Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems

Thierry Declerck
DFKI GmbH, LT-Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbruecken
declerck@dfki.de

Piroska Lendvai
Research Institute for Linguistics
Hungarian Academy of Sciences
Benczúr u. 33., H-1068 Budapest
piroska@nytud.hu

Tobias Wunner
DERI, NLP unit
NUIG
Newcastle Rd, IE-Galway
tobias.wunner@deri.org

Abstract

We investigate the benefits that can result from the formal representation of linguistic and semantic features of natural language expressions that are used as terms in labels of knowledge representation systems, like taxonomies and ontologies. We show that such a representation can support Human Language Technologies and Semantic Web applications, especially in the context of ontology-based information extraction, since it gives a basis for specifying mapping strategies between the restricted natural language used in taxonomies and ontologies and the unrestricted language used in documents processed by information extraction or semantic annotation tools.

1 Ontology-based Information Extraction

In the last decade, we have been witnessing changes in the field of Information Extraction (IE) due to the emergence of a significant amount of semantic resources available in the form of taxonomies and ontologies. These Knowledge Representation (KR) systems have been gradually replacing the predefined templates, which were formerly used for specifying IE applications, and are now often building the background against which texts are processed in order to extract relevant information for specific applications. In those cases, we speak of Ontology-based Information Extraction (OBIE)[1].

In the Description Logic (DL) approach, KR systems are viewed as consisting of two components, the T-Box (Terminological axioms) and the A-Box (Assertion axioms).[2] We adopt here this terminology (*T-Box*, *A-Box*), even if not all the KR systems we are dealing with are modeled using the DL representation language, and in fact we are dealing in this short paper only with examples taken from a complex taxonomy modeled in XML.

A main issue for OBIE tasks is to establish an accurate mapping between the classes and properties described in a T-Box and the natural language expressions occurring in unstructured textual documents. Fortunately, most KR systems come equipped with a label feature associated with their elements; these include natural language expressions that are meant to "provide a human-readable version of a resource's name"[3] and that act very often as domain specific terms.

It is an empirical issue whether linguistic and semantic analysis of the formal description and machine-readable representation of such labels would support the task of associating classes and properties of KR systems with (fragments of) textual documents. If an OBIE application detects information that corresponds to T-Box elements, this information can be marked as their related A-Box *instances*. Ontology Population (OP) then consists in storing all instances of taxonomy or ontology classes and properties we can extract from text in a knowledge base.

The work described in this paper is closely related to the "LexInfo' (Buitelaar et al.2009), (Declerck and Lendvai2010) and to the "lemon" (lexi-

[1] See also (Buitelaar et al.2008) for more details

[2] See (Baader2009) for more details.

[3] http://www.w3.org/TR/rdf-schema/#ch_label

con model for ontologies)[4] models that all work towards the goal of describing and representing lexical and linguistic properties of the textual content of taxonomy and ontology labels. On this basis, we started to analyze the textual content of labels encoded in XBRL taxonomies (see section 2 below) in order to see if this type of text can be used for supporting the task of finding corresponding information in related textual documents, like for example annual reports of companies. We discuss in detail some examples below after having briefly introduced the XBRL framework.

2 XBRL

XBRL, eXtensible Business Reporting Language[5], is an XML-based mark-up language for the exchange of business information, including financial reporting. XBRL specify the semantics of business data, its presentation, its calculation, and associated business rules, which are called formulas. XBRL also has its own special terminology and comes up in the form of a taxonomy, that is used for modeling various types of international standards[6] and national or regional legislations for financial reporting[7]. An XML document that contains concrete values for a number of XBRL concepts, like name of the company, period of the reporting and concrete values for financial items is called an instance document[8].

3 Examples of Terms in Labels and in Text

In section 3.1 four examples are given of textual content of labels in the IFRS taxonomy encoded in XBRL. Section 3.2 illustrates the typical content of a financial table of an annual report (in this case from the Deutsche Bank company, in German). In section 3.3 a short, partial segment of an explanatory note, in German, of a financial report (the company Bayer AG) is displayed.

It can be observed that neither the vocabulary of financial reports, nor the grammatical realizations of the concepts is harmonized with that used in labels. Our goal is to automatically assign the relevant concepts of the IRFS-XBRL taxonomy to (segments) of the two types of financial reports, and to transform (parts) of those documents onto an XBRL instance document with high precision.

3.1 Examples from the IFRS-XBRL Taxonomy

In each example below we have the name of the concepts (in italics within brackets) and both the corresponding English and German labels.[9]

1. Reconciliation of minimum finance lease payments payable by lessee / Überleitungsrechnung der vom Leasingnehmer im Rahmen von Finanzierungs-Leasingverhältnissen zu zahlenden Mindestleasingzahlungen (*ReconciliationOfMinimumFinanceLeasePaymentsPayableByLesseeAbstract*)

2. Reconciliation by end of reporting period / Überleitungsrechnung am Abschlussstichtag (*ReconciliationByEndOfReportingPeriodAbstract*)

3. End of period not later than one year / Bis zu einem Jahr bis zur Ende der Periode (*EndOfPeriodNotLaterThanOneYearAbstract*)

4. Minimum finance lease payments payable, at present value, end of period not later than one year / Im Rahmen von Finanzierungs-Leasingverhältnissen zu zahlende Mindestleasingzahlungen, zum Barwert, bis zu einem Jahr bis zum Ende der Periode (*MinimumFinanceLeasePaymentsPayableAtPresentValueEndOfPeriodNotLaterThanOneYear*)

3.2 Example from a Financial Table

```
Finanzleasingverpflichtungen
     275  25  46  60  144
```

This particular line is about the value of to be paid finance leases for the next periods: the total amount is 275 million euros and the periods are 1 year, 1-3 years, 3-5 years, more than 5 years.

[4] see: http://www.isocat.org/2010-TKE/presentations/Monnet-slides.pdf

[5] See http://www.xbrl.org/Home/

[6] Like the International Financial Reporting Standards (IFRSs), see http://www.ifrs.org/Home.htm

[7] For example the so-called General Accepted Accounting Principles (GAAP) of different countries, like Germany or the United States of America. The IFRS, the German and the US GAAPs, among others, can be browsed at http://www.abrasearch.com/ABRASearch.html

[8] Examples of these can be retrieved among others at the U.S. Securities and Exchange Commission (SEC, http://xbrl.sec.gov/) or at the Belgian National Bank (BNB, http://euro.fgov.be/.

[9] As an addtional information: The four concepts are in a sub-class relation in the taxonomy: $4 > 3 > 2 > 1$.

3.3 Example from an Explanatory Note

This (partially reproduced) note is describing the policy of the company with respect to finance leases.

"Ist der Bayer-Konzern Leasingnehmer in einem Finanzierungsleasing, wird in der Bilanz der niedrigere Wert aus beizulegendem Zeitwert und dem Barwert der Mindestleasingzahlungen zu Beginn des Leasingverhältnisses ... Die Mindestleasingzahlungen setzen sich im Wesentlichen aus Finanzierungskosten und dem Tilgungsanteil der Restschuld zusammen. ... Ist ein späterer Eigentumsübergang des Leasinggegenstands unsicher, ... Die zu zahlenden Leasingraten werden nach der Effektivzinsmethode aufgeteilt ... . Ist der Bayer-Konzern Leasinggeber in einem Finanzierungsleasing, werden in Höhe des Nettoinvestitionswerts Umsatzerlöse erfasst und eine Leasingforderung angesetzt. ...."

4 Our Approach to the Linguistic and Semantic Enrichment of Labels

We follow a multi-layered approach, starting with layout analysis, on the top of which linguistic and semantic analysis are proposed.

4.1 Segmenting and Tokenizing the Terms

In a first step, we segment the terms used in the labels (as listed in Section 3.1). For this one can make use of IFRS guidelines on the terminology used in the taxonomy, e.g. some punctuation signs explicitly mark term/sub-term segments (e.g. the commas segment term (4) in Section 3.1 into three subterms).

This approach is being consolidated by checking if the suggested sub-terms are themselves used as full terms in the labels of other concepts. In the given case we verify that this holds for only two subterms, but not for `zum Barwert` (*at present value*). From the linguistic point of view, we can tentatively associate the "consolidated" subterms with a status similar to an "arguments" of a functional term (to be established still).

4.2 Linguistic Analysis of the Terms

Subsequently, lemmatisation of the words used in the terms is performed in order to detect and link all possible forms of e.g. `Finanzierungs-Leasingverhätnissen` (*finance lease*) – its current inflection is dative plural, but the same term with other inflectional suffixes can be present in other labels of the taxonomy, or in external documents.

Next, we propose performing PoS tagging and complex morphological analysis, including derivation and compounding. This allows for example to detect in texts related terms such as `Finanzierungskosten` (occurring in the example in Section 3.3) and `Finanzleasingverpflichtungen` (occuring in the example of Section 3.2).

A chunking and a dependency analysis are also proposed, following the approach described in (Declerck and Lendvai2010), but refraining from showing the linguistic annotation due to limitations of space. Dependency analysis allows for detecting head nouns in terms. We can then compare labels sharing at least one identical head noun (its lemma) and thereby establish lexical semantic relations across concepts , taking into account the different linguistic contexts in all those labels.

Lemmas of head nouns are also considered as anchors for starting the search of relevant segments in textual documents. This strategy is motivated by the fact that in the taxonomy labels mainly nominal phrases are present.

4.3 Semantic Enrichment

Semantic annotation of subterms is recommended in case they represent temporal information (*end of reporting period*). Semantic enrichment can further be proposed on the basis of information that is either internal or external to the taxonomy.

An example for the internal case: as we noted in Section 3.1 the concept listed under (4) is a subclass of the concept listed under (2). We observe that none of the words used in the German label of the subclass occurs in the label of the superclass. But in both cases there is a subterm that can be annotated as a temporal expression (`Bis zu einem Jahr bis zur Ende der Periode` and `am Abschlussstichtag`). Between those expressions one can thus assume a semantic relation (the one containing in duration the other one, but we can also infer a lexical semantic is-a relation between *Minimum finance lease payments* and *Reconciliation*).

An additional semantic information we can in-

fer from internal information is about the semantic roles: the payments, which are a reconciliation, have a *lessee* and a *lessor*. This information is distributed over two classes, which are both at the (local) highest level in the taxonomy. This information helps to detect in text the corresponding concepts. But differently, depending if the document basis is a table or a free text. In the first case the semantic role *lessee* has to be infered as being the author of the document (the company providing for the annual report), since in tables nthe name of the company is normally not mentioned. In the second case both roles can be found, and here the use of Named Entity recognition tools is required.

With external enrichment we mean the use of resources like WordNet or FrameNet etc. for "importing" into the ontology labels additional lexical-semantic information.

We have to note here that with this issue the "classical" annotation of the terms with the means of XML, as proposed by (Declerck and Lendvai2010) comes to its limit. We plan therefore to test the lemon model[10] for encoding the linguistic and semantic enrichment of the labels of the taxonomy. It will be interesting to see if the resulting network of linguistic and semantic information, on the basis of the analysis of the "human-readable version" of the taxonomy is still comparable with the original concept-based taxonomy.

5 Conclusion and future work

We described in this short paper actual work on enriching taxonomy and ontology labels with linguistic and semantic information. With this approach we follow two goals: Improving the effectiveness and quality of ontology-based information extraction and possibly suggesting re-organizing the actual model of the domain of consideration.

In the case of XBRL taxonomies we see a large potential for getting not only a more compact but also a more complete model of the domain under consideration. While we are still using an XML annotation schema for this enrichment work, we plan to move to the RDF model proposed by lemon in order to support an ontological organization of the linguistic and semantic enrichment of the labels.

We are currently implementing a unification-based approach for comparing the linguistic and semantic features of the labels in KRs and of the result of the processing of the textual documents. This allows to make use of underspecification in the matching of information included in both sides, while requiring identity in the values of the "lemma" features.

We note finally that since the size of the taxonomy is limited and that many sub-terms are repeated in various concept labels, we can imagine a manually supervised annotation of the labels, this in order to ensure a high quality result of this task.

[10] As a reminder, see: http://www.isocat.org/2010-TKE/presentations/Monnet-slides.pdf

Acknowledgments

The ongoing research described in this paper is part of the RD project Monnet, which is co-funded by the European Union under Grant No. 248458 (see http://www.monnet-project.eu/). The conribution by Piroska Lendvai is co-funded by the European project CLARIN (www.clarin.eu).

References

Baader, F. (2009). Description logics. In *Reasoning Web: Semantic Technologies for Information Systems, 5th International Summer School 2009*, Volume 5689 of *Lecture Notes in Computer Science*, pp. 1–39. Springer–Verlag.

Buitelaar, P., P. Cimiano, A. Frank, M. Hartung, and S. Racioppa (2008). Ontology-based information extraction and integration from heterogeneous data sources. *International Journal of Human-Computer Studies* (11), 759–788.

Buitelaar, P., P. Cimiano, P. Haase, and M. Sintek (2009). Towards linguistically grounded ontologies. *The Semantic Web: Research and Applications*, 111–125.

Declerck, T. and P. Lendvai (2010). Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In *LREC 2010- The seventh international conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10), May 19-21, Valetta, Malta*. ELRA.