Extraction of Multilingual Term Variants in the Business Reporting Domain

Thierry Declerck[1] and Dagmar Gromann[2]

[1] DFKI GmbH, Language Technology Department,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
declerck@dfki.de
[2] Vienna University of Economics and Business,
Nordbergstrasse 15, 1090 Vienna, Austria
dgromann@wu.ac.at

Abstract. Within the context of the European research project "Monnet", which implements among other activities ontology-based multilingual information extraction, we tackle the the issue of recognizing variants of concept labels in business reports that guide the information extraction process. In this short paper, we describe two related experiments in finding variants of multilingual taxonomy labels used in business reporting – across distinct reporting legislations and languages. A core taxonomy developed by the XBRL-Europe Association provides a starting point, as we map multilingual term variant candidates we extract from the web presence of relevant players in the field of business reporting to its labels.

Keywords: Terminology extraction, variants, ontology, multilingualism, business reporting

1 Introduction

Within the context of the European research project "Monnet"[3], which implements among other activities the ontology-based extraction of multilingual information to be used in the field of business reporting, we face the challenge of detecting relevant terms and their variants in a variety of document types. Afterwards, these terms and variants as well as their associated data have to be transformed into domain facts that can be stored as instances of classes of an integrated financial and reporting ontology.

In the European context the fact that each country is marked by different legislations as regards the description of information companies have to provide represents a particular challenge as well as the fact that the corresponding financial statements to be reported are mainly based on so-called national General Accepted Accounting Principles (GAAP). Fortunately, most of these GAAPs

[3] See http://www.monnet-project.eu for more details.

are nowadays encoded using a standard representation language, called XBRL[4], which provides relatively harmonized taxonomies listing the main concepts and associated natural language labels (using the xml:lang attribute) containing the official reporting terminology. A simplified example from the taxonomy of the Belgian National Bank is provided below.

```
<loc xlink:label="Assets_loc" xlink:type="locator"
   xlink:href="pfs-2011-04 01.xsd#pfs_Assets"/>
<labelArc xlink:from="Assets_loc" xlink:to="Assets_lab"
   xlink:type="arc"xlink:arcrole="http://www.xbrl.org/2003/arcrole/concept-label"/>
<label xlink:label="Assets_lab" xlink:type="resource"
   xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="fr">Total de l'actif</label>
<label xlink:label="Assets_lab" xlink:type="resource"
   xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="nl">Totaal van de activa</label>
<label xlink:label="Assets_lab" xlink:type="resource"
   xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="de">Summe der Aktiva</label>
<label xlink:label="Assets_lab" xlink:type="resource"
  xlink:role="http://www.xbrl.org/2003/role/label" xml:lang="en">Total assets</label>
```

[Simplified excerpt from the Belgian taxonomy for reporting: The concept `pfs_Assets` with labels in four languages.]

Although the representation of such information in XBRL, also allowing machine readability of the data, already marks a substantial progress towards more transparency in financial reporting, the cross-country and cross-lingual comparison still continues to be an issue. A working group of the XBRL-Europe Association started investigating this problem, and developed a core taxonomy, called xEBR (eXtended European Business Registers)[5], which connects concepts used in different legislations with common concepts, using also SKOS descriptors to indicate whether the mappings are exact, broad, or narrow, as can be seen in the code example below.

```
pfs_GainLossPeriod              exactMatch      xebr_ProfitLossForThePeriodTotal
pfs_FormationExpenses           narrowMatch     xebr_FixedAssetsTotal
pfs_AccumulatedProfitsLosses    broadMatch      xebr_ProfitLossForThePeriod
```

[Example of xEBR mappings from concepts of the Belgian National Bank, indicated by using the namespace "pfs", to the core concepts of xEBR.]

This work on the core taxonomy constitutes a very valuable step towards conceptual interoperability across reporting legislations. The xEBR core taxonomy has been semantically "upgraded" in our project to become an ontological module in a set of ontologies describing a class hierarchy and related properties in the broader financial domain. The labels of the core taxonomy (only available in English) are encoded in our ontology by means of the `rdfs:label` annotation property. Results of the information extraction procedure applied to local XBRL

[4] XBRL stands for "eXtensible Business Reporting Language, see `http://www.xbrl.org/` for more details.

[5] This taxonomy, which has not been published yet, is briefly described at: `http://www.monnet-project.eu/Monnet/Monnet/English/Navigation/XBRLEuropeanBusinessRegisterxEBR`.

instance documents are transformed into xEBR and stored as instances of the classes of this ontology.

Nevertheless, the aspect of multilingual terminology has not been resolved. It would be nice to offer a financial analyst not only the concept IDs (and the associated English labels) of the core xEBR taxonomy we can identify in business reports, but also the terms as they are used both in the source taxonomies and in the corresponding documents.

2 Linking Labels of National Taxonomies on the Basis of xEBR

On the basis of conceptual mappings, as displayed in Table 2, we implemented a procedure that extracts all the labels associated to national concepts from national taxonomies as a first step. Thereby, we achieve a mapping between the terms in these labels that is similar to the mapping between national taxonomies and xEBR. So if we, for example, detect the (Belgian) concept *pfs_Intangible FixedAssets* in an XBRL instance document of the Belgian National Bank, this concept is mapped to the xEBR concept *xebr_IntangibleFixedAssetsTotal*. However, in addition to the xEBR English label *Intangible fixed assets*, our procedure delivers all Belgian labels (*Immobilisations incorporelles@fr*, *Immaterielle Anlagewerte@de*, etc.)[6], and interlinks these labels using the SKOS descriptors applied to the corresponding concepts. Thus, we are not only able to deliver the combined xEBR and Belgian National Bank terminology, but we can also automatically link to other national legislations. Our current work focused on the relation between the Belgian and Spanish taxonomies as mediated by xEBR. Our tool also delivers the Spanish correspondences for the "IntangibleFixedAssets" example, both at a conceptual and terminological level, as can be seen in Table 3:

```
"concept" => "pgc-07-c-bs_ActivoNoCorrienteInmovilizadoIntangible"
"prefLabel" => "I. Inmovilizado intangible"
"altLabel" => "Activo no corriente inmovilizado intangible"
```

[The concept in the Spanish taxonomy corresponding to the xEBR concept *xebr_ IntangibleFixedAssetsTotal* with two associated labels in the Spanish language.]

Analysts can submit an instance XBRL document encoded in the Spanish taxonomy to our tool and receive both the xEBR concepts with the associated

[6] Due to limited space, we do no display all labels here. We just mention that the national taxonomies distinguish between labels and verbose labels, which we encode then as *prefLabel* vs *altLabel*, using RDF and SKOS for encoding this information:

<http://www.xbrl.org/xbrl_be.owl#pfs_hasIntangibleFixedAssets>
<http://www.xbrl.org/skos.owl#exactMatch>
<http://www.xbrl.org/xebr.owl#hasIntangibleFixedAssetsTotal> .
<http://www.xbrl.org/xbrl_be.owl#pfs_hasIntangibleFixedAssets>
<http://www.xbrl.org/skos.owl#prefLabel> "Immaterielle Anlagewerte"@de .

English labels as well as the Belgian concepts with the associated labels in four languages. Consequently, we have built an integrated terminological repository, generated on the basis of officially accepted terminologies in different business reporting legislations in Europe. This multilingual term base allows for a semantic processing of instance documents generated by national banks or by business registers, which use these taxonomies as their primary source of knowledge.

3 Extracting Multilingual Term Variants from Web Sources

Our second experiment is dedicated to the extension of the term base we generated from the official taxonomies with automatically detected term variants in on-line sources, which have been automatically extracted as structured or semi-structured data. For the time being, we consult company information on the bilingual web presence of the DAX Index of the German Stock Exchange (deutsche-boerse.com)[7], on the monolingual page of the Bundesanzeiger[8], and in annual reports published directly by companies. The annual report published by the company BASF SE serves as an example herein. In this case, we consult the bilingual, i.e., English and German, PDF reports of BASF manually, contrary to the other sources, from which the data has been extracted automatically.

Concentrating on various reports in various languages for one company for a specific year allows for the additional use of a simple heuristics in order to detect multilingual term correspondences: the financial positions associated with terms have the same values. We are well aware of the fact that this heuristics cannot be applied to all financial positions in reports. For example, the monetary value of *Total assets* and *Total equity and liabilities* should be identical, as can be seen in Table 1, however no equivalence relation can be established as they are no variants of each other. Nevertheless, the taxonomy indicates possible positions of terms in specific parts of tables, which provides us with a precise context for the application of our heuristics.

Some results for the BASF example are summarized in Table 1, which exemplifies that equivalences among monolingual business reporting concepts can be established on the basis of previously normalized financial figures. Thus, a synonymy relation between *Langfristiges Fremdkapital* and *Langfristige Verbindlichkeiten* in German or between *short term assets* and *current assets* in English can be established. As regards the bilingual level, a relation can be established between, for example, the German terms *Langfristiges Fremdkapital* and *Langfristige Verbindlichkeiten* and the English term *Longterm Liabilities.*

[7] See for example the bilingual DAX pages on the company BASF: `http://www.boerse-frankfurt.de/de/aktien/basf+se+DE000BASF111/kennzahlen` and `http://www.boerse-frankfurt.de/en/equities/basf+se+DE000BASF111/key+figures`.

[8] The "Bundesanzeiger" is the official institution for company reporting in Germany. `https://www.bundesanzeiger.de/ebanzwww/wexsservlet`.

Table 1. Monolingual term variants and bilingual term correspondences established by comparing different financial reports for the same company in the same period

German	Figure	English	Source
Umsatzerlöse	63.873	Sales	BASF
Umsatz	63.873		Bundesanzeiger
Umsatz	63.873	Sales	DAX
Langfristige Vermögenswerte	34.532	Long-term assets	BASF
Langfristiges Vermögen	34.532		Bundesanzeiger
Anlagevermögen insgesamt	34.532	Total Capital Assets	DAX
Kurzfristige Vermögenswerte	24.861	Short-term assets	BASF
Kurzfristiges Vermögen	24.861		Bundesanzeiger
Umlaufvermögen	24.861	Total Current Assets	DAX
Langfristiges Fremdkapital	21.168	Long-term liabilities	BASF
Langfristiges Fremdkapital	21.168		Bundesanzeiger
Langfristige Verbindlichkeiten	21.168	Total Longterm Liabilities	DAX
Gesamtkapital (Passiva)	59.393	Total equity and liabilities	BASF
Gesamtvermgen (Aktiva)	59.393	Total assets	BASF
Gesamtkapital (Passiva)	59.393		Bundesanzeiger
Gesamtvermgen (Aktiva)	59.393		Bundesanzeiger
Bilanzsumme	59.393	Total Liabilities and Equity	DAX

Mediated by the corresponding xEBR concepts, these German and English term variants can also be linked to other languages, re-using the mechanisms described in section 2, so that the German term variants *Kurzfristige Vermögenswerte*, *Kurzfristiges Vermögen* and *Umlaufvermögen* can be linked – via the xEBR concept *xebr_CurrentAssetsTotal* – to the Spanish labels *B) ACTIVO CORRIENTE* and *Activo corriente*, which are associated to the concept *pgc-07-c-bs_ActivoCorriente* of the Spanish taxonomy.

4 Conclusion and Future Work

We have described completed and ongoing work in the building of an integrated term base in the domain of standardized business reporting. The starting point is a core taxonomy that maps reporting and financial concepts from various European taxonomies. We integrated this taxonomy in our set of financial and reporting ontologies, proposing at the same time a multilingual extension of the labels with all the terms officially introduced in the national taxonomies. As a second step, we automatically extract term variants for the extended labels of xEBR concepts from on-line sources, also in a multilingual fashion, thus augmenting the term base that supports the information extraction task applied to financial reporting documents.

Acknowledgements. This work has been supported under the Seventh Framework Programme of the European Commission through the Monnet project

(Multilingual ONtologies for NETworked knowledge) co-funded by the European Commission with Grant No. 248458.

References

1. Declerck,T., Krieger, H.U., Thomas, S.M., Buitelaar, P., O'Riain, S., Wunnder, T., Maguet, G., McCrae, J., Spohr, D., Montiel-Ponsoda, E.: Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe. In: Proceedings of MONTIFIC/ECQA Conference, September, Budapest (2012)
2. Wunner, T., Buitelaar, P., O'Riain, S.: Semantic, Terminological and Linguistic Interpretation of XBRL. In: Proceeding of EKAW, October, Lisbon (2010)
3. Declerck, T., Lendvai, P., Wunner, T.: Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems. In: Proceedings of ISA6, ISO/ACL-SIGSEM Workshop, January, Oxford (2011)
4. McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de- Cea, G., Cimiano, P.: Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In: Proceedings of the Fifth workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5), Portland (2011)
5. Aguado-de-Cea, G., Montiel-Ponsoda, E.: Term variants in ontologies. In: Proceedings of the 30th International Conference of AESLA, pp. 19-21 April, Spain, Lleida (2012)