Concept and Relation Extraction in the Finance Domain

Mihaela Vela
DFKI Saarbrücken
Mihaela.Vela@dfki.de

Thierry Declerck
DFKI Saarbrücken
Thierry.Declerck@dfki.de

Abstract

In this paper, we describe the state of our work on the possible derivation of ontological structures from textual analysis. We propose an approach to semi-automatic generation of domain ontologies from scratch, on the basis of heuristic rules applied to the result of a multi-layered processing of textual documents.

1 Introduction

In the context of the MUSING R&D European project[1], which is dedicated to the development of Business Intelligence (BI) tools and modules founded on semantic-based knowledge and content systems, we are investigating among others the integration of semantic web and human language technologies for enhancing the technological foundations of knowledge acquisition and reasoning in BI applications.

In the first phase of the project many efforts have been dedicated to the manual creation of domain related ontologies and their integration in upper level ontologies. The creation of those ontologies was guided by domain experts, who were submitting so-called competency questions to the ontology engineers in order to support they work on ontology design and implementation. Since this approach to ontology creation is very time and resource consuming, we wanted to investigate the possible automation of ontology building directly from textual documents, with a special focus on the financial domain.

2 The Approach

We consider the semi-automatic ontology derivation from text as a linguistic rule-based approach, which on the basis of lexical and syntactic properties

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

Proceedings of the 8th International Conference on Computational Semantics, pages 346–350,
Tilburg, January 2009.

can suggest potential ontology classes and properties that can be used for building an ontology.

We suggest 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 (POS) 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: at the shallow level it is possible to detect possible classes and relations, which can then be consolidated, refined or rejected at further stages of textual analysis, with the help of domain experts and ontology engineers.

Our final goal 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 first with a corpus of economical news articles from the German newspaper *Wirtschaftswoche*.

3 String-Based Processing

As a first step in the task of extracting ontology concepts and relations from (German) textual documents, we decided to look in the corpus for words occurring alone (and starting with a capital letter) and in the context of larger strings (which we assume to be mostly nominal compounds). We call the words that show this property "anchor-words". We consider them potential labels of ontology classes and the compounds, in which they occur, as expressing potential relations for the labels of ontology classes.

Take for example the anchor-word `konzern` (*corporation*), which is also occurring as part of the compound `medienkonzern` (*media corporation*). At this very shallow and pattern-based processing level, we tentatively derive that from the compound construction `PREFIX + ANCHOR` we can extract `medienkonzern ISA_SUBCLASS_OF konzern`. Another example of compound is *konzernverwaltung* (*corporation management*). Here we derive from the compound construction `ANCHOR + SUFFIX` the relation: `konzern HAS` verwaltung (*corporation HAS management*);

Although the examples demonstrate that a string analysis can to some extent propose some guidelines for ontology extraction, there are for sure major limitations, due to the lack of well-defined domain and range specifications in the proposed relations, the constraint relative to the number of extracted relations and classes, the multiple appearance of morphological variations, the lack of textual context etc.

In order to reduce the limitations just mentioned, we started by looking for alternative formulations of the compounds, which can help in establishing some filters and validation steps for relation extraction. So for example the

expression *Chef vom Konzern* (*chief of the corporation*) is validating the property relation `Konzern HAS Chef` (*corporation HAS chief*), due to the meaning we can associate with the preposition *von* (*part-of, belonging-to*).

Concerning compound reformulations expressed by genitive post- modification, like *mitarbeiter einer deutschen bank* (*employees of a german bank*), we can see that they validate the relation extracted from the corresponding compounds, since the genitive constructions have here a part-of/belonging-to meaning.

4 Morphology and Lexical Semantics for Ontology Derivation

A way to reduce some of the limitations described in Section 3 lies in the use of morpho-syntactic information. So for example the word *Firmenchef* (*the boss of the firm*) would be analyzed as follows:

(1) <W INFL="[17 18 19]" POS="1" STEM="chef" COMP="firmenchef" TC="22">Firmenchef</W>

This annotation is to read like this: the word *Firmenchef* has the stem *chef*, has POS `noun`, is the result of combining the word *Firmen* and the word *Chef*, and has certain morphological properties (here encoded with numbers). We can then describe certain morphological constraints for filtering out some suggested relations from Section 3. For example, *Chemiekonzern* is introducing a subclass relation between *Chemiekonzern* and *Konzern*, whereas for *Grosskonzern* (*large corporation*) the subclass relation between *Grosskonzern* and *Konzern* does not apply. The constraint proposed for solving this kind of ambiguities is: the compound should consists of two nouns.

Lexical semantics can also improve the quality of relation extraction. For the compound *Chefdenker* (*chief thinker*), we want to ensure that no `HAS`-relation between an ontology class labeled by *chief* and *thinker* is suggested. For this purpose we use lexical semantic resources, like WordNet, and formulate a rule that states that if the word occurring in the prefix position of a compound is a person, and the second part of the compound is also denoting a person, then the `HAS`-relation can not be derived.

Despite of the improvements made possible by morphology and lexical semantics a major limitation remains: ontology extraction is proposed only on the basis of word analysis and not on the basis of phrases and sentences, which offer more context.

5 Syntactic Information for the Generation of Ontologies

By combining the processing steps described above, we were able to extract possible relevant ontology classes, relations and properties. For further improvement we need to consider both the linguistic context and some information available in ontologies so far. The syntactic analysis of the sentence below is a good example to show how a larger linguistic context can help improving the method described in this paper.

(2) [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 learning is being currently investigated and implemented.

6 Conclusions and Further Work

In this paper we have been describing a multi-layer textual analysis strategy that can help in building up ontologies from scratch, or integrate new suggested ontology classes (or relations and properties) into existing ontologies. Since this is work in progress we also intended to get a clearer picture on what kind of ontological knowledge can be extracted from the different layers of textual processing.

For the "shallowest" parts of our suggested approach we could see that proposed labels for ontology classes and relations seem to be appropriate. For sure, some evaluations of this work has to be done. Nevertheless, we see a big potential in a combination of suggestions generated by linguistic analysis, domain experts and ontology engineers.

References

Massimiliano Ciaramita, Aldo Gangemi, Esther Ratsch, Jasmin Saric, and Isabel Rojas. Unsupervised learning of semantic relations for molecular biology ontologies. In Paul Buitelaar and Philipp Cimiano, editors, *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 91–107. IOS Press, Amsterdam, 2008.

Philipp Cimiano, Siegfried Handschuh, and Steffen Staab. Towards the self-annotating web. In *Proceedings of the 13th World Wide Web Conference*, 2004.

Thierry Declerck and Mihaela Vela. A generic nlp tool for supporting shallow ontology building. In *Proceedings of LREC*, Genoa, May 2006.

Thierry Declerck, Hans-Ulrich Krieger, Bernd Kiefer, Marcus Spies, and Christian Leibold. Integration of semantic resources and tools for business intelligence. In *International Workshop on Semantic-Based Software Development held at OPSLA 2007*, 2007.

Roberto Navigli and Paola Velardi. *From glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions*, pages 71–91. IOS Press, 2008.

Patrick Pantel and Marco Pennacchiotti. *Automatically Harvesting and Ontologizing Semantic Relations*, pages 171–199. IOS Press, 2008.