

## **Research Internship**

### **Ontology Reasoning for Text Processing**

#### **Supervision**

Yue Ma

*firstname.lastnom@lipn.univ-paris13.fr*

#### **Context**

In recent decades, the study of « meaning » (or « semantics ») has gotten a high-speed progress because of the dramatic expansion of documents within enterprises and on the web. To this end, Semantic Web [1] has been proposed as the next generation of web, where semantics plays a key role. It has been gaining momentum driven by World Wide Web Consortium (W3C, <http://www.w3.org>) since 2001. A typical research under Semantic Web is to study ontology languages [2], such as OWL which is standardized by W3C and has Description Logics [3] as its formal semantic underpinning. It is remarkable to see that many scalable reasoners are implemented for different profiles of OWL.

Advanced or domain-oriented text processing systems can benefit significantly from the access to ontologies [4]. Among many, three important advantages are as follows: 1) Unlike word-based systems, ontology can consist of concepts not specifically found in a document; 2) Ontology has the well-defined semantics which can express information without ambiguity such that data are machine-readable; 3) Benefit from the growing of Semantic Web research, numerous ontology-oriented tools are available, so we can flexibly access ontology reasoning techniques. In all, the study of ontology reasoning for text processing systems is valuable.

#### **Description of work**

Supported by the Quaero program, the goal of this internship is to study how advanced ontology reasoning techniques can contribute to text processing systems. One of the main topics is to study and evaluate the ontology reasoning based metrics to improve semantic annotations on texts, where semantic annotation is to tag fragments of texts by suitable ontological elements, which makes texts machine processible via the semantics of the ontology.

This work concerns two layers of techniques: ontology reasoning (logics) and text processing (information extraction). Relevant existing approaches are mostly based on the assumption that the information extraction is a black-box and not interactive even if errors may be detected in ontological annotations. We are interested in getting over this disadvantage in our semantic annotation platform. It will contain several steps: 1) Analysis of domain corpus on which semantic annotations are made; 2) Checking the inconsistency of those annotations by ontology reasoners; 3) Design metrics to measure the quality of the annotation rules which are used to generate semantic annotations. 4) Evaluation of this approach.

The benefits of this work are to associate a certainty degree of reliability to each annotation and to exclude bad annotation rules for further annotation procedures.

The continuing progress of this work is promising and can include the following aspects, such as inconsistency handling techniques for handling content conflicts in texts; text-based ontology matching technique; and developing extra reasoning services for text processing.

## **Bibliography**

1. Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):35-43, 2001.
2. Ian Horrocks: Ontologies and the semantic web. *Commun. ACM* 51(12): 58-67 (2008)
3. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
4. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Ronen Feldman, James Sanger, Cambridge university press, 2007.

## **Desired background**

Bac + 5 in Computer Science, able to work in English.

Skills in Computer Science and in NLP will be appreciated when studying applications.

## **Conditions**

Six months internship, supported by a project.

The internship is located in LIPN ([www-lipn.univ-paris13.fr](http://www-lipn.univ-paris13.fr)).

In case of success, the internship might be continued by a PhD supported by a project.