

# Ontology Learning in Multimedia Information Extraction from Product Catalogues

Roberto Bartolini, Emiliano Giovannetti, Simone Marchi and Simonetta Montemagni  
{roberto.bartolini, emiliano.giovannetti, simone.marchi, simonetta.montemagni}@ilc.cnr.it  
Istituto di Linguistica Computazionale, CNR – via Moruzzi 1 - 56124 Pisa – Italy

Claudio Andreatta and Roberto Brunelli  
{andreatta, brunelli}@itc.it  
Istituto Trentino di Cultura (ITC-irst) – Via Sommarive 18 - 38050 Povo - Italy

Rodolfo Stecher and Claudia Niederée  
{stecher, niederee}@ipsi.fraunhofer.de  
Fraunhofer IPSI – Integrated Publication and Information Systems Institute - Dolivostrasse 15 – 64293 Darmstadt – Germany

Paolo Bouquet and Stefano Bortoli  
{bouquet, bortoli}@dit.unitn.it  
Dept. of Information and Communication Technologies, University of Trento Via Sommarive 14 – 38050 Trento – Italy

We propose a methodology for extracting multimedia information from product catalogues empowered by the synergetic use and extension of a domain ontology. The use of domain ontologies in this context additionally opens up innovative ways of catalogue use. The method is characterized by incrementally feeding and exploiting the ontology during an information extraction process, implemented by the semantic annotation of the analysed document, and by providing support for detecting existing similar ontologies to enable reuse of (parts of) them.

Technical topics: knowledge driven multimedia analysis, ontology learning, semi-automatic content annotation tools.

The field of semi-automatic information extraction from textual corpora is central for overcoming the so-called “knowledge acquisition bottleneck”. Multimedia sources of information, such as product catalogues, contain text (captions) and images (pictures of the products) thus requiring information extraction approaches combining several different techniques, ranging from Natural Language Processing to Image Analysis and Understanding. In our approach we have three main aspects to consider: 1) the information extraction per se, 2) the ontology, its use and creation, and 3) the usage of the ontology in the information extraction process and the synergy between different kinds of extraction processes (Fig 1).

The development of adequate ontologies itself is one of the knowledge acquisition bottlenecks: the use of (semi-) automatic tools for semantic information extraction from multimedia corpora is very promising but, to be efficiently exploited, must have access to a formal representation of a certain domain, i.e., an ontology. We support the ontology creation process in two different and complementary ways, ontology learning and reuse of existing ontologies. The ontology learning approach takes advantage of the results of the extraction to enrich the ontology, and the reuse support

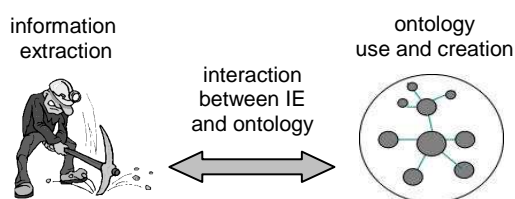


Fig.1 The three main activities involved in the process

provides methods and tools to re-use already existing ontologies which capture the target domain under a similar modelling perspective as the one of interest for the extraction task. This (apparent) *vicious* circle (between the need of having the domain represented in the ontology for an extraction process and the enrichment of the ontology based on the results obtained from the extraction) can be turned to a *virtuous* circle if the necessary conditions are set to let the evolving ontology and the information extraction tool interact in a synergetic way. From the textual information extraction point of view, for example, an initially semantically ambiguous sentence can be interpreted correctly (and thus reverted to be a source of information) as soon as the appropriate knowledge (coming, for instance, from the analysis of the recurrent patterns found inside the catalogue) is added to the ontology. The disambiguated sentence can itself contribute to provide new knowledge to be added to the ontology and so on. The information conveyed by a multimedia document is analyzed and extracted on two different levels: the document level, in which the document (geometrical) layout is investigated considering both text and images, and the image level, in which pictures are examined in order to describe their visual content and to recognize the depicted objects. The main difficulty in the description of image content is the lack of information about the kind and the number of objects possibly present. Supporting the information extraction process with a domain ontology permits the development of context-aware strategies in order to guide and focus the multimedia semantic analysis. The other way round content based image analysis allows the acquisition and exploitation of similarity relations among multimedia entities thus allowing to refine and enrich the knowledge representation modelled in the domain ontology.

We believe the methodology we propose to be strongly related to BOEMIE project, focusing, as stated onto the BOEMIE 2006 web page, “on the automation of the process of knowledge acquisition from multimedia content, using evolving multimedia ontologies which will be used for the extraction of information from multimedia content in networked sources”.

The methodology we present is developed inside the Vikef project (Virtual Information and Knowledge Environment Framework, IST-2002-507173 - <http://www.vikef.net/>), which creates an advanced software framework for enabling the integrated development of semantic-based Information, Content, and Knowledge (ICK) management

systems, as part of the Advanced Semantic Annotation of Italian Text task applied to trade fair catalogues. Apart from the scientific and academic interest related to these fields of research we have also registered a growing need from industrial parties for automated knowledge elicitation tools to be applied to their commercial resources, such as product catalogues.