Integrating Ontological and Linguistic Knowledge for E-Learning

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

Abstract

In computer-aided and Web-based learning a critical role is played by the methods and frameworks for Information Access. First of all, intelligent methods for retrieval and synthesis of information are crucial for making available timely and correct information during the training process. Moreover, the level of interaction between the target user and the retrieval subsystem constitutes an important quality factor for the learning process. The abstraction supported by the linguistic generalization adopted in the interaction is an inherent component of the student development: the higher is the linguistic level, the faster and more effective is the training. In order to ensure the proper abstraction level, the retrieval component should make use of capabilities for generalization throughout a number of phases: indexing, retrieval, organisation and presentation. All these tasks require thus an underlying concept-oriented approach usually relying on ontological resources. In particular tasks like indexing and presentation are also faced (in both directions of input/recognition and output/production) with linguistic data: source texts and dialogue/interaction sessions, respectively. In both cases, i.e. text understading and natural language generation, a non trivial process of semantic recognition is involved. All the above implies that strong assumptions about the conceptualisation of the underlying knowledge domain are usually made in e-learnig. However, building domain conceptualisations from scratch is a very complex and timeconsuming task. Traditionally, the reuse of available domain resources, although not constituting always the best, has been applied as an accurate and cost effective solution. This paper presents a method to exploit sources of domain knowledge (e.g. a subject reference system as a controlled language for document indexing and classification) used to build a linguistically motivated domain concept hierarchy. Because in the specific perspective of supporting linguistic inference in Information Extraction and Retrieval (IE/IR) for e-learning, the use of domain taxonomies as ontological resources is not straightforward. We discuss here how a method for integrating the taxonomical domain knowledge and a general-purpose lexical knowledge base (like WordNet) can be used for improving the accuracy and flexibility of IE.

1 Introduction

In the text understanding process, such as the one underlying Information Extraction (IE) or Question Answering (QA) systems, strong assumptions on the conceptualisation of the knowledge domain are made. The explicit representation of the 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 [8] suggests that richer semantic representation in IE may result in more accurate co-reference resolution (see the IE system described in [6]).

On the other hand, concept hierarchies are very expensive resources. Lexical databases such as WordNet [7] are currently widely used in NLP applications (e.g. in Question Answering [4] or in automatic hyperlinking [1]). However, they required huge efforts and large investments. Moreover, in light of the limited 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 pattern) seem better performing than deeper ones based on weaker conceptualisations. The quality of the available domain conceptualistion is a key issue for the accuracy of the underlying NLPtask.

Building domain conceptualisations from scratch is a very complex and time-consuming task. Traditionally, there-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 an 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. 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).

In [2], we investigated the exploitation of domain knowledge (e.g. a subject reference system) in the design of a linguistically motivated domain concept hierarchy. The limitation connected with the use of domain taxonomies as ontological resources will be here discussed in the specific light of IE (Sec. 2). The method for integrating the taxonomical domain knowledge and a general purpose lexical knowledge base, like WordNet (Sec. **??**) can be here applied to e-learning. A case study, i.e. the integration of the MeSH, Medical Subject Headings, andWordNet, is presented in [2] as a proof of the effectiveness and accuracy of the overall approach.

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

2 Topic taxonomies and linguistic knowledge for Information Extraction

In a variety of applications, topic taxonomies are often seen as conceptual vocabularies for text processing and retrieval. When forms of text understanding are required (as in IE), the kind of needed semantic information often relates to a linguistically principled concept hierarchy. This latter may have many usages. It is often used as the dictionary for describing selectional restrictions for the syntax-semantics mapping. For instance, when information related to drug and diseases is the target, a functional relation like *treat(drug,disease)* may be reflected in a matching rule such as the following:

treat(SUBJ(**human**),
OBJ(**human**,MOD(with,**disease**))
MOD(with,**drug**))

It can be thus used to match relevant information from a text fragment like: *We treated 17 patients with cutaneous sarcoidal granulomas with hydroxychloroquine (2 to 3 mg/kg/day) in an open clinical trial.*
Here *hydroxychloroquine* **treats** *cutaneous sarcoidal granulomas* (and someone has tried to apply this remedy) is obtained if some generalizations are supported by the domain conceptualisation: *hydroxychloroquine*, *cutaneous sarcoidal granulomas* and *patients* are kind of **drug**s, **disease**s and **human**s respectively. Other text related functionalities are supported by domain concept hierarchies. Semantic dictionaries can drive more informed rules for co-reference resolution (as suggested in [6]). This is another critical activity in IE. For instance the relation
lag screw fixation ***treats*** *scaphoid nonunion*
can be suggested by the following text fragment:
We report our experience in 42 patients, using lag screw fixation for un-united scaphoid fractures. ... We recommend the operation for the treatment of scaphoid nonunion, ... (1)
only if the co-reference link between *lag screw fixation* and *operation* is decided. In medical texts, such a kind of coreferences (i.e. the nominal anaphors) is very frequent [3]. In the example possible coreferent of *operation* are *experience*, *patient*, *lag screw fixation* and *un-united scaphoid fractures*. Notice that the conceptual hierarchy can be here very useful. Legal candidates could be limited to those having a common conceptualisation (i.e. shared ancestor in the concept hierarchy) with the target (i.e. *operation*). In the example, WordNet suggests *activity* as a common generalization of *operation* and *lag screw fixation*. A semantic preference can be thus given to the latter with respect to the other candidates for which no common ancestor (or only a weaker one) can be found. For triggering the above inference a sense disambiguation step has been undertaken, as one out the 10 possible senses for the word *operation* has been selected initially. It is thus to consider that a domain specific use of a general-purpose semantic model likeWord- Net has always problematic aspects. However, the contribution of domain information serves straightforwardly the purpose of reducing the overall word sense ambiguity and is a promising way to harmonisation.

Domain topic taxonomies although interesting repositories of denotations for domain concept differs from lexical knowledge bases. These semantic thesauri are in fact controlled vocabularies (i.e. indexes) as categorization systems for textual databases. The meaning of a thesaurus entry has not the suitable referential properties required in linguistic interpretation: it has not a direct reference in the world. It rather denotes a class of referred entities

involved in a complex process or topics. For example, *Tissue* is a category in MeSH and *follicular dendritic cell* is one of its descendants. Now, it is not quite true that *Tissue* denotes "*an aggregate of cells having a similar structure and function*" (as WordNet states). It rather serves in MeSH as an index for medical articles that deals, among other things, with tissues. Moreover, it is also false that *follicular dendritic cell* is a "kind of" *tissue*, as it is rather a *cell* and not a *tissue* in general.

Considering therefore a topic taxonomy such as MeSH as the only concept hierarchy is unsatisfactory for several reasons. First, the knowledge embodied in MeSH 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 example, the *Cardiovascular System* sub-tree in MeSH (rooted at category A07) is mainly partitioned in the *Blood vessels* and *Heart* classes. Under the *Heart* sub-hierarchy, very different concepts as *Heart Atrium*, *Fetal Heart*, or *Heart Conduction System* can be found. All these three concepts have very different meanings and will appear in different linguistic context with different roles. For instance, a *Fetal Heart* will be probably affected by diseases that are quite different from the ones affecting the adult hearth. Moreover, the notion of *Heart Conduction System* implies a functional meaning that is not directly reflected in the two others. Finally, other strictly related concepts, such for instance *Blood*, are in MeSH represented under different topics sub-trees, thus being left totally unrelated from the previous ones. Different correlations are required for linguistic inferences and are usually found in a lexical knowledge base. InWordNet, *Cardiovascular System* is a subclass of *Vascular System*, and just *Fetal Circulation* is among its sub-classes. *Heart, Blood Vessel, Bloodstream, Lymph, Lymph node, Veins* and *Liver* are in the *meronimy* relationship with it. Furthermore, *Lymph* and *Blood* are also classified as kind of *Body Fluid*. As a result, MeSH concepts alone, although strongly representative in the medical domain, cannot drive several useful inferential processes, as their cohesion is only postulated in terms of Narrower/Broader relations. These relations are not systematic and cannot properly support text analysis. Beside 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. When dealing with *lymph* and *blood*, it is in fact likely to find text fragments such as:

Unlike the circulatory systems, the lymphatic system lacks any central heart like organ to ***pump lymph*** *throughout the lymph vessels. Because the left ventricle cannot* ***pump blood*** *adequately out to the body, the Norwood*

procedure allows the right ventricle to ***pump blood*** *to both the lungs and ...* The fact that both can be pumped derives from their liquid nature that is defined in WordNet. Such a conceptualisation is then better suited for writing syntactic-semantic interfaces since selectional preferences for prototypical text fragment can be more expressively defined. The availability of linguistically principled semantic primitives as explanations of MeSH knowledge has two effects. First, it helps in better explaining fragments of medical texts during analysis, indexing or retrieval. On the other side, it supports a more expressive semantic model including more predictive rules about the (functional) behaviour of concepts able to deal with missing pieces of knowledge in MeSH. Predictive rules may enable text mining and the discovery of new concepts and relations. This second aspect is important as topic taxonomies may have a rather limited coverage of concepts relevant for text analysis. This is the case of MeSH in the medical domain. Several neglected linguistic forms have to be considered in the domain concept hierarchy as they are needed during analysis. For example, a concept like *operation* is so general in the medical domain that it is not reflected in any MeSH topic. However, as the above example shows it is very useful during analysis of medical texts. In the example (1), the coreference can be successfully resolved only if *operation* is mapped into a referent node within the concept hierarchy. The *coverage* issue is very relevant. In the first 10,000 documents of the OhSUMED collection [5] under the *Disease* sub-hierarchy, about 32% of words in object position with verb *undergo* are not found in the MeSH category system. Specific words (e.g. *adrenalectomies* or *lobectomy*) as well as too general words (like *operation* and *exploration*) are not foreseen. On the contrary, the 95% of the uncovered words have at least one interpretation in WordNet. The above observations suggest that lexical information is useful for supporting several inferences otherwise made impossible by domain taxonomies. However, a generalpurpose resource (e.g. WordNet) is almost neutral with respect to the domain: this means that there is usually a manyto- many mapping between linguistic expressions (e.g. terminology) in the sub-language and concepts. For instance, the word *Heart* has in general different interpretations (10 in WordNet), while it refers to only one specific concept (*the muscular organ located behind the sternum and between the lungs*) in medical texts. A domain provides a strong bias on the possible interpretations that is usually absent from general-purpose resource. The idea of this paper is that this bias should be enforced *a posteriori*. The target resource can work as an augmented lexical KB where domain preferences (as domain concept labels) are attached to word senses. In WordNet, for

example, significant synsets s could be labelled by (lists of) MeSH category names, m: these express domain concepts m in which senses s enter for semantically different motivations. These labels try to capture both paradigmatic and functional semantic information typical of the domain. In [2] we proposed and experimented a method to build a semantic dictionary (i.e. a domain concept hierarchy) for text understanding via the integration of domain knowledge and a general-purpose lexical resource (such as WordNet). As a systematic (i.e. linguistically principled) level of semantic interpretation has been proofed to be obtainable, our next target is to employ it as a common conceptual framework in the retrieval (and question answering) phases as well as in the presentation layer of Web-based e-learning platforms.

3 Concluding Remarks

Domain knowledge for semantic interpretation is a relevant source of information. However, the integration of domain specific thesauri within a text processing task is not straightforward as the primitives available in such resources have an unclear semantic status. Whenver a method to harmonise a domain concept hierarchy with a lexical knowledge base is made available then Information Extraction/ Retrieval and even laguage-driven dialogue, as used in several phases of e-learning processes, are given a suitable (and linguistically adequate) level of abstraction.

The method suggested here 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. The results obtained by the application of the proposed method within a medical knowledge domain are more than promising (see [2]). 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 bene- fits. More in depth analysis of the impact of the method within a knowledge based e-learning system is still needed. First, more work is necessary to assess the effectiveness of the method

within domains different from the medical one. Moreover, the implications of the above procedure in the semantic interoperability problems within Web learning scenarios applications will be the target of further research in the near future.

References

[1] 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.

[2] R. Basili, V. Michele, and Z. Fabio. Integrating ontological and linguistic knowledge for conceptual information extraction. In *Proceedings of the Web Intelligence Conference 13- 16 October*, Alberta, Canada, 2003.

[3] U. Hahn and M. Romacker. Text structures in medical text processing: empirical evidence and a text understanding prototype. In *Proceedings of the 1997 AMIA Annual Fall Symposium (formerly SCAMC)*, Nashville, TN, 1997.

[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] W. Hersh, C. Buckley, T. Leone, and D. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[6] 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.

[7] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, Nov. 1995. [8] MUC-7. Proceedings of the seventh message understanding conference(MUC-7). In *Columbia, MD*. Morgan Kaufmann, 1997.