CA-Manager: a middleware for mutual enrichment between information extraction systems and knowledge repositories

Hacène Cherfi*, Martin Coste*, and Florence Amardeilh*

*Mondeca SA, 3 cité Nollez - 75018 Paris
{firstname.lastname}@mondeca.com,
http://www.mondeca.com

Abstract. Knowledge enrichment aims at bridging the large gap between structured knowledge and the large volumes of unstructured text data that companies and people need to deal with daily. Alas, the process is very laborious and error-prone, even when performed semi-automatically. The two key steps in this process -semantic annotation and ontology population- still hold outstanding challenges although they are actively studied by researchers. While there exists a large number of tools, many of them lack compliance with Semantic Web standards, but more important, they lack the flexibility to customise the entire knowledge acquisition workflow. In this paper, we present the Content Augmentation Manager (CA-Manager) framework which plays a middleware role between Information Extraction (IE) tools and knowledge repositories (KR)s. CA-Manager allows us an easy plug-in of various types of components leading to create a virtuous cycle within the annotation workflow.

1 Introduction

One of the main challenges for the large adoption of Semantic Web technologies is to get semantic data in order to be able to develop smarter applications to search, browse, publish, infer, etc. Even Google understood this by buying Freebase [1], an online graph-based knowledge base of thousands of interconnected entities. Google can now build its new semantic search engine, called Knowledge Graph [2]. The exponential growth of semantic data published through the Linked Open Data Initiative is another important marker of the actual technological shift that we are going through. But everyone is not Google nor have semantic datasets ready to be publicly exposed (or not) to build the innovative services/applications of tomorrow. It is absolutely necessary to provide tools to support the creation of such knowledge repositories.

The first step is the creation of an ontology to represent the knowledge of the concerned domain. An ontology has been defined as a formal conceptualization of a model, composed of concepts, properties (attributes and relations) and axioms. It can be understandable by machines, used for sharing and re-using knowledge and permitting reasoning thanks to the semantics explicitly represented in the ontology. This issue alone is a major research field of

1. http://www.freebase.com
2. http://mashable.com/2012/02/13/google-knowledge-graph-change-search