Linguistically analyzed labels of knowledge objects: How can they support OBIE? Lessons learned from the Monnet and Trend-Miner projects

Thierry Declerck
DFKI GmbH
thierry.declerck@dfki.de

1 Abstract

We are investigating the use of natural language expressions included in Knowledge Organization Systems (KOS) for supporting Ontology-Based Information Extraction (OBIE), in a multi- and cross-lingual context.

Very often, Knowledge Organization Systems include so-called annotation properties, in the form of *labels*, *comments*, *definitions*, etc, which have the purpose of introducing human readable information in the formal description of the domain modelled in the KOS.

An approach developed in the Monnet project, and continued in the TrendMiner project, consists in transforming the content of annotation properties into linguistically analysed data. Natural language processing of such language expressions, also called sometimes *lexicalisation* of Knowledge Organisation Systems, are thus transforming the unstructured content of annotation properties into linguistically structured data, which can be used in comparing language data included in a KOS with linguistically annotated texts. If some match of linguistic features between those two types of documents can be established, corresponding segments of the textual documents can be semantically annotated with the elements of the KOS the content of the annotation property is associated with. Evidently, this semantic annotation procedure can be of great help for OBIE, relating text segment to relevant parts of thesauri, taxonomy or ontologies.

But looking in more details at the language data contained in annotation properties, we can see that this data very often has to be modified in order to be better used in the context of OBIE. Also there is a need for a formal representation of such linguistically annotated language data in order to ensure interoperability with semantic data available in the Linked Data Framework.

The talk will expand on those issues.

2 Short Bio

Thierry Declerck is senior consultant at DFKI's LT lab and he was leading the DFKI contribution to the European Project MONNET[1]. Before this he was in charge of the DFKI contribution to the Integrated Project MUSING[2], which finished in April 2010, and till March 2009 he was involved as well in the European Network of Excellence "K-Space" (*Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content*). In the field of standardization of language resources, Thierry Declerck was involved in the eContent "Lirics" (Linguistic Infrastructure for Interoperable Resources and Systems) project (see http://lirics.loria.fr/) and was leading the MUMIS project on the Indexing and Search of Multimedia data. He was also in charge of the ACL Natural Language Software Registry, which is now integrated in lt-world. Since Mai 2004, he was conducting the DFKI contribution of the eTen WINS project. Thierry Declerck is also actively involved in ISO TC37/SC4/ (on language resources management).

[1] http://www.monnet-project.eu/
[2] http://www.musing.eu/

Proceedings of the Joint Workshop on NLP&LOD and SWAIE, page 49,
Hissar, Bulgaria, 12 September 2013.