A Semantic-based Approach to Interoperability of Classification Hierarchies: Evaluation of Linguistic Techniques

Bernardo Magnini and **Manuela Speranza** and **Christian Girardi**
ITC-irst
Via Sommarive, 18 Povo
38050 Trento,
Italy,
magnini@itc.it, manspera@itc.it, cgirardi@itc.it

Abstract

Classification Hierarchies (CHs) are widely used to organize documents in a way that makes their retrieval easier. Common examples of CHs are Web directories, marketplace catalogs, and file systems. In this paper we discuss and evaluate CtxMatch, an approach to interoperability that discovers mappings among CHs considering the semantic interpretation of their nodes. CtxMatch performs a linguistic processing of the labels attached to the nodes, including tokenization, Part of Speech tagging, multiword recognition and word sense disambiguation. We present an evaluation of the overall performance of the approach over Web directories as well as a systematic analysis of the linguistic modules involved.

1 Introduction

Classification Hierarchies (CHs) are taxonomic structures used to organize large amounts of documents. Documents 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; in the directories of Web portals, documents are pointers to Web pages; in the marketplace, catalogs organize either product cards or service titles.

CHs are now widespread as knowledge repositories and the problem of their integration is acquiring a high relevance from a scientific and commercial perspective. In this paper we present CtxMatch, an algorithm that takes as input the labels attached to two nodes belonging to different partially overlapping CHs and returns a mapping relation (i.e. equivalence, more general, more specific) between them. Unlike previous approaches to interoperability, CtxMatch does not consider the content of the documents classified in the CHs; rather, it relies both on the semantic interpretation of the labels describing the nodes, which is obtained through a linguistic analysis, and on the structure of the CH itself. The contribution of the paper is in two main directions: (i) we address the linguistic processing required for the semantic interpretation of CH labels; in our knowledge there are no previous attempts that systematically apply NLP tools and resources to this task; (ii) we report on a large-scale evaluation of the performance of such tools over real CHs. The results we obtained are a useful benchmark, available for future work in this area.

In the attempt to carry out a semantic interpretation over CH nodes, at least the following issues seem to be crucial (examples are taken from Figure 1, in which a small subsection of Google Web Directories is reported):

Splitting and contextual interpretation. Information is split on several levels; a single node provides only partial information, so that the interpretation process has to consider a larger scope. As an example, *Players* in Figure 1 refers to billiard players, not to players in general.

Redundancy. Information can be partially repeated at different levels of a CH. For instance, if *ACL-02* is placed under *Papers-2002*, the fact that *ACL-02* refers to a conference of the year 2002 is implicitly represented at two levels.

Linguistic complexity of the labels. Labels can be arbitrarily complex: they may include abbreviations, multiwords (e.g. *United States* in Figure 1), coordinated expressions, proper names (e.g. *Comaneci, Nadia*), etc.

Ambiguity and Synonymy. Labels may have different meanings and need to be disambiguated within their context. On the other hand, different labels may have the same meaning (e.g. *Papers* and *Articles*). In order to deal with these aspects of language we have used WordNet as a repository of senses, and we have designed word sense disambiguation techniques specifically tuned for CHs.

Lack of linguistic context. The interpretation of a label is necessarily based on a limited linguistic context. As a consequence, the applica-

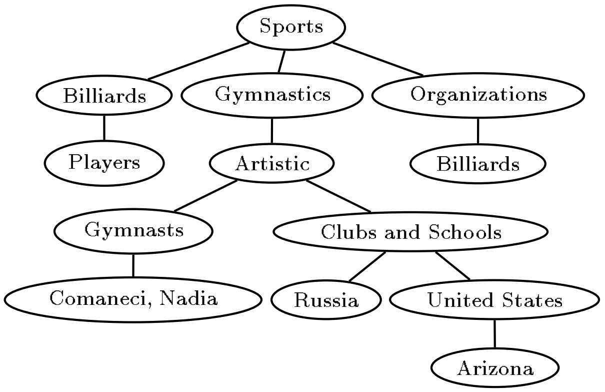

Figure 1: Example of a classification hierarchy (from Google Web Directories).

tion of NLP techniques (e.g. PoS-tagging, word sense disambiguation, etc.) opens up the problem related to the use of tools usually developed for texts. For instance, we re-trained the PoS-tagger on a specific CH corpus.

Relation to world knowledge. CHs implicitly reflect the world knowledge of a specific domain, but they also reflect the subjective criteria adopted for organizing documents. World knowledge and subjective criteria may interact in subtle ways.

The paper is structured as follows. In Section 2 we review the relevant approaches to interoperability among CHs, outlining the main differences with respect to the semantic-based approach we propose. In Section 3 we describe the CtxMatch algorithm. In Section 4 we report on the results of the evaluation experiments where CtxMatch is applied to the Web Directories of Yahoo! and Google. Finally, in Section 5 we draw some conclusions.

2 Approaches to CH Interoperability

In our view, the problem of the interoperability among different CHs can be roughly stated in this way: given a node N_s in a source CH and a node N_t in a target CH, the algorithm has to discover a relation between N_s and N_t. Although there can be differences in the definition of the task itself (Agrawal and Srikant, 2001; Madhavan et al., 2002), and considering that this is a relatively new challenge, approaches to CH mapping can be grouped into four classes, according to the kind of information used: (i) approaches which consider the content of the documents belonging to the CH; (ii) approaches based on the classification of the documents; (iii) approaches that exploit the structure of the CH; and (iv) approaches that attempt a semantic interpretation of the CH labels. In the rest of this Section we will briefly review the first three approaches, while the semantic-based approach will be introduced in more detail in Section 3.

Mapping based on document content. These approaches rely on the content of the documents classified in a CH. As an example, the GLUE system (Doan et al., 2002) employs machine learning techniques to discover mappings among CHs. The idea consists of training a classifier using documents of the source CH, and then apply that classifier to documents of the target CH, and vice-versa. The major drawback of this approach is that it requires textual documents, which prevents its usability when such documents are of a different nature (e.g. images) or they are not available at all.

Mapping based on document classifications. An improvement with respect to the content-based approach has been proposed by Ichise et al. (2003), who address the mapping problem by computing a statistical model of the classification criteria of the CHs. Such a statistical model attempts to determine the degree of similarity between two categorization criteria considering the number of documents in common to nodes of different CHs. The advantage over the content-based approach is that the analysis of the documents is not necessary. However, it is required that the source and the target CHs share a certain amount of documents, which is hard to obtain in most of the concrete application scenarios.

Mapping based on structural information. These approaches attempt to discover mappings independently of the number and the type of the classified documents. For instance, Daude et al. (2000) exploit a constraint satisfaction algorithm (i.e. relaxation labeling) for discovering relations among ontologies. It first selects candidate pairs using lexical similarities (i.e. concepts with the same label) and then considers a number of structural constraints among nodes (e.g. connections between their hypernyms) to increase or decrease the weights of the connection. Although the approach has been experimented and evaluated to map two versions of WordNet, achieving high accuracy, our impression is that mapping CHs is a sensibly harder task, due to the highly idiosyncratic way in which CHs may organize their content.

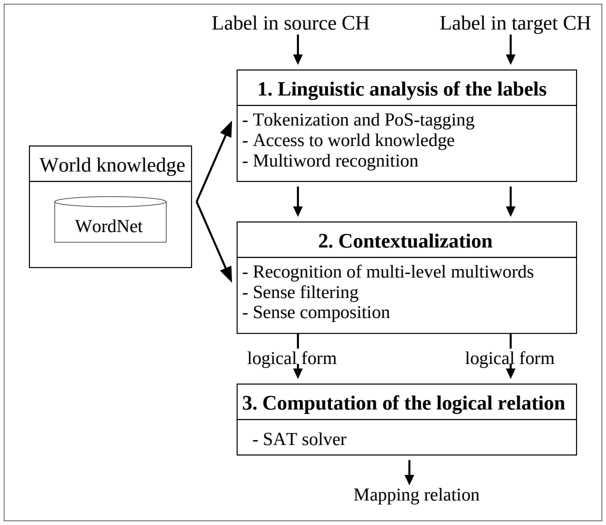

Figure 2: The architecture of CtxMatch.

3 CtxMatch Algorithm

CtxMatch is a particular implementation of an approach to *semantic coordination* recently proposed in Bouquet et al. (2003) and Magnini et al. (2003). The main difference between CtxMatch and other approaches to schema matching (see Section 2) is that in order to interpret a node of a hierarchy it considers the *implicit information* derived from the *context* where the node occurs, i.e. the structural relations with the other nodes of the hierarchy.

CtxMatch consists of three main phases (see Figure 2): (i) linguistic analysis of the labels, (ii) contextualization, and (iii) computation of the logical relation.

Linguistic analysis of the labels. In this phase, nodes are interpreted as stand alone objects, i.e. independently of their context and position in the hierarchy.

Words in a label are first tokenized, lemmatized and tagged for Parts of Speech. We use TokenPro and LemmaPro, both developed at IRST, and the TNT tagger (Brants, 2000) with a tag set reduced to the four categories that are significant for accessing WordNet (i.e. nouns, adjectives, adverbs, verbs), and a generic category 'other'. Then, we access a multilingual version of WordNet developed under the Meaning Project (Rigau et al., 2002). When a lemma is found, all the senses provided for the syntactic category selected by the PoS-tagger are attached to the lemma. In the case of *United States* in Figure 1, for instance, the WordNet senses of both the adjective and the noun are added to the label (1).

(1) $[\mathtt{united*}_a \; \mathtt{state*}_n]$[1]

When a group of words in a label are contained in WordNet as a single expression, the corresponding senses are selected and the senses of the single lemmas are replaced with those of the multiword. The multiword recognizer we have developed first retrieves the multiwords containing at least two adjacent words of a label and then selects those containing the highest number of words. For instance, 'United States' is recognized as a WordNet multiword, so this information is added to the label (2).

(2) $[\mathtt{United\_States*}_n]$

Then, we transform each label into a formula in description logic (Baader and Nutt, 2002) representing a first approximation of the meaning of the node, where the node is considered a stand alone object.

As a general rule, a label consisting of more than one word is interpreted as the conjunction of its elements, since the documents classified under a node with a certain label should be concerned with all the words contained in that label; for instance, the label *Laser Games* found in Google Web Directories under *Sports* is interpreted as $[\mathtt{laser*} \sqcap \mathtt{game*}]$.

Other rules are based on the linguistic material provided in the labels: coordinating conjunctions and commas are interpreted as a disjunction; prepositions are interpreted as a conjunction; expressions denoting exclusion, like 'except', are interpreted as negations. For example, *Clubs and Schools* in Figure 1 is interpreted as a disjunction (3), since under that node there might be both documents about clubs and documents about schools.

(3) $[\mathtt{club*}_n \; \sqcup \; \mathtt{school*}_n]$

Contextualization. In the second phase of CtxMatch we contextualize the interpretation of a node, i.e. we take into consideration its ancestors in order to generate a logical form representing its meaning.

Intuitively, we define the *focus* of a node as the part of the hierarchy that a user is required to visit in order to understand whether a document is under that node. More precisely, given a node N_j in a classification hierarchy H, the focus of N_j include all the ancestors of N_j and all their direct descendants in H.

[1] We use the following notation: state* denotes the disjunction of all the senses of 'state' in WordNet, while state#1 indicates sense number 1.

The logical form of a node is built combining the logical form of the node with the logical form of its ancestors through intersection. For example, the logical form of the root of the CH in Figure 1 is simply [**sport***], while the logical forms of *Billiards* and *Players* contain conjunctions, as shown in 4a and 4b respectively (the label attached to the node to which the logical form refers is highlighted in bold type).

(4a) $[\texttt{sport*}] \sqcap [\textbf{billiards}]$
(4b) $[\texttt{sport*}] \sqcap [\texttt{billiards*}] \sqcap [\textbf{player*}]$

The recognition of multiwords can also be performed on different contiguous levels. For instance, in WordNet there is a multiword 'billiard player', so in our example (4b), the intersection between billiards and player is substituted by the senses of the multiword (5).

(5) $[\texttt{sport*}] \sqcap [\texttt{billiard\_}\textbf{player*}]$

The focus of a concept is taken into consideration to perform sense filtering: the senses of Nj that are not compatible given the senses belonging to its focus are deleted. As an example, two senses are attached to *Arizona*, denoting respectively a state in the USA and a snake, and two senses are attached to *United_States*; since there exists a part-of relation between Arizona#1 and United States#1, and United States#1 belongs to the focus of Arizona#1, Arizona#2 and United_States#2 can be discarded.

The next step is sense composition, where we address possible inconsistencies between the hierarchical structure and the world knowledge provided in WordNet. For example, Google Web Directories has *Sociology* and *Science* as sibling nodes under *Academic Study of Soccer*, which admits two conflicting interpretations: from the point of view of the world knowledge provided in WordNet, sociology#1 is a second level hyponym of science#2 (which means that sociology is a science); on the other hand, from the point of view of the hierarchical structure, the sets of documents classified under the two nodes are disjoint. In order to combine the two information sources, *Science* has to be interpreted as if it were *Science except Sociology.*

Computation of the logical relation. We check whether a mapping relation, i.e. an equivalence, a more general or a less general relation, holds between the logical forms k and k' representing the meaning of the input nodes. To this aim, the task of finding a relation is transformed into a problem of propositional satisfiability (SAT), and then computed via a standard SAT solver. The SAT problem is built in two steps. First, the algorithm selects the portion T of the background theory relevant to the two logical forms, namely the semantic relations involving the WordNet senses that appear in them. Then, it computes some of the logical relations which are implied by T. The background theory T relevant for computing the relation between two formulas k and k' is obtained by transforming the WordNet hierarchical relations between senses appearing in k and k' into a set of subsumptions in description logic according to the following rules:

- $\texttt{c\#i} \rightarrow \texttt{c\#j}$ (if c#i is a hyponym of c#j);
- $\texttt{c\#j} \rightarrow \texttt{c\#i}$ (if c#i is a hypernym of c#j);
- $\texttt{c\#i} \equiv \texttt{c\#j}$ (if c#i and c#j are synonyms).

The equivalence relation between k and k' (and thus between the nodes whose meanings are represented by the logical forms) is checked by verifying that $k \sqsubseteq k'$ and $k' \sqsubseteq k$ are both implied by T. Similarly, the less [more] general relation between k and k' is checked by verifying that $k \sqsubseteq k'$ [$k' \sqsubseteq k$] is implied by T. For example, the mapping between the source node *Clubs and Schools* in Figure 1 and the target node *schools* classified under *athletics/acrobatics/artistic* in a different CH is one of inclusion. The logical forms of the nodes (6, 7) and the logical relations implied by the background theory (8, 9) are given to SAT.

(6) $[\texttt{sport\#1}] \sqcap [\texttt{gymnastics\#1}] \sqcap$
$\sqcap [\texttt{artistic\#1}] \sqcap [\texttt{club\#2} \sqcup \texttt{school\#1}]$

(7) $[\texttt{athletics\#1}] \sqcap [\texttt{acrobatics\#1}] \sqcap$
$\sqcap [\texttt{artistic\#1}] \sqcap [\texttt{club\#2}]$

(8) $\texttt{sport\#1} \equiv \texttt{athletics\#1}$

(9) $\texttt{acrobatics\#1} \rightarrow \texttt{gymnastics\#1}$

Through SAT we check for satisfiability the union of all the propositions (e.g. 8 and 9) and the negation of the implication between the logical forms 6 and 7. Since the check fails, a more general relation is computed between the two nodes; otherwise a similar procedure is followed for the other mapping relations.

4 Evaluation of CtxMatch

In this Section we present an experiment performed on the Web Directories of Yahoo! (2003) and Google (2003) where the outputs of the individual tools and modules we have developed or adapted have been systematically evaluated against a manually tagged gold standard.

The Web Directories of Yahoo! and Google have respectively fourteen and fifteen main cat-

	Yah. Arc.	Goo. Arc.	Yah. Med.	Goo. Med.
# labels	149	413	706	675
# tokens	218	947	1344	931
Tokens/label	1.5	2.3	1.9	1.4
Multiwords/label	0.12	0.09	0.08	0.14
# lexical words	207	700	1137	889
# lemmas not in Wn	7	324	51	30
Wn lemmas coverage	97%	54%	95%	97%
# polysemic lemmas	117	129	672	508
Average polysemy	4	3.7	5.2	3.6

Table 1: Analysis of the 'Architecture' and 'Medicine' directories in Yahoo! and Google.

egories, each of which can be considered as the root of a CH. For the evaluation of CtxMatch we have selected the 'Medicine' and 'Architecture' sub-hierarchies, whose sizes range from one hundred to seven hundred labeled nodes (see Table 1). Labels are generally short (on average 1.5-2.3 tokens per label) but nonetheless the occurrence of multiwords is significant (on average, a multiword every ten labels).

WordNet's coverage with respect to lemmas is generally very high (between 95% and 97% of the lexical words, e.g. nouns, adjectives, verbs and adverbs, are found in WordNet), with the exception of Google 'Architecture' where it falls to 53.7% (this is due to the fact that more than half the labels consist of names of architects that are not provided in WordNet). Polysemic lemmas (both single words and multiwords) have on average between 3.6 and 5.2 senses, which makes the need for word sense disambiguation very important.

The evaluation took into consideration (i) tokenization and PoS-tagging; (ii) multiword recognition; (iii) sense filtering; and (iv) logical relation computation. Every phase has been evaluated independently of the errors which occurred in the previous phases, since at every step the algorithm was fed with the correct input built from the gold standard.

4.1 Tokenization and PoS-tagging

The performance of the tokenizer was calculated in terms of accuracy with respect to labels: for every label, the output of the tokenizer was evaluated against the gold standard (recall is not significant as the tokenizer always provides an answer). The results (see Table 2) show that the performance of the tokenizer is not penalized by the lack of context as, in most cases, we obtained an accuracy of 100%. Only in Google 'Architecture', did the tool make some mistakes (e.g some middle initials were treated as single letters followed by a full stop).

The performance of the lemmatizer and the PoS-tagger are presented in terms of accuracy with respect to single tokens (again, recall is not significant). [2] The evaluation of both tools is not influenced by tokenization errors as the tokens given as input were taken from the gold standard. Accuracy was satisfactory both for lemmatization and PoS-tagging (with rates in the ranges of 97-99% and 90-97% respectively). In most cases, if the selected lemma is wrong, the assigned part of speech is also wrong; however, the cases where the lemma is assigned correctly and the PoS is not (e.g, the plural noun 'States', correctly lemmatized as 'state', but erroneously tagged as verb) are more frequent than the reverse, which explains the slightly better performance in lemmatization.

4.2 Multiword Recognition

The performance of the multiword recognizer were more than satisfactory, both in terms of precision (correctly retrieved/retrieved) and in terms of recall (correctly retrieved/relevant): in total, only three multiwords were missed by the algorithm and three others were misidentified. For example, in the label *Online Databases* in Google 'Medicine', the algorithm did not recognize the multiword `on-line_database` because WordNet provides only the hyphenated version and the algorithm does not handle this kind of linguistic variation. In *Gropius, Walter* and *Jefferson, Thomas* (in Google 'Architecture'), the algorithm did not recognize `Walter_Gropius` and `Thomas_Jefferson` since it depends on word order (giving up this strict connection to word order would increase recall but would decrease precision).

On the other hand, some false positives occurred because the multiword recognizer does not take into consideration any information about dependency structure and semantics. For example, the multiword `city_state` identified by the algorithm in *Traverse City State Hospital* (Google 'Architecture') is wrong in the context of the State Hospital of Traverse City (Michi-

[2] Multiple tags were not admitted in the gold standard, so a literal interpretation was preferred; for example, 'New' in 'New York' was tagged as an adjective (only after multiword recognition was it considered as part of the noun multiword `New_York`).

	Yahoo! Architecture			Google Architecture			Yahoo! Medicine			Google Medicine		
Tokenization (Accuracy)	1			.98			1			1		
Lemmatization(Accuracy)	.97			.98			.99			.98		
PoS-tagging(Accuracy)	.96			.90			.97			.90		
Multiword rec. (Pr, Re, F)	.95	1	.97	.95	.95	.95	1	1	1	1	.99	.99
Sense filtering (Pr, Re, F)	.72	.24	.36	.68	.26	.38	.66	.04	.08	.73	.35	.47

Table 2: CtxMatch results on the linguistic analysis of 'Architecture' and 'Medicine' directories.

gan) and so are `art movement` in *Arts and Crafts Movement* (Yahoo! 'Architecture') and `Andrew_Jackson` (the US president) in *Downing, Andrew Jackson* (Google 'Architecture'),

4.3 Sense Filtering

The performance of sense filtering is satisfactory as far as precision is concerned: we obtained precision rates varying between 66% and 73%. As an example of wrong sense filtering, in the label *Employment* (placed directly under the root *Medicine*), the algorithm erroneously removes the sense with the meaning of *job* and retains `employment#4` (defined in WordNet as 'the act of using') because of the WordNet relation between `employment#4` and `optometry#1` (which occurs in the focus of *Employment*).

Since sense filtering strictly depends on the relations found in WordNet, recall is sensibly lower. In most cases we obtained satisfactory results, i.e. in the range from 24% to 35%, with a resulting F-measure ranging from 36% to 47%. In the case of Yahoo! 'Medicine', on the other hand, we obtained a recall of 4%. The algorithm actually identified a very low number of WordNet relations (around hundred) which mainly involved monosemic lemmas (for which no sense filtering is required) and so, in total, sense filtering was applied only to 27 lemmas. This can be explained by the fact that this particular hierarchy contains words which are not much interrelated from the semantic point of view.

4.4 Logical Relation Computation

Since it was not feasible to create a manual mapping between all possible pairs of nodes, the logical relations computed by CtxMatch have been evaluated considering the URLs classified in the CHs. The underlying assumption is that, given a source node and a target node belonging to different hierarchies, the higher the number of the documents (i.e. URLs) shared by the nodes, the higher the similarity between them. The fact that the URLs in Google and Yahoo! Web directories have been classified manually guarantees both that these classifications are of high quality and that they represent a good approximation of human judgment.

The evaluation was performed in four steps: (i) we identified the set D of documents classified in both CHs and selected the nodes containing at least one document belonging to this set; (ii) we established a correlation between the proportion of documents shared by source node and target node and the logical relation existing between them. The methodology for this was taken from Doan et al. (2002), who propose three formulas for calculating the similarity between nodes of CHs; (iii) we ran CtxMatch on the selected nodes; and (iv) evaluated the mapping relations computed by CtxMatch.

Equivalence relation. The evaluation of the `equivalence` relation is based on the similarity (calculated with the cosine measure) between two sets of documents: the set A of documents belonging to the common set of documents D classified under the source node, and the set B of documents belonging to D classified under the target node. According to (10) the similarity between the two sets is 1 if they contain the same documents and 0 if they are disjoint. Since in Yahoo! and Google Web directories the number of documents shared by pairs of nodes is low and there can be different classifications of the same document due to human disagreement, we introduced an approximation factor ϵ, so that an `equivalence` relation is judged as correct if the similarity measure ranges between 1 and (1 - ϵ), where ϵ is empirically set to 0,1.

(10) $SIM(A, B) = A \cap B/\sqrt{(A * B)}$

More [less] general relation. The *most-specific-parent* [*most-general-child*] measure (11) takes a value in the range [0,1] when a node subsumes the other, so a `more [less] general` relation is correct if it ranges between

		Pr.	Re.	F
Arc.	equiv.	.33 (.25)	.04 (.04)	.07 (.07)
	more g.	.92 (.93)	.42 (.44)	.58 (.60)
	less g.	.88 (.90)	.62 (.41)	.73 (.56)
Med.	equiv.	.27 (.25)	.07 (.05)	.11 (.08)
	more g.	.91 (.95)	.48 (.45)	.63 (.61)
	less g.	.83 (.86)	.61 (.54)	.70 (.66)

Table 3: CtxMatch (and baseline) results on the mapping of Google and Yahoo! 'Architecture' and Google and Yahoo! 'Medicine'.

0 and 1.

$$(11)\, MSP(A,B) = \begin{cases} P(A|B) & if P(B|A) = 1 \\ 0 & otherwise \end{cases}$$

The results of the experiment are reported in Table 3, in terms of precision, recall, and F-measure obtained for the mapping relations returned by CtxMatch. A baseline for the experiment was defined by considering a simple string match comparison among the labels placed on the path spanning from a concept to its root in the CH (the results of the baseline are reported in bracket). The results show that both the baseline and the CtxMatch algorithm perform quite well. Not surprisingly, the baseline reveals itself as very precise, while CtxMatch outperforms it with respect to recall. This confirms an important strength of CtxMatch, namely that a content-based interpretation of contextual knowledge allows the discovery of non-trivial mappings. As an example, the equivalence between *Pharmacology/Psychopharmacology/Psychiatry* and *Psychiatry/Psychopharmacology* is found thanks to the WordNet hyponymy relation between *Pharmacology* and *Psychopharmacology*. A mapping of inclusion (source concept is less general than target concept) between *History/Periods_and_Styles/Gothic/Gargoyles* and *History/Medieval* is computed thanks to the WordNet relations between *Medieval* and *Gothic*.

5 Conclusions and Future Work

In this paper we have faced the linguistic processing of Classification Hierarchies, a task which is receiving an increasing interest in view of Semantic Web applications. Two main directions have been addressed: (i) the linguistic processing required for the semantic interpretation of CH labels, and (ii) the design of an evaluation methodology. We have presented CtxMatch, an algorithm that discovers mappings among overlapping CHs through a semantic interpretation of the labels. Although this work represents a first step within a long term plan and the results we obtained are subject to improvements, they can be considered as a benchmark for future work in this area. A preliminary conclusion is that the employment of linguistic tools and resources is crucial for this task. In the future we plan to refine the evaluation methodology, e.g. by applying it to CHs with different features, such as marketplace catalogs.

References

R. Agrawal and R. Srikant. 2001. On integrating catalogs. In *Proc. of WWW-2001*, Hong Kong, China, May.

F. Baader and W. Nutt. 2002. *Description Logic Handbook.* Pages 47-100, Cambridge University Press.

P. Bouquet, L. Serafini, and S. Zanobini. 2003. Semantic coordination: A new approach and a application. In *Proc. of ISWC-03*, Sanibel Island, Florida, USA, October.

T. Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proc. of ANLP-2000, 6th Applied NLP Conference*, Seattle, April-May.

J. Daude, L. Padro, and G. Rigau. 2000. Mapping wordnets using structural information. In *Proc. of ACL-2000*, Hong Kong, October.

A. Doan, J. Madhavan, P. Domingos, and A. Halevy. 2002. Learning to map between ontologies on the semantic web. In *Proc. of WWW-2002*, May, Honolulu, Hawaii.

Google. 2003. http://directory.google.com/.

R. Ichise, H. Takeda, and S. Honiden. 2003. Integrating multiple internet directories by instance-based learning. In *Proc. of IJCAI-2003*, Acapulco, Mexico, August.

J. Madhavan, P. Bernstein, P. Domingos, and A. Halevy. 2002. Representing and reasoning about mappings between domain models. In *Proc. of AAAI-2002*, Edmonton, Alberta.

B. Magnini, L. Serafini, and M. Speranza. 2003. Making explicit the semantics hidden in schema models. In *Proc. of the Workshop on 'HLT for the Semantic Web and Web Services', ISWC-2003*, Sanibel Island, Florida.

P. Rigau, B. Magnini, E. Agirre, P. Vossen, and J. Carrol. 2002. Meaning: A roadmap to knowledge technologies. In *Proc. of the workshop 'A Roadmap for Computational Linguistics', COLING-2002*, Taipei, Taiwan.

Yahoo! 2003. http://uk.yahoo.com/.