Integrating ontological and linguistic knowledge for Conceptual Information Extraction

Roberto Basili, Michele Vindigni, Fabio Massimo Zanzotto
Department of Computer Science
University of *Roma, Tor Vergata*, Roma (Italy)
{basili,vindigni,zanzotto}@info.uniroma2.it

Abstract

Text understanding makes strong assumptions about the conceptualisation of the underlying knowledge domain. This mediates between the accomplishment of the specific task at the one hand and the knowledge expressed in the target text fragments at the other. However, building domain conceptualisations from scratch is a very complex and time-consuming task. Traditionally, the re-use of available domain resources, although not constituting always the best, has been applied as an accurate and cost effective solution.

In this paper, we investigate the possibility of exploiting sources of domain knowledge (e.g. a subject reference system) to build a linguistically motivated domain concept hierarchy. The limitation connected with the use of domain taxonomies as ontological resources will be firstly discussed in the specific light of IE, i.e. for supporting linguistic inference. We then define a method for integrating the taxonomical domain knowledge and a general-purpose lexical knowledge base, like WordNet. A case study, i.e. the integration of the MeSH, Medical Subject Headings, and WordNet, will be then presented as a proof of the effectiveness and accuracy of the overall approach.

1. Introduction

In text understanding processes, such as the one underlying Information Extraction (IE) or Question Answering (QA) systems, strong assumptions are made on the conceptualization of the knowledge domain. An explicit representation of key domain concepts and relationships helps in explaining the mapping between the specific task (e.g. event matching in IE) and the analysed text fragments. When domain concept hierarchies are available more principled information extraction patterns may be written or, in a complementary fashion, induced from textual collections. Moreover, specific subtasks (e.g. the resolutions of anaphoric references) can rely on simpler models with clearer linguistic explanations. For example, the evaluation in [7] suggests that richer semantic representation in IE may result in more accurate co-reference resolution (see the IE system described in [5]).

Concept hierarchies are very expensive resources. Lexical databases such as WordNet [6] or Euro-WordNet [9] are currently widely used in NLP applications (e.g. in Question Answering [4] or in automatic hyperlinking [2]). However, they required huge efforts and large investments. Moreover, in light of the narrow domains sought by IE applications, these resources are overly general and may even amplify dangerous phenomena, e.g. semantic ambiguity. In information extraction, domain and task specific approaches (e.g. shallow and fully lexicalised IE patterns) seem to perform better than deeper ones based on weaker conceptualizations. The quality of the available domain conceptualization is a key issue for the accuracy of the underlying NLP task.

Building domain conceptualizations from scratch is a very complex and time-consuming task. Traditionally, the re-use of available domain resources, although not constituting always the best, has been applied as an accurate and cost effective solution. Pre-existing resources such as domain ontologies or topical taxonomies are in general not suited for linguistic tasks. There is in fact no clear separation between concepts, their lexical realization (i.e. category names as referential expressions) and their conceptual properties. For example, text classification schemes, such as the Medical Subject Headings (MeSH) or the IPTC Subject Reference System[1], provide a taxonomic organization of bodies of knowledge made explicit via linguistic definitions, i.e. labels of the defined categories like $Tissues$ in MeSH. Topic labels are here used to denote complex domain concepts while the hierarchical structure suggest taxonomic relationships among concepts.

[1] Details can be found respectively in www.nlm.nih.gov/mesh/meshhome.html and in www.iptc.org

However, the use of these subject reference systems as domain conceptualisations is not as straightforward as it is too often assumed. This is particularly true when these latter have a role in the interpretations of textual material (e.g. in IE). These reference systems are in fact devoted to hierarchically organise documents in classes and the referential properties of class labels are very complex. Labels denote here not just one concept, but rather a set (or better a system) of world concepts that enter into a given topic, i.e. a phenomenon, discussed by a class of documents. This has almost nothing to do with linguistic denotations and inferences used to explain or predict natural language structures within the actual documents. Other knowledge organisations (e.g. general purpose lexical databases such as WordNet) derive from a fully different design and are better suited to deal with language understanding (e.g. disambiguation phenomena).

Consider the MeSH topic taxonomy. This is unsatisfactory mainly for two reasons: firstly, the nature of the broader/narrower relation is not clear and, secondly, the knowledge embodied is not linguistically principled. It has not in general a direct explanation in terms of language constituents so nodes do not work as selectional primitives for activities such as co-reference resolution. As an example, consider the term *dendrite* appearing in three points of the MESH hierarchy, i.e. "*Dendrites → Neurons → Nervous System*", "*Dendrites → Cell Surface Extensions → Cellular Structures → Cells*", and "*Dendrites → Neurons → Cells*". It may be observed that the nature of the arrows changes between *part_of* and *is_a*, e.g. "*Dendrites is_a Cell Surface Extensions*" or "*Dendrites part_of Neurons*". As a result, these classifications of the word *dendrite* do not help in predicting and, therefore, interpreting text fragments as:

None of the dendrites were cut.
...but the dendrites and axons are often cut.
Researchers don't know why, but for some reason, the ends of dendrites tangle and knot.

where the fact that dendrites may be cut, tangled, or knotted strictly depends on the properties they inherit being *fibres*. This is better represented in WordNet where the generalisation chain "*dendrite → nerve fiber, nerve fibre → fiber, fibre*" is postulated. Beside the considerations about the quality of WordNet as a valid semantic model for the medical knowledge and terminology, its psychologically principled organization better capture the meanings expressed through language. Such a conceptualisation is better suited for writing syntactic-semantic interfaces since selectional preferences for prototypical text fragment can be more expressively defined. An integration of the two resources is then desirable.

In this paper, we want therefore to investigate the exploitation of domain knowledge (e.g. a subject reference system) in the design of a linguistically motivated domain concept hierarchy. We then define a method for integrating the taxonomical domain knowledge and a general purpose lexical knowledge base, like WordNet (Sec. 2). A case study, i.e. the integration of the MeSH, Medical Subject Headings, and WordNet, will be then presented as a proof of the effectiveness and accuracy of the overall approach (Sec. 3).

2. Building a Semantic Dictionary

The process of building a semantic dictionary aims to detect of a suitable subset of semantic primitives able to represent promising and effective generalizations of linguistic expressions in the domain. The domain specific resource here is a topic taxonomy, hereafter referred to as *domain concept hierarchy* (**DCH**). The need for linguistically consistent knowledge requires the availability of language oriented "is_a" hierarchies to model and explain textual phenomena in the domain. We will hereafter refer this latter knowledge as the domain-independent lexical knowledge base **LKB**. An example of such information could be the hyponomy/hyperonimy taxonomy in WordNet [6]. Such an overall framework requires a suitable mapping strategy between LKB and DCH. Notice how this mapping deals with a general many-to-many correspondence between the DCH ontological primitives (like *Tissue* in MeSH) and the LKB word senses. For example, the word "*tissue*" has two senses in WordNet while corresponding to just one MeSH topic category.

In the rest of the section we need to discuss concepts (in DCH), word senses (in LKB) and several significant implications. We hereafter introduce more formally a set of useful definitions. Any concept C in the domain hierarchy (DCH) is characterized by its linguistic label hereafter noted as t_C. This label t corresponds either to a singleton word or to a multiword expression. This information can be used as a reference within the lexical knowledge base LKB. We will denote LKB entries by means of Greek letters, e.g. α. Those LKB senses that correspond to possible linguistic meanings of label t will be denoted as α_t. Sometimes α_t may not exist for technical concepts as they are not present in the domain independent LKB [2]. In general, a label t will correspond to more than one sense.

As DCH and LKB have both an internal structure some other useful properties can be introduced. First of all we will call *linguistic extension* of a DCH concept C, denoting

[2]Notice that in this case we could relax the search of the multiword expression, e.g. *Common Hepatic Duct*, and try to match senses for subexpressions obtained by neglecting some modifier, e.g. *Hepatic Duct*. The longest expressions corresponding to one LKB entry would be retained as a possible linguistic interpretation.

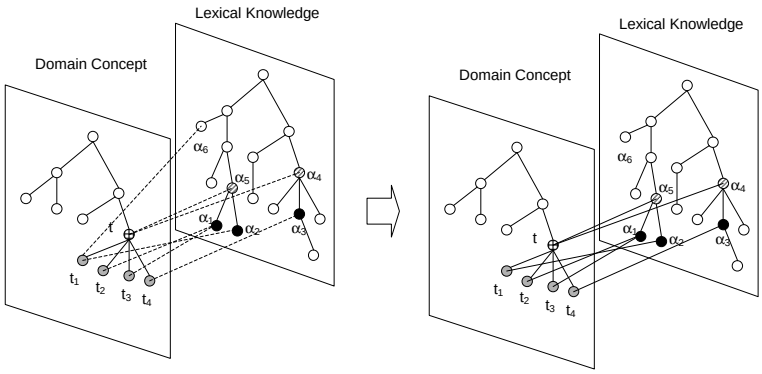

Figure 1. Integration of domain and generic knowledge: WordNet and MeSH

it as $ext(C)$ the set of the labels for C or for one of its descendants C' as follows:

$$ext(C) = \{t_{C'} | C \text{ subsumes } C' \text{ according to } DCH\} \quad (1)$$

For example the linguistic extension of *Tissues* in MeSH includes words and terms like "*Articular Cartilage*", "*Corneal Endothelium*".

Given its extension, a DCH concept C can be interpreted in LKB via its *linguistic generalization set*, that is the set of generalizations, α_t, in LKB for the labels $t \in ext(C)$. It will be denoted by $lgen(C)$ that is defined as

$$lgen(C) = \{\alpha \in LKB | \exists t \in ext(C) \text{ and } \alpha_t \text{ is subsumed by } \alpha \text{ in } LKB\} \quad (2)$$

Due to language ambiguity the generalization set $lgen(C)$ includes more senses in LKB than those needed for represent C as each sense α_t for a given $t \in ext(C)$ is retained. In the next two sections the model to constraint generalizations in LKB by means of DCH information will be defined aiming to reduce the overall ambiguity and detect the correct LKB sense assignments(s) to DCH elements.

2.1. Inspiring principles

One of the aims of the proposed integration is to constraint the search for word sense assignment (i.e. navigation in the LKB) through information provided by the domain resource. Vice versa the LKB structure will be used to bias the search of DCH meanings, i.e. explain linguistically the nature of the DCH primitives: for example *Cardiovascular System* has just one sense in WordNet, under the "*body_part*" sub-hierarchy; however, as a MeSH topics, it is also related to functionalities and physiological processes not coded as "*body_part*"s.

Cross-corresponding concepts between a DCH and a lexical model LKB can be detected by exploiting in combination both constraints. We will rely on the following two principles:

(**P1**) (*Extensional Nature of DCH*). Given a domain concept hierarchy DCH, whatever the nature of its basic unit is, *subsumption throughout the hierarchy has always an extensional interpretation*, i.e. for each couple of concepts C' and C'' subsumed by a common ancestor C in DCH, there is always a linguistically consistent concept $\alpha \in LKB$ such that the linguistic expressions $t' = t_{C'}$ and $t'' = t_{C''}$ have senses $\alpha_{t'}$ and $\alpha_{t''}$ both subsumed by α in LKB[3].

(**P2**) (*Intentional strength in LKB*). A set of linguistic denotations $W = \{w_i\}$ [4] whose senses are all subsumed by a given $\alpha \in LKB$ has an *intentional strength* for W that is a function of the senses of w_i and of their distribution in the LKB sub-hierarchy dominated by α. α represents the trade-off between the generalization required to represent all the denotations w_i and their specialization, i.e. the capability of separating the individual different senses of the w_i's. Any monotonic non-decreasing function of such a trade-off is a valid

[3] The extensional interpretation α may not be unique. In fact, given a bipartite set of C descendants $\{C'_1, ..., C'_n, C''_1, ...C''_m\}$, then two (or more) concepts may exist, $\alpha' \neq \alpha''$, such that $\forall i = 1, ..., n \quad t_{C'_i}$ generalizes in α' and $\forall j = 1, .., m \quad t_{C''_j}$ generalizes in α''

[4] The use of w_i here emphasizes the difference with respect to the previously adopted notion of t_i. w_i are linguistic symbols that independently from any domain are referential in the world. t_i are terminological labels of DCH concepts and their semantics is NOT exhaustively determined on a linguistic ground. Principle **P1** focuses on the interpretation of domain symbols t_i by means of the DCH hierarchy. As **P2** focuses on purely linguistic information determined by LKB, a different notation is required.

Table 1. MeSH Headings mapped in WordNet synsets

MeSH	Category	WordNet Synset	CD Score	Term Coverage
A10	Tissues	body_part	981980	85%
A10	Tissues	epithelium	359.066	3%
A10	Tissues	scar cicatrix cicatrice	0.971	2%
A10	Tissues	object physical_object	0.575	7%
A10	Tissues	cell	0.085547	3%
B02	Vertebrates	mammal_family	1.06E+19	6%
B02	Vertebrates	taxonomic_group taxon	1.74E+18	31%
B02	Vertebrates	vertebrate craniate	7.68E+14	60%
B02	Vertebrates	life_form organism being living_thing	8.07E+14	2%
B02	Vertebrates	object physical_object	291.277	1%
C11	Eye Diseases	symptom	3264.270	27%
C11	Eye Diseases	visual_impairment visual_defect vision_defect	722.276	16%
C11	Eye Diseases	condition status	149.956	34%
C11	Eye Diseases	cognition knowledge	246.031	5%
C11	Eye Diseases	obstruction	0.946	2%
C11	Eye Diseases	hole	0.828	3%
C11	Eye Diseases	happening occurrence natural_event	0.535	8%
C11	Eye Diseases	membrane tissue_layer	0.312	5%
C11	Eye Diseases	physical_phenomenon	0.019	2%
D07	Reproductive Control Agents	contraceptive preventive preventative contraceptive_device prophylactic_device birth_control_device	579.461	50%
D07	Reproductive Control Agents	hormone internal_secretion	0.157	50%

measure of the intentional strength of α with respect to words w_i.

Notice that DCH nodes C are in a many-to-many mapping to LKB senses. As a consequence sets $W = ext(C)$ may not receive a unique $\alpha \in LKB$ but are usually covered by more than one generalization (i.e. there is not α that is a common ancestor for all $t \in W$, implying that the intentional strength is 0). In this case an alternative can be found by partitioning W in more coherent (and possibly overlapping) subsets W_i. These will give independently rise to common generalizations, α_i: each one is a trade-off as the higher in the hierarchy is α_i, the larger is the size of the corresponding W_i.

2.2. Mapping domain concepts to lexical senses

The semantic dictionary that will result from the harmonization of DCH and LKB is a lexical hierarchy extended with the domain concepts (see Fig. 1), as metafeatures. In fact, each useful sense α in LKB will be augmented with references (e.g. the labels t) to domain concepts C linguistically interpretable as α. In terms of the properties **P1** and **P2** domain concepts are mapped to word senses in the following way. A domain concept C receives the minimal set of word senses in $lgen(C)$ with the *maximal intentional strength* as subsumers of non-trivial subsets of the linguistic extension of C, i.e. $ext(C)$. Usually specific entries in DCH (e.g. *Tissues*) are mapped into one ore more LKB senses (e.g. *'body_part'* and *'epithelium'* in WordNet). Vice versa one sense may be tagged by several DCH primitives (e.g. *'body_part'* as *'Digestive System'*, *'Cardiovascular System'*, *'Tissues'*, ...).

In our model the notion of *conceptual density* (cd), as introduced by [1], is used as a measure of intentional strength (principle **P2**). The conceptual density aims to state why and how much a set of words may be considered similar according to a reference lexical hierarchy, LKB[5]. Given a set W of words (eventually with multiple senses) and a specific node α in the lexical hierarchy dominating at least one sense for each $w \in W$, the conceptual density $cd(W, \alpha)$ is a real value associated to the common ancestor α: it is proportional to the number of covered senses of $w \in W$ and inversely proportional to the size of sub-tree rooted at α. Therefore, the smaller the sub-tree (i.e. the more specific is α as a generalization of the senses of w's), the higher is the cd value. Although its application in the ontology engineering framework proposed in this paper is new, this measure has been widely applied to word sense disambiguation problems: technical details are also discussed in [1].

The model we propose here requires that a triggering set T of DCH concepts C (with category labels t_C) has been previously selected. This set will drive the application of

[5] WordNet has been used as the underlying reference taxonomy for the definitions and experiments related to the conceptual density, [1].

the principle **P1** and **P2** over the DCH. Then, for each concept $C \in T$, the corresponding set of linguistic expressions ($ext(C)$ in Eq. (1)) is determined by DCH. $ext(C)$ and the conceptual density are then used to derive an *optimal* set of LKB senses within the linguistic generalizations of C (i.e. $lgen(C)$ in Eq. (2)): this set is optimal as it is made of the intentionally strongest senses α_i that generalize all the expressions of $t \in ext(C)$. By means of a greedy technique, the generalizations α_i of non-trivial subsets $W_i \subset ext(C)$ are selected according to decreasing values of conceptual density until the entire set is not completely covered. In this way, each $C \in DCH$ is mapped to an $\alpha_C \in lgen(C)$ characterized by the highest intentional strength (i.e. $cd()$).

The algorithm that maps the DCH into the LKB is triggered by the subset of concepts T and is sketched in the following. Its discussion will make reference to the example in Fig. 1.

procedure merge(DCH,LKB,T)
for each $C \in T$
(Step 1) Determine the linguistic extensions $lgen(C)$
in DCH made of all descendants of C
(Step 2) Compute the optimal mapping $G(C) \subset lgen(C)$,
by a greedy selection that maximizes conceptual density
(Step 3) Attach t_C to senses in $G(C)$
(Step 4) **for each** $t \in ext(C)$
Attach t to $\alpha \in LKB$ iff:
α is a sense for t in LKB and
$\exists \beta \in G(C) | \beta$ subsumes α in LKB

The subset T of the domain concepts in DCH is therefore an input parameter. For example, the top levels of DCH can be retained as T[6]. Then, for each $C \in T$ the process depicted in fig. 1 is carried out: first linguistic expressions $t_i \in ext(C)$ of C descendants in the DCH hierarchy are determined in (Step 1), e.g. $ext(C) = \{t_1, ..., t_4\}$ in fig. 1. Linguistic descriptions t_i are analysed against the lexical semantic hierarchy LKB. Different subsets are derived $W_1 = \{t_1\}, W_2 = \{t_2, t_3\}$ and $W_3 = \{t_4\}$ as they receive different interpretations, i.e. activate senses $\alpha_1, ..., \alpha_6$. All elements t_i are "somehow" more specific of T_i. (Step 2) allows to select the optimal generalizations $G(T_i)$ as word senses. These are the valid generalizations of subsets of $ext(C)$ having the higher cd and covering the entire set $ext(C)$. In the example, $G(C) = \{\alpha_5, \alpha_4\}$ as they are enough general to represent all t_i and enough specific to refuse some senses. The DCH concept C is thus used to annotate senses α_5 and α_4 (Step 3). Finally, the linguistic labels of DCH concepts are attached to the related senses in the LKB (Step 4). It is easy to see that this information reduces the ambiguity of t_j. For instance, the interpretation α_6 of t_1 is discarded: its conceptual density is too low and other senses are sufficient to cover the entire set $\{t_1, ..., t_4\}$.

The resulting of the above process is a *Concept Hierarchy* that integrates denotations of domain concepts with their linguistic counterparts: the former will support disambiguation in language processing, while the latter will favour linguistically consistent generalizations of general (i.e. non domain-specific) surface forms.

As previously noted, some labels in the DCH are not represented in the LKB: they are possibly too much specific and are uncovered by LKB. As an example, *Common Hepatic Duct* of MeSH has no counterpart in WordNet, although the sub-term *Hepatic Duct* has a unique sense. Partial forms of terms $t \in ext(C)$ are also used in the above mapping algorithm as they bring useful information for determining suitable interpretations (α_C) and their conceptual density. As the term head is the semantic carrier of multiword expressions, terms not covered in LKB are processed by backing off to the sub-terms obtained via incremental elicitation of modifiers: e.g. $w_1 w_2 ... w_n$ is reduced to the longest covered sub-term $w_i ... w_n$ that has a sense in the LKB. Unfortunately this approximation may introduce noise in the mapping process as sub-terms are usually more polysemic than complete terms.

3. Mapping MeSH to WordNet: a case study

To investigate the features of the proposed method, we applied it for mapping the Medical Subject Headings (MeSH) in WordNet. MeSH has been therefore partitioned according to its first level, i.e. the 111 main index categories. This subset of concepts referred as T in the algorithm description triggers the interpretation of the MeSH concepts in WordNet.

The first observation is that only the 24% of MeSH terms are fully represented by WordNet, while domain specific complex terms are almost absent. As discussed in Sec. 2.2, we allowed for degraded interpretations of these complex terms, by discarding modifiers in hierarchical order until a correspondence with WordNet is found. For instance, trying to assign WordNet interpretation to *cancerogenic blood cell*, as this term is unknown to WordNet, we drop the more external modifier *cancerogenic* and repeat the test with *blood cell*. Table 2 summarizes the results of this activity: note that MeSH terms that have a direct interpretation in WordNet are generally unambiguous (polysemy=1.2), while terms that result from pruning modifiers become less specialised (polysemy=2.74). In fact, for most of these terms, we are selecting a hyperonym as term representative. This loss of selective information will also affect the clusters inspected by the method we propose, as it will be applied to less specialized senses and, therefore, resulting senses will be more likely selected in the upper part of

[6] A limited semantic dictionary for which wide extensional evidence is available can improve the mapping accuracy

Table 2. MeSH vs WordNet relevant features

MeSH Terms	20603
MeSH Categories	37864
1st Level MeSH Headings	111
MeSH Terms in WordNet	4960
Partial MeSH Terms in WordNet	9345
Unrepresented MeSH Terms	6298
Partial Terms Average Polysemy	2.74
Full Terms average polysemy	1.2

the WordNet hierarchy.

As expected, the *cd* combined with the exploitation of the properties **P1** and **P2** enables to evaluate semantic cohesion of the candidates in the DCH, allowing to focus over *particularly dense* regions and to select good senses in WordNet.

For instance, if all the terms covered by *Cardiovascular System* in MeSH were taken alone, several possible generalizations would have been admitted. A trivial set of generalisations would be the union of all the topmost WordNet concepts of the activated senses. Properties **P1** and **P2** with the *cd* allow evaluating how well the different senses are representative of the knowledge underlying *Cardiovascular system* (i.e. generalize this knowledge while staying sufficiently specific). As a result, we obtain in this case only a small number of interesting senses: *body_part* (with $cd = 104.60$) covering 67% of the original material, *object physical_object* (with $cd = 0.57$), covering a remaining 20%, and a tail of other senses cumulating a 0.001 score for the rest (13%). These spurious interpretations are manly due to noise deriving either by WordNet polysemy, or by terms that are only partially represented in WordNet.

The fact that medical terminology is only partially represented gives a further evidence of the distance existing between this domain and the language. Most of the terms convey an implicit meaning that is only accessible through strong background knowledge. In a significant number of cases, the hypothesis to use term heads as an approximation of complex nominals could result in a strong noise source, as sometime syntactic heads alone fail to convey the intended meaning. After applying the proposed method, we obtain a set of WordNet senses that represents, from a Machine Learning perspective, an extensional linguistic definition of each MeSH category in terms of WordNet concepts. More examples of the resulting mapping are shown in Table 1.

It is worth noticing that the score synsets receive by Conceptual Density allows to drastically reduce the impact of polysemy as in the estimation of the cluster cohesion sparse senses (i.e. spurious interpretations suggested by the resulting topology) are filtered out receiving lower scores in weighting with respect to their frequencies. Thus for instance, even if 7% of terms of *Tissues* (A07) (see row 4) category are best generalized with *object physical object* synset, while only 3% by *epithelium* (see row 2), this last receives a greater score, as the internal cohesion of its terms is stronger. Moreover, complex terms missing in WordNet, as for instance *Gastric Chief Cell* whose only the head *cell* is found, even if strongly ambiguous (*cell* has 6 WordNet senses), become unambiguously assigned to specific senses (*body part* in the example).

Finally, the chosen a set of primitive types in WordNet allows covering and explaining novel situations, bearing as a side effect the rest of the hierarchy into the analysis. By generalizing lexical phenomena in the corpus with respect to this model we are now able to find a linguistic definition for more words in text analysis.

3.1. Ontology Engineering based on the created semantic dictionary

The previous section has described the derivation of linguistic explanations for medical concepts. MeSH topics have thus been mapped into a dictionary of Wordnet senses in a many-to-many mapping. This knowledge DCH+LKB resource is involved in several tasks. The first is *text analysis* as topic labels t mapped into Wordnet senses support disambiguation in sentence understanding. MeSH is thus translated into a large-scale terminological resource for any NLP process insisting on a medical corpus. *Semantic indexing* is also enforced as topics labels t have now interpretations $\alpha_t \in LKB$. These latter can drive the interpretation of text portions and suggests for them domain labels t. These are ontological indexes by which texts are mapped to domain primitives.

An important activity, that in fact includes also the already mentioned ones, is *Ontology engineering*. Here processes of *domain-specific lexical learning* can be run on texts and used within a conceptual information extraction framework. From one side, the extracted knowledge can be interpreted linguistically (according to the LKB) and its ontological counterpart (supported by DCH+LKB) may be used to populate/refine the domain knowledge (i.e. DCH).

In a perspective of lexical learning, the following inductive phases are useful to the ontology engineering enterprise: (1) acquisition of a domain terminology to integrate/extend the DCH; (2) acquisition of linguistic patterns that typically express concepts and relations in the domain (e.g. relationships among new medicines towards pathologies); (3) generalization of the detected patterns in new ontological relations or in instances of known relations.

The resource built with the method described in this paper can support all the above phases. When new terminology is available (as in phase (1) above) it can be linguis-

tically interpreted according to syntagmatic and semantic principles of LKB. Traversing the LKB hierarchy is used to find interpretations, α_t. Mapping towards concepts C is then used to locate terminological information in DCH. In phase (2) patterns can be acquired as linguistic structures (e.g. sub-graphs) connecting systematic domain phenomena (e.g. named entities and/or terminological expressions acquired in (1), see e.g. [3, 8]). Linguistic patterns made of domain concept labels t_C can be here built while consistent generalization are allowed via traversing the DCH as well as the LKB hierarchy. The material observed in the corpus can here be generalized (or refused as noise) when validated by DCH and/or LKB. This evidence improves the learning accuracy and can be effectively used to build the ontological interpretations (i.e. relations in DCH) from the linguistic elements of these patterns (e.g. constituents in grammatical structures). Classes of generalized relations can be then mapped back to DCH to extend existing relations or populate DCH of new relation instances.

4. Conclusions

Domain knowledge for semantic interpretation is a relevant source of information. However, the integration of domain specific resources within a text processing task is not straightforward as available primitives have an unclear semantic status. In this paper a method to harmonise a domain concept hierarchy with a lexical knowledge base has been defined. The method tries to keep separate the information provided by a taxonomic organization of concepts and the linguistic counterpart. Linguistic information here first seen as an extensional definition (i.e. an explanation) of domain concepts through the hypothesis (i.e. their descendants) provided by the taxonomy. Then a measure of the representativity of each linguistic interpretation (sense) is proposed as a function of the concept labels as well as of the lexical hierarchy. Finally, an augmented lexical knowledge base is released as a semantic network annotated by domain concepts. Several linguistic inferences are discussed and can be improved by such an extended resource. The results obtained by the application of the proposed method within a medical knowledge domain are more than promising. A significant reduction of the average ambiguity in the interpretation of domain labels uncovered by the lexical knowledge base is a first achievement. The interpretation of term labels for newly discovered terms and the potentials opened for the correct interpretation of textual phenomena are two further benefits. More in depth analysis of the impact of the method within a knowledge based information extraction system is still needed. More work is necessary to assess the consistency of the method hypothesis within domains different from the medical one as well as to reproduce the accurate results obtained in this first experiments. Implications of the above procedure in the semantic interoperability problems within Web applications will be the target of consistent research in the near future.

References

[1] E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*, 1996.

[2] R. Basili, R. Catizone, L. Padro, M. T. Pazienza, G. Rigau, A. Setzer, N. Webb, Y. Wilks, and F. M. Zanzotto. Multilingual authoring: the namic approach. In *Proceedings of the WORKSHOP ON HUMAN LANGUAGE TECHNOLOGY AND KNOWLEDGE MANAGEMENT, held jointly with ACL'2001 Conference*, 2001.

[3] R. Basili, M. T. Pazienza, and M. Vindigni. Corpus-driven learning of event recognition rules. In *Proceedings of the Workshop on Machine Learning for Information Extraction, held jointly with ECAI 2000*, Berlin, Germany, 2000.

[4] S. Harabagiu, D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, and P. Morarescu. Falcon: Boosting knowledge for answer engines. In *Proceedings of the Text Retrieval Conference (TREC-9)*, 2000.

[5] K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of sheffield: Description of the LASIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conferences (MUC-7)*. Morgan Kaufman, 1998.

[6] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, Nov. 1995.

[7] MUC-7. Proceedings of the seventh message understanding conference(MUC-7). In *Columbia, MD*. Morgan Kaufmann, 1997.

[8] E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, Portland, Oregon, 1996.

[9] P. Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.