DAFOE: A Platform for Building Ontologies from Texts

Sylvie Szulman
CNRS/LIPN et Université Paris 13, France ;
ss@lipn.univ-paris13.fr

Nathalie Aussenac-Gilles
CNRS/IRIT et Université de Toulouse, France ;

Adeline Nazarenko
CNRS/LIPN et Université Paris 13, France ;

Henry Valéry Teguiak
LISI-ENSMA et CRITT-Informatique ;

Eric Sardet
LISI-ENSMA et CRITT-Informatique ;

Jean Charlet
INSERM, France.
jean.charlet@spim.jussieu.fr

ABSTRACT

Although text-based ontology engineering gained much popularity in the last 10 years, very few ontology engineering platforms exploit the full potential of the connection between texts and ontologies. We propose DAFOE, a new platform for building ontologies with a terminological component using different types of linguistic entries (text corpora, results of natural language processing tools, terminologies or thesauri). DAFOE supports knowledge structuring and conceptual modelling from these linguistic entries as well as ontology formalization. DAFOE outputs models with two main original features: an ontology articulated with a lexical component and a connection with the text or linguistic entry that motivated their definition.

Categories and Subject Descriptors

D.2.11 [**Software Architectures**]; I.2.4 [**Knowledge Representation Formalisms and Methods**]; H.2.1 [**Database Management**]: Logical DesignData models

General Terms

Design

Keywords

Ontology Building, Ontology Editor, Meta-Modelization, Data Model

1. INTRODUCTION

DAFOE[1] is a new platform for building ontologies using different types of linguistic entries (text corpora, results of natural language processing tools, terminologies or thesauri). DAFOE supports knowledge structuring and conceptual modelling from these linguistic entries as well as ontology formalization. DAFOE outputs models with two main original features: an ontology articulated with a lexical component and a connection with the text or linguistic entry that motivated their definition. The requirements of the platfom and its development focus 1) on integrating various kinds of tools currently used within a single modelling platform, 2) on guaranteeing persistence and traceability of the whole ontology building process, and 3) on developing the platform in an open source paradigm with possible plugin extensions.

[1] http://dafoe4app.fr

EKAW 2010 Lisbon, Portugal

2. TEXT-BASED ONTOLOGY ENGINEERING

There is a growing interest for ontologies and related tools, including Ontology Engineering Environments. Many of them are pure ontology editors that support the development of formal ontologies but do not assist the tasks of knowledge acquisition or structuring. Knowledge engineers are supposed to have a first of ontology draft before using such tools. Since 2003, a significant shift occurred. Firstly, a parallel has been established between ontology population from text and semantic (textual) annotation. Secondly, many projects have proved the benefit brought by Human Language Technologies (HLT), including NLP, Information Extraction, Knowledge Discovery or Text Miming, for complementarity activities such as ontology learning from text and ontology population. The diversity and richness of existing HLT tools as well as the complexity of the ontology development tasks underlined the need for tool suites and platforms where the knowledge engineer can define its modelling strategy.This challenge is also one major motivation of the DAFOE project but the platform targets more ambitious goals: a better interoperability, a higher robustness and an easier combination of HLT and ontology technologies.

The goal of DAFOE is both to extend the variety of HLT that can be used and to support scalable ontology engineering. It claims that there are several ways to get an ontology, and that tools and processes must be selected according to each ontology case-study. DAFOE will propose tools similar to those of Text2Onto, but human supervision will play a major role for selecting tools, validating their results and conceptualizing. Knowledge conceptualization requires that a human selects and organizes properly concepts and relations, but this process can be guided. The result of DAFOE will typically be a termino-ontological resource where the ontology is connected to a lexical component.

3. DATA MODEL

DAFOE data model has to take into account various ontology building strategies, whatever information source (texts, terminologies, thesauri or human expertise) is used.

3.1 Overall Architecture

The data model is based on a valid methodology for building ontologies from texts, which has inspired tools such as Terminae [2] or Text2Onto [3]. This methodology takes into account the whole process of ”transforming” textual data into ontologies and split it into different phases, which correspond to various input levels if one wants to start with a thesauri rather than text, for instance. This methodologies relies on two main ideas: 1/ textual data are an important information source to build ontologies, especially if the ontology is to be used to annotate textual documents but 2/ textual data cannot be mapped directly into an ontology and the transformation must be mediated. The data model is therefore structured into four layers as represented in Figure 1. Each one corresponds to a specific methodological step.

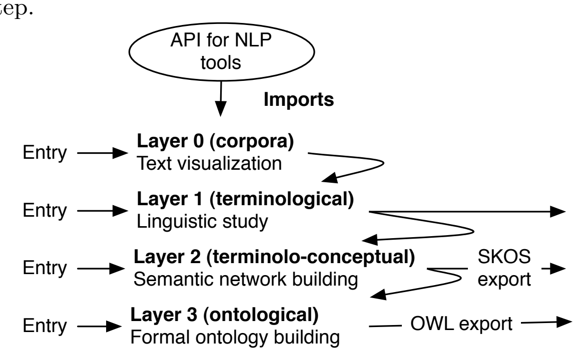

Figure 1: Data model architecture.

3.2 Corpora Layer

The corpora layer is useful for the knowledge engineer willing to build an ontology from text. He/she can build a working corpus by selecting different source documents and browse that corpus, either as plain documents or as segmented ones. In the data model the corpus is represented as a sequence of sentences, each one having a unique identifier.

3.3 Terminological Layer

The terminological layer gives a view over the domain specific lexicon of the corpus. It gathers the terms of the domain and their relationships. Terminological knowledge is traditionally produced by NLP tools such as term extractors applied on the working corpus. The underlying assumption is threefold: text analysis can extract term candidates that are relevant for a given domain, those terms are likely to be turned into ontology concepts and the distribution of these terms reflects their semantics [4]. DAFOE visualizes results of NLP tools such as YaTeA term extractor [1]. NLP results are given to DAFOE through an API. The data model is extensible and may be adapted to different NLP tools.

3.4 Termino-Conceptual Layer

This layer represents a semantic structure of unambiguous termino-concepts (TC) and termino-conceptual relations (RTC). The knowledge engineer may build that layer by importing a preexisting termino-conceptual resource such as a thesaurus or out of the analysis of the terminological layer. In that case, he/she analyses the meaning of terms and relations that appear at the terminological layer with respect to each other and by looking at their occurrences. The termino-conceptual layer is pivotal for transforming linguistic elements into conceptual ones and tracing the ontology back to the linguistics.

3.5 Ontology Layer

The ontology data model allows to formalize TCs and RTCs in a formal language equivalent at OWL-DL. Concepts are described as classes, individuals as instances of classes, properties between classes as object properties and properties between a class and a value as data properties or attributes. An automatic process will translate TCs and RTCs into formal concepts in a hierarchy with inherited properties as usual subsumption in description language. This translation exploits the structure of the semantic network represented in the termino-conteptual layer and the differential criteria associated with TCs and RTCs.

4. CONCLUSION

A prototype of the DAFOE platform has been implemented. DAFOE is intended to provide a variety of ontology engineering methods. As such a diversity can not be managed in a unique and static model, we adopted an extended Ontology-Based Database (OntoDB) architecture that supports model management and plugins. The strength of DAFOE approach is i) a precise definition of the various steps by which one can design a formal ontology; ii) a data model guaranteeing persistence and traceability of the whole ontologie building process; iii) the supply of flexible methodological guidelines that support the knowledge engineer without constraint; iv) an architecture based on the MOF model and plugins adaptability to ensure extensibility of the model and processes around a core tool; v) the specification of various modelling strategies based on different input/output of the platform; vi) the final production of an ontology which is associated to a terminological component.

5. REFERENCES

[1] S. Aubin and T. Hamon. Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, number 4139 in LNAI, pages 380–387. Springer, August 2006.

[2] N. Aussenac-Gilles, S. Despres, and S. Szulman. The Terminae method and platform for ontology engineering from texts. In P. Buitelaar and P. Cimiano, editors, *Bridging the Gap between Text and Knowledge: Selected Contributions to Ontology learning from Text.* IOS Press, 2008.

[3] P. Cimiano and J. Volker. Text2onto - a framework for ontology learning and data-driven change discovery. In A. Montoyo, R. Munoz, and E. Metais, editors, *Proc. of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238, Alicante, Spain, 2005. Springer.

[4] Z. Harris. *Mathematical Structures of Language.* Interscience Publishers, 1968.