Heuristics for Automated Text-Based Shallow Ontology Generation

Mihaela Vela
DFKI Saarbrücken
Stuhlsatzenhausweg 3
66123 Saarbruecken, Germany
Mihaela.Vela@dfki.de

Thierry Declerck
DFKI Saarbrücken
Stuhlsatzenhausweg 3
66123 Saarbruecken, Germany
Thierry.Declerck@dfki.de

ABSTRACT

In this poster, we present the actual state of our work on the possible derivation of ontological structures from textual analysis. We propose an approach to the extension of existing domain ontologies or even to the semi-automatic ontology generation of such ontologies from scratch, on the base of heuristic rules applied to the result of a multi-layered processing of textual documents. This ongoing investigation is being pursued in the context of the MUSING R&D project[1], which is dealing with the use of semantic technologies for Business Intelligence applications.

1. INTRODUCTION

MUSING is an R&D European project dedicated to the development of Business Intelligence (BI) tools and modules founded on semantic-based knowledge and content systems. In the first phase of the project many efforts have been devoted to the manual creation of domain related ontologies and their integration in upper level ontologies (see [2]). These manually crafted ontologies can in the next future serve as a base for the evaluation of the MUSING task dedicated to ontology generation/acquisition from text. We can then compare ontology structures suggested on the base of text processing with the existing ontologies. Some work has already been described in the past on the role that can be played by textual analysis for supporting the acquisition of (shallow) ontologies (see among others [3]). One can apply statistical methods to linguistic annotation stored in a large database, and one can propose a set of (heuristic) rules for deriving ontology classes from linguistic annotation. The new classes can then be added to already existing ontologies or be proposed as the base for a new ontology. We discuss in more details in this paper an approach deriving ontology structures from scratch, on the base of heuristic rules applied to the results of textual processing.

2. OUR APPROACH

We consider in our experiment ontology generation (extraction or acquisition) as a linguistically rule-based approach to discovering and suggesting potential ontology classes and properties out of lexical and syntactic properties that can be detected in text and offer a pre-structuring for the detection of relevant semantic properties in the text. We propose a multi-layered approach, which starts with a very shallow analysis of certain lexical properties of words or very short combination of words, going from there to Part-of-Speech Tagging and morphological analysis, before using, in a next step, deeper syntactic analysis and taking into account larger parts of text, up to the level of sentences or even paragraphs. The idea behind this approach: at the shallow level it is already possible to detect possible classes and relations, which can then be consolidated, refined or rejected at further stages of analysis. We present in the following section a first prototype of automated ontology generation. At the actual stage of development our aim is to clearly state what kind of ontological resource can be extracted from financial documents (annual reports of companies, financial newspapers) at various level of textual processing. As a data source we work now with a corpus of economical news article from the German Newspaper "Wirtschaftswoche" from the year 1992.

2.1 String-Based Processing

We ran first a very simple algorithm on the whole corpus, looking for a small set of relevant words, and classifying them according to their string properties. We look whether the word occurs alone, or in the context of a compound word (as prefix or a suffix of the compound word). For example the German word "Konzern" (corporation) can appear in the following compounds, where we consider "Konzern" as being the keyword:

(1) Der größte deutsche Chemiekonzern
the largest German Chemical corporation
(2) PKI erstellte erstmals einen Konzernabschluss
PKI generated the first time a corporation report
(3) Der 75 jährige Konzernchef
The 75 year old chief of the corporation
(4) beim amerikanischen Johnson-Konzern
by the American Johnson corporation

In those cases we can already extract a lot of information that can be used as the basis of a proposal for an ontology. The compounded sequence *named_entity hyphen keyword* leads to the definition of an instance of an ontology class that could have "Konzern" as its label (or an alias); the compounded sequence "Konzern" *word* leads to a has-relation associated to the ontology class that could have "Konzern" as its label (or an alias): "Konzern" has "Chef"; the compounded sequence *word* "Konzern" leads to a subclass of the ontology class that could have "Konzern" as its label (or an alias). Here: "chemical corporation" is a subclass of the class "corporation". As attractive as this very simple approach might appear, serious drawbacks have to be expected concerning the generalization of rules (for

[1] see www.musing.eu for more details

the keyword "Chef" and the compound "Chefdenker" ("chief-thinker") we cannot enounce that there is a has-relation between ontology classes labeled "Chef" ("chief"), respectively "Denker" (thinker").

2.2 Using Morpho-Syntactic Information

A way to reduce the drawbacks of the approach described in section 2.1 lies in the use of morpho-syntactic information, as this one is typically delivered by a combination of a stem, a POS-Tagger and a morphological analyser. So for example the word "Gewinnkurve" ("curve of benefits") would be analyzed as the following:

(5) <W STEM="kurve" STTS_POS="NN" COMP="gewinn kurve" INFL=" [2 3 4 5]">Gewinnkurve</W>

This annotation is to read like this: the word "Gewinnkurve" has the stem "Kurve", has POS "noun", is the result of combining the word "Gewinn" and the word "Kurve", and has certain morphological properties (here encoded with numbers). One advantage of this approach is that no one has to give a list of words to be looked for as the basis of the procedure for ontology extraction. The whole corpus is being analyzed and all the compounds are recognized (to the level of accuracy authorized by the used tools), on the top of which classes, relations and properties can be tentatively extracted. Despite of the improvements made possible by this approach based on basic (and shallow) linguistic analysis, a major drawback remains: ontology extraction can be proposed only on the basis of word analysis and not of phrases and sentences. This way we can not extract relations out of German texts which are not expressible in compound terms, like the succession of states, the anchoring of states in temporal and locative context.

3. INTERMEDIATE RESULTS

On the base of the combination of the approaches described above, we are able to extract already quite a lot of possibly relevant ontology classes, relations and properties. So for example we found in the corpus strings like:

(6) Er soll im Konzern Finanzchef Gerhard Liener folgen (He should in the corporation take the place of financial chief Gerhard Liener).

Here we can infer that a financial chief is in fact a position in a corporation, so that we know that the relations between "Finanz" and "Chef" on the one hand and "Konzern" and "Chef" on the other hand can not be on the same level. But clearly this level has to be defined from other sources in the text. Or it is already stated in the (extracted) ontology that a corporation has a financial department. Then we can extract the information that a department has a chief. At this level of processing, we need to consider both the linguistic context and some possibly available information in the ontology so far. Just to give an idea of how a larger linguistic context can help, let us look again at the sentence we discussed just above, and the kind of linguistic annotation it gets through the analysis by our tools:

[NP-Subj Er] [VG soll] [PP im Konzern] [NP-Ind-Obj Finanzchef [NE-Pers Gerhard Liener]] [VG folgen]

Through the syntactic structuring of the sentence, we can semantically group the items, so that we can extract the fact that a "financial chief" is "within a corporation", since the description of job succession is within a corporation (marked by the prepositional phrase "im Konzern"). This aspect of ontology generation is being currently investigated and implemented. Interesting here for the ontology population case, is the fact that the sentence doesn't tell us who is taking the job (the subject of the sentence is realized by a pronoun). We thus have to look at a broader textual context to find the referent of the pronoun "he". The same remark is valid for the name of the "corporation". Here we just have a definite description "in the corporation". So we know that our tools have to depict the exact reference somewhere else in the text (normally in a former sentence, or even in the title of the newspaper article).

4. CONCLUSIONS AND FURTHER WORK

We described an approach that can support the semi-automation of either building up ontologies from scratch or suggesting extension to existing ontologies, relying on heuristic rules applied to linguistic annotation attached to textual documents by natural language processing. We presented first implementation steps toward this goal, and saw that we already identified to certain extents the type of ontological resources that can be generated from text on the base of a multi-layered processing strategy.

There is still a problem with this approach: how do we know if the extracted classes (and relations and properties) do not already exist somewhere else in the existing ontologies of the MUSING framework. This is a problem related to ontology mapping: when are two or more ontologies to be considered identical or distinct? While this issue will be tackled later in the project, a first pragmatic approach for us lies in looking if we can populate existing ontologies with the information extracted from text. In this case, we assume for now that the identified candidate classes are not really new and they are filtered out.

5. REFERENCES

[1] M. Ciaramita, A. Gangemi, E. Ratsch, J. Saric, and I. Rojas. *Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies*, pages 91–107. IOS Press, 2008.

[2] T. Declerck, H.-U. Krieger, B. Kiefer, M. Spies, and C. Leibold. Integration of semantic resources and tools for business intelligence. In *International Workshop on Semantic-Based Software Development held at OPSLA 2007*, 2007.

[3] T. Declerck and M. Vela. A generic nlp tool for supporting shallow ontology building. In *Proceedings of LREC*, Genoa, May 2006.

[4] R. Navigli and P. Velardi. *From glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions*, pages 71–91. IOS Press, 2008.

[5] P. Pantel and M. Pennacchiotti. *Automatically Harvesting and Ontologizing Semantic Relations*, pages 171–199. IOS Press, 2008.