Ontology-Based Information Extraction for French Newspaper Articles

Nebhi, Kamel

Ontology-Based Information Extraction for French Newspaper Articles

Kamel Nebhi

LATL, Department of linguistics
University of Geneva
Switzerland
kamel.nebhi@unige.ch

Abstract. In this paper, we describe a rule-based approach to perform automated semantic annotation of named entities in a corpus of newspaper articles. The originality of our system is in the fact that it establishes a connection between the French named entity, the DBpedia ontology and the DBpedia databank. We present our system, discuss its architecture and report the first evaluation results.

Keywords: Ontology-based Information Extraction, Semantic Web, Linked Data.

1 Introduction

The goal of the Semantic Web, as described by Tim Berners-Lee [1], is to bring meaning to the Web, creating an environment where software agents can readily carry out sophisticated tasks of users. Thus, the realization of this Web of data on a large scale implies the widespread annotation of Web documents with ontology-base knowledge markup.

In this paper, we present an Ontology-based Information Extraction (OBIE) system for French newspaper articles using a rule-based approach. Our system establishes relation between named entities in a text, the ontological standardized semantic content of the DBpedia ontology and the DBpedia databank.

This article is structured as follows : section 2 defines Ontology-based Information Extraction; section 3 describes the proposed system architecture. In section 4, we present the first evaluation results. We conclude and give some perspectives in section 5.

2 OBIE

Information Extraction (IE) is a key NLP technology to introduce supplementary information and knowledge into a document. The term "Ontology-based Information Extraction" has been conceived only a few years ago and has recently emerged as a subfield of IE. OBIE is different from traditional IE because it finds type of extracted entity by linking it to its semantic description in the

B. Glimm and A. Krüger (Eds.): KI 2012, LNCS 7526, pp. 237–240, 2012.

formal ontology. The task of OBIE has received a specific attention in the last few years [9] with many publications that describe systems. Several of these systems have not been integrated in the general schema of Semantic Web and are essentially developed for English documents. To solve this problem, we propose an OBIE system for French that uses *Linked Data* such as DBpedia databank.

3 System Description

Our OBIE system is built on GATE [3] to annotate entities in text and relate them to the DBpedia ontology[1] where appropriate. The DBpedia ontology is a shallow, cross-domain ontology, which has been manually created based on the Wikipedia projects. The ontology organizes the knowledge according to a hierarchy of 320 classes and 1650 different properties.

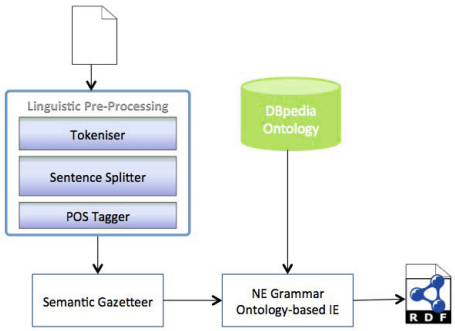

Fig. 1. Ontology-based Information Extraction Architecture

Figure 1 describes the architecture of our OBIE system. The source data is a set of newspaper articles from *LeMonde.fr*. Semantic annotation is performed by GATE with respect to the DBpedia ontology. The GATE application consists of a set of processing resources executed in a pipeline over a corpus of documents. The pipeline consists of 4 parts :

- Linguistic pre-processing
- Gazetteer (used to identify entities directly via look-up)
- Rule-based semantic annotation
- Final output creation

The linguistic pre-processing phase contains GATE components such as tokenisation and sentence splitter. It also contains specific tools like TreeTagger for French part-of-speech tagging. The gazetteer lookup phase comprises combination of default gazetteer lists from ANNIE[2] and some newly gazetteer lists

[1] `http://wiki.dbpedia.org/Ontology`

[2] GATE is distributed with an IE system called ANNIE (A Nearly-New IE system). It comprises a set of core processing like tokeniser, sentence splitter, POS tagger, Gazetteers, JAPE transducer, etc.

extract from Wikipedia and DBpedia. The grammar rules for creating semantic annotation are written in a language called JAPE [4] which is a finite state transducer. The rules are based on pattern-matching using several informations taken from the gazetteer or the part-of-speech tags. In total, the application contains approximately 100 grammar rules.

For example, the rule of the Figure 2 is used to identify a city directly via look-up in gazetteer. So the string "Liverpool" found in the text might be annotated with the features :

```
class : http://dbpedia.org/ontology/City
inst. : http://dbpedia.org/ontology/#Liverpool
linked-data : http://dbpedia.org/data/Liverpool
```

```
Rule: CityLookup
({Lookup.minorType == city}):cityLabel
-->
:cityLabel {
    String city = stringFor(doc, cityLabelAnnots);
    String baseUri = "http://dbpedia.org/"
    newFeatures.put("class", baseUri + "ontology/City");
    newFeatures.put("inst", baseUri + "ontology/#" + city);
    newFeatures.put("linked-data", baseUri + "data/" + city)
        ;
}
```

Fig. 2. An example of a JAPE rule

4 Experience

Traditional IE systems are evaluated using Precision, Recall and F-Measure. These measures are inadequate when dealing with ontologies. In order to take ontological similarity into account our OBIE system was evaluated using the Balanced Distance Metric [6]. To evaluate the performance of the system we applied the processing resources on the evaluation corpora of 40 newspaper articles

Table 1. Results

	F_1	BDM_F_1
Location	0.92	0.94
Organization	0.91	0.95
Person	0.90	0.94
Total	0.91	0.94

of *LeMonde.fr*. We manually annotated these documents with the concepts of the DBpedia ontology. Then, we compare the system with the gold standard. For the evaluation, we only use Person, Organization and Location named entity categories. In table 1, the system achieved a traditional F-Measure of 91% and an augmented F-Measure of 94%.

5 Conclusion - Further Work

In this paper we have presented an Ontology-based Information Extraction system for French newspaper articles. We have successfully integrated the system in the general schema of Semantic Web using *Linked Data*. As our evaluation shows, performance measured through BDM look promising.

In future work, we intend to provide deeper linguistic processing with the Fips analyzer [8]. We also try to integrate the application into a ReSTful Web service [7].

References

1. Berners-Lee, T., Fischetti, M.: Weaving the web: The original design and ultimate destiny of the World Wide Web by its Inventors. Harper, San Francisco (1999)
2. Brewster, C.: Natural Language Processing as a Foundation of the Semantic Web. Now Publishers Inc., Delft (2009)
3. Cunningham, H., et al.: Text Processing with GATE (Version 6). University of Sheffield (2011)
4. Cunningham, H., Maynard, D., Tablan, V.: JAPE: a Java Annotation Patterns Engine. Technical report, University of Sheffield (2000)
5. Handschuh, S., Staab, S.: Annotation for the Semantic Web. IOS Press, Amsterdam (2003)
6. Maynard, D., Peters, W., Li, Y.: Evaluating Evaluation Metrics for Ontology-Based Applications: Infinite Reflection. In: Proc. of 6th International Conference on Language Resources and Evaluation (LREC), Marrakech (2008)
7. Richardson, L., Ruby, S.: RESTful Web Services. O'Reilly (2007)
8. Wehrli, E.: Fips, a deep linguistic multilingual parser. In: ACL 2007 Workshop on Deep Linguistic Processing, Prague, Czech Republic (2007)
9. Wimalasuriya, D.C., Dou, D.: Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches. Journal Inf. Science (2010)