[]
getConcepts(term: String): Collection
getConcept(ConceptID: String): Concept
«interface»
FlatLR
getPotentialSynonyms(Lexical: String): String[]
«import»
«import»
Concept
getConceptRepresentation(): String
«import»
«import»
«interface»
TaxonomicalLR
getDirectAncestors(c: Concept): Collection
getAncestors(c: Concept): Collection
getDirectDescendants(c: Concept): Collection
getDescendants(c: Concept): Collection
«interface»
LRWithGlosses
getConceptGloss(c: Concept): String
BilingualLinguisticInterface
sourceLanguage: String
getSourceLanguage(): String
isBidirectional(): boolean
«interface»
BidirectionalTranslator
«interface»
BidirectionalTranslatorFlatLR
getInverseTranslationPotentialSynonyms(term: String): String[]
«interface»
BidirectionalTranslatorConceptualizedLR
«interface»
ApproximateSearchToggling
«interface»
MatchCaseToggling
setMatchCaseSearchStatus(bool: boolean)
«interface»
WholeWordToggling
setWholeWordSearchStatus(bool: boolean)

Fig. 1. The Linguistic Watermark

Maintaining expressions (and even concepts) belonging to different cultures and idioms is important to preserve and respect the cultural heritage of every country; at the same time, the role of multilinguality is to make ontological information explicitly accessible in different languages, thus reducing the cost of mediating its content across different domains and idioms.

3.2 The Linguistic Watermark: a uniform interface for accessing Linguistic Resources

Along with the analysis of a general interface for linguistic resources, it emerged the logical independence which it could maintain with respect to its possible embedding applications. Our experience pointed out usefulness in diverse natural language related applications like Ontology Mapping, Question&Answering and Information Extraction, where support for multilinguality and a wider linguistic awareness could be, if not necessary, at least useful for improving performances. Moreover, the interface could also act as a sort of unique fingerprint for describing the underlying resource for which access is provided, its information being exploitable in many application-dependant contexts.

For this reason, we introduced the notion of Linguistic Watermark (LW), as the series of characteristics and functionalities which distinguish a particular resource inside our framework. As we can observe from the Class Diagram in Fig. 1, we sketched a sort of classification of *linguistic resources*, with the addition of operational aspects. Linguistic resources are in fact structured and described in terms of their features and how their lexical information is organized; the structure has then been completed with query methods for accessing resource's content. We thus implemented this operational ontology as a java package on its own, which can externally be imported by any application willing to exploit natural language resources like lexicons and terminologies. The core of the package is composed of an Abstract Class, named `LinguisticInterface`, which is both the locus for a formal description of a given linguistic resource and a service-provider for exposing the resource specific methods. The other abstract classes and interfaces in the package, which can be implemented or not, depending on the profile of the resource being wrapped, provide instead the signatures for known interface methods.

We have currently developed several implementations of the Linguistic Watermark. Two of them, the Wordnet Interface and the last DICT Interface, being related to freely available resources, have been made publicly available on the Linguistic Watermark site[1].

The first one is an almost totally complete implementation of the Linguistic Watermark. The Wordnet Interface is in fact a `ConceptualizedLR`, because its linguistic expressions are clustered upon the different senses related to the each term. These senses – "synsets", in Wordnet terminology – have been implemented through the *Concept* interface, which we see bounded by the import statement in the class diagram. Wordnet is a `LRWithGlosses`, as glosses are neatly separated from synonyms and organized in a one-to-one relation with synsets. Finally, Wordnet Interface implements `TaxonomicalLR`, as its indexed word senses are organized in a taxonomy of more specific/more generic objects.

The other one, DICT Interface, is based on the Dictionary Server Protocol (DICT) [15], a TCP transaction based query/response protocol that allows a client to access dictionary definitions from a set of natural language dictionary databases. The DICT interface is *conceptualized* too, though its word senses are not indexed as in Wordnet (that is, it is not possible to correlate senses of two different terms upon the same meaning). DICT Interface is also a `BilingualLinguisticInterface`, as its available wordlists provide translations for several idioms.

Other available interface classes denote *Flat* resources (as opposed to `Conceptualized` ones), which contain flat lists of linguistic expressions for each defined term, and `BidirectionalTranslators`, which represent a further specialization of

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

Bilingual Linguistic Interfaces providing bidirectional translation services. Other interfaces (`ApproximateSearchToggling`) are not directly related to the characteristics of the wrapped LR, but to search functionalities which have been provided for it.

As previously mentioned, we defined two classes of methods for browsing LRs: those defined in advance in the interfaces, which can thus be exploited inside automatic processes, and other very specific resource-dependent methods, which are loaded at run-time when the LR is interfaced to some browsing application (e.g. Ontoling). Two methods available in `LinguisticInterface`: `getLexicalRelationList` and `getConceptualRelationList` act thus as service publishers, the former providing different methods for exploring lexical relations among terms or relating terms to concepts, the latter reporting semantic relations among concepts. Through these methods, the Wordnet Interface makes available to the user all the semantic relations contained in Wordnet.

4 Ontoling

Ontoling is a tool dedicated to assist ontology developers and users in the process of enriching ontologies with information coming from available linguistic resources. It has been developed as a plug-in for the popular ontology editing tool Protégé [6].
The architecture of the Ontoling plug-in (Fig. 2) is based on three main components:

1. the GUI, characterized by the Linguistic Resource browser and the Ontology Enrichment panel
2. the external library *Linguistic Watermark*, which has been presented in the previous section, providing a model for describing linguistic resources
3. the core system

and an additional external component for accessing specific linguistic resources. This

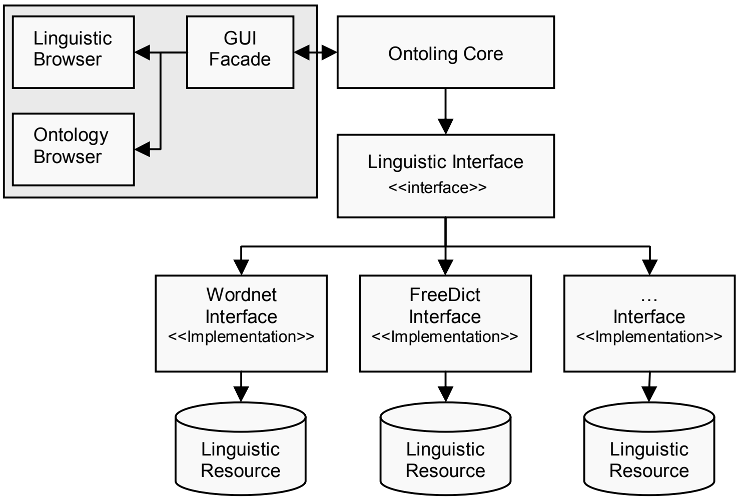


Fig. 2. Ontoling Architecture

component, which can be loaded at runtime, must implement the classes and interfaces contained in the LW library, according to the characteristics of the resource which is to be plugged. In the following sections we provide details on the above components.

4.1 Ontoling Core Application

The core component of the architecture is responsible for interpreting the Watermark of linguistic resources and for exposing those functionalities which suit to their profile. Moreover, the behavior of the whole application is dependant on the nature of the loaded resource and is thus defined at run-time. Several methods for querying LRs and for exposing results have been encapsulated into objects inside a dedicated library of behaviors: when a given LR is loaded, the core module parses its Linguistic Watermark and assigns specific method-objects to each GUI event.

With such an approach, the user is provided with a uniform view over diverse and heterogeneous linguistic resources, as they are described in the LW class diagram, and easily learns how to interact with them (thus familiarizing with their peculiarities) by following a policy which is managed by the system.

For example, with a *flat* resource, a search on a given term will immediately result in a list of (potential) synonyms inside a dedicated box in the GUI; instead, with a *conceptualized* resource, a list of word senses will appear in a results table at first, then it will be browsed to access synonymical expressions related to the selected sense. Analogous adaptive approaches have been followed for many other aspects of the Linguistic Watermark (mono or bidirectional Bilingual Translators, presence of glosses, Taxonomical structures and so on…) sometimes exploding with combinatorial growth.

Future development of Ontoling will go in the direction of considering supervised techniques for automatic ontology enrichment; selecting and modeling the right strategies for the adopted LRs is another task the core module is in charge for.

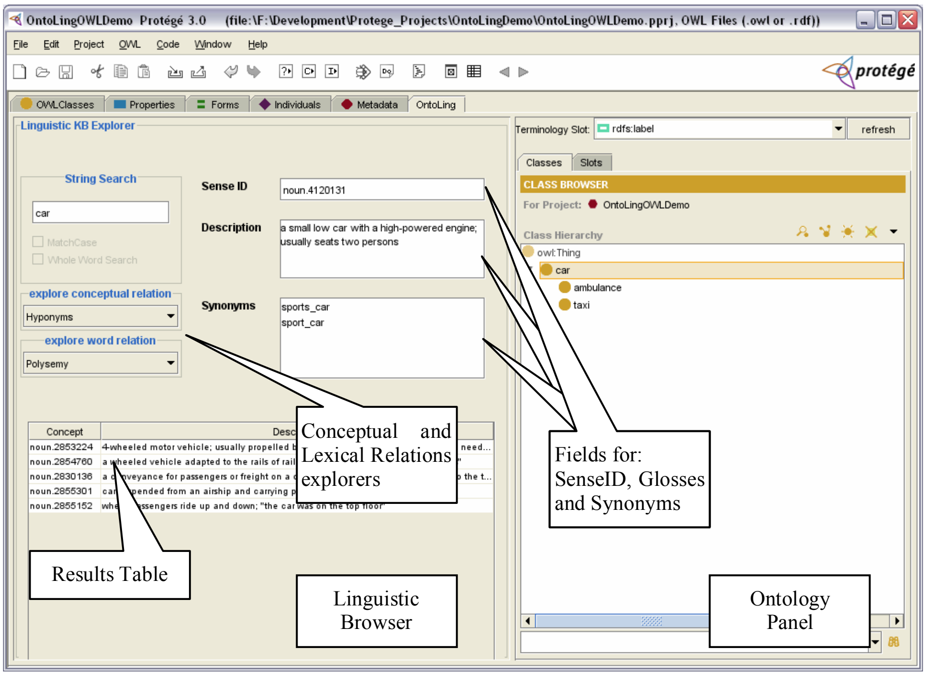
4.2 Ontoling User Interface

Once activated, the plugin displays two main panels, the Linguistic Browser on the left side, and the Ontology Panel on the right side (see Fig. 3).

The Linguistic Browser is responsible for letting the user explore the loaded linguistic resource. Fields and tables for searching the LR and for viewing the results, according to the modalities decided by the core component, are made available. The menu boxes on the left of the Linguistic Browser are filled at run time with the methods for exploring LR specific Lexical and Conceptual relations.

The Ontology Panel, on the right, offers a perspective over ontological data in the classic Protégé style. By right-clicking on a frame (class, slot or instance), the typical editing menu appears, with some further options provided by Ontoling to:

1. search the LR by using the frame name as a key
2. change then name of the selected frame to a term selected from the Linguistic Browser



String Search: car

explore conceptual relation: Hyponyms

explore word relation: Polysemy

Sense ID: noun.4120131

Description: a small low car with a high-powered engine; usually seats two persons

Synonyms: sports_car
sport_car

Terminology Slot: rdfs:label

For Project: OntoLingOWLDemo

- owl:Thing
  - car
    - ambulance
    - taxi

Concept	Desc
noun.2853224	4-wheeled motor vehicle; usually propelled b need...
noun.2854760	a wheeled vehicle adapted to the rails of rail .
noun.2830136	a c nveyance for passengers or freight on a c the t...
noun.2855301	car pended from an airship and carrying p
noun.2855152	whe ssengers ride up and down; "the car was on the top floor"

Fig. 3 A screenshot of the Ontoling Plugin

3. add terms selected from the Linguistic Browser as additional labels for the selected frame
4. add glosses as a description for the selected frame
5. add IDs of senses selected from the linguistic browser as additional labels for the frames
6. create a new frame with a term selected from the Linguistic Browser as frame name (identifier)
7. only in class and slot browser: if the LR is a *TaxonomicalLR*, explore hyponyms (up to a chosen level) of the concept selected on the Linguistic Browser and reproduce the tree on the frame browser, starting from the selected frame, if available

These functionalities allow not only for linguistic enrichment of ontologies, but can be helpful for Ontologists and Knowledge Engineers in creating new ontologies or in improving/modifying existing ones.

Note how functionality 5 has not a rigid linguistic motivation, but is indeed dedicated to those willing to build an artificial controlled vocabulary which contains direct references to the semantic elements of a particular resource; this functionality is ideal for performing the semantic enrichment task described in section 2.

5 Automatic Linguistic Enrichment of Ontologies: mapping between ontologies and Linguistic Resources.

We're currently experimenting different techniques for performing automatic linguistic enrichment of ontologies. These techniques will be included in the next release of Ontoling, which will feature a completely new interface for letting Ontology Developers switch between prompted suggestions for enriching concepts and/or creating new ones, and a completely automatic process for feeding an ontology with linguistic content.

In line with our interface-based approach, which maximizes modularization and openness towards new techniques and resources, our aim is to deploy a whole range of specific techniques suitable for resources exposing different Linguistic Watermarks.

At present time, we have developed an algorithm for performing Semantic Enrichment of ontologies, which exploits LRs implementing the `ConceptualizedLR`, the `TaxonomicalLR` and (but not necessarily) `LRWithGlosses` interfaces of the LW. The Semantic Enrichment task, which we described in section 2, is very similar to an ontology mapping process, in that ontological concepts are mapped with elements from the semantic structure of a given LR (e.g. WordNet synsets).

Our technique for semantic enrichment foresees a first discovery phase in which semantic elements from the considered LR are selected as candidates for enriching ontological concepts. In a second phase, semantic and linguistic evidences are considered to verify candidates' suitability to enrich their target concepts.

During the discovery phase, 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 second phase. 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.

In the second phase, the algorithm exploits different techniques for producing the evidences which will then be used for choosing the best candidate for enriching concepts. These techniques combine analysis of both LR and ontology taxonomies for discovering possible alignments between concepts and candidate elements from the LR, together with statistical approaches which look at occurrences of labels, used to denote ontological concepts, into glosses of the LR. In particular, for each ontology concept *C* and for every candidate semantic element *S*, these gloss-based techniques explore:

- Presence of words into the gloss of *S*, identifying concepts related to *C*
- Presence of words into the gloss of *S*, identifying a concept which is present in the ontology
- Presence of uncommon words into both gloss of *S* and into glosses of other semantic elements which are candidates for concepts other than *C*.

The *related to* in the first technique means “is in the context of”. For each of these techniques, different contexts of *C* are in fact examined, considering objects bound to *C* by subclass/superclass relationships, linked to *C* by properties (or pointed by *C*, if *C* is itself a property), by restrictions over properties etc...

Notice that the second kind of gloss-based evidences is implied by the first one, and is in fact a less restrictive condition, which may however prove to be ideal in situations where presence of very few semantic relationships requires extension of context to the whole ontology, while the small size and/or specialization towards a given domain of the ontology, guarantees that every concept in it may be considered as somewhat related to each other.

The produced evidences are then projected in a feature space and evaluated according to the following formula:

$$p(C,S)=\frac{p_0+\left(1-\prod_{i=1}^{n}\left(1-\rho_i^+(C,S)\right)\right)\cdot\left(1-p_0\right)}{1+\left(1-\prod_{i=1}^{m}\left(1-\rho_i^-(C,S)\right)\right)\cdot\left(\frac{1}{p_0}-1\right)}$$

where *p(C,S)*, with $0 < p(C,S) < 1$, represents the plausibility that the candidate semantic element *S* be mapped to the ontological concept *C*. p_0 is an initial assignment for this plausibility which mainly depends upon the ambiguity (wrt the considered *LR*) of the label associated to *C*, while the various $\rho^+(C,S)$ and $\rho^-(C,S)$ denote respectively positive and negative evidences for plausibility of *(C,S)* pair.

6 Linguistic-based coordination for automatic discovery of ontology mappings

Aspects of Semantic Coordination of ontological resources can be roughly categorized according to two research areas:

1. Development of specific techniques for ontology mapping and alignment
2. Specification of basic communication modalities between semantic peers (Agents, Semantic Web Services etc…) to establish a meaning negotiation

In the first area, a number of linguistic-based methods and algorithms for performing automatic ontology mapping is emerging in literature. CTXMATCH [3,8] discovers mappings among Classification Hierarchies considering the semantic interpretation of their nodes. Similarities discovery is performed through linguistic processing of labels attached to hierarchies’ nodes, including tokenization, Part of Speech tagging, multiword recognition and word sense disambiguation. Exploitation of WordNet resource also helps in discovering synonymical expressions and in identifying ambiguous words. Recently, in [7] results from Google snippets were combined with other linguistic based techniques for learning/verifying subclass relations among concepts from different ontologies. In [14] a new metric for computing string-matching distances has been developed and evaluated specifically for ontological names: the assumption behind that work is that labels in ontologies bear common morphological properties upon which dedicated metrics can be tuned to obtain better results.

All of the above works concentrate on specific techniques for exploiting resources, while no role is given on how and when LRs need to be used, how they should be made available and which kind of servicing they would require: all of these specifications pertain to the second area of research on Semantic Coordination.

Our approach to ontology mapping methods exploits the environment for accessing LRs' content which is provided by the Linguistic Watermark interface, and is based on the same feature-based approach which we described for semantic enrichment of ontologies, with some exceptions:

- It is no more an asymmetric process, in which an ontology must be enriched with references to an LR. In this case mappings between two ontologies need to be established, where completeness of the process is evaluated against full coverage of both resources' content.
- 1-to-1 identity relationships are no more sufficient. Discovery of super-class/subclass relationships, as well as more complex mappings between complex ontological constructs (e.g, adopting a DL formalism, to state that `RedCar` from one ontology, is equivalent to the complex statement `Car ⊓ color ∋ red` from the other ontology), is thus expected. Again, analysis of context can help in finding multiple lexical anchors whose combination can give raise to complex mappings.

We do not dive further into the specific features adopted in our ontology mapping methods, and stress instead the role played by linguistic resources in the whole framework.

If one of the main aspects of the Semantic Web is to rely on semantic coordination performed automatically by distributed agents and Semantic Web Services (service discovery, content negotiation), it is important for them to recognize what are the needs, in terms of resources and skills, for performing this task under the best possible conditions.

Following our past works [12, 13] on "linguistic agents communication", we thus define a Linguistic Watermark even for ontologies. The role of this ontological LW is to provide, for every ontology, information about the (natural) language adopted for

```
<Linguistic Watermark> ::= { ontology enriched_by: <linguistic_resource> }

<linguistic_resource> ::=
              <linguisticResourceURI>,
              [ semantic_enrichment | linguistic_enrichment,<language> ]
              <enrichment_modality>
              <coverage>

<enrichment_modality> ::= [ supervised_enrichment | automatic_enrichment ]

<coverage> ::=  <conceptual_coverage>, <linguistic_coverage>
```

Fig. 4 Specifications for Ontological Linguistic Watermark

describing its content, as well as evaluation of the "linguistic expressiveness" with which this content is exposed. Its objective is to make agents able to immediately recognize the context of their communication and invoke proper services for coordinating with other Semantic Peers.

Explicitly reference to the linguistic resources which have been adopted to enrich ontological content should thus be completed with quantitative data about how these LRs contributed to enrich the ontology and on the modalities followed for performing the enrichment process. In Fig. 4, using BNF notation, specifications for the ontological Linguistic Watermark are reported: `<language>` is a tag denoting the natural language expressed by the LR; `<conceptual_coverage>`and `<linguistic_coverage>` are instead expanded in three data reports:

- Percentage of terms common to both ontology and linguistic resource wrt total number of terms (for the same language) in the ontology. This information is useful for knowing how much the specific resource participated to the enrichment of the ontology
- Average number of terms per concept, which belong to the linguistic resource
- Percentage of ontology concepts which are represented by at least a term from the linguistic resource.

With this data, agents can get precise information on the context of communication (that is, linguistic information about the knowledge of the agents they are communicating with) and follow the best strategy for negotiate the meaning of their ontological content. Different scenarios are possible, and mapping techniques must take into considerations cases where the involved ontologies present a strong enrichment (possibly with different natural languages) as well as cases where conceptual knowledge presents no linguistic enrichment at all. In general, when strong linguistic knowledge is associated to full coverage of ontological data on both sides of communication, agents may simply inspect each other's LWs and agree on which language(s) rely to negotiate the meaning of their concepts, much in the way humans do when in need of communicating with other people. Otherwise, assistance may be requested to other agents offering linguistic services (translation facilities, wider knowledge about a same language, specific ontology mediation skills etc…).

7 Conclusions

In this work we proposed solutions to the problem of ontology mapping, and of meaning negotiation in general, which require a new perspective on the way knowledge representation is handled nowadays. The attention paid to formal concept representation in the Semantic Web is in fact not being matched by an equivalent interest on how this information will be made easily accessible by humans, and by machines not sharing a shared semantic commitment.

Adherence to nowadays standards for ontology representation has been itself a limit for our research, where a more structured and close bridging between conceptual and linguistic knowledge would be expected. The link we establish in this work between conceptual knowledge and its associated linguistic representation is characterized by simple references between concepts and labels (being them terms or semantic descriptors), while more sophisticated relationships are required to address the complex constructs which characterize a significant fraction of every ontology.

Nonetheless, our approach aims to define a very scalable and heterogeneous framework where no strong assumptions are required a-priori on the way ontological data need to be exposed. This facilitates adherence of the framework to nowadays realistic scenarios, while leaves open new possibilities for more efficient solutions, as those proposed in our full 3-step methodology.

The Linguistic Watermark for LRs defines a set of functionalities for accessing content of existing Linguistic Resources in a uniform manner (thus favoring pluggability of new resources to the same framework, without need of specific ad-hoc solutions), and at the same time represents a model for characterizing different aspects of a LR. Its prototype may thus be exploited by tools and applications, like the presented Ontoling, willing to support ontology enrichment in a scenario characterized by a plethora of heterogeneous resources and languages. The ontological LW follows the same paradigm; it describes the degree of linguistic expressiveness of ontologies and thus facilitates semantic coordination of agents and web services by providing information on the context of their communication.

One of the major benefits of this approach is not only represented by the proposed techniques and methodologies for supporting knowledge interoperability, but the fact that it also guarantees more intellectual and cultural independence to the ontology development process, trading strict and limitative forms of semantic commitment with the adoption of a universally agreed form of communication: natural language.

References

1. 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
2. Benjamins, V. R., Contreras, J., Corcho, O., and Gómez-Pérez, A.: Six Challenges for the Se-mantic Web. SIGSEMIS Bulletin, April 2004
3. Bouquet, P., Serafini, L. and Zanobini, S.: Semantic Coordination: A new approach and a application. In Proc. of ISWC-03, Sanibel Island, USA, October, 2003
4. M. Dean, D. Connolly, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language 1.0 Reference, W3C Working Draft 29 July 2002, http://www.w3.org/TR/owl-ref/.
5. Fellbaum, C: WordNet - An electronic lexical da-tabase. MIT Press, (1998).
6. Gennari, J., Musen, M., Fergerson, R., Grosso, W. Crubézy, M., Eriksson, H., Noy, N. and Tu, S: The evolution of Protégé-2000: An environment for knowledge-based systems development". International Journal of Human-Computer Studies, 58(1):89–123, 2003.
7. van Hage, W. R., Katrenko, S. and Schreiber. A. Th.: A Method to Combine Linguistic Ontology-Mapping Techniques, 4th International Semantic Web Conference (ISWC-2005) Galway, Ireland, November, 2005
8. Magnini, B., Speranza, M., Girardi, M.: A Semantic-based Approach to Interoperability of Classification Hierarchies: Evaluation of Linguistic Techniques. In: Proceedings of COLING-2004, Geneva, Switzerland, August 23 - 27, 2004.
9. Miles, A., Brickley, D.: SKOS Core Guide. W3C Working Draft, 2005

10.Pazienza, M. T., Stellato, A.: The Protégé Ontoling Plugin - Linguistic Enrichment of Ontologies in the Semantic Web, in Poster and Demo Proceedings of the 4th International Semantic Web Conference (ISWC-2005) Galway, Ireland, November, 2005

11.Pazienza, M.T., Stellato, A., Vindigni, M., Zanzotto. F.M.: XeOML: An XML-based extensible Ontology Mapping Language. Workshop on Meaning Coordination and Negotiation, in 3rd International Semantic Web Conference (ISWC-2004) Hiroshima, Japan, November 8, 2004

12.Pazienza, M.T., Vindigni, M.: Language-based agent communication. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), Sigueza, Spain (2002)

13.Pazienza, M.T., Vindigni, M: Agent based Ontological Mediation in IE Systems. Information Extraction in the Web Era: Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents, papers from SCIE 2002, Frascati (Rome, Italy), July 2002 Springer 2003

14.Stoilos, G., Stamou, G., Kollias, S.: A String Metric for Ontology Alignment

15.www.dict.org/bin/Dict