Human Language and Semantic Web Technologies for Business Intelligence Applications

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

[1] Language Technology Lab, DFKI GmbH
[2] NLP Group, Department of Computer Science, Sheffield University
[3] Digital Enterprise Research Institute, Universität Innsbruck
{declerck,krieger}@dfki.de, Saggion@dcs.shef.ac.uk, Marcus.spies@deri.at

Abstract

In this LangTech 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.
Index Terms: language technology, semantic web, business intelligence

1. Introduction

MUSING is an R&D European project[1] 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.

So for example imagine that you have an annual report from a company, which specifies the name of the CEO (and of other members of boards). The described CEO-relationship between the named person and the company is valid in this case only for the reporting period. But very often, the final report is published 2-3 months after the end date of the covered period. In the meantime it can be that the CEO position is now in the hand of another person, and this has been announced in press articles. We need here accurate information extraction (IE) systems, that detect in the news this change of name of the CEO, and overwrites the information that might have been extracted from the annual report, for the period following the temporal coverage of the annual report until the publication date of this report. This period has to be specified, also on the base of temporal information extracted from the article (a combination of the publication date and temporal information).

As the example above shows, work on IE and ontology population in MUSING is highly depending on temporal information associated with both the publication date of documents and the content of the document itself. So for example the date of publication of an annual report doesn't coincide with the end of the reporting period, and we have to extract the values for the starting and the ending time of the reporting period from the document itself. Additionally, the temporal information associated with certain quantitative information contained in the annual reports can be of two types, whereas this is not explicitly mentioned in the report: duration or instant (for example the number of employees given is valid for a specific instant in time, whereas the value of certain financial indicators is valid for a specific period). This distinction is formulated in available background semantic resource (like XBRL taxonomies) and has to be made explicit in our semantic representation of the information extracted from relevant documents in the MUSING applications.

As a summary of our needs with respect to temporal information in the concrete tasks described above, we learned that we can not work only with synchronic relationships, but

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

rather that we need some means to deal with the intrinsic diachronic aspects of entities and relations.

We describe in the following the actual state of development of MUSING ontologies, including our proposal for temporal representation. We give then 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.

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.

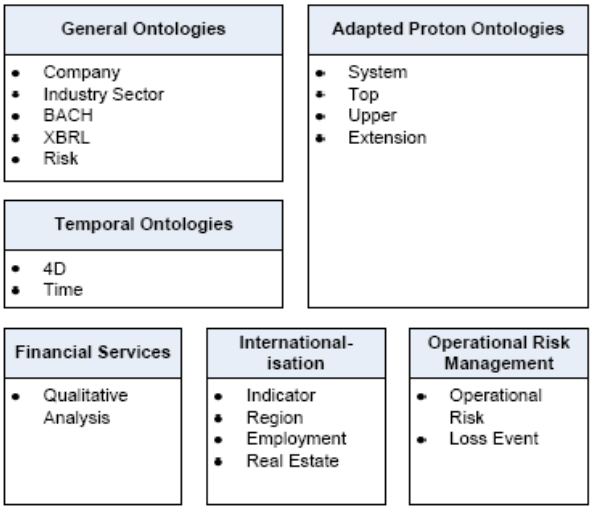


Figure 1: The various types of MUSING ontologies

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 (see the box "Adapted Proton Ontologies" in the figure above).

"The System module of PROTON, http://proton.semanticweb.org/2005/04/protons, provides a sort of high-level system- or meta-primitives, which are likely to be accepted and even hard-coded in tools that may use PROTON. It is the only component in PROTON that is not to be changed for the purposes of ontology extension." (Terziev et al. 2005).

The contained Top-Level classes, 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 (Terziev et al. 2005).

The PROTON upper module, http://proton.semanticweb.org/2005/04/protonu, 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 (the three boxes at the bottom of the figure above).

Besides the time ontology developed within MUSING, there are currently five domain ontologies, which are not assigned to any particular application (the "General Ontologies" in the figure above). 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. For instance, *Jürgen Schrempp* (JS) is a perdurant that comes up with several time slices, talking about his CEOship with Daimler Chrysler (DC), his resignation as CEO of DC (end of 2005), his membership for a certain time within the supervisory board of Allianz and Vodafone, etc. All facts are associated with a temporal dimension, even if they are instants, i.e., having an infinitely-small extension. This kind of representation is encoded in the "4D" ontology (see Figure 1). 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.

There is only one ontology module specific to the Financial Services applications: "Qualitative Analysis". This ontology describes in fact what can be the result of different kind of questionnaires used in this applications field. All the quantitative information relevant for Financial Risk management are covered by the General ontologies of MUSING, since this information plays also a role in the other application domains considered in MUSING.

The set of ontologies for the Internationalization applications in MUSING contains four ontologies. The most important ontology is the one defining the indicators used to measure properties of political regions. It is based on a list of 162 indicators grouped into 14 categories.

In the ItOperational Risk applications of MUSING, we introduce two ontologies, which deal with processes and IT infrastructure, on the one hand, and operational risk in general and IT operation risk in particular, on the other hand.

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" (Gruninger, M., & Fox, M., 1994) 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.

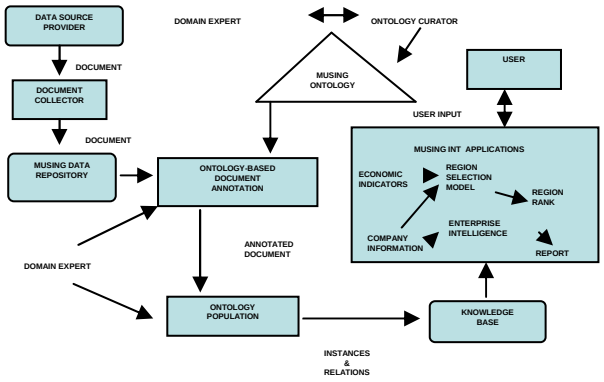
3. Ontology-based Information Extraction in MUSING

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[1] 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.

In the following figure, the reader can see the interaction between textual analysis, information extraction and the semantic resources (ontologies) in MUSING, whereas the examples of classes presented in the box on the right side of the figure are taken from the Internalisation ontologies developed in MUSING:



.

3.1. (Temporal) Ontology population from News articles in the financial domain

Temporal information is given in news articles at two levels: date of publication and within the article reporting on issues related to the financial domain (we concentrate here on this application domain of MUSING).

We can make use of the typical layout of electronically available newspapers, with a title, an abstract, the names of authors and the publication date, etc., for extracting information with high accuracy from this "structured" part of the document (see for example RSS feeds). Important is here the publication date, which is offering an anchoring temporal index for the interpretation of temporal expressions present in the article itself.

In the content part of news articles, the reader can find a large variety of temporal expressions. This variety is partly motivated by the habits of journalistic style to use as much as possible variances in their formulations on facts. So writing about a CEO of a company, the journalists will use maximally one time the "dry" and precise formulation "Mr. X, Ceo of Y limited", and will in the text refer to this person mentioning other properties, like her/his nationality, age, region or city of birth, or even character properties, like "the ambitious Swiss manager …". This is similar with temporal expressions, where journalists also make use of metaphors and says. One thing being quite sure: authors will very seldom use the ISO standard for writing down a date in their article!

In the following we focus on some examples of natural language expressions in newspapers articles, and show how MUSING NLP components deal with those expressions. We take here as example news articles, from different providers, reporting on one person, Sergio Ermotti.

In one article, the title of the news and the publication date are:

Title: Ermotti statt Jentzsch (*Ermotti instead of Jentschz*)
Pubdate: 16.12.2005

The date is detected (in our case with the DFKI Information Extraction tools "SProUT), and annotated as a temporal_unit:

```
<F name="TEMPORAL-UNIT"><FS type="temporal unit"></FS></F>
<F name="YEAR"><FS type="2005"></FS></F>
<F name="MONTH"><FS type="12"></FS></F>
<F name="DAY"><FS type="16"></FS></F>
<F name="MUC-TYPE"><FS type="date"></FS></F>
```

This (simplified) feature structure representation tells us that the system found a "temporal unit" (equivalent to the MUC-TYPE="date") and the temporal expressions is being split into its components (Day, Month and Year). We associate then this temporal information with the "PubDate" feature.

Concerning the title of the article; the Information Extraction system also detects that both Ermotti and Jentzsch are person names, on the base of either the use of a gazetteer, the MUSING knowledge base or heuristics:

```
<F name="SURNAME">
      <FS type="Ermotti"></FS></F>
<F name="SURNAME">
      <FS type="Jentzsch"></FS> </F>
```

Further information about those persons, like nationality and age can be detected by SProUT in the text (if the information is present at all), or is available in the knowledge base (or in a gazetteer). But the Named Entity Detection system will not get the particular relation between both persons as this relation is expressed in the title: "Ermotti instead of Jentschz". We need here some lexical semantic information about "statt" (*instead*) in order to detect that there is a kind of competition between the two persons,

So far we got the information that at a certain date it is reported on a decision taken between two persons. The abstract of the article is giving us more information:

"Die italienische Großbank Unicredit hat Sergio Ermotti mit sofortiger Wirkung zum Chef der Sparte internationale Großunternehmen und Investmentbanking ernannt. Ursprünglich war der Posten für den ehemaligen HVB-Manager Stefan Jentzsch vorgesehen." (*The large Italian Bank Unicredit has named Sergio Ermotti with immediate effect as the head of the Branch "Large International Companies and Investment Banking". Originally was this job foreseen for the HVB Manager Stefan Jentzsch.)*

Two temporal expressions are included in this abstract: "mit sofortiger Wirkung" (*with immediate effect*) and "ursprünglich" (*originally*). And here our approach consists in first in linking those underspecified time expressions to the publication date. So that we have the relation:

[1] But to be quite correct here, we should mention that the "knowledge bases" are constituted by the ontologies together with their sets of instances.

t = 2005-12-16: <Ermotti, headOf, Unicredit Branch "Large International Companies and Investment Banking">

The actual value of *t* might not be the exact date at which Ermotti is starting (we expect that some time interval will exist between the event itself and its announcement by a press organ). But it is a very good approximation.

In the main text of the article, then more details are given, for example about Ermotti:

"Ermotti arbeitete früher kurz für den weltgrößten Finanzkonzern Citigroup und danach 17 Jahre lang bis 2004 für 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 "$[_{NP\text{-}SUBJ}$ X] works $[_{PP}$ for $[_{NP\text{-}IOBJ}$ 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 the pseudo formal representation below, we use for ease of presentation, the reader can see how the tools put the information together[1]:

SUBJ(Ermotti) arbeitete TEMP_ADV1(früher) TEMP_ADV2(kurz) für-IOBJ1 (Citigroup)
==> SUBJ-related(job_Position) to IOBJ1
==> INTERVAL1 = TEMP_ADV1 (here OpenIntervalLeft)
==> INTERVAL2 = TEMP_ADV2 (here OpenInterval) & INTERVAL2 ⊆ INTERVAL1

SUBJ(Ermotti) arbeitete TEMP_NP-TEMP1(17 Jahre) PP-TEMP (bis 2004) für IOBJ2 (MerrilLynch)
==> SUBJ-related(job_Postion) to IOBJ2
==> INTERVAL1 = NP-TEMP (here OpenInterval)
==> INTERVAL2 = PP-TEMP (here OpenIntervalLeft, DATE_TYPE = year representation/date representation) & INTERVAL2 = RightParanthesis of INTERVAL1

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 this 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 of the Handelsblatt, published on the 2006-12-06, one year later.

4. Conclusion

In this short paper, we have been showing how Human Language Technology, in close collaboration with Semantic Web resources and tools, can help in creating knowledge bases in the field of Business Intelligence applications, "upgrading" thus the actual strategies implemented in this field, building on quantitative information and statistical models, towards a new generation of semantically driven Business Intelligence methods and tools. We concentrated on the representation and extraction of temporal information, since this is a crucial topic for the applications within the MUSING Project.

5. Acknowledgements

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

6. Short References

[1] Ivan Terziev, Atanas Kiryakov & Dimitar Manov. D1.8.1 Base upper-level ontology, Guidance1, EU-IST Project IST-2003-506826 SEKT, WP1, D1.8.1, 2005,

[2] Theodore Sider. Four Dimensionalism. Philosophical Review 106, 197–231, 1997.

[3] Gruninger, M., & Fox, M. (1994). *The role of competency questions in enterprise engineering.* Paper presented at the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice, Trondheim, Norway.

[1] „OpenIntervalLeft" and similar expressions in the example are classes of our time ontology.