Ontologies as a Source for the Automatic Generation of Grammars for Information Extraction Systems

Thierry Declerck and Paul Buitelaar

DFKI GmbH, Language Technology Lab,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
declerck@dfki.de
Unit for Natural Language Processing, DERI,
National University of Ireland, Galway
paul.buitelaar@deri.org

Abstract. Grammars for Natural Language Processing (NLP) applications are generally built either by linguists – on the basis of their language competence, or by automated tools applied to existing large corpora of language data — using either supervised or unsupervised methods (or a combination of both). Domain knowledge usually played just a little role in this process. The increasing availability of extended knowledge representation systems, like taxonomies and ontologies, is giving the opportunity to consider new approaches to the (automated) generation of processing grammars, especially in the field of domain-oriented Information Extraction (IE). The reason for this being that most of the taxonomies and ontologies are equipped with natural language expressions included in ontology elements like labels, comments or definitions. These de facto established relations between (domain) knowledge and natural language expressions can be exploited for the automatic generation of domain specific NLP and IE grammars. We describe in this paper steps leading to this automation.

Keywords: Ontology-based Information Extraction, Grammar Generation, Business Reporting Standards

1 Introduction

In the last 10-15 years we have experienced a huge increase of available knowledge sources of various types, like taxonomies or ontologies, which are also available online. The more recent establishment of the linked (open) data framework[1] has further boosted this development, making available a tremendous amount of (interlinked) knowledge objects in the web. Some formal and logic-based knowledge representation languages like RDF(s) and OWL[2], which are used for encoding

[1] See http://linkeddata.org/

[2] RDF(s) stands for "Resource Description Framework (schema)" and OWL for "Web Ontology Language". See http://www.w3.org/TR/rdf-schema/ and http://www.w3.org/TR/owl-features/ respectively.

these knowledge objects, have foreseen various possibilities to include natural language expressions.

These expressions can be part of RDF URI references, identifying ontological resources (e.g. natural language string used in `rdf:ID`), a fragment (e.g. natural language string in `rdf:about` statements) or marking empty property elements (kind of leaf nodes in a graph, using the `rdf:resource` statement). Examples of the use of such reference elements in ontologies are given below[3]:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:ex="http://example.org/stuff/1.0/"
         xml:base="http://example.org/here/">
  <rdf:Description rdf:ID="snack">
    <ex:prop rdf:resource="fruit/apple"/>
  </rdf:Description>
</rdf:RDF>
```

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Seq rdf:about="http://example.org/favourite-fruit">
    <rdf:_1 rdf:resource="http://example.org/banana"/>
    <rdf:_2 rdf:resource="http://example.org/apple"/>
    <rdf:_3 rdf:resource="http://example.org/pear"/>
  </rdf:Seq>
</rdf:RDF>
```

Natural language expressions can also be used in taxonomies and ontologies as the content of RDF annotation properties, like `rdfs:label` and `rdfs:comment`, as this is exemplified below[4]:

```
<rdf:Property ID="hasAccessTo">
  <rdfs:label xml:lang="en">has access to</rdfs:label>
  <rdfs:comment xml:lang="en">Relates an Access Rule
     to the resources to which the rule applies.
     The inverse relation is 'accessedBy'</rdfs:comment>
  <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
  <rdfs:domain rdf:resource="#ResourceAccessRule"/>
  <rdfs:isDefinedBy resource="http://www.w3.org/2001/02/acls/ns#"/>
</rdf:Property>
```

In this paper, we focus on the content of annotation properties since they contain "real" natural language expressions. And additionally, labels and comments locally support multilingualism by means of language tags of RDF literals, i.e. `xml:lang`, whereas this is not the case for RDF URI references.

[3] Examples are taken from the (revised) RDF/XML Syntax Specification, a W3C Recommendation from 2004/02/10, see `http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/`.

[4] The slightly modified example is taken from `http://www.w3.org/2001/Talks/0710-ep-grid/slide21-0.html`.

2 Labels as a Source for Ontology-Based Information Extraction

The use of labels in knowledge representation systems is nowadays widely supported, as can be seen for example in the XBRL[5] taxonomies representing different legislations for business reporting, in the FMA ontology[6] for human anatomy or in the RadLex ontology[7] encoding radiology terms. Figure 1 shows an example from the XBRL taxonomy of the Belgian National Bank, where the reader can see the use of labels, also including the xml:lang feature, in a multilingual setting, relating natural language expressions to the concept `FixedAssets`.

```
<loc xlink:label="FixedAssets_loc" xlink:type="locator"
 xlink:href="pfs-2011-04-01.xsd#pfs_FixedAssets" />
<labelArc xlink:from="FixedAssets_loc" xlink:to="FixedAssets_lab"
 xlink:type="arc"
 xlink:arcrole="http://www.xbrl.org/2003/arcrole/concept-label" />

<label xlink:label="FixedAssets_lab" xlink:type="resource"
 xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="fr">
 Actifs immobilisés</label>
<label xlink:label="FixedAssets_lab" xlink:type="resource"
 xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="nl">
 Vaste activa</label>
<label xlink:label="FixedAssets_lab" xlink:type="resource"
 xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="de">
 Anlagevermögen</label>
<label xlink:label="FixedAssets_lab" xlink:type="resource"
 xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="en">
 Fixed assets</label>
```

Figure 1: Example of multilingual labels in the taxonomy for business reporting of the Belgian National Bank.

This combination of domain-specific knowledge and terms suggests that the main task of Ontology-Based Information Extraction (OBIE) is consisting now in the mapping or unification of language data available in ontologies/taxonomies and language data as available in running text. While the labels in the example displayed in Figure 1 are very simple and finding matches in running text would be straightforward, let us just give another example for showing the complexity of the task of matching labels in taxonomies and relevant expressions in free text, considering another concept of the Belgian National Bank taxonomy (displaying only the English label):

---

[5] XBRL (eXtended Business Reporting Language) "is a language for the electronic communication of business and financial data", see `http://www.xbrl.org/GettingStarted`.

[6] See `http://sig.biostr.washington.edu/projects/fm/`.

[7] See `http://bioportal.bioontology.org/ontologies/42801`.

```
<label xlink:label="GuaranteesGivenByThirdPartiesBehalfEnterprise_lab"
 xlink:type="resource" xlink:role="http://www.xbrl.org/2003/role/label"
 xml:lang="en">Guarantees given by third parties on behalf of the
 enterprise</label>
```

Figure 2: Example of a more complex label used in the taxonomy of the Belgian National Bank.

The IE task consists in mapping the text "Guarantees given by third parties on behalf of the enterprise" to a running text, in which the choice and the order of words is not corresponding to the term used in the label. And also one needs to find the naming of an enterprise in the text and so to provide for an instance to the abstract notion of "enterprise" in the term. In this, OBIE extends the typical task of IE in the fact that the extraction of information from documents is not done any more on the base of language knowledge and pre-specified domain templates, but that it is now relying on a central semantic dimension – an ontology or a taxonomy, which can have dynamic properties. OBIE needs thus both access to representations of linguistic units of interest (i.e. names, terms, phrases, relation-expressing verbs or nouns, prepositions, etc.), as sued in labels, and to semantically structured representations of object classes of interest (i.e. domain concepts and properties).

3 The *lemon* Model for the Representation of Language Data in Ontologies

In order to support the mapping of language data in ontology labels and the language data in running text, there is the need to render the language data comparable and thus to represent it in a standardized way. For this we propose as a first step the lexicalization of the labels of ontologies, associating to them typical linguistic information, like lemma, morphology, syntax. But there is also a need to state the particular relation of such lexicalized labels to the ontology elements they are associated with. We opted for solving this representation issue for the *lemon* (LExicon Model for ONtologies) model, which is designed to represent lexical information about linguistic units (e.g. words, phrases, terms) relative to an ontology.[8] *lemon* uses ontological elements, which are termed 'references', identified by URIs, to represent the semantics of a linguistic unit, i.e. by specifying a class or a property that can be linked to. The lexical information for the linguistic unit will then be expressed relative to this semantic 'reference', e.g. the fact that this ontology element will be typically realized by a noun "asset" with plural form "assets", modified by the adjective "fixed" (considering here our example in Figure 1). Similar for the more complex term "Financial liabilities at amortised cost", for which the *lemon* representation will encode the fact that the prepositional head has to be "at". *lemon* is based on a clear separation between semantics and morpho-syntax, while at the same time enabling the

[8] See [6].

precise specification of their correspondences. A simplified lemon entry ("asset") is shown in Figure 3.

```
:asset_noun lemon:canonicalForm [ lemon:writtenRep "asset" ] ;
    lemon:altForm [ lemon:writtenRep "assets"  ;
                lexinfo:number lexinfo:plural ] ;
    lexinfo:partOfSpeech lexinfo:noun .
    lemon:sense [ lemon:ref <http://www.ebr.org/xbrl#Asset> ] ;
```

Figure 3: A simplified representation of the lemon entry for *asset*. The semantics of the word (its "lexical sense"') is encoded by a reference to an XBRL concept.

Given this rich structure and content of *lemon* ontology-lexicons there is much potential for exploiting them in the generation of ontology-specific extraction grammars that capture semantic as well as linguistic aspects of the lables used in the domain ontology.

4 Exploiting Structural Information from the Ontology

Additionally to this representation of the lexical and linguistic content in RDF encoded *lemon* entries, we need to derive from the ontology element the *lemon* entry is referring to some structural information, so for example the class-hierarchy in which the element is introduced, and information about associated properties. So for example the xEBR[9] ontology we are working with is stating that " hasFixedAssetsTotal" is a property with domain class "FixedAssetPresentation", and has as range a monetary value, which we specified as an xsd type, which states that a monetary value consists of both an amount and a currency. The owl representation of this property is shown below in Figure 4:

```
<owl:DatatypeProperty
 rdf:about="http://www.dfki.de/lt/xebr.owl#hasFixedAssetsTotal">
 <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#monetary"/>
 <rdfs:domain
  rdf:resource="http://www.dfki.de/lt/xebr.owl#FixedAssetsPresentation"/>
 <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
 <rdfs:label xml:lang="en">fixed assets [total]</rdfs:label>
</owl:DatatypeProperty>
```

Figure 4: The owl-xml representation of the property "hasFixedAssetsTotal".

[9] xEBR is a core taxonomy for various national XBRL taxonomies. See http://www.monnet-project.eu/Monnet/Monnet/English/Navigation/XBRLEuropeanBusinessRegisterxEBR. In the project Monnet (www.monnet-project.eu/) , we have "upgraded" eXBR onto an ontology, and interfacing this ontology with other financial ontologies, derived from examples from stock exchange pages (Xetra, Euronext). eXBR, as well as Xetra and Euronext, are multilingual. See [5] for more details. Here we are dealing with English and German.

Since the class "FixedAssetPresentation" is in a part-of relation to the class "AssetsPresentation", being itself a subclass of "KeyBusinessFigureReport", which is defined for a specific duration, this temporal value is inherently valid for the property "hasFixedAssetsTotal"'. In this case the IE system knows that if in a document an expression is found that could be attached to the ontology element "FixelAssetTotal", the system also has to find as minimal condition, in relevant segments of the processed documents, expressions for both a monetary and a temporal value, in order to allow the system to populate the ontology with adequate values for the property.

5 Generating Grammars for the OBIE system

Combining both the *lemon* representation of the textual content of an ontology element label and the structural information about this element, a simple extraction grammar rule is being generated as follows[10]:

```
[fixed asset]
     N-plu & monetary-value & temporal-value
     => xebr:hasFixedAssetsTotal
```

The simplified rule specifies that if in a certain textual context (sentence, clause, or in a financial table) a natural language expression can be matched to the term "fixed assets"', while in the textual context also both a monetary and a temporal expressions can be found, then the whole construction can be marked as being of type "xebr:hasFixedAssetsTotal". The monetary and temporal values found in the text are used then for populated the ontology elements related to the report of the company under consideration. The compound noun"fixed asset" is represented here in its ground from (lemma), with the additional information that it should be used in the plural form. This rule can be applied to the following sentence from a financial report as follows, where we mark-up by in-line annotation the elements that lead to the detection of the XBRL concept in text and to the subsequent ontology population:

```
[In [2011](period) the Kuehne + Nagel Group invested
[CHF 207 million](monetary) in fixed assets] (xebr:hasFixedAssetsTotal)
```

A more complete version of this example involves the generation of an extraction rule that captures the relation between the actual monetary value mentioned in this sentence (CHF 207 million) and the property xebr:hasFixedAssetsTotal. This corresponds to the (simplified) following rule[11]:

[10] Those rules, partly derived from the *lemon* representation, are encoded in the NooJ formalism (`www.nooj4nlp.net/`). We present here a more readable pseudo-code. See also [3] for an overview on how ontologies can interact with NooJ linguistic data.

[11] The rule gives the impression that a fixed order for the semantic elements is reuird, but NooJ foresees a mechanism (the feature "ONCE") that allows to find the elements at most one time in a certain fragment, in no specific order.

```
[fixed asset]N-plu  .* [X](monetary) .* [Y](date)
                => xebr:hasFixedAssets[range = X]
```

Obviously this version of the analysis is also far from complete, as we need to represent also the linguistic and semantic aspects of a monetary value, of intermediate words and phrases (invested (date) in), and names (Kuehne + Nagel Group), in order not allow only semantic annotation of the text with relevant ontology elements, but also to populate the ontology, with the information that the company Kuehne + Nagel Group has a report covering the year 2011, with the additional information about the specific amount for fixed assets.

6 Actual Experiments

As a test for the derived IE grammars, we started to process the so-called "Portrait" of companies displayed in the bi-lingual web presence of the German Börse. Our ontology background is the MFO integrated ontology, as descibed in [5], and our main goal is to extract eXBR (Business Reporting) information, but the tools are also extracting information about activitiy fields of companies, etc. In a first experiment, we focus on the 30 companies, as they are listed in DAX. Once the system has been adjusted to the specific type of input, we will test the system on the 130 companies listed in the related stock exchanges SDAX, MDAX and TecDAX. First preliminary results are encouraging, since our system can for example extract the following information from the Portait of the DAX-listed company "adidas".[12]:

```
<DATE><DAY>31</DAY> <MONTH>12</MONTH> <YEAR>2011</YEAR></DATE>
<XEBR+NetTurnover+13,3 Mrd>
.....
<DATE>Zum  <DAY>31.</DAY>
 <MONTH><NP HEAD>Dezember</HEAD></NP></MONTH> <YEAR>2011</YEAR></DATE>
<EMPLOYEE+46.000>
```

The two sentences from which this information is extracted are: "Im Jahr 2011 erzielte die adidas Gruppe einen Konzernumsatz in Höhe von 13,3 Mrd." (*In 2011, the adidas Group generated net sales in an amount of 13.3 billion*) and "Zum 31. Dezember 2011 beschäftigte die adidas Gruppe mehr als 46.000 Mitarbeiter." (*Effective December 31, 2011, the adidas Group employed more than 46,000 people.*)

The most difficult part is to detect in the running text term variants for the labels of the xEBR labels. We are working on automated term extraction for supporting this task (see [4]). Please also note that we can specify the exact date for the xEBR concept, although the text just specify "2011", since in the Business Reporting domain the end of the year is the typical date. In other cases, the full date is always specified in the text.

[12] See `http://www.boerse-frankfurt.de/en/aktien/adidas+ag+DE000A1EWWW0/unternehmensdaten` for the English version and `http://www.boerse-frankfurt.de/de/aktien/adidas+ag+DE000A1EWWW0/unternehmensdaten` for the German version.

7 Conclusion

In this short paper we have described the process of generating automatically from ontologies grammar rules for information extraction systems. A pre-requisite for this is to perform a lexicalization of the labels of ontology elements and to represent this information using for example the *lemon* model. A *lemon* lexicon provides a very rich description of the language data used in labels and includes a reference to a semantic element (a class or a property in an ontology), a term structure (decomposition of the label into sub-terms), and linguistic knowledge (lexical features of word forms used in the term). Combined with structural information about the "location"' of the ontology element in the whole ontology, an efficient definition and even automatic derivation of extraction grammars is being proposed and currently tested.

Acknowledgments. This work is supported in part by the European Union under Grant No. 248458 for the Monnet project.

References

1. Aggarwal, N., Wunner, T., Arcan, M., Buitelaar, P., O'Riain, S.: A Similarity Measure based on Semantic, Terminological and Linguistic Information. In: Proceedings of the 6th InternationalWorkshop on Ontology Matching collocated with the 10th International Semantic Web Conference (ISWC-2011) Bonn, Germany (2011)
2. Declerck, T., Lendvai, P., Wunner, T.: Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems. In: Proceeedings of ISA6, ISO/ACL-SIGSEM Workshop, January 11-12, Oxford (2011)
3. Declerck, T. Lendvai, P., Váradi, T., Koleva, K.: Integration of Ontological Semantic Resources in NooJ. In: Proceedings of the 2011 International NooJ Conference. Cambridge Scholars Publishing (2012)
4. Federmann, C., Gromann, D., Declerck, T., Hunsicker, S., Krieger, H.U., Budin, G. Multilingual Terminology Acquisition for Ontology-based Information Extraction. In: Proceedings of the 10th Terminology and Knowledge Engineering Conference, Pages 166-175, Madrid, Spain (2012)
5. Krieger, H.K., Declerck, T., Nedunchezhian, A.K.: MFO - The Federated Financial Ontology for the MONNET Project. In: Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain (2012)
6. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In: In Proceedings of the 2011 Extended Semantic Web Conference (2011)
7. Wunner, T., Buitelaar, P., O'Riain, S.: Semantic, Terminological and Linguistic Interpretation of XBRL. In: Proceedings of EKAW, October 11-15, Lisbon (2010)