CA Manager Framework: Creating Customised Workflows for Ontology Population and Semantic Annotation

Danica Damljanovic, Florence Amardeilh, Kalina Bontcheva

CA Manager Framework: Creating Customised Workflows for Ontology Population and Semantic Annotation

Danica Damljanovic
Department of Computer Science
University of Sheffield
Sheffield, UK
D.Damljanovic@dcs.shef.ac.uk

Florence Amardeilh
Mondeca
Paris, France
florence.amardeilh@mondeca.com

Kalina Bontcheva
Department of Computer Science
University of Sheffield
Sheffield, UK
K.Bontcheva@dcs.shef.ac.uk

ABSTRACT

We present the Content Augmentation Manager Framework for creating various adapted workflows for ontology population and semantic annotation based on Semantic Web recommendations and UIMA precepts. This framework supports ontology population from text (semi)automatically, by allowing easy plug-in of various types of components including information extraction tools, customised domain ontologies, and diverse semantic repositories. Our evaluation reveals that the framework offers flexibility, without compromising on precision and recall of the constituting components.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: representation languages, miscellaneous

General Terms

Design

1. INTRODUCTION

Gartner reported in 2002[1] that for at least the next decade more than 95% of human-to-computer information input will involve textual language. They also report that by 2012 taxonomic and hierarchical knowledge mapping and indexing will be prevalent in almost all information-rich applications. There is a tension here: between the increasingly rich ontology-based, semantic models on the one hand, and the continuing prevalence of human language materials on the other. This process may be characterised as the dynamic creation of interrelationships between *ontologies* and unstructured textual content in a bidirectional manner.

However, transforming huge number of unstructured text into the semantically interlinked knowledge space is a big challenge. Two key parts of this process are *semantic annotation* and *ontology population*. While there are numerous tools to support these processes in isolation, many lack compliance with recent standards, but more importantly, lack the flexibility to customise and link them together.

In this paper, we present the Content Augmentation (CA) Manager Framework which is capable of performing and controlling the process of semantic annotation and ontology population by means of *consolidation algorithms*. CA Manager allows easy plug-in of various types of components including Information Extraction (IE) tools, customised domain ontologies, and diverse semantic repositories.

2. THE CA MANAGER FRAMEWORK

The core philosophy of the CA Manager is to bridge the gap between the content augmentation tools, and the semantic repository tools. It is conceived as a middleware, capable of controlling the quality and the validity of IE results against an ontology, matching them with existing resources (e.g. the application's knowledge base or repositories from the Linked Open Data initiative (`linkeddata.org`), and enriching them. To achieve that goal, the CA Manager relies on the recommendations formulated by the Semantic Web community (RDF/OWL languages, Service Oriented Architecture) combined with the UIMA-based infrastructure which have been enriched and customised to address the specific needs of semantic annotation and ontology population tasks.

The CA Manager proposes a list of logical steps arranged in a workflow, see Figure 1: a) *Extracting* the valuable knowledge and annotating the content; b) *Con-*

[1] www3.gartner.com/DisplayDocument?id=379859

K-CAP'09, September 1–4, 2009, Redondo Beach, California, USA.

solidating knowledge with regards to the ontology model and the semantic repository; c) *Serialising* the type-system output in various formats and *storing* it in the semantic repository.

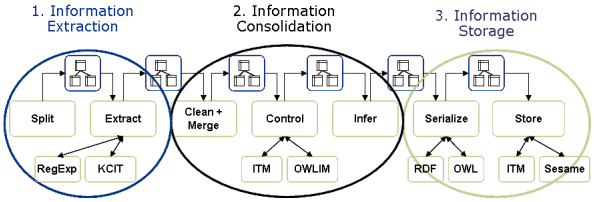

Figure 1: Specialised UIMA processing pipeline

The pipeline described above is exposed as a web service, and a testing web client application is available from `62.210.155.132/ca-test`. The CA Manager framework is open-source and available from: `sourceforge.net/projects/scan-ca-manager`. More details about CA Manager is available in [1].

3. EVALUATION

CA Manager was developed in the course of the TAO project (`www.tao-project.eu`), but is now used in several others. We implemented different workflows (a combination of different ontologies, semantic annotation tools, semantic repositories and corpora) for evaluating the *flexibility* and the *scalability* of the CA Manager framework, see Table 1. IE tool can either call a natural language processing tool or any kind of knowledge provider such as the semantic databases available within the Linked Open Data community. This particularly shows how the CA Manager is flexible to adapt its generic framework to all kinds of application needs.

Table 1: CA Manager workflows

ontology	corpus	CA tool	Repository
Architectural ontology (3D objects)	3D objects	DBPedia and Geonames web services	ITM
Adverse Drug Effect ontology	PubMed abstracts	Luxid (Temis)	ITM
Tourism ontology	Touristic web sites	TimeFrame (Modyco, Univ Paris X)	ITM
MiRNA ontology	PubMed articles	FunGen Discovery (INSERM)	Sesame

One of the workflows trialed during the TAO project was used to evaluate consolidation algorithms [1] of CA Manager. Namely, we used ontology-based IE tool KCIT [3], but one semantic repository being the ITM (see `http://mondeca.com/`) and the other one being the Sesame RDF repository. Using KCIT, we annotated 20 documents about GATE [2], with regards to the domain ontology (`gate.ac.uk/ns/gate-kb`). As consolidation is entirely dependent on the quality of the IE output, we have first measured precision and recall values for KCIT in isolation, by comparing automatically processed with human-annotated corpus. Then we calculated the performance of the consolidation algorithms based on the same settings.

The CA Manager obtains 100% recall hence there is no loss of information after processing the KCIT annotations to produce the final semantic annotations and knowledge instances. The CA Manager consolidation algorithms that deal with merging, need to be improved in order to eliminate duplicated terms with different orthographic labels such as "datastore", "data store", and "DATASTORE". On the other hand, the consolidation algorithms that control the ontology model are performing nicely which is not very difficult in the GATE case study as we mostly refer to the class instances. There are no annotations which refer to the relations between knowledge instances for example. In future, we plan to improve the linguistic analysis to support that feature.

4. CONCLUSIONS

We presented the CA Manager framework which serves as a mediator between semantic annotation and ontology population, and is capable of consolidating and controlling this process while allowing human annotators to be involved, if required. We created various workflows to evaluate the flexibility and scalability of this framework, which is based on Semantic Web standards and UIMA concepts. The main contribution of the CA Manager in comparison to other similar tools is that it allows easy plug-in of any IE tool, semantic repository, ontology or corpus, while also applying its own consolidation algorithm in order to link IE phase with ontology population.

5. ACKNOWLEDGMENTS

This research was supported by the EU-funded TAO (FP6-026460), MUSING (FP6-027097), and ServiceFinder (FP7 215876) projects.

6. REFERENCES

[1] F. Amardeilh. Semantic annotation and ontology population. In J. Cardoso and M. Lytras, editors, *Semantic Web Engineering in the Knowledge Society*. Idea Group Publishing, 2008.

[2] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[3] D. Damljanovic, V. Tablan, and K. Bontcheva. A text-based query interface to owl ontologies. In *6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, May 2008. ELRA.