Engineering a Domain Ontology in a Semantic Web Retrieval System for Pathology

Robert Tolksdorf[1], Elena Paslaru Bontas[2]
[1] *research@robert-tolksdorf.de, http://www.robert-tolksdorf.de*
[2] *paslaru@inf.fu-berlin.de*
Freie Universität Berlin Institut für Informatik AG Netzbasierte Informationssysteme
Takustr. 9, D-14195 Berlin Germany

Abstract: Telepathology allows pathologists to acquire, analyze and exchange high quality digital histological images for diagnostic and teaching purposes. Concrete applications in this area have the essential drawback that they restrict their retrieval capabilities to automatic picture analysis, ignoring corresponding medical reports. In this paper we propose a Semantic Web based retrieval system, which uses domain ontologies and a rule engine, as well as ontology-driven NLP algorithms to overcome these deficiencies.

1 Introduction

Digital Pathology and Telepathology focus on the acquisition and transmission of digital medical images between remote sites for diagnostic, prognostic, quality control, research and education purposes. While the importance of telepathology systems is accepted in modern medicine, current implementations have not gained wide acceptance in routine pathology, mainly due to the technical problems related to the management of the huge sizes of image data and the limitations of image-based retrieval. Most of the approaches restrict retrieval capabilities to automatic picture analysis, ignoring corresponding medical reports or patient records. Therefore, they have the major drawback of operating exclusively on structural or syntactical image parameters like color, texture and basic geometrical forms while ignoring the content and meaning of the pictures. As every digital image corresponds to a medical report (in textual form), the reports capture *implicitly* the actual semantics of what the pictures represent. The meaning of the textual content can be extracted and represented *explicitly* with ontology-driven text processing algorithms.

In this paper we propose a *semantic* retrieval system for the domain of lung pathology, which brings both text and image information together and offers advanced content-based retrieval services for diagnosis, differential diagnosis and teaching tasks. The system that we are building consists of a NLP component and a Semantic Web knowledge component, working closely together. The knowledge component gathers ontological domain knowledge, rules describing key tasks and processes in pathology and concrete medical reports. The NLP component annotates the textual reports with concepts from the domain ontology and enriches this ontology by discovering possible ontology extensions. The usage of *Semantic Web* standards and well-established medicine thesauri facilitates the realization of a distributed infrastructure for knowledge sharing and exchange.

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 lung pathology domain. For this purpose we first annotated a large archive of case reports, turning them into a valuable resource for diagnosis and teaching. By correlating the information contained within the case reports produced by experts (the pathologists) with the accompanying images (of the tissue samples), the system produces semantic annotation both for reports and for digital images. The annotation process is supported by a Semantic Web knowledge component, consisting of domain ontologies and rules describing the diagnosis processes. The search within the case archive is content-based in that it can make use of semantic relationships between search concepts and those occurring in the report. Queries asking for "differential diagnosis" (different findings with similar symptoms) – tasks which normally require consultation of textbooks – can be processed using diagnosis rules. These search capabilities are useful both for pathology-specific tasks and for case-based teaching, by making interesting examples and reference cases available to students. Another use case is quality control during input of new reports, which are analyzed on-the-fly for potential inconsistencies w.r.t. the background domain knowledge. During the development phase of the system, we are using this feature for an incremental generation of the knowledge base. In the following we focus on the realization of this knowledge base, leaving issues like search and retrieval, as well as details about text-processing for another paper [SSPB04].

The core of the retrieval system is a domain knowledge base formalized with Semantic Web technologies. It puts together pathology-specific ontological knowledge, generic ontologies, rules and medical reports and adapts this information to the requirements of our concrete application domain "lung pathology". Domain ontologies are used for the machine-processable representation of pathologic-characteristic knowledge, while generic ontologies capture common sense knowledge useful in knowledge intensive tasks. The necessity of using this second category of ontologies has been emphasized in several similar projects which analyzed the quality and usage challenges of UMLS in building knowledge bases [SH01, GPS99]. While ontologies model the background knowledge of the pathologists, the rules are used to describe the decision processes using this knowledge: diagnostics, microscope analysis, observations, differential diagnosis etc. The acquisition of such rules will be accomplished during an intensive collaboration with domain experts.

As input for the medical ontology we use UMLS[2], as the most complex medical thesaurus currently available. UMLS as in the current release contains over 1,5 million concepts from over 100 medical libraries and is permanently growing. Due to the complexity of the thesaurus and the limitations of current Semantic Web tools we need to customize the available medical collection w.r.t. two axes: a) the identification of relevant UMLS libraries and concepts corresponding to "lung pathology" from UMLS and, b) their adaption to the particularities of language and vocabulary of the case reports.

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

[2]Unified Medical Language System: http://nlm.nih.gov/research/umls.

2.1 Identifying relevant knowledge in UMLS

UMLS brings together concepts from over 100 medical libraries and thesauri and integrates them to a common data format. Basically it consists of the Semantic Network and the Metathesaurus. The Semantic Network defines semantic types and relations and acts as a meta-ontology for the Metathesaurus, which contains the concrete medical concepts (UMLS concepts) and their linguistic variants. Every UMLS concept is related to at least one semantic type from the Semantic Network. Concepts are connected through relationships, which in turn reference semantic relationships of the Semantic Network.

Due to the close correlation between the two levels within UMLS a first step in building our pathology ontology was to incorporate the Semantic Network in the knowledge base. For the identification of the relevant Metathesaurus concepts, we started by analyzing the features of the available UMLS Knowledge Server[3], which provides the MetamorphoSys tool and an additional API to tailor the complete thesaurus to specific application needs. However, both allow mainly syntactical filtering methods (e.g. exclude complete UMLS Sources, exclude languages or term synonyms) and do not offer means to analyze the semantics of particular libraries or to use only relevant parts of them. We adopted two approaches to overcome this problem.

2.1.1 Top-down Approach

The aim of the top-down approach was to restrict the huge amount of medical information from UMLS to the domain “lung pathology”. For this purpose, we consulted a team of domain experts, who identified UMLS libraries potentially relevant to “lung and “lung diseases”. However, the complexity and content heterogeneity [4] of the particular libraries made a manual identification time-consuming and inefficient. Approximately 50 percent of the UMLS libraries have been selected as possibly relevant for lung pathology, containing more than 500000 concepts. Managing an ontology of such dimensions with Semantic Web technologies is related to unsolved issues w.r.t. scalability and performance of the system. Because of the application-oriented character of the system, such parameters play a crucial role for further considerations. Building the knowledge base automatically requires also a subsequent manual adaptation of the content, performed by domain experts. They should therefore be able to evaluate and modify the ontology. Apart from the technical drawbacks, a very-large ontology can not be used efficiently by humans as well.

2.1.2 Bottom-up Approach

In the second approach we used the case reports archive to identify concepts, which actually occur in medical reports (i.e. are really used by pathologists while writing down their observations and therefore will also occur as search parameters). For this purpose we used a retrieval engine mapping a lexicon representing the vocabulary of the reports archive to the content of the UMLS sources. The lexicon containing the most frequent nominal phra-

[3] UMLS Knowledge Server: http://umlsks.nlm.hin.gov.

[4] Most of the UMLS libraries contain concepts belonging to different medicine specialities.

ses was the result of the lexical analysis of the medical reports (in German). Due to the restricted set of German terms within UMLS (e.g. from 500000 concepts only 12000 have corresponding German translations in the actual version 2003AC of UMLS), the lexicon was subsequently translated to English and compared to UMLS. The result of this task was a list of 10 UMLS libraries, still containing approximately 350000 different concepts. The size of the concept set can be explained if we consider the fact that the UMLS knowledge is concentrated in several major libraries (e.g. MeSH [5],SNOMED98[6]), which cover important parts of the complete thesaurus and therefore contain the most of the concepts in our application lexicon. To differentiate among the derived libraries we mapped in a second step 10 central concepts in lung anatomy and extract similar or related concepts from UMLS sources. A total of approximately 1000 concepts describing the anatomy of the lung and lung diseases served as initial input for the domain ontology.

2.2 Adapting the ontology to the application domain

The content-related limitations of UMLS have been revealed by comparing it against the archive-based lexicon. The lexicon demonstrated also a need for a coherent and detailed model for certain non-medical terms (e.g. properties of physical objects like solid, color, length, diameter, spatial relations). Generic concepts, though modelled to a certain extent in the Semantic Network, which is also part of our ontology, deserve a special attention and are subject of current work. Several ontologies describing the structure of pathology reports, common terms used in diagnostic tasks in routine pathology completed the ontology library.

2.3 Modeling

Medicine ontologies though containing a huge amount of concepts or terms have seldom been developed for machine processing, but rather as controlled vocabularies and taxonomies for specific tasks in medicine [SH01]. From a strict Semantic Web point of view they proved to be deficiently designed and incomplete. Apart from the absence of a Semantic Web compatible representation language, UMLS adopts an error-prone modelling style, which is characterized by few semantic relations among concepts and an ambiguous way to interpret such relations (e.g. concepts of the UMLS Metathesaurus are connected through relations like “related”, “broader”, “narrower”, “similar”). A typical example is the usage of the relation “is-a” for both instantiation and specialization/generalization, the usage of a unique “part-of” relation with different meanings (“functional part”, “component”, “substance”) or the usage of one of these relations instead of the other. Mathematical properties of the same semantical relation (e.g. transitivity) are not fulfilled for each pair of concepts connected by the relation and the “is-a” relation between two concepts does not always guarantee the inheritance of the properties of the parent concept to its children.

We generated a core domain ontology[7] in OWL based on the original UMLS knowledge base. From a modelling perspective, we mapped each UMLS concept to an OWL class, saved associated definitions and alternative concept names with language specification

[5]Medical Subject Headings: http://nlm.nih.gov/mesh/meshhome.htm.
[6]Snomed International: http://snomed.org.
[7]The ontology can be found at http://nbi.inf.fu-berlin.de/research/swpatho/owldata/swpatho1.owl

and related it to the corresponding UMLS sources. We also translated UMLS relations with a specified meaning to range restrictions on the corresponding concepts and cumulate fuzzy relations like “synonyms”, “related”, “other-related” etc. to a generic “related_to” relationship. After an automatic discovery of the (logical) inconsistencies of the model, the next step was be the manual adaptation of the OWL ontology in order to correct these errors and to include pathology-specific knowledge not covered by UMLS(see 2.2).

3 Related work

Medicine is one of the application domains, where the utility of ontologies is widely accepted and therefore medicine ontologies have already been deployed at large scale. However available ontologies (UMLS inclusively), actually used as common vocabulary in medical applications, cover particular domains of medicine to different granularities and cannot be directly used for the Semantic Web. This issue has been addressed in project GALEN[8], where the authors developed a special representation language, tailored for the particularities of the (English) medical vocabulary. However, the usage of a proprietary representation makes the ontological knowledge difficult to be extended by third parties or exchanged on the Web or in a Semantic Web setting. Several important research initiatives analyze the usage of UMLS as input for building knowledge bases for medicine [BO01, SH01, GPS99, CM93, GPG+00]. They prove the ontological commitment of UMLS in order to use it in knowledge intensive tasks (e.g. the ONIONS [GPS99] methodology for ontology merging is exemplified on UMLS and the MEDSYNDIKATE [SH01] project uses it for knowledge discovery from texts). Both projects offer valuable experiences and facts concerning UMLS and medical ontologies generally, but they do not use Semantic Web technologies to facilitate knowledge share and reuse, a significant feature of ontologies. Besides, our focus is application-oriented. We intend to build a ontology for lung pathology *in our application setting*. Despite standards and tools for the main technologies, making Semantic Web applications a reality, its potential and acceptance at a broader scale is still a challenging issue for the Semantic Web research community.

Literatur

[BO01] Burgun, A. und O.Bodenreider: Mapping the UMLS semantic network into general ontologies. In: *Proc. of the AMIA Symposium*. 2001.

[CM93] Carenini, G. und Moore, J.: Using the UMLS semantic network as a basis for constructing a terminological knowledge base: A preliminary report. In: *Proceedings of 17th Symposium on Computer Applications in Medical Care (SCAMC ’93)*. 1993.

[GPG+00] Gu, H., Perl, Y., Geller, J., Halper, M., Liu, L., und Cimino, J. Representing the UMLS as an OODB: Modeling issues and advantages. 2000.

[GPS99] Gangemi, A., Pisanelli, D. M., und Steve, G.: An overview of the ONIONS project: Applying ontologies to the integration of medical terminologies. *Data Knowledge Engineering*. 31(2):183–220. 1999.

[SH01] Schulz, S. und Hahn, U.: Medical knowledge reegineering - converting major portions of the UMLS into a terminological knowledge base. *International Journal of Medical Informatics*. 2001.

[SSPB04] Schlangen, D., Stede, M., und Paslaru Bontas, E.: Feeding OWL: Extracting and representing the content of pathology reports. In: *to appear in Proc. NLPXML 2004*. 2004.

[8] GALEN Ontology: http://opengalen.org