OLE – Automatic Ontology LEarning and Processing Platform

NLP Laboratory
Faculty of Informatics
Masaryk University
Brno, Czech Republic

by Vít Nováček, xnovacek@fi.muni.cz

Introduction

OLE is a system for automated bottom-up generation and merging of domain specific ontologies. Uncertain knowledge representation implementation efforts are incorporated. An inherent part of the project is also a development of a reasoning engine and a question-answering system for the resulting ontologies.

Motivation

The main objective of the project is to implement a system that is able to create and update a domain specific ontology for any given domain. All phases of the process are nearly automatic with minimum need of human assistance. The further utilisation of resulting ontologies can be summed up in the following general points:

- complex and concise **empiric representation** of the related domain
- **visualisation** of the domain's conceptual structure
- **querying** among the domain representation
- **reasoning** among the concepts and relations stored in the ontology

Design Considerations

The design of OLE has been influenced by the need for autonomy, efficiency and precision of the resulting platform. The following list summarizes the major requirements:

- The tool should support the user-friendly interactive way of ontology acquisition, but also the fully automatic process of knowledge mining that can run without any human assistance.
- The precision is preferred over the recall. Even if the number of the extracted conceptual structures can be relatively low (compared to the number of relations a human can identify in the same resource), it will be balanced by the extensive quantity of resources available.
- The relations between concepts stored in the resulting ontology need not to be precise – the explicit uncertain knowledge representation is one of the essential parts of OLE. The loss of exactness is balanced by the increased fuzzy precision of the whole process.

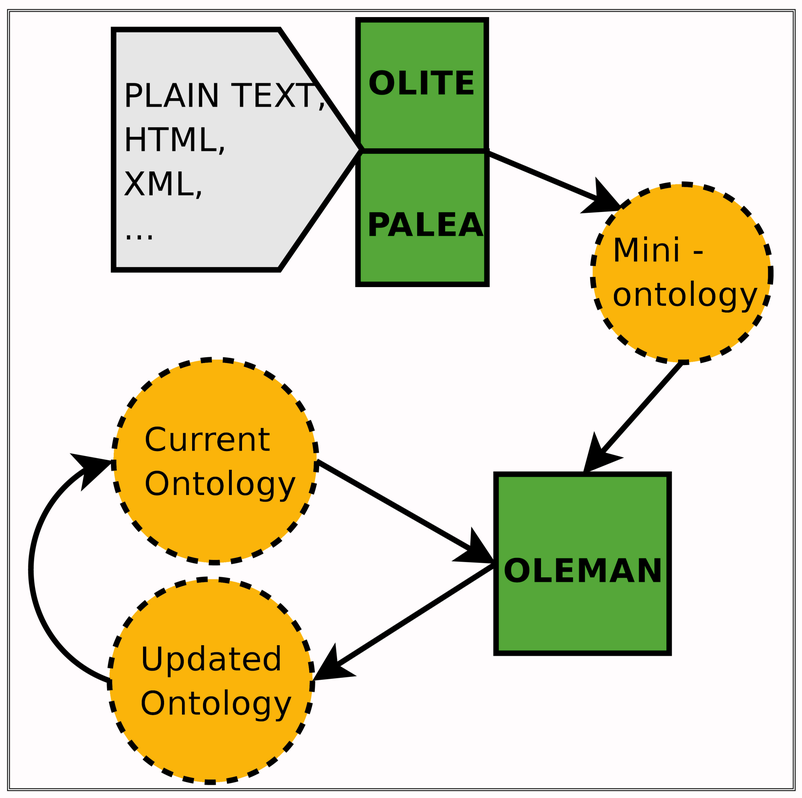
Architecture

The OLE system consists of three main parts:

- **OLITE** – extractor of concepts and their relations from various (mainly structured or unstructured natural language) resources; creates a simple miniontology from a resource, using a plug-in for particular extraction technique (pattern driven extraction, FFCA and so forth)
- **PALEA** – the module is responsible for learning new patterns for extraction of semantic relations
- **OLEMAN** – as the most important part, the module can integrate the ontologies produced by **OLITE** into the current domain ontology; **OLEMAN** contains the question-answering system, inference engine and implements basic interactive ontology management functions as well

Ontology Creation Cycle

The domain specific ontology managed by OLE is continuously updated in the cycle depicted on the figure below:

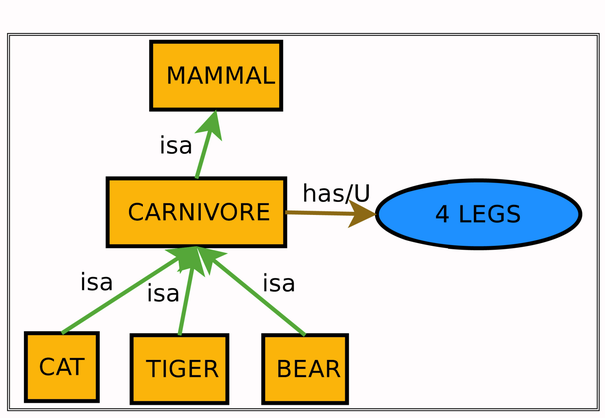


Examples

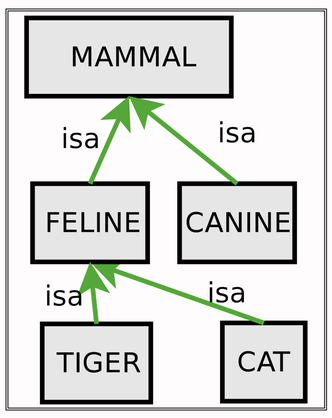
Consider the following text in natural language:

Cat is a carnivore. We know other carnivores, such as tiger or bear. Carnivores are mammals and have usually four legs.

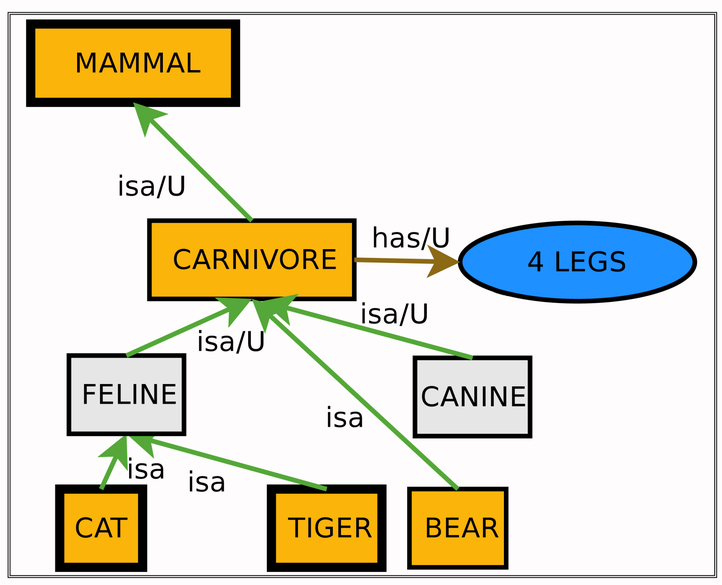
Here is an example of the resulting miniontology structure:



Suppose we have a domain specific ontology with the following fragment:



An example of uncertain integration of the miniontology into the above domain specific ontology is given here:



Notes:

- the boxes represent concept classes and the ovals concept properties
- the arrows correspond to a semantic relation according to the provided caption (*isa*, *has*, ...)
- the */U* suffix in a relation's caption represents uncertain weight of the respective relation (fuzzy measure or continuous probability value), which is supposed to be assigned by various novel algorithms employed in the ontology acquisition and integration process

Current State and Future Directions

The extraction module is tested and refined now with real data from a collection of about 12000 plain text articles related to the computer science domain. The preliminary results of one of the extraction methods (pattern driven extraction of hyponymy relation) applied on the BNC corpus seem to be promising – almost 90 percents of relations identified by human can be extracted this way.

Quite a challenging work still remains – the merging of ontologies is currently implemented only very rudimentarily. A universal format of uncertain knowledge representation has to be specified and inference engine implemented at the same time. These tasks are under intensive research now and their proper implementation is planned in the nearest future.

Acknowledgements

The project is partly related to my diploma thesis at Faculty of Informatics of Masaryk University in Brno, Czech Republic. However, my work is a part of and supported by Academy of Sciences of the Czech Republic, 'Information Society' program, the research grant T100300419.