A semi-automatic approach for building ontologies from a collection of structured web documents

Mouna Kamel, Nathalie Aussenac-Gilles, Davide Buscaldi, Catherine Comparot

A semi-automatic approach for building ontologies from a collection of structured web documents

Mouna Kamel
Université de Toulouse,
IRIT, CNRS UMR 5505
118 route de Narbonne
31062 Toulouse - France
kamel@irit.fr

N. Aussenac-Gilles
Université de Toulouse,
IRIT, CNRS UMR 5505
118 route de Narbonne
31062 Toulouse - France
aussenac@irit.fr

Davide Buscaldi
LIPN-Univ. Paris Nord
av.Jean-Baptiste Clément
93430 Villetaneuse - France
davide.buscaldi@lipn.univ-paris13.fr

Catherine Comparot
Université de Toulouse,
IRIT, CNRS UMR 5505
5 Allées Antonio Machado
31100 Toulouse - France
comparot@irit.fr

ABSTRACT*

Many collections of structured documents are available on the web. The collection generally describes the characteristics of entities from a single type, where each page describes one entity. These documents are adequate knowledge sources for building ontologies. As they benefit from a strong and shared layout, they contain less well written text than plain text files but their architecture is very meaningful. Classical linguistic-based methods for identifying concepts and relations are no longer appropriate for analyzing them. The approach we propose in this paper exploits various properties of such documents, combining layout/formatting analysis and linguistic analysis, and using semantic annotation.

General Terms

Algorithms, Design, Experimentation.

Keywords

Ontology building, ontology enrichment, document structure, document layout..

1. INTRODUCTION

Among the large quantity of documents that can be found in the web, some of them are well structured and organized in browsable collections, describing the characteristics of entities from a single type, where each page describes one entity. In order to make this knowledge accessible to a wide audience, filled forms or data sheets are often used. These documents have the property to present information in a synthetic manner. Their layout plays a crucial role for their meaningfulness. Classical methods for identifying concepts and relations are no longer appropriate for this kind of documents. Current implementations of these methods only work when syntactic parsers produce relevant analyses [1,2]. Our goal is to take advantage of a document layout and structure to get additional clues for knowledge extraction, to improve the information extraction process just like Role and Rousse [3] and O'Connor et al. [4] did it for XML documents. Therefore, we propose an approach for building an ontology from a collection of structured web documents such as filled forms. It is based on prerequisites about the content of the documents in the collection:

- all documents describe entities of a single domain;
- each of the documents describes one entity, which is a subset or a member of a more general domain concept.

We assume that the document layout is such that:

- all the documents in the collection exhibit a large degree of regularity that may be exploited for transforming them into files compliant with a same model (DTD, XML schema).
- most of the properties are specified in the form of <attribute, content> pairs.

The ontology building process takes place into three main steps: (1) pre-processing the document, (2) building a kernel ontology, and (3) automatically enriching the kernel with concepts and restrictions on relations.

2. THE ONTOLOGY BUILDING PROCESS

A filled form of interest generally describes one entity. It contains at least the denomination of this entity (*title* document section) and information about its properties (as a list of *fields*). Each field is composed of an attribute and some content. The domain described by the set of documents is formalized with a concept termed *main concept*, and each entity described within each document is formalized with a concept termed *pivot concept*.

2.1 Document Pre-Processing

The purpose of this two-stage step is to provide a unified XML representation of each document of the collection that will make it easier to extract knowledge from each document. *Identification of lay-out semantic features*: an expert analyses the document to define the set of typo-dispositional and lexical markers [6] that characterize titles, field names and field contents. He then defines a set of extracting rules (patterns or XSLT transformations) including these markers. *Field content segmentation* into Elementary Text Units (ETUs) which make sense: sentence chunking is used to identify different relations or values in a single field. We consider as an ETU a chunk delimitated by punctuation marks or by lexical markers such as conjunctions *or* and *and*. To extract them, the expert may use a set of patterns we defined in [5] to isolate the different elements of a list.

2.2 Building a Kernel Ontology

The expert defines sets of fields according to the type of knowledge to be found in each of them (the same field may belong to several sets):

- $\mathcal{F}_O$ as the set of fields holding relations linking individuals to individuals; these fields will provide objectProperties in the OWL representation of the ontology.

- $\mathcal{F}_D$ as the set of fields holding relations linking individuals to data values; these fields will provide dataTypeProperties in the OWL representation of the ontology.
- $\mathcal{F}_L$ as the set of fields holding terms denoting the *pivot* concept of the document. $\mathcal{F}_L$ will contain at least one field name, i.e. the <Title> field. They will be represented with the standard rdfs:label property.

2.3 Enriching the Kernel Ontology

The enrichment process requires to define new concepts and to add relation restrictions to the ontology. Let's first give the following definitions: $\mathcal{D}$ is a set of pre-processed documents; f is a field name defined in the document XML schema; $\mathcal{T}_d^f$ is the set of ETUs extracted from field f in document d; $\mathcal{T}_d$ is a set of ETUs extracted from document d; c_p^d is the pivot concept of document d; c_m is the main concept of the collection; $\mathcal{R}_f$ is the set of relations held by field f; *is-a* (c_1, c_2) is the hypernymy relation between two concepts (c_1 is a direct sub-concept of c_2).

In the enrichment algorithms, we use the following functions: *createConcept* (c, Id) creates a new concept c identified with Id; *subsumes* (c_1, c_2) is true if c_2 is a sub-concept of c_1; *build* (f, d) creates $\mathcal{T}_d^f$, a list of ETUs from f in d; *addLabel* (c, l) adds a new label l to an existing concept c; *addRelation* $(r\ (c_1, c_2))$ adds a new relation r between concepts c_1 and c_2; *domain* (r) returns the domain of the relation r; and *range* (r) returns the range of the relation r.

The following algorithm describes how new concepts are added to the ontology:

```
for each d ∈ D
   for each f ∈ F_L
        T_d^f ← build (f, d)
        T_d ← T_d ∪ T_d^f
   end for each
   createConcept (c_p^d, l1) where l1 is the label extracted from the
                              <title> field and belonging to the T_d list
   addRelation(is-a (c_p^d, c_m))
   for each l ∈ T_d, addLabel (c_p^d, l) end for each
end for each
```

One *pivot concept* is created for each document, which is linked to the *main concept* with the relation *is-a*. All the terms extracted from the fields in $\mathcal{F}_L$ are added as labels of this new concept.

The following principle is used to extract a relation r from document d:

$\forall f \in \mathcal{F}_{O,}\quad \forall c$ annotating f, $\forall r \in \mathcal{R}_f$

is-a(c_p^d, domain(r),) and subsumes (range(r), c)

$\rightarrow$ addRelation(r (c_p^d, c))

3. EVALUATION

In order to carry out an evaluation of the enrichment process, we decided to compare ontologies enriched according to this process and manually. The two approaches start from the kernel ontology. A random set 20 documents from the botanic encyclopedia "Jardin! L'encyclopédie" [†] has been processed by an ontologist to build a reference ontology, which has been compared with the result of the automatic enrichment process applied to the same kernel ontology and using the same text collection.

The assessment we made on these 20 documents gives the following results: 248 restrictions have been correctly detected, 76 restrictions have not been found, and 62 restrictions have been wrongly detected. We obtain a Recall value of 0.76 and a Precision value of 0.8. The most recurring linguistic problem is the negation problem, which we do not take into account for the time being.

4. CONCLUSION AND FURTHER WORK

Our contribution to relation extraction for ontology engineering explores some of the gains brought by the use of textual layout, and the semantics that it conveys, to identify relations that would have been missed by the analysis of the language in text. As a further work, we plan to improve the NLP chain to better process negations and intervals, by integrating existing work carried out in our group [6]. We plan also to evaluate our work on a larger data set in the context of the BioNLP challenges. For instance, the GRO task[‡] of the 2013 challenge offers both an ontology and a scientific corpus.

Acknowledgements

We thank our colleagues and MOANO[§] partners from LIUPPA (M.-N. Bessagnet, A. Royer and C. Sallabery) for their valuable feed-back when building and enriching the ontology.

5. REFERENCES

[1] Navigli, R., Velardi P. 2006. Ontology enrichment through automatic semantic annotation of online glossaries. In S. Staab et V. Svàtek (Eds.), *15th International Conference EKAW 2006, Volume LNCS 4248*, 126–140. Springer.

[2] Schutz, A., Buitelaar P. 2005. Relext: A tool for relation extraction from text in ontology extension. In *4th International Semantic Web Conference (ISWC 2005),* Volume 3729, 593–606. Springer: Berlin.

[3] Role, F., Rousse G. 2006. Construction incrémentale d'une ontologie par analyse du texte et de la structure du document. *Document numérique* 9(1), 77–91.

[4] O'Connor, J., K. M., Das A. 2011. Acquiring owl ontologies from xml documents. *International Conference on Knowledge Capture KCAP 2011.*

[5] Kamel M., Aussenac-Gilles N., Laignelet M. 2010. Correction d'ontologies construites à partir de la structure de documents. In *Journées Francophones d'Ingénierie des Connaissances (IC 2010),* Nîmes (France), S. Despres (Eds.).

[6] Benamara F., Chardon C., Mathieu Y., Popescu V., Asher N. 2012 How do Negation and Modality Impact on Opinions? *Extra-propositional aspects of meaning in computational linguistics - Workshop at ACL 2012, Jeju Island, Korea*, 8-17.

[†] http://nature.jardin.free.fr/
[‡] http://nlp.sce.ntu.edu.sg/wiki/projects/bionlpst13grotask/
[§] http://moano.liuppa.univ-pau.fr