Making Explicit the Semantics Hidden in Schema Models

Bernardo Magnini, Luciano Serafini, and Manuela Speranza

ITC-irst Istituto per la Ricerca Scientifica e Tecnologica,
Via Sommarive 18 - Povo,
38050 Trento, Italy
{magnini, serafini, manspera}@itc.it

Abstract. Most of the data stored in the Semantic Web is organized in schema models, which can be represented as labeled graphs where labels are short natural language expressions. Examples of schema models include ER-schema automata, ontologies, taxonomies, and Web Directories. The semantics of schema models is not explicit but is hidden in their structures and labels. To obtain semantic interoperability we need to make their semantics explicit by taking into account both the interpretation of the labels and the structures described by the arcs. We propose a methodology for interpreting schema models on the basis of the taxonomic relations and the linguistic material they contain. We rely on a set of linguistic repositories, such as WordNet, and explore a number of crucial linguistic issues such as disambiguation of polysemous words, multiwords, and coordinations. The Web Directories of Google and Yahoo! have been chosen as an evaluation set. We show that there is a considerable amount of information to be made explicit and discuss the performance of an implementation of our analysis.

1 Introduction

Hierarchical classifications are taxonomic structures used to organize large amounts of documents. The most typical examples of hierarchical classifications are file systems, marketplace catalogs, and the directories of Web portals. Documents of a hierarchical classification can be of many different types, depending on the characteristics and uses of the hierarchy itself. In file systems, documents can be any kind of file (e.g. text files, images, applications, etc); in the directories of Web portals, documents are pointers to Web pages; in the marketplace, catalogs organize either product cards or service titles.

Hierarchical classifications are quite useful for document classification and retrieval. Users browse hierarchies of concepts and quickly access the documents associated with the different concepts. The content of a concept is typically described by a label, but it also depends on the concepts at higher levels in the hierarchy, even though the relations between concepts are usually not explictly labeled.

Hierarchical classifications are now widespread as knowledge repositories and the problem of their integration and interoperability is acquiring a high relevance from a scientific and commercial perspective. A typical application of hierarchical classification interoperability occurs when a set of companies want to exchange

products without sharing a common product catalog. In these cases the best solution is to find mappings between their catalogs [17]. In [5], we have proposed an algorithm that finds semantic relations between the nodes of different hierarchical classifications. This algorithm strongly relies on a linguistic analysis of the labels contained in the classifications. The main difference between this algorithm and other approaches to schema matching such as [2], [7] and [4] is that in order to interpret a node of a hierarchy we do not limit ourselves to a linguistic analysis of its labels. Instead, we extend this analysis by considering the *implicit information* deriving from the *context* where the node occurs, i.e., the structural relations with the other nodes of the hierarchy.

Indeed, one of the most evident peculiarities of hierarchical classifications (and in general of schemas) is that the meaning of a node depends not only on the label of the node, but also on the position of the node in the hierarchy. Indeed, like databases and ontologies, concept hierarchies are built on taxonomic relations between concepts, but such relations are implicit and have to be interpreted. Like plain texts, hierarchical classifications contain linguistic material, i.e. labels that can be analyzed with NLP techniques; the context provided for a label, however, is not a sentence or a paragraph, but a set of concepts placed at different levels. Consequently, the interpretation is performed in two steps: first, each individual concept is analyzed separately from the others and is associated with a basic logical form. Then, on the basis of its position in the hierarchy and of its relations with other nodes, we build a full logical form for each concept.

For instance we can have a hierarchical classification of documents about sports that contains a node labeled with *Sports Organizations*, and a second hierarchical classification on the same topic, containing a node labeled with *Organizations* which is a child of a node labeled with *Sport*. Clearly the semantics of the node "Sports Organizations" in the first hierarchy coincides with the semantics of the node "Organizations" in the second hierarchy . This equality however cannot be discovered by simply analyzing the labels. One has to discover that the arc connecting "Sports" and "Organizations" is a specification arc. This interpretation is very ambiguous and context dependent. Consider the example of a node labeled with *Schools* with a descendant node labeled *United States*, in this case the hierarchical relation between the two nodes has to be interpreted as a location relation. This interpretation is based on the semantics that is hidden in the labels and in the hierarchical structure.

The aim of this paper is to describe a method to analyze the implicit knowledge hidden in hierarchical classifications and to make it explicit in order to provide a correct interpretation of its concepts. In particular, we describe an algorithm that, given a concept hierarchy, returns an interpretation of each node of a hierarchy in terms of a logical formula of a description logic [3]. The ideal output of our algorithm for the concept hierarchy of Figure 1 is reported in Figure 2. This algorithm does not consider the documents classified under the nodes, so that it can be used in situations where such information is partially available or not available at all.

The paper is structured as follows. In Section 2 we provide a formal definition of concept hierarchy based on intuitions on how documents are classified by humans. In Section 3 we describe the analysis of the concepts performed without considering the hierarchical structure of the context. In Section 4 we describe the interpretation of the concepts based on the structural relations of the concept

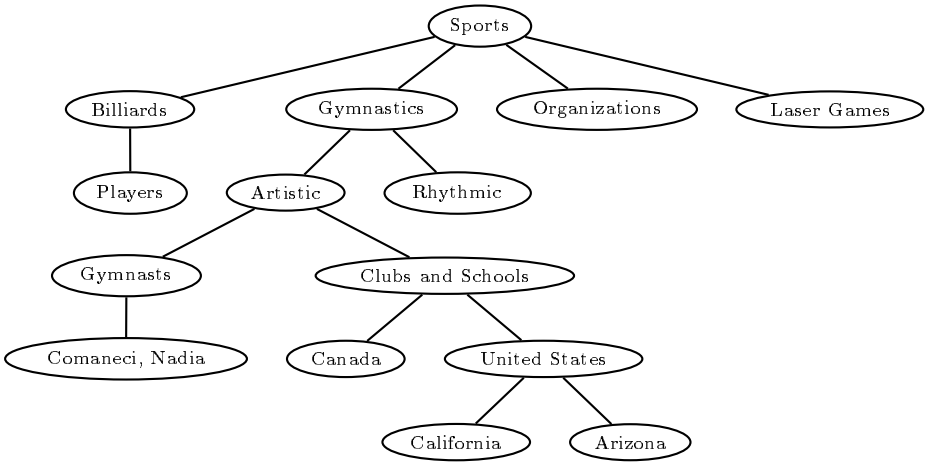

Fig. 1. Example of concept hierarchy (from Google Web Directories).

hierarchy. In Section 5 we describe the interpretation of implicit negations and disjunction. In Section 6 we discuss the results of an evaluation experiment where the procedure is applied to the Web Directories of Yahoo! and Google. Finally, Section 7 reports some relevant related work.

2 Concept Hierarchies

Here we introduce more formally the terms of our problem (see [18] for a more detailed description).

Definition 1 (Concept hierarchy) *We define a* concept hierarchy *as a triple* $H = \langle C, E, l \rangle$ *where* C *is a finite set of nodes,* E *is a set of arcs on* C*, such that* $\langle C, E \rangle$ *is a rooted tree, and* l *is a function from* C *to a set* L *of labels expressed in natural language.*

An example of concept hierarchy is provided in Figure 1, where a small part of the category 'Sport' in the Web Directories of Google is represented.

Definition 2 (Hierarchical classification) *A* hierarchical classification *of a set of documents* Δ *in a concept hierarchy* $H = \langle C, E, l \rangle$ *is a function* $\mu : C \rightarrow 2^{\Delta}$.

Classifications guide users in retrieving documents from the whole set Δ. The common procedure for seeking documents in a hierarchical classification is by entering the hierarchy from the root node, and, at each node, by choosing the child node under which the document is more likely to be classified. This choice is based on a semantic interpretation of the labels associated with the nodes, so that in most cases users do not need to check the content of the documents.

Consider, for instance, the concept hierarchy of Figure 1. In order to find documents about Romanian artistic gymnasts, a user would start from the root labeled with *Sports*, would first select *Gymnastics*, then *Artistic*, and then *Gymnasts*, and would finally retrieve the documents classified under this node. Users' choices are guided by the following facts:

- Understanding of the meaning of the labels attached to the nodes encountered during the navigation; in this case, *Sports, Billiards, Gymnastics, Organizations, Artistic, Rhythmic, Gymnasts* and *Clubs and Schools*.
- Knowledge of the fact that Romanian gymnasts are artistic gymnasts, and that gymnastics is a sport.
- Assumption that a document about artistic gymnasts is much more likely to be classified under the sub-tree rooted at *Sports/Gymnastics* than under the ones rooted at *Sports/Billiards* and *Sports/Organizations*. Similar assumptions are related to the choice between the children of *Sports/Gymnastics*, and so on.
- Awareness that the node *Gymnastics* is the most specific node about the topic 'Romanian gymnasts', as there is no node *Romania* available under *Gymnastics*.

In order to be useful for a user, a classification a should therefore respect a number of classification criteria, which can be summarized as follows:

M1 Each concept $c \in C$ has a *meaning* $m(c)$, which is some entity of a world domain.

M2 The meaning of a concept c depends only on the labels associated with a finite set of nodes $F(c) \subset C$ called the *focus of* c.

C1 A document $\delta \in \Delta$ is classified under a descendant node c' of c, i.e., $\delta \in \mu(c)$, only if δ is concerned with $m(c')$.

C2 A document δ is classified under the node c if c does not have any descendant c' such that δ is concerned with $m(c')$.

Criterion M1 guarantees that the hierarchical classification is done with a domain model in mind which is assumed to be shared by the users. Criterion M2 guarantees that the meaning of the concepts can be determined by visiting a finite (and possibly small) subset of the whole classification. Criteria C1 and C2 are standard classification criteria which can be found for instance in Yahoo! or in the Open Directory Project.[1] These two criteria provide the connection between the meaning of the labels and the set of documents classified under the node.

To formalize criteria M1-2 and C1-2, we provide a notion of *meaning* of a concept in a real world. In order to express meanings we adopt a logical approach, i.e. meaning are expressed in terms of formulas of a description logic

[1] Two foundamental criteria which are stated in many guidelines for Web Directories classification are the "Get specific" criterion and the "Look familiar" criterion.

Get Specific: When you add your document, get as specific as possible. Dig deep into the directory, looking for the appropriate sub-category. You can't submit your company to a top level category. [...] Dig deeper.

Look Familiar: Armed with the above knowledge, browse and search your way through the hierarchy looking for the appropriate category in which to add your company. Look for categories that list similar documents.

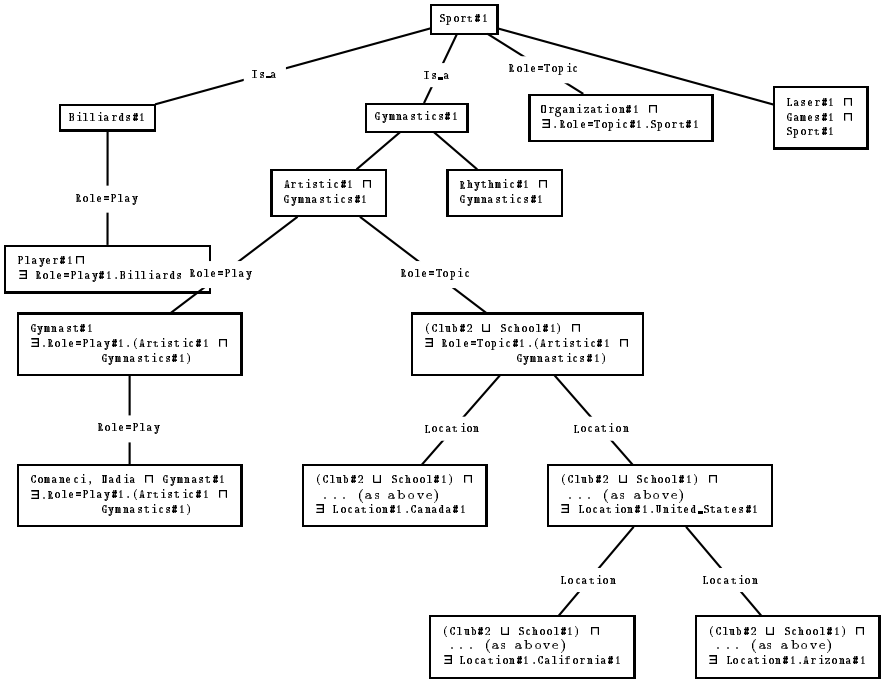

Fig. 2. The concept hierarchy of Figure 1 enriched with the meanings, and the relation types.

$\mathcal{DL}$, so that a label of a node can be associated either with a concept expression, or with a role description, or with an individual constant of $\mathcal{DL}$. The $\mathcal{DL}$ we adopt in this work is the description logics containing $\{\top, \bot, \sqcap, \sqcup, \neg, \exists R.$ and $\forall R.\}$. Furthermore, in the current work a concept c is associated with a meaning $m(c) = \phi$ which is a concept (roles and constants are not yet considered). For instance, the node *Sports* in Figure 1 corresponds to the primitive concept `sport`, while the node labeled with *Organizations* is associated with the concept $\texttt{Organization} \sqcap \exists \texttt{role} = \texttt{topic.sport}$, as it refers to sports organizations whose core business is sport.

This analysis is performed by using the semantic information provided in WordNet [8], which has been adopted because it is the largest repository of word senses and semantic relations currently available. The primitive concepts and roles of $\mathcal{DL}$ are selected among WordNet senses, on the basis of the labels occurring in $f(c)$. For instance, for the label *Sports*, WordNet provides 5 senses, which represent 5 different concepts of $\mathcal{DL}$. Again, for the word 'topic', WordNet provides different senses; for the role, we chose `topic#1`, meaning 'subject, theme'; another example of role is `location#1`, which describes spatial

relations between concepts. The process of finding the WordNet senses composing the label interpretation for a concept c is based on a linguistic analysis of $l(c)$. In order to minimize the ambiguity a further filtering of such senses is performed.

We build a basic label interpretation for the single concepts (see Section 3), and then we construct the contextual interpretation of every concept $c \in H$ by combining the basic interpretation of the concepts and their ancestors (see Section 4). The best we can obtain through the contextual interpretation of a concept hierarchy is an interpretation of each node such that the two criteria C1 and C2 are satisfied. For instance, the ideal output for the concept hierarchy in Figure 1 is reported in Figure 2.

3 Basic Label Interpretation

Concepts in concept hierarchies are described by labels, which in turn are composed by words and, possibly, separators between them. Labels are taken from a wide variety of linguistic expressions and can be single common words, such as *Dictionaries* and *Archaeology*, proper nouns, such as *Johann Sebastian Bach* and *California*, complex noun phrases, such as *Research Centers* and *Local Currency Systems*, prepositional phrases, such as *Sociology of Religion*, verb phrases, etc. More complex labels can also contain conjunctions, e.g. *Ecological and Environmental Anthropology*, punctuation, e.g. *Clubs, Teams, and Societies*, and acronyms, e.g., *GIS*.

In the first phase, i.e. basic label interpretation, we linguistically analyze the labels attached to the nodes and generate a formula in $\mathcal{DL}$ representing a first approximation of the meaning of the node.

Definition 3 (Basic label interpretation) *The* basic interpretation *is a function $lm : L \rightarrow \Phi_{\mathcal{DL}}$, that maps each single label $l \in L$ in a concept description $lm(l)$ of the description logic $\mathcal{DL}$ whose concepts are taken from the set of* WordNet *senses.*

Intuitively, a basic interpretation provides an interpretation of the concept label as a stand alone object. For instance, the basic interpretation of the node labeled with *Organizations*, occurring under *Sports*, would be equal to the basic interpretation of a node *Organizations* occurring under *Billiards*.

The first step of the procedure consists of text chunking, i.e. dividing each label into syntactically correlated parts of words. For this we run the Alembic chunker [6], developed by MITRE Corporation as part of the Alembic extraction system [1]. For example, with the label *Science Fiction and Horror*, the chunker first selects a part of speech for each word ('Science', 'Fiction', and 'Horror' are nouns, 'and' is a conjunction); then, it identifies two noun groups (NGs), i.e. 'Science Fiction' and 'Horror' (the syntactic head is marked in small capitals), and a coordinating conjunction between them (1a).

$$(1) \quad [(Science)_{nn}(\textsc{Fiction})_{nn}]_{NG}(and)_{cc}[(\textsc{Horror})_{nn}]_{NG}$$

The output of the chunker is used to transform each label into a basic logical form. A noun group consisting of more than one word is interpreted as the conjunction of the head and all its modifiers. For instance, $[(Science)_{jj}(\textsc{Fiction})_{nn}]$

is interpreted as [Science ⊓ Fiction], the reason being that the documents classified under a node with such label should be concerned both with 'science' and with 'fiction'.

The relations between different noun groups are interpreted on the basis of the linguistic material connecting them:

- coordinating conjunctions and commas are interpreted as a disjunction;
- prepositions, like 'in' or 'of', are interpreted as a conjunction;
- expressions denoting exclusion, like 'except' or 'but not', are interpreted as negations.

For example, *Science Fiction and Horror* is interpreted as a disjunction (2a), since under that node there might be both documents about 'science fiction' and documents about 'horror'; on the other hand, *Professional Photographers of America* and *Garments except Skirts* are examples of conjunction (2b) and negation (2c) respectively.

(2) *a*. [Science ⊓ Fiction] ⊔ [Horror]
b. [Professional ⊓ Photographers] ⊓ [America]
c. [Garments] ⊓ ¬ [Skirts]

The interpretation of proper nouns requires a process of named entities recognition (NER). The output of a chunker is passed to a rule-based NER system [13] which recognizes named entities and classifies them into one out of five categories (person, organization, location, measure, date). As an example, *J.S. Bach* is analyzed as in (3).

(3) <BNAMEX TYPE=PERSON *J.S.Bach* ENAMEX>

WordNet. In order to perform the semantic interpretation of the labels we access WordNet. We use a multilingual version of WordNet initially developed in the framework of the EuroWordNet Project [19] and currently in further development under the Meaning Project [16]. Five languages (English, Italian, Spanish, Catalan, and Basque) are aligned and additional semantic information, such as top ontology concepts, domains, selectional preferences, and distinctions between classes and instances, is provided. Moreover, we rely on the work carried on by [9], which aims at introducing formal distinctions in the WordNet framework. In particular, we make use of the following meta-level categories [10] associated with the synsets: TYPE, for synsets representing rigid properties (e.g. person#1), FORMAL ROLE, for synsets representing anti-rigid properties (e.g. student#1), and ATTRIBUTION, for synsets representing possible values of attributes (e.g. red#1, an attribute-value of color). We will use this meta-level information to construct appropriated logic forms for relations between concepts (see 'arc interpretation' in Section 4).

When a word is found in WordNet, all the senses of that word are selected and attached to the basic logical form. In the case of *Science Fiction and Horror*, for instance, WordNet provides all the three nouns contained in the label, and

so in the logical form we have the conjunction of the sets of senses of the three lemmas (4).

$$(4) \quad [\texttt{science}* \sqcap \texttt{fiction}*] \sqcup [\texttt{horror}*]$$

We use the following notation: `trade*` denotes the disjunction

$$\texttt{trade\#1} \sqcup \texttt{trade\#2} \sqcup \ldots \sqcup \texttt{trade\#n}$$

of all the senses of 'trade' in WordNet; `trade#3` indicates sense 3 of 'trade', while `trade#[2,4]` indicates the disjunction of senses 2 and sense 4.

Multiwords. When two or more words in a label are contained in WordNet as a single expression (i.e. a multiword), the corresponding senses are selected and, in the basic logical form, the intersection between the two words is substituted by the multiword. For instance, 'Science Fiction' is provided in WordNet as a single expression, so the logic interpretation is substituted by the senses of the multiword (5).

$$(5) \quad [\texttt{science\_fiction}*] \sqcup [\texttt{horror}*]$$

Word Sense Disambiguation. Since multiwords are much less polysemous than single words, the recognition of the multiwords provided in WordNet is a first step towards word sense disambiguation. A second step is performed by exploiting the relations between senses provided in WordNet. The label *maple tree*, for instance, is first transformed into [`maple*` $\sqcap$ `tree*`], since 'maple' and 'tree' are polysemous words and 'maple tree' is not provided in WordNet as a multiword. However, `maple#2` (defined as 'any tree or shrub of the genus Acer') is a second level hyponym of `tree#1` (i.e. 'tree' as a woody plant), and so the meaning of 'tree' as a diagram (i.e. `tree#2`) can be discarded to obtain the disambiguated basic logical form [`maple#2` $\sqcap$ `tree#1`].

4 Contextual Interpretation

An interpretation of a concept in a hierarchical classification as a stand alone object, however, is partial, as the meaning of a node depends on the context where the node occurs (see criterion M2). Intuitively, the focus $f(c, H)$ of a node c belonging to H is the part of H that the user is required to visit in order to understand whether a document is in c. The contextual interpretation of a node c gives a meaning to the node on the basis of the meaning of the nodes belonging to its focus (i.e. the ancestors of c with their direct descendants).

Let $\mathcal{H}$ be the class of concept hierarchies, and $\mathcal{C}$ the class of nodes occurring in some $\mathcal{H}$.

Definition 4 (Contextual interpretation) *A* contextual interpretation *is a function* $m : \mathcal{C} \times \mathcal{H} \rightarrow \Phi_{\mathcal{DL}}$*, where* $\Phi_{\mathcal{DL}}$ *is a concept of a description logics* $\mathcal{DL}$*, such that* $f(c, H) = f(c', H')$ *implies that* $m(c, H) = m(c', H')$*.*

Once we have associated a contextual meaning $m(c)$ to a concept c, we can define the class of documents classified under c to be the set $\Delta(c)$ such that for each $\delta \in \Delta(c)$ containing the document δ satisfying the following conditions:

C1: One of the main topics of δ is a concept ϕ, and ϕ subsumes $m(c)$, i.e., $\phi \sqsubseteq m(c)$;
C2: For any descendant c' of c, ϕ does not subsume $m(c')$, i.e., $\phi \not\sqsubseteq m(c)$.

Given the concept hierarchy H, the main task described in this Section is to find a proper contextual interpretation $m(c, f(c))$ by combining the linguistic analysis of the labels associated with c and $f(c)$, with the information provided by the structure of $f(c)$. In the following we describe how we deal with word sense ambiguity, with multiwords and arcs interpretation, taking advantage of the context provided by the hierarchical classification.

Multiwords in context. The recognition of multiwords can also be performed on different contiguous levels. For instance, in WordNet there is a multiword 'billiard player', so in a hierarchy where *Players* has *Billiards* as a parent node, its basic logic form is substituted by the senses of the multiword.

Word Sense Disambiguation in context. The context of a concept is taken into consideration to perform further disambiguation of the concept itself. We perform word sense disambiguation by taking into consideration both structural relations between labels and conceptual relations between words belonging to different labels.

Let L be a generic label and L^1 either an ancestor label or a descendant label of L and let `s*` and $\mathtt{s}^1\mathtt{*}$ be respectively the sets of WordNet senses of a word in L and a word in L^1. If one of the senses belonging to `s*` is either a synonym, a hypernym, a holonym, a hyponym or a meronym of one of the senses belonging to $\mathtt{s}^1\mathtt{*}$, these two senses are retained and all other senses are discarded.

As an example, imagine *Apple* (which can denote either a fruit or a tree) and *Food* as its ancestor; since there exists a hyponymy relation between `apple#1` (denoting a fruit) and `food#1`, we retain `apple#1` and discard `apple#2`.

Arc interpretation. The intuition underlying the methodology we propose for arc interpretation is that it depends on the ontological features of the concepts or instances connected. Arcs connecting two nodes admit different interpretations on the basis of the meta-level categories (i.e. TYPE, FORMAL ROLE, ATTRIBUTION, and INSTANCE) of such nodes. Table 1 defines the description logics interpretation for all the possible combinations of the meta-level categories we have used.

According to these rules the arc connecting `gymnastics#1` (belonging to the category TYPE) and `artistic#1` (an ATTRIBUTE), for example, is interpreted as a ROLE relation (6a). In the case of `sport#1` and `gymnastics#1`, which are both types, we base the interpretation of the arc on WordNet; since a hyponymy relation between the two concepts is provided in WordNet, we interpret the arc as the intersection of the two concepts (6b).

(6) *a.* $\mathtt{gymnastics\#1} \exists ROLE.\mathtt{artistic\#1}$
b. $\mathtt{sport\#1} \sqcap \mathtt{gymnastics\#1}$

ARC	Description Logics Interpretation	Examples
$T_1 \rightarrow T_2$	a WordNet relation between T_2 and T_1, if available; $T_2 \sqcap T_1$, otherwise	Gymnastics $\rightarrow$ Sports
$T \rightarrow R$	$R \sqcap \exists$ ROLE.T	Organ $\rightarrow$ Organists
$T \rightarrow I$	$I \sqcap \exists$ ROLE.T	Cantatas $\rightarrow$ Bach
$T \rightarrow A$	$A \sqcap \exists$ ROLE.T	Clubs and Schools $\rightarrow$ Artistic
$R \rightarrow T$	$T \sqcap \exists$ ROLE.R	Organists $\rightarrow$ Organ
$R_1 \rightarrow R_2$	a WordNet relation between R_2 and R_1, if available; $R_2 \sqcap R_1$, otherwise	Composers $\rightarrow$ Artists
$R \rightarrow I$	$I \sqcap \exists$ ROLE.R	Organist $\rightarrow$ Bach
$R \rightarrow A$	$A \sqcap \exists$ ROLE.R	Gymnasts $\rightarrow$ Artistic
$I \rightarrow T$	$T \sqcap \exists$ ROLE.I	Canada $\rightarrow$ Clubs and Schools
$I \rightarrow R$	$I \sqcap R$	Comaneci, Nadia $\rightarrow$ Gymnasts
$I_1 \rightarrow I_2$	a WordNet relation between I_2 and I_1, if available; $I_2 \sqcap I_1$, otherwise	California $\rightarrow$ United States
$I \rightarrow A$	$A \sqcap \exists$ ROLE.I	Bach $\rightarrow$ Famous
$A \rightarrow T$	$T \sqcap \exists$ ROLE.A	Artistic $\rightarrow$ Gymnastics
$A \rightarrow R$	$R \sqcap \exists$ ROLE.A	Famous $\rightarrow$ Players
$A \rightarrow I$	$I \sqcap \exists$ ROLE.A	Young $\rightarrow$ Bach
$A_1 \rightarrow A_2$	$A_2 \exists$ ROLE.A_1	International $\rightarrow$ Rhythmic

Table 1. Interpretation of the arcs on the basis of the metalevel ontological categories of the concepts connected by the arcs, where the abbreviations 'T', 'R', 'I', and 'A' stand for TYPE, FORMAL ROLE, INSTANCE, and ATTRIBUTE respectively, and where the arrow means 'classified under' ('I $\rightarrow$ T' represents the arc between a TYPE and the INSTANCE classified under that TYPE).

Finally, in order to build the contextual interpretation of the nodes in a hierarchical classification, we combine the interpretation of the labels with the interpretation of the arcs. For example, the different contextual interpretations of `gymnastics#1`, `artistic#1`, and `gymnasts#1` are represented in 7a, 7b, and 7c respectively.

(7) *a.* $\texttt{sport\#1} \sqcap \texttt{gymnastics\#1}$
b. $\texttt{sport\#1} \sqcap (\texttt{gymnastics\#1} \sqcap \exists ROLE.\texttt{artistic\#1})$
c. $\texttt{sport\#1} \sqcap (\texttt{gymnastics\#1} \sqcap \exists ROLE.\texttt{artistic\#1}) \sqcap \exists ROLE.\texttt{gymnast\#1}$

5 Implicit Disjunctions and Negations

Implicit Disjunctions. As explained in Section 3, the presence of a coordinating conjunction makes the disjunction between noun groups within a label explicit , but we can also have implicit disjunction between elements placed at different levels of the hierarchy (concepts with a disjoint descendant).

As an example, let's take a concept hierarchy with the root *Soccer*, a descendant *Leagues*, and a further descendant *Clubs*, which admits two conflicting interpretations: from the point of view of the hierarchical structure, *clubs* denotes a subset of *leagues* (being a child of it); on the other hand, from the point of view of the world knowledge provided in WORDNET, [`club#2`] (defined as 'a formal

association of people with similar interests') and [`league#1`] (defined as 'an association of sports teams'), can be considered as disjoint because they have the same hypernym, i.e. `association#1`. In order to combine the two information sources, *Leagues* has to be reinterpreted as if it were *Leagues and Clubs* (8a).

$$(8) \quad [\texttt{soccer}*] \sqcap [[\textbf{league\#1}] \sqcup [\textbf{club\#2}]]$$

When two concepts in a path are disjoint, i.e. when a concept is disjoint from a concept that is either an ancestor or a descendant, the meaning of the ancestor has to be reinterpreted. More formally:

Let c and c' be two concepts, and let `c#i` and `c'#j` be two senses of c and c' respectively. We apply the following rule:

- replace the sense `c#i` with `c#i` *and* `c'#j`, if `c'#j` is disjoint from `c#i` and c is an ancestor of c'[2]

Implicit Negations. Similarly, the negation can be marked by expressions like 'but not' or 'except', but can also be implicit in the case of elements belonging to different labels (inclusion relation between two siblings). For instance, in Google Web Directories we have *Sociology* and *Science* as sibling nodes classified under *Academic Study of Soccer*; from the point of view of world knowledge, sociology is a science (and in fact in WordNet `sociology#1` is a second level hyponym of `science#2`). As a consequence, the node labeled with *Science* has to be interpreted as if it were *Science except Sociology*.

Whenever a concept in a label has a part-of relation or an is-a relation with a concept in another label on the same level, it is necessary to re-interpret the meaning of the more general concept. More formally:

Let c and c' be two concepts, and let `c#i` and `c'#j` be two senses of c and c' respectively. We apply the following rule:

- replace the sense `c#i` with `c#i` $-$ `c'#j`, if `c#i` is either a hyponym or a meronym of `c'#j` and c and c' are siblings

6 Experiments

As a test set for the evaluation of the algorithm for the interpretation of concept hierarchies, which has been implemented in Java, we have focused on the Web Directories of Yahoo! and Google, where documents are represented by millions of Web page URLs'.

Yahoo! and Google Web Directories have respectively fourteen and fifteen main categories (e.g. 'Computer & Internet', 'News & Media', 'Recreation & Sport', 'Health', 'Society & Culture', 'Arts & Humanities', 'Science', 'Social Science' in Yahoo! and 'Arts', 'Computer', 'Health', 'News', 'Recreation', 'Science' and 'Society' in Google) consisting of a number of nodes ranging between a few thousand and tens of thousands nodes each. Each of these categories can be considered as the root of a sub-hierarchy.

[2] `s#k` is disjoint from `t#h` if `s#k` belongs to the set of opposite meanings of `t#h` (if `s#k` and `t#h` are adjectives) or, in the case of nouns, if `s#k` and `t#h` are different hyponyms of the same synset.

	Yahoo! Arch.	Google Arch.	Yahoo! Med.	Google Med.
# Concepts	105	312	703	1,023
Average label repetition	1.0	1.3	2	1.8
# Words	170	521	1,231	1,549
# Words/label	1.6	1.7	1.8	1.5
WordNet's coverage	95.5%	91.5%	88.7%	91.4%
Average polysemy	3.8	3.7	4.6	3.2
# Multiwords	11	45	51	116
# Disjunctions	10	51	99	58
# Conjunctions	41	109	239	325

Table 2. Analysis of the Architecture and Medicine sub-directories in Yahoo! and Google.

A preliminary analysis has been performed on two sub-hierarchies, i.e. 'Architecture' (under the main category 'Art') and 'Medicine' (under the main category 'Health'), whose sizes range between one hundred and one thousand nodes (see Table 2). The labels attached to the nodes are generally short, with an average of 1.5-1.8 words per label. Labels can be repeated more than once since the same label can be attached to different nodes in different places of the hierarchy; in fact, the bigger a hierarchy is, the higher is the average repetition of the labels (there are no repetitions in Yahoo! 'Architecture', while in Yahoo! 'Medicine' a label is repeated on average two times).

These two sub-hierarchies have been chosen among those where WordNet's coverage was highest (in fact, between 88.7% and 95.5% of the words and lemmas occurring in the labels are found in WordNet). It has been found that each lemma has on average between 3.2 and 4.6 senses, which makes the need for word sense disambiguation very important.

A manual and partial evaluation of the disambiguation process (we have checked manually the WordNet senses suggested by the algorithm for the 312 labels in Google 'Architecture') has shown that the procedure is very precise, with a precision rate ranging between 69 and 75%, but has a low recall, which is due to the fact that WordNet contains only hyponymy and meronymy relations and no other kinds of relations, like role or location relations.

As for the presence of multiwords, between 11.3% and 14.4% of the labels contain a multiword, which is remarkable, if we consider that between 54.8% and 62.6% of them consist of one single word. The good recognition of multiwords contributes to reduce the polysemy of concepts as most of them (almost 80% of the multiwords recognized in the experiment) are monosemous, with an average polysemy rate around 1.2 senses per multiword.

As far as negations are concerned, the hierarchies under analysis do not contain expressions denoting exclusion, while many implicit negations have been discovered. Our procedure works quite well with nouns denoting concrete objects like buildings. For example, in Google 'architecture', the node *Architecture/History/Periods and styles/Romanesque* has two descendants, *Churches* and *Cathedrals*; since cathedrals are actually churches (and an hyponymy relation between them is provided in WordNet), *Churches* is reinterpreted as if it were

Churches except Cathedrals. On the other hand, performance is not so good with abstract nouns, like states and events.

A limitation of the current system is that the use of the Alembic chunker does not permit the resolution of coordination ambiguities involving nominal compounds (and neither would the use of a more sophisticated parser, as this problem has received relatively little attention [15], when compared to other aspects, like for instance prepositional phrase attachment). To make an example, noun phrase coordinations with the form *n1 and n2 n3* admit two structural analysis, one in which *n1* and *n3* are the two syntactic heads being conjoined (9a) and one in which the conjunction is between the modifiers *n1* and *n2* (9b).

(9) *a* $(Nightclubs)\ and\ (Dance\ Halls)$
b $(Food\ and\ Drug)\ Administration$

Most of the times Alembic suggests the first analysis, which is correct in the case of *Nightclubs and Dance Halls* (10a), but is incorrect in many other cases, like *Food and Drug Administration* (10b) , which should be analyzed as in 10c. We plan to refine our analysis of coordinations involving nominal compounds by introducing rules based on number agreement as in [15].

(10) *a* $[(\textsc{Nightclubs})_{nn}]_{NG}(and)_{cc}[(Dance_{nn}(\textsc{Halls})_{nn}]_{NG}$
b $*[(\textsc{Food})_{nn}]_{NG}(and)_{cc}[(Drug)_{nn}(\textsc{Administration})_{nn}]_{NG}$
c $[((Food)_{nn}(and)_{cc}(Drug)_{nn})(\textsc{Administration})_{nn}]_{NG}$

7 Related Work

Our use of linguistic analysis to enrich hierarchical classifications with semantic information is related to the work presented in [20] on conceptual indexing. In that work the conceptual structure of phrases is analyzed using semantic relationships between words to establish connections between the terms in a conceptual taxonomy. Even if similar methodologies are applied, our approach aims at interpreting already existing taxonomies in order to make explicit a number of semantic relations, while in conceptual indexing the starting point is the extraction of terminology from documents.

WordNet is also used in [14] to give semantic interpretation to complex terms that have been automatically extracted from texts; relations between synsets are then exploited in order to organize concepts into trees. In our approach, on the other hand, the hierarchies represent the starting point, and structural information is used together with semantic information in order to interpret the labeled nodes.

Contextual interpretation of Web Directories headings has also been suggested in [12], but with a different aim. Here the knowledge embedded in the structure of the Directories is used to obtain labeled training data for Information Extraction from Web documents with limited human effort, while we aim at interpreting the Web Directories headings in order to discover the content of the classified documents without looking them up.

The problem of allowing for the interoperability of concept hierarchies, and in particular of catalogs, has been addressed also in [2]. Their approach is based

on the use of document classification algorithms; our methodology, on the other hand, does not use the documents associated with the nodes of the conceptual hierarchy, since it is based on the interpretation of labels and of the relations between them.

Finally, research related to the linguistic analysis of multi-word expressions and terminology has been conducted by Jacquemin and Morin in [11], who describe a framework for organizing multi-word candidate terms with the help of automatically acquired links between single-word terms derived from WordNet.

8 Conclusions

We have provided a formal semantics for hierarchical classifications and then used that formal framework to explore a number of linguistic issues crucial for interpreting the knowledge implicitly represented in such classifications.

The methodology we have proposed, based on a linguistic interpretation of the labels provided in the hierarchy, takes as input a concept hierarchy and returns the interpretation of each label. The process of interpreting a label coincides with the progressive construction of a logical form in description logics, where predicates are WordNet senses. It is performed in two steps: on the basis of the output of the chunker, basic logic forms are first built for each single concept independently of the others; then, full logical forms are built by combining the basic logic form of each concept with the basic logical forms of the nodes belonging to its focus.

In the future we plan to work on a systematic analysis of the performance of the methods with respect to the different steps and on the realization of a module for the discovery of the different kinds of relations between concepts, such as role, location, etc.

References

1. Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P. and Vilain, M.: MITRE: Description of the Alembic System as Used for MUC-6. Proc. of the Sixth Message Understanding Conference (MUC-6), Columbia, Maryland, November, 1995
2. Agrawal, R. and Srikant, R.: On Integrating Catalogs. Proc. of the Tenth International World Wide Web Conference (WWW-2001), Hong Kong, China, May, 2001
3. Baader, F. and Nutt, W.: Description Logic Handbook. Pages 47-100, Cambridge University Press.
4. Bergamaschi, S., Guerra, F. and Vincini, M.: Product Classification Integration for E-Commerce. Proc. of WEBH-2002, Second International Workshop on Electronic Business Hubs. Aix En Provence, France. September, 2002
5. Bouquet, P., Magnini, B., Serafini, L. and Zanobini, S.: A SAT–based Algorithm for Context Matching. To appear in: Proc. of the Fourth International and Interdisciplinary Conference on Modeling and Using Context
6. Day, D.S. and Vilain, M.B.: Phrase Parsing with Rule Sequence Processors: an Application to the Shared CoNLL Task. Proc. of CoNLL-2000 and LLL-2000. Lisbon, Portugal, September, 2000

7. Doan, A., Madhavan, J., Domingos, P. and Halevy, A.: Learning to Map between Ontologies on the Semantic Web. Proc. of WWW-2002, 11th International World Wide Web Conference, Honolulu, Hawaii, May, 2002
8. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, US, 1998
9. Gangemi, A., Guarino, N. and Oltramari, A.: Restructuring WordNet's Top-Level: The OntoClean Approach. Proc. of ONTOLEX 2002 (Workshop held in conjunction with LREC 2002), Las Palmas, Canary Islands, Spain, May, 2002
10. Guarino, N.: Some Ontological Principles for Designing Upper Level Lexical Resources. Proc. of LREC 1998, Granada, Spain, 1998
11. Jacquemin, E. and Morin, E.: Projecting Corpus-Based Semantic Links on a Thesaurus. Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June, 1999
12. Kavalec, M. and Svatek, V.: Information Extraction and Ontology Learning Guided by Web Directory. Proc. of OLT-02 (Workshop on ML and NLP for Ontology Engineering) held in conjunction with ECAI 2002, Lyon, France, July 2002.
13. Magnini, B., Negri, M., Prevete, R. and Tanev, H.: A WordNet-Based Approach to Named Entities Recognition. Proc. of the Workshop SemaNet'02: Building and Using Semantic Networks, at COLING-02, Taipei, Taiwan, 2002
14. Missikoff, M., Navigli, R. and Velardi P.: Integrated Approach for Web Ontology Learning and Engineering. IEEE Computer, November, 2002.
15. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 2001
16. Rigau, P., Magnini, B., Agirre, E., Vossen, P. and Carrol, J.: MEANING: a Roadmap to Knowledge Technologies. Proc. of the workshop "A Roadmap for Computational Linguistics", COLING-02, Taipei, Taiwan, 2002
17. Schulten, E., Akkermans, H., Botquin, G., Drr, M., Guarino, N., Lopes, N. and Sadeh, N.: Call for Participants: The E-Commerce Product Classification Challenge. IEEE Intelligent Systems, 16-4, 2001
18. Serafini, L., Bouquet, P. and Donà, A.: CTXML Context Markup Language. Technical report, ITC-irst, 2002
19. Vossen, P.(Ed.): Special Issue on EuroWordNet, Computers and Humanities, 32, 1998
20. Woods, W.A.: Conceptual Indexing: A Better Way to Organize Knowledge. SUN Technical Report TR-97-61, 1997