Ontology-Driven Human Language Technology for Semantic-Based Business Intelligence

Thierry Declerck[1] and **Hans-Ulrich Krieger**[2] and **Horacio Saggion**[3] and **Marcus Spies**[4]

Abstract. In this poster submission, we describe the actual state of development of textual analysis and ontology-based information extraction in real world applications, as they are defined in the context of the European R&D project "MUSING" dealing with Business Intelligence. We present in some details the actual state of ontology development, including a time and domain ontologies, which are guiding information extraction onto an ontology population task.

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. MUSING integrates Semantic Web and Human Language technologies for enhancing the technological foundations of knowledge acquisition and reasoning in BI applications. The impact of MUSING on semantic-based BI is being measured in three strategic domains:

- Financial Risk Management (FRM), providing services for the supply of information to build a creditworthiness profile of a subject – from the collection and extraction of data from public and private sources up to the enrichment of these data with (semantic) indices, scores and ratings;
- Internationalization (INT), providing an innovative platform, which an enterprise may use to support foreign market access and to benefit from resources originating in other markets;
- IT Operational Risk & Business Continuity (ITOpR), providing services to assess IT operational risks that are central for Financial Institutions – as a consequence of the Basel-II Accord - and to asses risks arising specifically from enterprise's IT systems – such as software, hardware, telecommunications, or utility outage/disruption.

Across those development streams of MUSING, there are some common tasks, like the one consisting in extracting relevant information from annual reports of companies and to map this information into XBRL (Extended Business Reporting Language). XBRL is a standardized way of encoding financial information of companies, but also the management structure, location, number of employees, etc. (see www.xbrl.org). This is mostly "quantitative" information, which is typically encoded in structured documents, like financial tables or company profiles etc. But for many Business Intelligence applications, there is also a need to consider "qualitative" information, which is most of the time delivered in the form of unstructured text, which one can find in textual annexes to the balance sheets in annual reports or in news articles. The problem is here how to accurately integrate information extracted from structured sources, like the periodic reports of companies, and the day to day information provided by news agencies, mostly in unstructured text form. The detection and interpretation of temporal information in structured and unstructured documents is also a central focus of our attention in MUSING. We describe in the following the actual state of development of MUSING ontologies, including our proposal for temporal representation. Due to lack of space, we can not show here examples of the kind of temporal expressions we encounter in applications of MUSING, and how our IE and Ontology Population tools deal with those expressions in the light of our representation of temporal information, aiming also at supporting temporal reasoning in various applications. But those examples will be available on the poster.

[1] DFKI GmbH, Germany, email: declerck@dfki.de
[2] DFKI GmbH, Germany, email:kriegger@dfki.de
[3] University of Sheffield, UK, email: H.Saggion@dcs.shef.ac.uk
[4] Semantics Technology Institute, Austria, marcus.spies@sti2.at

2 STATE OF MUSING ONTOLOGIES

In MUSING we decided to use as the upper level ontology the PROTON ontology (http://proton.semanticweb.org), on the base of which domain-specific extensions can be easily defined.

The species of the model of the PROTON Upper module is OWL Full. The MUSING version available contains mostly the same information as the original one but is slightly changed to fulfill the OWL Lite criteria. The System module of PROTON, http://proton.semanticweb.org/2005/04/protons, provides a sort of high-level system- or meta-primitives. It is the only component in PROTON that is not to be changed for the purposes of ontology extension." The Top-Level classes in PROTON, http://proton.semanticweb.org/2005/04/protons, represent the most common definition of world knowledge concepts. These can directly be used for knowledge discovery, metadata generation and to interface intelligent knowledge access tools. The PROTON has also an upper module, http://proton.semanticweb.org/2005/04/protonu, which adds sub-classes and properties to the Top-module super classes to the concepts other than "Abstract, Happening and Object" from the original PROTON Top ontology. The "Extension" ontology in MUSING has been designed as a single contact point between upper and MUSING application specific ontologies. In MUSING we also developed a general time ontology, which is also added to the upper module. Besides the time ontology, there are currently five domain ontologies, which are not assigned to any particular application. They cover the following areas: Company, Industry sector, BACH (Standard for a harmonization of financial for harmonizing accounts of companies across countries), XBRL (Standard language for "Business Reporting") and Risk. In the time ontology of MUSING, temporally-enriched facts are represented through time

slices, four dimensional slices of what Sider (1997) calls a space-time worm (we only focus on the temporal dimension in MUSING). These worms, often referred to as perdurants, are the objects we are talking about. The time ontology itself contains the conceptualization of temporal objects that are relevant in MUSING. In fact, any time ontology can be combined with the "4D" ontology. The other other ontologies are domain and applications specific. As a concluding remark about the ontologies, we would like to mention that they have been built by hand, most of them on the base of "compentency questions" addressed by domain experts. But it is also planned in MUSING to investigate the topic of (semi-)automatic ontology learning or creation, on the base of information and knowledge extracted from the analyzed data.

The poster presentation will mainly visualize the interconnections of the ontologies, and the integrated reasoning component that has been designed for acting on the ontologies and the knowledge bases of MUSING.

3 ONTOLOGY-BASED INFORMATION EXTRACTION IN MUSING

In the former chapter, we presented in some details the different types of MUSING ontologies, and the way they interact (mainly via the "Extension" ontology). This model of the relevant concepts for a set of Business Intelligence applications has to be filled (or populated) with real data, so that the applications can make use of the semantic capabilities of such an ontology infrastructure. We call this task "ontology population", which in a sense is Information Extraction (IE) guided by ontologies, the results of IE not being displayed in the form of templates, but in knowledge representation languages, e.g. OWL in the case of MUSING. The information stored in this way is considered as "instances" of the concepts and relations introduced in the ontology. The set of instances is building the knowledge base for the applications, and this knowledge base is supporting for example credit institutes on their decision-making procedures on credit issuing issues. As we mentioned in the introduction, a substantial amount of the needed information for the development of semantic business intelligence applications is to be found in unstructured textual documents, so that the automatic ontology population task is relying on natural language processing in general and Information Extraction in particular.

It is important to note here that all the instances of the ontologies, populated by means of the IE tools, are automatically "enveloped" within temporal information, which turns every entity or event into a perdurant In case termporal information is not available, or has not been found, this can be left underspecified in the representation of the instances, and filled by information generated from other resources, or by the temporal reasoning engine, also implement in MUSING.

As an example we can look at the following sentence, we took from a newspaer:

"Ermotti arbeitete frueher kurz fuer den weltgroessten Finanzkonzern Citigroup und danach 17 Jahre lang bis 2004 fuer die Investmentbank Merrill Lynch." (Ermotti have worked before for a short time for the world largest financial concern, Citigroup, and afterwards for 17 years, till 2004, for the investment bank Merrill Lynch.)

This is a quite interesting sentence, since it contains a lot of temporal expressions (actually a quite normal fact in news articles). The first two expressions ("before" and "a short time") are again very vague. So here we assume that the before is actually "before the pubdate". The next temporal expressions are "for 17 years" and "till 2004". In those two expressions we get now more precise information: The relation "Ermotti works_at Merrill Lynch" is first associated with the duration of 17 years, and in a second step we can calculate the starting point of this relationship since an ending point is given: 2004 (we allow for such under-specification in the time ontology, having introduced a class called "yearDate"). In order to extract this information and to populate the ontology we need here a deeper linguistic analysis.

We extract with the help of syntactic analysis (and more specially dependency analysis) that there is a working relationship between Ermotti (as the subject of the first part/clause of the sentence) and Merril Lynch. We can associate the time code to this relationship on the base of the dependency analysis of the two temporal expressions as linguistic expressions that "modify" the main verb "arbeitete" (worked). The name of the company for which Ermotti is working is included in a prepositional phrase (PP). The linguistic pattern "[NP-SUBJ X] works [PP for [NP-IOBJ Y]]"is a very good candidate for a mapping into a relation <X is_employed_by Y>. But clearly the constraints that apply to both "X" and "Y" are, that the first is an instance of a person and the second an instance of a company (domain and range of the relation). In this example, the reader could see how the constituent analysis of text, coupled with named entity detection, some lexical semantics and dependency relations, is guiding the ontology population.

In this example we can also see that there are at least three syntactic ways to express temporal information; as an Adverb, an NP and a PP. First the textual analysis gives a linguistic structure to the unstructured text, on the base of which we define a mapping, which associates the name of the person to the person ontology and the name of the company to the company ontology. The relationship "<Errmotti, is_employed_by, Merril Lynch>" can then be associated to the time slice "1987-2004". From the individual news article under consideration we can not extract information about activities of Ermotti in the time between 2004 and 2005-12-16, but we assume that he had an activity in the banking domain. We can thus automatically query for documents telling us something about "Ermotti" and "Year 2005", in order to "fill the temporal gap" in the information card about Ermotti. The already extracted information and the temporal ontology of MUSING are structuring the semantic content of the query. On this base we found for example an article published on the 2006-12-06, one year later.

The poster presentation will visualize in details the interconnections of the ontologies and the NLP and IE tools in order to populate the ontologies.

4 Conclusion

In this poster, we show how we combine Semantic Web resources and tools with Language Technologies, in order to help in creating knowledge bases in the field of Business Intelligence applications, "upgrading" thus the actual strategies implemented in this field, building on quantitative and qualitative information automatically extracted from various types of documents, towards a new generation of semantically driven Business Intelligence methods and tools.

ACKNOWLEDGEMENTS

The research described in this paper has been partially financed by the European Integrated Project MUSING, with contract number FP6-027097.