PATExpert: Semantic Processing of Patent Documentation

Leo Wanner, Sören Brügmann, Barrou Diallo, Mark Giereth, Yiannis Kompatsiaris,
Emanuele Pianta, Gautam Rao, Pia Schoester, and Vasiliki Zervaki

Abstract— **PATExpert is a recently started "Specific Targeted Research Project" funded by the EC in FP 6, IST priority. PATExpert's goal is to change the paradigm currently followed for patent processing from textual to semantic. We are about to develop a semantic multimedia content representation based on Semantic Web technologies for selected technology areas and to investigate some central topics from the semantic representation angle: patent retrieval and classification, content extraction, generation of multilingual user comprehensible patent information, visualization of and navigation in patent content spaces, and patent valuing and technology area assessment.**

Index Terms— **semantic representation, (multimedia) ontology, OWL-DL, SUMO, PULO, patent processing techniques.**

I. Introduction

Patents belong to the few types of public information that have a big impact on the European economy, and whose proper monitoring, retrieval, representation, interpretation, and assessment so clearly depend on the access to its content, and, thus, on advances in semantics-based techniques. However, research and development in the area of patent processing still focuses on selected traditional tasks such as text retrieval, classification, and shallow linguistic analysis. Recent initiatives that target the automatic access to content of patents attempt to cover ALL knowledge areas. This forces them to rely on term frequency, term co-occurrence and grammatical term categories. I.e., despite the use of a Semantic Web-based formalism, the resulting representation is not a real content representation. As a consequence, tasks that ultimately require knowledge-based multimedia techniques (content-oriented search, assessment, abstracting, etc.) are still, to a major extent, carried out manually. The overall goal of the PATExpert project, which started 01.02. 2006 and is funded by the EC (FP6, IST-028116, http://www.patexpert.org), is to change the paradigm currently followed for patent processing from textual (viewing patents as text blocks enriched by "canned" picture material, sequences of morpho-syntactic tokens, or collections of partial syntactic structures) to semantic (viewing patents as multimedia knowledge objects) processing. PATExpert is about to develop a multimedia content representation formalism based on Semantic Web technologies for selected technology areas and to investigate some central topics from the semantic representation angle: patent retrieval and classification, content extraction, generation of multilingual user comprehensible patent information, visualization of and navigation in patent content spaces, and patent valuing and technology area assessment, taking into account the information needs of all user types as defined in a user typology. PATExpert's technological goal is to develop a showcase that demonstrates the viability of PATExpert's approach to content representation for real applications.

II. Semantic Representation of Patents

The semantic representation of patent documentation must cover, on the one hand, propositional, multimedia and metadata information and, on the other hand, lingustic knowledge—first of all the characteristic text structures encountered in patent documentation and the lexical information. We developed an initial working schema of the knowledge representation (KR) in PATExpert, with OWL-DL as the KR language.

As a rule, the content in patent documentation makes reference to knowledge of three levels of abstraction: (a) common sense knowledge, (b) patent-specific knowledge and terminology, and (c) domain-specific knowledge that refers to technology area details. PATExpert focuses on the ontologies of two technology areas: optical recording media and mechanical engineering tools.

Traditionally, common sense knowledge representation is dealt with by core ontologies and domain-specific knowledge representation by domain-ontologies. As core ontology, we use SUMO [2]. Following a recent trend in semantic web representation technologies (see, e.g., the *Midlevel Ontology*, MILO by *Teknowledge Knowledge Systems Group* [3]), we capture patent-specific knowledge and terminology by a *midlevel ontology*, called *Patent Upper Level Ontology*, PULO. PULO aims to bridge the gap between the high level concept descriptions in SUMO and the detailed domain ontologies.

Multiple media (in particular images such as photographs, diagrams, flow charts, drawings, etc.) come into play in the representation of patent documentation and must thus be modelled by a multimedia ontology. For linguistic knowledge representation, we foresee a patent document structure ontology and lexical (word level) ontologies. As lexical ontologies, we use WordNets [1]. Patent-related meta information (such as the patent holder, inventor and his current affiliation, etc.) are captured in a separate metadata ontology. Figure 1 shows the initial working schema of the knowledge representation in

L. Wanner is with ICREA and Foundation Barcelona Media, email: leo.wanner@upf.edu; S. Brügmann is with Industrie Software J. Brügmann, email: soeren.bruegmann@isjb.de; B. Diallo is with the European Patent Office; M. Giereth is with the University of Stuttgart, email: mark.giereth@vis.uni-stuttgart.de; Y. Kompatsiaris and V. Zervaki are with the Aristotle University; E. Pianta is with the Istituto Trentino di Cultura, email: pianta@itc.it; G. Rao is with IALE, email:rao@iale.es, and Pia Schoester is with the Fraunhofer Gesellschaft, email: pia.schoester@pst.fhg.de

PATExpert. It will be revised as needed when the work on the tasks that make use of the semantic representation progresses.

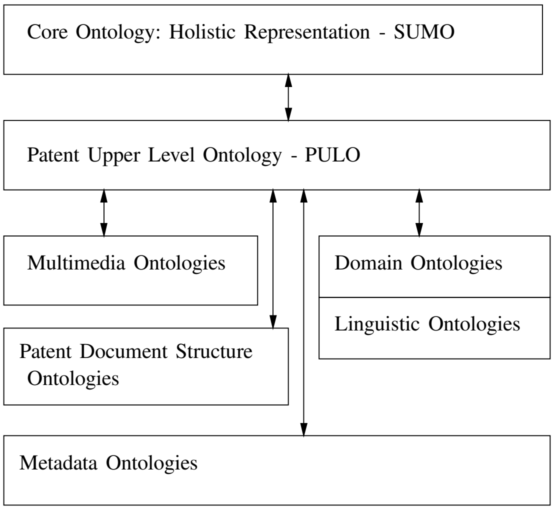
Core Ontology: Holistic Representation - SUMO
Patent Upper Level Ontology - PULO
Multimedia Ontologies
Domain Ontologies
Linguistic Ontologies
Patent Document Structure Ontologies
Metadata Ontologies

Fig. 1. Knowledge representation modules in PATExpert

III. Processing Patent Documents

The state of the art in a number of central patent processing areas suffers from the lack of an adequate representation of the content and content structure of patent documentation. With the KR-schema presented above at hand, PATExpert addresses these areas. The most central of them are listed below. In the poster presentation, more details will be given on each topic.

A. Information Extraction from Patents

Extraction of content information (e.g., composition and function of the invention) and meta information (e.g., the productivity of an inventor) from patent documentation is one of the burning issues in patent processing. In PATExpert, this topic is approached from two angles: as a stand alone task, and as a way to populate the knowledge base. Strategies using partial syntactic and semantic analysis, information extraction techniques, inference mechanisms are being explored.

B. Patent Retrieval and Classification

The (multi) media content representation of patent documentation will allow us to develop patent retrieval strategies that go considerably beyond the state of the art patent retrieval techniques. In particular, it will facilitate semantic retrieval, i.e., search for patents that describe inventions with specific content features, image-based retrieval and document similarity-based retrieval. It will also allow for a classification and clustering of patent documents along semantic criteria.

C. Production of Multilingual Patent Information

The language style in patent documentation is very complex and repetitive. It is thus hard to comprehend by human readers. Our goal is to provide the reader with a comprehensible variant of text passages chosen by him in the language of his choice. Two topics are addressed: (a) paraphrasing of patent passages and (b) generation of multilingual gists of given passages. For both, shallow techniques and deep techniques that draw on content representation of the text passages in question (and that implement thus text generation proper) are being developed.

D. Visualization and Navigation in Patent Knowledge Spaces

Given the complexity of patent knowledge spaces, the availability of techniques for visualization of patent content material that is retrieved from the patent KB or selected by the user while browsing the patent KB is crucial. We develop techniques that make the complex content structures transparent (as, e.g., the IS-A, PART-OF, CAUSE, OPERATE, etc. relations between objects, semantic similarity / entailment links, etc.) and help navigate through such structures. The navigation techniques combine browsing mechanisms with advanced strategies that guide navigation taking the user's focus, the context and the discourse relations between knowledge objects into account.

E. Patent Valuing and Technology Area Assessment

Currently, high quality valuing of patents and patent applications and the assessment of technology areas with respect to their potential to give rise to patent applications is done mainly manually—which is very costly and time consuming. We are developing techniques that use statistical and semantic information from patent (applications/) as well as user based data for market aspects to prognosticate the value of a patent (application).

IV. The State of Affairs

After having developed the initial schema of the knowledge representation, we work on the topics sketched in Section III. The first prototypical implementations of the techniques are planned to be operational by the end of May 2007, some of them (e.g., the gist generation) already by the end of January 2007.

References

[1] C. Fellbaum (ed.), *WordNet. An Electronic Lexical Database*. Cambridge, MA: The MIT Press, 1998.

[2] I. Niles and A. Pease, "Towards a Standard Upper Ontology," in *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, C. Welty and B. Smith, Eds., Ogunquit, MA, 2001.

[3] I. Niles and A. Terry, "The MILO: A General-Purpose, Mid-Level Ontology," in *Proceedings of the 2004 International Conference on Information and Knowledge Engineering*, Las Vegas, NE, 2004.