Ontology Learning and Semantic Annotation: a Necessary Symbiosis

Emiliano Giovannetti, Simone Marchi, Simonetta Montemagni, Roberto Bartolini
Istituto di Linguistica Computazionale - CNR
via G. Moruzzi 1, Pisa
E-mail: emiliano.giovannetti@ilc.cnr.it, simone.marchi@ilc.cnr.it, simonetta.montemagni@ilc.cnr.it,
roberto.bartolini@ilc.cnr.it

Abstract

Semantic annotation of text requires the dynamic merging of linguistically structured information and a “world model”, usually represented as a domain-specific ontology. On the other hand, the process of engineering a domain ontology through semi-automatic ontology learning system requires the availability of a considerable amount of semantically annotated documents. Facing this bootstrapping paradox requires an incremental process of annotation–acquisition–annotation, whereby domain–specific knowledge is acquired from linguistically–annotated texts and then projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted. The presented methodology is a first step in the direction of a full “virtuous” circle where the semantic annotation platform and the evolving ontology interact in symbiosis. As a case study we have chosen the semantic annotation of product catalogues. We propose a hybrid approach, combining pattern matching techniques to exploit the regular structure of product descriptions in catalogues, and Natural Language Processing techniques which are resorted to analyze natural language descriptions. The semantic annotation involves the access to the ontology, semi-automatically bootstrapped with an ontology learning tool from annotated collections of catalogues.

1 Introduction

Quick, effective and customizable acquisition, organization, processing, use and sharing of the implicit knowledge embedded in existing huge electronic document repositories (web page archives, repositories of company files and scientific literature, public administration records, law document bases etc.) represent major competitive factors in the emerging knowledge economy and core technological challenges of the modern information society. Over the last fifteen years, such demands have provided growing impulse to the development of a wide range of so-called “Semantic Annotation Platform” (SAPs), aimed at tracking down and explicitly representing unstructured text information. In more recent years, considerable effort has been put into cutting down development costs and making SAPs more portable to a variety of different knowledge domains and text genres. Advanced SAPs of this kind must include:

- incremental and robust NLP software for automated multi-level linguistic text annotation;
- supervised stochastic classifiers trained on pre-annotated text materials;
- ontological standards for formal representation of domain-specific knowledge;
- unsupervised or minimally supervised knowledge bootstrapping techniques, dynamically integrating the stages of manual building and population of ontological models, which are inevitably time-consuming and prone to errors.

Semantic annotation requires the dynamic merging of linguistically structured information (made accessible through intermediate stages of increasingly abstract parsing) and a “world model”, represented as a domain-specific ontology. The purpose of assigning linguistic structure to a natural language sentence is to single out text-to-ontology “anchors”, that is word sequences and constructions, such as proper names, simple and complex terms, event designators etc., that play the role of linguistic pointers to ontological concepts and properties. However, a free natural language sentence is likely to be dramatically underspecified in respect to the ontology content: some references to entities or relations can be left implicit and spotting their semantic “counterparts” (concepts and properties) in the ontology can be very difficult. These forms of presupposition, typical of domain specific, natural language sentences, call for massive recourse to background knowledge and inference, under suitable linguistic constraints.

On the other hand, it is well known that the process of engineering an ontology is costly (Simperl et al., 2006). In order to alleviate the costs involved in the activity of engineering ontologies, several proposals for automatically learning ontologies from semantically annotated textual resources have emerged (Buitelaar et al, 2005).

More in general, technologies in the area of knowledge management and information access are confronted with a typical acquisition paradox. As knowledge is mostly conveyed through text, content access requires understanding the linguistic structures representing content in text at a level of considerable detail. In turn, processing linguistic structures at the depth needed for content understanding presupposes that a considerable amount of domain knowledge is already in place. Facing this bootstrapping paradox requires an incremental process of annotation–acquisition–annotation, whereby domain–specific knowledge is acquired from linguistically–annotated texts and then projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted.

Concerning semantic annotation, the vicious circle (between the need of having the domain represented in the ontology for the semantic annotation and the construction of the ontology based on the results obtained from the annotation) can be turned to a virtuous circle if the necessary conditions are set to let the evolving ontology and the semantic annotation platform interact in symbiosis.

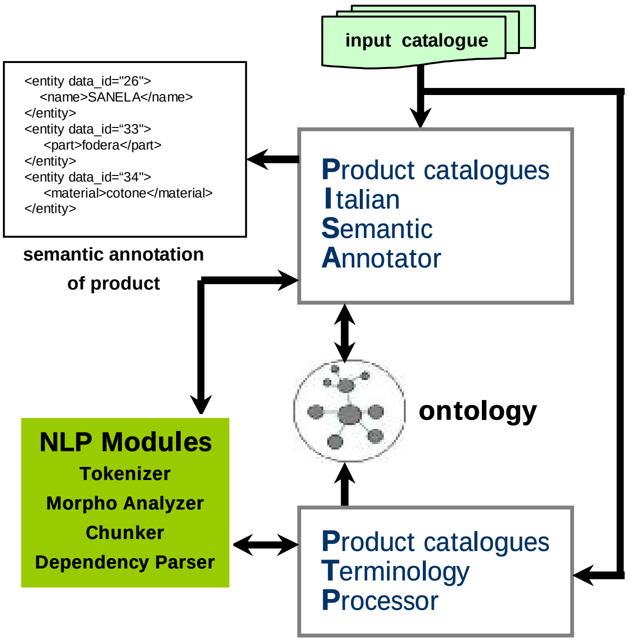
The importance of this mutual interaction has also emerged in the field of Information Extraction (IE). In

(Nédellec et al, 2005) the symbiosis between IE and ontologies has been investigated:

- Ontology is used for Information Extraction: IE needs ontologies as part of the understanding process for extracting the relevant information;
- Information Extraction is used for populating and enhancing the ontology: texts are useful sources of knowledge to design and enrich ontologies.

It is argued that these two tasks can be combined in a cyclic process: ontologies can be used for interpreting the text at the right level for IE to be efficient and IE can extract new knowledge from the text, to be integrated in the ontology.

In this paper we report the results of a case study for this paradigm which has been carried out on a peculiar text type, namely furniture product catalogues. Automatic extraction of knowledge from product catalogues appears to be a complex task. Catalogues do not contain continuous and linguistically sound text (i.e. typical sentences are constituted by nominal descriptions): this fact often discourages the recourse to traditional Natural Language Processing (NLP) techniques (Šváb et al., 2004). On the other hand, product descriptions appear as semi-structured texts where product names, prices, and other features appear in a regular order: unfortunately, this is generally not the case. Semantic annotation of product catalogues appears therefore as a challenging task requiring ad hoc solutions, i.e. the combination of different types of evidence and techniques. This fact notwithstanding, we believe that the results of this case study can be profitably extended to other text types.



```
<entity data_id="26">
   <name>SANELA</name>
</entity>
<entity data_id=“33">
    <part>fodera</part>
</entity>
<entity data_id=“34">
    <material>cotone</material>
</entity>
```

Figure 1. From product catalogues to semantically annotated texts: interaction of semantic annotation and ontology learning in the implemented prototype.

In this case study, a prototype has been designed and implemented for the bootstrapping of ontological information from document collections and for the semantic annotation of texts on the basis of the bootstrapped ontology (Fig. 1).

2 The Methodology

Semantic annotation of catalogue texts is carried out in two steps: in the first step, the catalogue collection is annotated, at a syntactic level, with no ontology support. The resulting linguistic annotation is then used as the basis for the ontology learning process, this latter producing the application ontology (Studer et al., 1998) of reference. For this objective, a component for the semi-automatic construction of ontologies was developed starting from an existing ontology learning tool: T2K (Text-to-Knowledge), a hybrid system combining linguistic technologies and statistical techniques jointly developed by CNR-ILC and Pisa University (Bartolini et al., 2005).

In the second step, the catalogue collection is annotated at the semantic level by also exploiting the semi-automatically bootstrapped ontology. To deal with the peculiarities of product catalogues, a hybrid approach is proposed, combining pattern matching techniques to exploit the regular structure of product descriptions in catalogues, and Natural Language Processing techniques which are resorted to analyze natural language descriptions. In particular, pattern matching techniques are used for isolating individual product descriptions within the textual flow and for identifying their basic building blocks (e.g. the product name, its price, as well as its natural language description). For each identified product, the natural language description is then processed by a battery of NLP tools for the analysis of Italian texts called AnIta (Bartolini et al., 2002) on top of which a semantic annotation component operates in charge of identifying relevant entities (e.g. colours, materials, parts of a given product) and the relations holding between them (which can be referred either to the product itself or to individual parts). The process of semantic annotation of product descriptions is driven by the application ontology bootstrapped from texts in the first step: in particular, ontological information is used for the recognition of semantically relevant terms occurring in the free text part of the product descriptions, and for the semantic interpretation of syntactic ambiguities emerged during the linguistic analysis process.

3 The System

The implemented prototype system (Fig. 1) includes two main components, the Product catalogues Terminology Processor (henceforth, PTP) and the Product catalogue Italian Semantic Annotator (henceforth, PISA), both exploiting the battery of NLP modules for the analysis of general Italian texts.

3.1 The ontology learning component

PTP (Fig. 2) was developed for bootstrapping terminological and ontological knowledge from the first-step annotation of the catalogue collection. PTP carries out the ontology learning task in two different steps:

1) extraction of domain terminology, both single and multi-word terms, from the annotated catalogues;
2) organization and structuring of the set of acquired terms into
 a) fragments of taxonomical chains, and
 b) clusters of semantically related terms.

Domain terms need to be recognized whatever their linguistic form in the documents is: term extraction thus requires some level of linguistic pre-processing of texts. In this case, term extraction is carried out starting from the syntactic annotation of the texts. Candidate terms may be one word terms or multi-word terms.

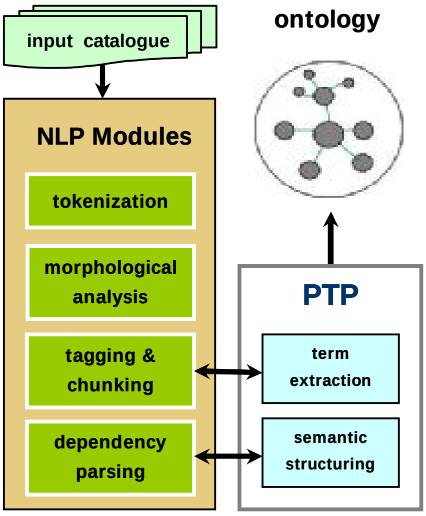


Figure 2: The general architecture of PTP.

The acquisition strategy differs in the two cases. Potential single terms are extracted from the syntactically chunked text, in particular from the nominal heads of different chunk types (typically, nominal and prepositional chunks). Candidate terms are purely identified on a frequency basis (after excluding stop-words). The acquisition of multi-word terms follows a two stage strategy: first, the chunked text is analysed on the basis of a mini-grammar for the extraction of potential complex terms; second, the list of acquired potential complex terms is ranked according to the log-likelihood ratio association measure (Dunning, 1993), which assesses the strength of the association between the words heading the chunks covering the candidate complex term.

In the second step, proto-conceptual structures involving the terms in the TermBank are identified. Since this represents a more complex task, the starting point is no longer the chunked text, but rather a dependency-annotated text enriched with the multi-word terminology acquired at the term extraction stage.

Furthermore, the identified terms are organized into fragments of taxonomical chains, which are reconstructed starting from the internal linguistic structure of terms. For instance, *gambe in acciaio* (steel legs) and *gambe regolabili* (adjustable legs) can be seen as hyponyms of a general single term *gambe* (legs). PTP also performs the identification of clusters of semantically related terms (henceforth, RTs) which is carried out on the basis of distributionally-based similarity measures as illustrated in (Allegrini et al., 2003). For each term, a set of semantically related terms is identified: given that typical sentences in catalogues are constituted by nominal descriptions, the clustering of semantically similar terms was grounded on complement relations governed by nominal heads. In this case, semantic relatedness of words (typically nouns) is inferred by their occurring in identical nominal contexts. For instance, words like *betulla* (birch) and *acciaio* (steel) in the domain of furniture catalogues appear to be semantically related words due to their occurrence in similar contexts like *struttura in betulla* (frame in birch), *struttura in acciaio* (frame in steel), *gambe in betulla* (birch legs), *gambe in acciaio* (steel legs), etc. For instance, to the term *plastica* (plastic) the following set of related terms has been associated: *piuma* (feather), *lattice* (latice), *cotone* (cotton), etc., thus identifying a set of different kinds of materials.

MATERIALS		
Manually built class starting from RTs	Union of RTs with associated material terms	Material class built starting from seed terms
Acciaio (*Steel*)	Acciaio (*Steel*)	Acciaio (*Steel*)
Alluminio (*Aluminium*)	Alluminio (*Aluminium*)	Alluminio (*Aluminium*)
Betulla (*Birch*)	Betulla (*Birch*)	Betulla (*Birch*)
Cotone (*Cotton*)	Cotone (*Cotton*)	Faggio (*Beech*)
Faggio (*Beech*)	Faggio (*Beech*)	Lamina (*Leaf*)
Frassino (*Ash*)	Frassino (*Ash*)	Legno (*Wood*)
Lamina (*Leaf*)	Gambe (*Legs*)	Melammina (*Melamine*)
Lattice (*Latice*)	Lamina (*Leaf*)	Pino (*Pine*)
Legno (*Wood*)	Lattice (*Latice*)	Piuma (*Feather*)
Melammina (*Melamine*)	Melammina (*Melamine*)	Plastica (*Plastic*)
Ovatta (*Wadding*)	Pino (*Pine*)	Poliestere (*Polyester*)
Pino (*Pine*)	Piuma (*Feather*)	Rovere (*Durmast*)
Piuma (*Feather*)	Plastica (*Plastic*)	Schiuma (*Foam*)
Plastica (*Plastic*)	Poliestere (*Polyester*)	Vendita (*Selling*)
Poliestere (*Polyester*)	Poliuretano (*Polyurethane*)	Vetro (*Glass*)
Poliuretano (*Polyurethane*)	Rovere (*Durmast*)	
Rovere (*Durmast*)	Schiuma (*Foam*)	
Schiuma (*Foam*)	Vendita (*Selling*)	
Vetro (*Glass*)	Vetro (*Glass*)	

Table 1. Ontological classes of materials built on the basis of PTP results

Acquired fragments of taxonomical chains of terms together with the clusters of semantically related terms (RTs) were used to bootstrap the ontological classes (e.g. colours, materials, parts, etc.) to be exploited in the construction of the final application ontology. In particular, from the sets of RTs associated with the terms denoting materials we manually built the ontological class of materials; the same was done for colours and parts. In the first column of Table 1 we reported the class of terms denoting materials manually built starting from the list RTs and updated with other material terms occurring in the TermBank but not occurring as part of any of the RT set of material terms (this is the case of *ovatta* 'wadding').

We also evaluated how and to what extent this process could be automatically carried out: Table 1 documents the results of different experiments carried out in this direction. In particular, the second column reports the class of materials as resulting from the union of all RT sets associated with all acquired material terms where it can be noticed that only two are the spurious terms included in the list; the third column documents the results of yet another experiment trying to build the class of materials from the union of RT sets associated with a few (namely 5) selected prototypical material terms, so-called "seed terms". It is interesting to note that in both experiments documented in columns 2 and 3 the list of material terms is still rich; we thus believe that it is worth working in the direction to semi-automatically infer ontological classes from the RT sets identified by PTP.

Term	**Complex hyponyms**
acciaio (*steel*)	acciaio cromato (*chrome-plated steel*), acciaio galvanizzato (*galvanized steel*), acciaio inox (*stainless steel*), acciaio rivestito (*powder-coated steel*)
betulla (*birch*)	betulla massiccia (*solid birch*)
cotone (*cotton*)	cotone egiziano (*egyptian cotton*), cotone con imbottitura (*padded cotton*)
faggio (*beech*)	faggio massiccio (*solid beech*), faggio verniciato (*beech veener*)
frassino (*ash*)	frassino trattato (*treated ash*)
lamina (*leaf*)	lamina bianca (*white leaf*), lamina blu (*blue leaf*)
legno (wood)	legno massiccio (*solid wood*)
melammina (*melamine*)	melammina bianca (*white melamine*)
ovatta (*wadding*)	ovatta di poliestere (*polyester wadding*)
pino (*pine*)	pino massiccio (*solid pine*)
plastica (*plastic*)	plastica bianca (*white plastic*)
poliestere (*polyester*)	poliestere verde (*green polyester*)
poliuretano (*polyurethane*)	poliuretano espanso (*expanded polyurethane*)
rovere (*durmast*)	rovere massiccio (*solid durmast*)
schiuma (*foam*)	schiuma di poliuretano (*polyurethane foam*)
vetro (*glass*)	vetro normale (*normal glass*), vetro temprato (*tempered glass*), vetro trasparente (*clear glass*)

Table 2. Expanded ontological class of materials

As it can be noticed in Table 1 the ontological classes inferred starting from RT sets include simple terms only: this directly follows from the strategy adopted for the acquisition of related terms, based on complement relations governed by a nominal head. This could represent a problem for the semantic annotation process since the list of RTs acquired by PTP should also include materials denoted by complex terms such as *vetro temprato* (tempered glass) or *betulla massiccia* (solid birch). The ontological classes based on RT sets were thus automatically expanded to include complex terms by combining taxonomical and horizontal relations as reconstructed by PTP. Table 2 reports the results of this expansion process for the ontological class of materials which now includes 42 material terms against the 19 of the original core class reported in Table 1.

A fragment of the ontology built starting from the results of PTP is illustrated in Figure 3. The concepts of *steel* and *wood* (automatically grouped by PTP), for example, have been manually set as sub-concepts of *material*, as represented in the figure by solid arrows standing for *is_a* relations. The same has been done for "parts" and "colours", two other "top-level concepts". Dotted arrows, on the other hand, represent hierarchical relations between concepts automatically detected by PTP as previously described. Relations between top-level concepts (i.e. *material_of_part* and *colour_of_part*) have been added manually.

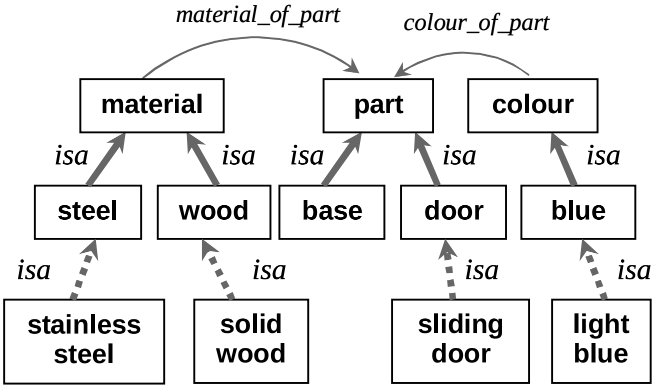


Figure 3. A fragment of the bootstrapped ontology

3.2 The semantic annotation component

The semantic annotation component (PISA) has a two-module architecture composed by: the Regular Expression Manager, performing pattern matching on the catalogue text to isolate individual product descriptions and to identify their basic building blocks, and the NLP Manager, in charge of the ontology-driven linguistic analysis of the free text descriptions.

From a procedural point of view, individual product descriptions are firstly extracted through pattern matching starting from a set of regular expressions. Once an individual description is identified, some of its parts can already be semantically classified and annotated: this is the case of entities like *name*, *type*, *price*, *dimensions* and *product id*, corresponding to subparts of the matching regular expression.

Consider as an example the catalogue fragment in Fig. 5 (upper box) which is relative to a product called "HOVET". Through pattern matching it is possible to extract its name, the type ("mirror"), the price, its dimensions and the product identifier, as well as the relations between this information and the product itself (i.e. name_of, price_of, etc).

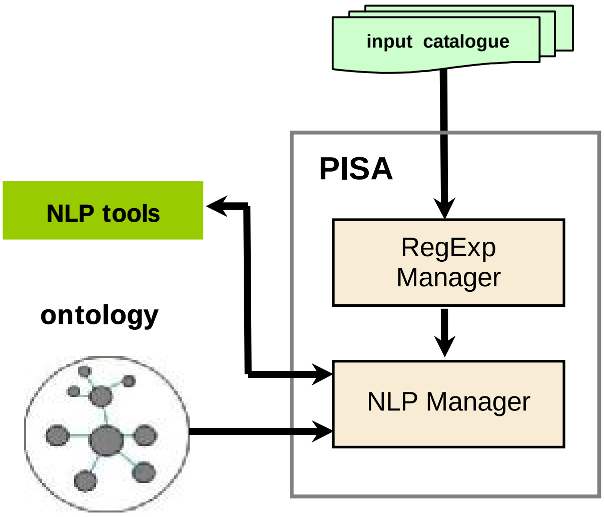


Figure 4: The general architecture of PISA.

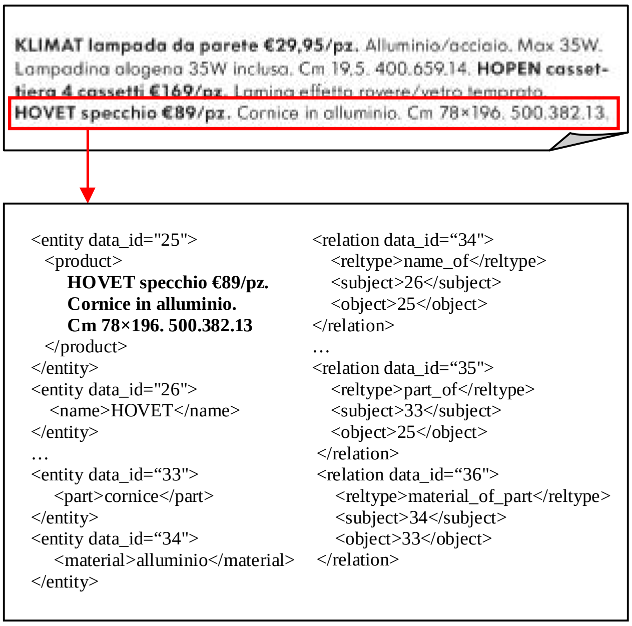
The linguistic analysis of product descriptions is carried out by AnIta, a battery of NLP tools consisting in an "assembly line" whose main components include: tokenization of the input text, morphological analysis (including lemmatisation) and syntactic parsing, the latter articulated in two different stages, i.e. chunking (which also includes morpho-syntactic disambiguation) and dependency analysis. Semantic annotation of product catalogues is ontology-driven and operates starting from the output of dependency analysis. The ontology accessed by PISA, built with the help of the PTP (see section 3.1), is used to:

- detect and (semantically) annotate relevant entities inside the free text of the product descriptions;
- detect and annotate relations between annotated entities;
- resolve possible ambiguities found during the process of semantic annotation.

Back to the example introduced above, once the natural language description has been isolated and identified by the RegExp component it can be passed to the NLP Manager in charge of acquiring, with the support of the domain ontology, further information about the product: in the example, the system detected a material ("aluminium") and a part ("frame"), as well as a relation holding between them ("material_of_part"). The final formalized product description is reported in the lower box of Fig. 5 where the different features of the product are listed, including the fact that the frame of the mirror is made of aluminium.

In the example, the NLP Manager exploits the ontology constructed in the first step by the PTP component in this way:

- it finds *cornice* (frame) as a (not necessarily direct) subclass of "Product Part";
- it finds *alluminio* (aluminium) as a (not necessarily direct) subclass of "Material";
- if finds out there is a relation holding between the classes "Part" and "Product" called "part_of";
- if finds out there is a relation between the classes "Material" and "Product Part" called "material_of_part".



KLIMAT lampada da parete €29,95/pz. Alluminio/acciaio. Max 35W. Lampadina alogena 35W inclusa. Cm 19.5. 400.659.14. **HOPEN cassettiera 4 cassetti €169/pz.** Lamina effetto rovere/vetro temprato. **HOVET specchio €89/pz.** Cornice in alluminio. Cm 78×196. 500.382.13.

```
<entity data_id="25">
  <product>
     HOVET specchio €89/pz.
     Cornice in alluminio.
     Cm 78×196. 500.382.13
  </product>
</entity>
<entity data_id="26">
   <name>HOVET</name>
</entity>
…
<entity data_id=“33">
    <part>cornice</part>
</entity>
<entity data_id=“34">
    <material>alluminio</material>
</entity>

<relation data_id=“34">
   <reltype>name_of</reltype>
   <subject>26</subject>
   <object>25</object>
</relation>
…
<relation data_id=“35">
   <reltype>part_of</reltype>
   <subject>33</subject>
   <object>25</object>
 </relation>
 <relation data_id=“36">
    <reltype>material_of_part</reltype>
    <subject>34</subject>
    <object>33</object>
 </relation>
```

Figure 5: An example of annotation.

On this basis, the system semantically annotates terms "cornice" and "alluminio" (respectively as a kind of "Product Part" and "Material") and the relations "part_of" and "material_of_part", the former holding between "cornice" and the whole product, the latter holding between "alluminio" e "cornice".

Furthermore, ontology can also be used to improve semantic annotation by resolving syntactic ambiguities found in product descriptions. Let's consider, for instance, the following product description:

- "Sedia in plastica con schienale regolabile" (*Plastic chair with adjustable back*)

In this case, the syntactic ambiguity occurs for what concerns the attachment of the prepositional phrase "con schienale regolabile" (*with adjustable back*), which can be governed either by the nominal head "sedia" (*chair*) or by "plastica" (*plastic*).

In the case at hand, the ontology can be usefully exploited in this way:

- "sedia" is a *Product*;
- "plastica" is a *Material*;
- "schienale" is a *Part*.

Since there is no property (directly) linking a Part to a Material, but there is one linking a Part to a Product (i.e. *part_of*), the correct interpretation is that "schienale regolabile" is a part of "sedia".

4 Evaluation of Acquired Results

A preliminary evaluation of both components, PTP and PISA, was carried out, distinguishing between annotations obtained through pattern matching and annotations performed thanks to the ontology-driven linguistic analysis.

A task-based evaluation was undertaken concerning the results of the ontology learning process performed by the PTP component, in terms of correctness of its role in supporting the semantic annotation process. Evaluation of PISA component was carried out with respect to the results obtained in the analysis of the corpus of reference, the italian furniture catalogue, IKEA 2006, where 793 product descriptions have been identified and processed.

First, we have created a "gold-standard" corpus of reference by randomly extracting and manually annotating 10% of the identified IKEA products.

Evaluation was concerned with name, type, dimensions, price, and id extracted by pattern matching and product material, product colour, product part, product part material, and product part colour extracted by the ontology driven linguistic analysis.

We have calculated precision (PRE), which measures the system output's accuracy, as:

$$PRE = \frac{COR + 0.5 \cdot PAR}{ACT}$$

The parameters introduced in the formula refer to:

- COR(rect): the number of annotations that are found to be correct after comparison with the gold-standard annotations for the same text span;
- INC(orrect): the number of annotations that are found to be incorrect;
- PAR(tially correct): the number of annotations that are partially correct after comparison with the gold-standard annotations (e.g. partial credit is given to the detection of "Birch" in relation to "Solid birch");
- ACT(ual): the total number of annotations, calculated as COR + INC + PAR.

To calculate recall, two additional parameters were considered:

- MIS(sing): the number of gold-standard annotations in the key that are not present in the system output;
- POS(sible): the total number of annotations in the gold-standard, computed as the sum of COR, PAR and MIS.

Recall was computed as follows:

$$REC = \frac{COR + 0.5 \cdot PAR}{POS}$$

The system has scored a precision of 0.99 for annotations obtained through pattern matching and 0.89 for those obtained through ontology driven linguistic analysis. Regarding recall, 0.94 for pattern matching and 0.70 for linguistic analysis.

To investigate the portability of the proposed methodology we have semantically annotated another italian furniture catalogues, Zanotta. As for IKEA, we have first randomly extracted and manually annotated a subset (20%) of the 135 analyzed product descriptions as a "gold-standard" corpus of reference. To correctly process the Zanotta product descriptions we just had to modify the set of regular expressions processed by the Regular Expression Manager component of the prototype.

Concerning evaluation of the results, both precision and recall are equal to 1 concerning the pattern matching analysis, and, respectively, to 0.86 and 0.50 for ontology driven NLP analysis.

The very high score obtained regarding the pattern matching step can be abscribed to the very regular structure of the product descriptions. On the contrary, the ontology driven natural language analysis result is not as good as the one obtained for the IKEA catalogue. As a matter of fact, whenever a term is found inside a product description but it is not related to any concept inside the domain ontology of reference, it cannot be correctly annotated. In other words, the worse results obtained concerning the second step ontology driven annotation of the Zanotta catalogue are mainly to be abscribed to the lack of ontology coverage.

5 Future perspectives

Concerning future works, in the short-term we will try to improve the system's performance by working on two different fronts:

- ontology coverage: the main cause for the relatively low recall is due to missing concepts in the application ontology and the consequent failure in detecting and annotating the relative entities and relations inside the free text description;
- ontology-driven linguistic analysis: another problem source turned out to be the adopted strategy for relation detection and annotation, which currently fails when facing unusual syntactic constructions.

In conclusion, the presented methodology can be considered as a first step in the direction of the full "virtuous" circle described in the introduction. The two-step interaction between the semantic annotation and the ontology learning modules provides encouraging results: future work will be devoted to "triggering the circle" by exploiting the second-step annotation results to adjust and enrich the domain ontology of reference.

6 References

Allegrini, P., Montemagni, S., Pirrelli, V. (2003). Example-Based Automatic Induction Of Semantic Classes Through Entropic Scores. *Linguistica Computazionale* XVI-XVII, pp.1--45.

Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V. (2002). Grammar and Lexicon in the Robust Parsing of Italian: Towards a Non-Naïve Interplay. In *Proceedings of Coling 2002-Workshop on Grammar Engineering and Evaluation*, Taipei, pp 1--7.

Bartolini, R., Giorgetti, D., Lenci, A., Montemagni, S., Pirrelli, V. (2005). Automatic Incremental Term Acquisition from Domain Corpora. In *Proceedings of the 7th International conference on Terminology and Knowledge Engineering (TKE2005)*, Copenhagen, Denmark. pp. 125--138.

Bartolini, R., Caracciolo, C., Giovannetti, E., Lenci, A., Marchi, S., Pirrelli, V., Renso, C., Spinsanti, L. (2006). Creation and Use of Lexicons and Ontologies for Natural Language Interfaces to Databases. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*. pp. 137--143.

Buitelaar, P., Cimiano, P., Magnini, B. (Eds.) (2005). *Ontology Learning From Text: Methods, Evalutation and Applications*. IOS Press.

Dunning, T. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), pp. 61--74.

Nédellec, C., Nazarenko, A. (2005). Ontology and Information Extraction: a Necessary Symbiosis. In P. Buitelaar et al. (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, pp. 155--170.

Simperl, E.P.B., Tempich, C., Sure. Y. (2006). OntoCom: A Cost Estimation Model for Ontology Engineering. In *Proceedings of the 5th International Semantic Web Conference (ISWC)*. pp. 625--639.

Studer, R., Benjamins, V.R., Fensel, D. (1998). Knowledge engineering: Principles, and methods. *IEEE Transactions on Data and Knowledge Engineering*, 25 (1-2), pp. 161--197.

Šváb, O., Labskỳ, M., Svátek, V. (2004). RDF-Based Retrieval of Information Extracted from Web. In *SIGIR'04 Semantic Web Workshop*, Sheffield.