Engineering a Semantic Web for Pathology

Robert Tolksdorf[1] and Elena Paslaru Bontas[2]

Freie Universität Berlin
Institut für Informatik, AG Netzbasierte Informationssysteme
Takustr. 9, D-14195 Berlin Germany
research@robert-tolksdorf.de, http://www.robert-tolksdorf.de
paslaru@inf.fu-berlin.de

1 Introduction

By using telepathology approaches like virtual microscopy, pathologists analyze high quality digitised histological images on a display screen instead of conventional glass slides at the common microscope. Currently, most of the applications in this domain restrict their retrieval capabilities to automatical picture analysis, ignoring corresponding medical reports or patient records. Therefore, they have the essential drawback that they operate exclusively on structural or syntactical image parameters such as color, texture and basic geometrical forms while ignoring the real content and the actual meaning of the pictures. Medical reports, as textual representations of the pictural represented *content* of the slides, capture *implicitly* the actual semantics of what the picture graphically represent, for example "a tumor" in contrast to "a red blob" or "a colocated set of red pixels". The meaning of the textual content (and of the digital slides), can be extracted and represented *explictly* by means of ontology-driven text processing algorithms. In this paper we propose a *semantic* retrieval system for the domain of lung pathology, which correlates both text and image information and offers advanced content-based retrieval services for diagnosis, differential diagnosis and teaching tasks. At the core of the system is a Semantic Web based knowledge base, gathering ontological domain knowledge, rules describing key processes in pathology and an archive of concrete medical reports. The usage of *Semantic Web* standards and medicine thesauri facilitates the realization of a distributed infrastructure for knowledge share and exchange. In the remaining of the paper we introduce the main features of the retrieval system we are building, with a focus on the construction of the knowledge base for lung pathology and summarize planned future work.

2 Building a Semantic Web for Pathology

The project "Semantic Web for Pathology"[1] aims to realize a Semantic Web-based text and picture retrieval system for the pathology domain. We foresee

[1] The project is funded by the Deutsche Forschungsgemeinschaft, as a cooperation among the Charité Institute of Pathology, the Institute for Computer Science at the FU Berlin and the Department of Linguistics at the University of Potsdam, Germany.

N. Koch, P. Fraternali, and M. Wirsing (Eds.): ICWE 2004, LNCS 3140, pp. 585–586, 2004.

several valuable uses of the planned system in routine pathology. First, it may be used as an assistent tool for diagnosis tasks. Since knowledge is made explicit, it supports new query capabilities for diagnosis tasks. Second, advanced retrieval capabilities may be used for educational purposes by teaching personnel and students. Currently, enormous amounts of knowledge are lost by being stored in data bases, which are behaving as real data sinks. They can and should be used for teaching, e.g. for case-based medical education. Third, quality assurance and checking of diagnosis decisions can be effectuated more efficiently because the system uses axioms and rules to automatically check consistency and validity. Finally, explicit knowledge can be exchanged with external parties like other hospitals. The representation within the system is already the transfer format for information. Semantic Web technologies are by design open for the integration of knowledge that is relative to different ontologies and rules.

At the core of the retrieval system is a domain knowledge base formalized with Semantic Web technologies. It puts together available medical knowledge sources from UMLS [2], generic ontologies, rules and medical reports and adapt this information to the requirements of our concrete application domain "lung pathology". The knowledge base/ontology coordinates the text processing and information extraction procedures. The case report archive in textual form is analyzed using ontology-based text processing algorithms and annotated with concepts from the ontology. Besides, new implicit knowledge from these texts is extracted and integrated in the ontology. As main input for the medical knowledge base we use UMLS, as the more complex medical thesaurus currently available. UMLS as in the actual release contains over 1,5 million concepts from over 100 medical libraries and is permanently growing. Nevertheless, UMLS libraries, though containing a huge amount of concepts or termini have seldom been developed for machine processing, but rather as controlled vocabularies and taxonomies for specific tasks in medicine. From a strict Semantic Web point of view they proved to be deficiently designed and incomplete. Therefore, the first step in generating an ontology based on the UMLS thesaurus was to specify a methodology to overcome these drawbacks. Further on, due to the huge amount of information within UMLS we had to identify the relevant UMLS libraries or concepts and integrate additional information, which is not covered by UMLS by now, but has proved to be relevant for our application domain.

We have generated a core domain ontology in OWL based on the original UMLS knowledge base. We leave additional details about the modelling primitives and the identification of application-relevant UMLS concepts to another paper. After an automatic discovery of the (logical) inconsistencies of the modell, the next step will be the manual adaptation of the OWL ontology in order to correct these errors and to include pathology-specific knowledge, like the concrete structure of case reports and frequently-used concepts from texts not supported by UMLS, contained in a lexicon generated by the lexical analysis of the corpus.

[2] http://umls.nlm.nih.gov