Multilingual Terminology Acquisition for Ontology-based Information Extraction

Christian Federmann, Dagmar Gromann, Thierry Declerck
, Sabine Hunsicker, Hans-Ulrich Krieger, and Gerhard Budin

DFKI GmbH, Language Technology Department,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
{cfedermann,declerck,Sabine.Hunsicker,krieger}@dfki.de
Vienna University of Economics and Business
Nordbergstrasse 15, 1090 Vienna, Austria
dgromann@wu.ac.at
ICLTT-Institut für Corpuslinguistik und Texttechnologie
Sonnenfelsgasse 19/3., 1010 Vienna, Austria
gerhard.budin@univie.ac.at

Abstract. We present current work on the automated acquisition of multilingual terms for labels of ontologies in the financial domain. The main approach consists in harvesting multilingual web pages of stock exchanges, and to extract the relevant data encoded in HTML feature structures from them. Out of these feature structures, we extract and align the multilingual vocabulary that can be used either in labels of classes or properties defined in ontologies, or as part of the value of properties. We also discuss the use of standardized terminological frameworks for improving and validating the results of the automated extraction of multilingual term candidates.

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

1 Introduction

Multilingual ontologies are a topic of active research, for instance within the European R&D Monnet project[1], in the context of which the experiments we describe in this paper are pursued. A use case in the Monnet project is dealing with the cross-lingual access to company business reporting information. This information can for example be used for recognizing trends across borders in the development of economic activity fields[2]. A prerequisite for this functionality is the ability to extract facts from business reporting and other financial information providers and to store them in a (RDF) repository, relating the facts to classes and properties of a corresponding ontology.

[1] See www.monnet-project.eu

[2] A topic being investigated in the European R&D project TrendMiner, which is also contributing to the development of multilingual ontologies in the financial domain. See www.trendminer-project.eu.

Aguado de Cea et al. (Eds.): Proceedings of the 10th Terminology and Knowledge
Engineering Conference (TKE 2012), pp.166-175. 19-22 June 2012, Madrid, Spain

Classes and properties should be equipped with labels in various languages, so that the user can query the system using her/his native language, the queries being mapped onto the labels stored in the knowledge base. Thus, a challenge consists in providing existing ontologies, which mostly have labels in English only, with the needed multilingual extension. This task, also named ontology localisation, is a central aspect of the Monnet project, and [11] discusses methods for ontology localisation and/or translation. In this paper we present a complementary approach to direct localisation/translation by searching candidate terms in various multilingual sources in the financial domain that can be used as labels of ontology classes and properties. The (combined) results of both approaches lead to the establishment of a coherent multilingual extension of ontologies only equipped with labels in one or two languages.

In the first part of the experiment described in this paper, we apply a terminology extraction tool developed for machine translation systems and adapt it to work with under-resourced data such as information extracted from multilingual websites, specifically from the financial domain. We then start to investigate the use of terminology principles and representation tools for validating and aligning the candidate multilingual terms we extracted automatically in the first phase.

2 Web as a Corpus for Extracting Multilingual Terms

Using the Web as corpus offers a valuable resource for building and contrasting comparable corpora on the same domain. With the enormous growth of the Information Society, the Web has turned into a reliable testbed of data for natural language processing, not only in terms of data size but also in terms of data type (e.g., multilingual data, linked data[3]). This has motivated many researchers to start considering the Web as a valid repository for Information Retrieval and Knowledge Acquisition tasks. However, the Web suffers from many problems that are not typically observed in the classical information repositories, such as:

- Web resources are presented in human oriented semantics (natural language) and mixed with a huge amount of information about visual representation;
- the amount of available resources can overwhelm the final user or information engineer that tries to search and access specific data, and it makes complex machine-based processing nonviable for extracting data in an automated way.

Despite of these shortcomings, the Web presents characteristics that can be interesting for knowledge acquisition: due to its huge size and heterogeneity it has been assumed that the Web approximates the real distribution of information in humankind. Moreover, its high degree of redundancy and the presence of publicly available search engines can be useful for developing reliable translation methods. In this context, we find documents that are aligned in various languages on the Web, for instance in the stock exchange domain, which can be reused for

[3] See [1] for more details.

both the semi-automatic creation of parallel corpora and as a source for bi- and multilingual terminology extraction.

One can look for example at the Web site of Euronext (www.euronext.com), which contains various types of information, including lengthy company profiles and company events in four languages: Dutch, English, French and Portuguese. Similar parallel, but bilingual, textual and data sources are available for German↔English on the web page of the German stock market (The DAX index at www.deutsche-boerse.com) and Spanish↔English on the web page of the Spanish stock market (www.bolsamadrid.es). The information contained in these stock exchange pages is available in both structured and unstructured form, whereas both types of information are encoded using clear meta-data, which in most cases are also available in the different languages covered by the web presence.

In order to make those multilingual, aligned terms available to the use cases of the Monnet and TrendMiner projects we need to transform the HTML encoded strings into another format, e.g. an XML encoded multilingual terminology database or directly into the *Lemon* format developed within the Monnet project[4]. However, before we can create these resources, we have to run a terminology extraction step that identifies interesting, parallel terms from the source data. This is a challenging task as the various sources come in different formats, with varying levels of structure and nearly no contextual information.

3 Linguistically Informed Terminology Extraction - TermEx

The TermEx tool was developed in the EuroMatrixPlus project[5] with the explicit goal to extract terminology lists suited for the extension of lexicons for rule-based machine translation (RBMT). Contrary to statistical machine translation systems, RBMT systems rely on their bilingual lexicon to find appropriate expressions in the target language for each translation unit in the source language. Besides the actual terms, the lexicon includes a lot of further information to ensure the appropriateness of the translation. For example, an entry in this lexicon also contains the grammatical gender, declension classes for nouns or subcategorisation frames for verbs.

Creating new entries for the lexicon is a time-consuming and tedious process. As it requires an expert linguist, RBMT suffers from lack of vocabulary coverage and wrong lexical selection. Statistical systems perform better in this area, as they can make use of large bilingual corpora to extend their translation models. The same corpora can and should be used to extend the lexicons of RBMT systems. Whereas statistical systems can rely on the surface forms alone, RBMT systems require that the additional linguistics information is extracted along with the matching surface forms.

[4] See [4] for more details.

[5] See www.euromatrixplus.net and [8] for more details.

3.1 Source Data

TermEx makes use of the following linguistic information which needs to be available for both source and target texts:

- surface form;
- lemma;
- part of speech;
- named entities;
- parse trees.

Parts of speech are used to assign a category to the extracted candidate term: the TermEx tool differentiates between noun, verb, adjective and adverbial terms. The candidates are converted to a dictionary form where the main component, i.e. the main noun, is in its dictionary form, but additional modifiers appear properly inflected based on their type: adjectives are inflected to suit the case and gender of the main noun, if applicable, and prepositional modifiers appear in the correct case. The lemma information is needed to ensure terms of good quality here.

Named entities are treated specially. They are separated from the general words, as named entities often follow specific rules. For instance, they have a semantical gender which needs to be annotated. TermEx can handle named entities which describe persons, locations or organisations—data which is especially important in the financial sector, e.g. in company profiles or stock exchange reports.

The parse trees are required to find candidate terms. This is explained in detail in the next Section.

3.2 Extraction Procedure

Currently TermEx uses parse trees as generated by a proprietary RBMT engine, but the possibility to use other formats will soon be given. The trees are aligned to the source text. When an *interesting* phrase, i.e. a noun phrase, is discovered in the tree, it is extracted along with the additional linguistic information available. Via a word alignment between the two sides of the bilingual corpus, the corresponding translation in the target text is located. As TermEx was designed to extend the coverage, phrases for which the translation by the RBMT system is identical to the reference translation are skipped.

These initial candidate terms are then subjected to heuristic filters to ensure a high precision in the quality of terms. Importing ungrammatical terms into the RBMT lexicon will be detrimental, as they will lead to incorrect translations and thus affect the overall translation quality.

Additionally, to improve the quality of the terms, terminology lists for both translation directions in a language pair are created and only those terms, which appear in both lists, are carried over to the final list.

Our experiment consisted then in applying the TermEx approach to small corpora derived from the multilingual Web presence of stock exchanges and to see whether and how we can acquire multilingual terms about company information as listed on the web pages.

4 Terminology Extraction for Ontologies

The Monnet project needs to apply such an extraction component to the various data sources described above, in order to get terminological equivalents to be used in the labels of multilingual ontologies, supporting also the task of ontology lexicalisation and localisation. The transformation into bi- and multilingual vocabulary XML data bases for the stock exchange pages mentioned above is already available, and a first schema in RDF is available for the German stock market example, which has been extended to an ontology. Parts of the extracted strings (or vocabulary) have been used for defining the T-Box and the R-Box of the ontology, together with the bilingual labels (using the `xml:lang` attribute), and parts of the extracted strings (textual or structured data) have been used for populating the ontology (A-Box)[6]. Current work is dedicated to extending this approach to other web resources and, of course, to improving the overall quality of the extracted string pairs and their alignment onto an accurate multilingual terminology, as described in the next Section.

4.1 Initial Experiments

The information extracted from bilingual websites is structured as feature-value pairs in an XML file. The first task consists of checking whether the information is parallel, i.e. whether source and target side are translations of each other or not. In our initial experiment we found this to be the case for the available feature-value data.

Most of the information is comprised of short phrases with only a few longer texts, mostly descriptions and short portraits. These texts need to be sentence-aligned to be used for TermEx, which was done by hand for our first experiment. The final corpus was made up of 46 sentences in German and English. We compiled the additional information including the parse trees.

We extract 92 term candidates from these 46 sentences, which results in 29 usable terms.

Table 1. Initial Term Candidates.

Type	Count
adverbs	3
adjectives	8
nouns	81
total	92

[6] Very briefly, one can see the T-Box of an ontology as the component introducing concepts and assertions about those, the R-Box being the component that defines related properties and their hierarchy, while the A-Box is specifying properties of individuals, as well as their class membership.

Table 2. Final Terms.

Type	Count
General terms	18
Named Entities	11

The high number of noun terms is not surprising, as there are only few longer texts and most of the information consists of short noun phrases. The term lists still contain quite a number of errors, many of which are due to incorrect word alignment. Since the sentences come from an already well-structured text, this information can be used to improve the alignment to fix especially the number of non-aligned words. Figure 1 shows an initial incorrect alignment: *Segment* remains unaligned.

Correcting this to the alignment shown in Figure 2, the quality of the extracted terms can be increased. We are currently investigating how the alignment can be automatically improved, which is a challenging task due to the shortness of the given terms and the lack of context[7].

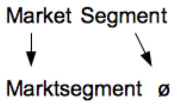

Fig. 1. Incorrect Alignment.

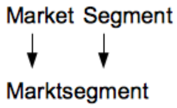

Fig. 2. Correct Alignment.

TermEx was previously only used on large quantities of text made up of full sentences. The data used in this experiment, however, is much closer to the data contained in a translation memory. Currently the usability of TermEx with translation memories is also examined by us.

[7] For sure the use of bilingual dictionaries can help here, if the words are partly covered by the dictionary.

5 Multilingual term alignment

As detailed in the previous section, multilingual terminological equivalents extracted from the stock exchange sources serve as a basis for labels of multilingual ontologies, which are the prerequisite for multilingual ontology-based information extraction. Extracting from comparable corpora on the Web additionally allows us to observe terms in use.

A terminology's full potential can only be explored if the resources are used in context, i.e. the original source and natural occurrence of the term can easily be consulted due to the details provided in the entry. Generally speaking we take a concept-oriented approach towards terminology, being interested in the modeling of domain knowledge. Concept orientation refers to the fact that each term entry contains the full terminological data for the respective concept [2]. Nevertheless, we refrain from considering the relation between concept and term as unequivocal, adding multilingualism further complicates matters.

The knowledge industry is increasingly interested in multilingualism, however, often sacrifices consistency across languages due to time constraints. This paper focuses on the facilitation of term consistency in multilingual ontology labels starting from term extraction and thoroughly validated term bases. Bi- and multilingual information extracted with TermEx from the German stock exchange page are transformed into the Terminological Markup Framework (TMF) [10] compliant TermBase eXchange (TBX) [3] format, as exemplified below.

```
<termEntry id="d20">
    <descrip type="subjectField">Master Data</descrip>
    <langSet xml:lang="en">
        <tig>
            <term>Transparency Standard</term>
            <termNote type="partOfSpeech">noun</termNote>
        </tig>
    </langSet>
    <langSet xml:lang="de">
        <tig>
            <term>Transparenzlevel</term>
            <termNote type="partOfSpeech">noun</termNote>
            <termNote type="grammaticalGender">neuter</termNote>
        </tig>
  <tig>
            <term>Transparenzstandard</term>
             <termNote type="termType">synonym</termNote>
            <termNote type="partOfSpeech">noun</termNote>
            <termNote type="grammaticalGender">masculine</termNote>
        </tig>
    </langSet>
</termEntry>
```

(Simplified TermBase eXchange (TBX) example)

Term consistency is one of the major principles in terminological analysis. Equivalence of the English *Transparency Standard* and the German *Transparenzlevel* could be fully verified. However, within the same set of extracted terminology, the designation *Current Transparency Standard* is matched to the German *Aktueller Transparenzstandard*, using a different German designation from the one above. Stock market resources use both terms interchangeably, assigning equivalent definitions and values to it. Thus, both terms need to be represented in the terminological resource, as the ontology only allows for one prescriptive designation in the `rdfs:label`.

Terminological analysis represents an important aspect in the process of building ontologies from extracted information, as it ensures the validity of natural language representations of concepts. The example above illustrates a reason for the incorporation of terminological resources in ontology-based tasks. Ontologies might not be the adequate resource for representing term variants and the multi-faceted nature of language, aspects that terminologies and lexicons cover for ontologies.

The advantage of using a concept-oriented term base format such as TBX is the clear separation of concept, language, and term level, facilitating its use for ontology building. Furthermore, data categories can be associated with each level, most commonly the standardized categories of ISO 12620 [12]. In order to improve the terms' quality, only terms available in both or all language pairs are included in the final term base. The multilingual vocabulary list needs to be aligned with the final term base encoded in the XML based TBX format.

Additionally, TBX represents an open standard and thus, facilitates re-usability of terms and resources. It allows for a clean and smooth comparison of for instance the Deutscher Aktien IndeX (DAX) terminology with other stock indexes and provides the necessary interoperability of terminological resources in ontology-based tasks.

One problem of a concept-oriented approach to term bases is the varying equivalence of terms across languages. Term variation within one stock exchange index can be considered minimal, as the terms are aligned. However, comparisons across globally acknowledged indexes alter the situation. For instance, the Siemens AG is classified in the sector of *Diversified Industrials*, which holds for the DAX as well as for the Industry Classification Benchmark (ICB) in English.

```
<termEntry id="DAX28">
    <descrip type="subjectField">Industrial</descrip>
    <langSet xml:lang="en">
        <tig>
            <term>Diversified Industrial</term>
            <termNote type="partOfSpeech">noun</termNote>
            <descrip type="definition">Companies with activities across various
            industrial sectors (including holding companies investing in
            different sectors)</descrip>
        </tig>
    </langSet>
    <langSet xml:lang="de">
```

```
        <tig>
            <term>Diverse Industrieunternehmen</term>
            <termNote type="partOfSpeech">noun</termNote>
            <termNote type="grammaticalGender">neuter</termNote>
            <descrip type="definition">Unternehmen, die in mehreren
            verschiedenen industriellen Bereichen ttig sind. Hierzu gehren
            auch Beteiligungsunternehmen, die in unterschiedlichen
            Branchen investieren</descrip>
        </tig>
    </langSet>
</termEntry>

<termEntry id="ICB2727">
    <descrip type="subjectField">Industrial</descrip>
    <langSet xml:lang="en">
        <tig>
            <term>Diversified Industrial</term>
            <termNote type="partOfSpeech">noun</termNote>
            <descrip type="definition">Industrial companies engaged in
            three or more classes of business within the Industrial industry
            that differ substantially from each other.</descrip>
        </tig>
    </langSet>
    <langSet xml:lang="de">
        <tig>
            <term>Diversifizierte Gewerbe</term>
            <termNote type="partOfSpeech">noun</termNote>
            <termNote type="grammaticalGender">neuter</termNote>
            <descrip type="definition">Industrieunternehmen, die in
            drei oder mehr, verschiedenen Geschftszweigen innerhalb einer
            Branche ttig sind.</descrip>
        </tig>
    </langSet>
</termEntry>
```

(Simplified TermBase eXchange (TBX) example of extracted data)

The process of analyzing and defining concepts and relations is crucial for each language in a multilingual terminology. For instance, the English designations of the concept above might be considered an exact match, lexically as well as conceptually, whereas the German version differs orthographically. However, taking its definitions into account, the German matching of concepts can be easily verified, especially since the ICB definition clearly refers to *Industrieunternehmen*.

This example further illustrates that despite of automatic term extraction, the actual identification and alignment of terms has to be at least verified manually. To automatically construct term alignments with varying degrees of equivalence across concepts and languages within one concept, i.e. term entry, seems rather challenging. Prerequisites for the task are basic lexical information, such as part of speech, lexical semantic relations, and linguistic structures, such as the head-modifier principle, constituency and dependency information, etc.

Naturally, relations among terms differ from relations in ontologies. Term bases utilize hierarchical, partitive, or associative relations. However, the hierarchy is flattened to a string in TBX. In contrast, the RDF-based SKOS format provides more elaborate means to representing hierarchical relations for terminologies. Nevertheless, SKOS constitutes a highly prescriptive approach to the classification of terms. Thus, instead of using SKOS we seek to render TBX in RDF in order to obtain ontology labels, such as the short ontology example below:

```
<owl:FunctionalProperty rdf:ID="transparencyStandard">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <rdf:type rdf:resource="#DiachronicProperty"/>
    <rdfs:label xml:lang="de">Aktueller Transparenzstandard</rdfs:label>
    <rdfs:label xml:lang="en">current transparency standard</rdfs:label>
    <rdfs:domain rdf:resource="#Company"/>
    <rdfs:range rdf:resource="#TransparencyStandard"/>
 </owl:FunctionalProperty>
```

(Example of DAX ontology)

6 Conclusions

We have described how a tool developed for use in machine translation, can be applied to terminology extraction that feeds into building multilingual ontologies. First results show that word and phrase alignment errors have a severe impact on the quality of the extracted term pairs. We are working on improvements of this, using linguistic methods. But considering the lessons we could draw from the experiment, and also due to the fact that multilingual labels in ontologies have to be very accurate, we see the need for a manual post-processing of any type of terminology extraction used for populating labels of multilingual ontologies. We suggest adopting terminological principles and frameworks for these tasks, such as TBX, the functionalities of which we are extending in porting it to RDF. In this scenario, we remain confident that by applying terminology extraction to bilingual websites from the financial domain, the creation of multilingual ontologies can be supported and their coverage and quality can be improved.

Acknowledgments

This work has been co-funded under the Seventh Framework Programme of the European Commission through the EuroMatrixPlus contract 231720, the Monnet (Multilingual ONtologies for NETworked knowledge) contract 248458 and the TrendMiner contract 287863.

References

1. Bizer, C., Heath, T., Berners-Lee T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS). 5:3, 1-22 (2009)
2. Bassey, A., Budin, G., Picht, H. Rogers, M., Schmitz, K.D., Wright, S.E.: Shaping Translation: A View from Terminology Research. Translators' Journal 50:4, 195–197 (2005)
3. ISO 30042 (2008): Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX), Geneva, ISO.
4. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. The Semantic Web: Research and Applications. Volume 6643 of Lecture Notes in Computer Science, 245-259. Springer, Berlin (2011)
5. Declerck, T., Lendvai, P. and Wunner, T.: Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems. Proccedings of the Sixth Joint ISO - ACL/SIGSEM Workshop on Interoperable Semantic Annotation 2011.
6. Miles, A., Bechhofer, S.: SKOS-Simple Knowledge Organization System Reference, W3C Recommendation, 18 August (2009)
7. Federmann, C. and Hunsicker, S.: Stochastic Parse Tree Selection for an Existing RBMT System Sixth Workshop on Statistical Machine Translation (2011).
8. Federmann, C. and Hunsicker, S., Wolf, P., and Bernardi, U.: From Statistical Term Extraction to Hybrid Machine Translation. Proceedings of the 15th Annual Conference of the European Association for Machine Translation, Leuven, Belgium, EAMT (2011)
9. Montiel-Ponsoda, E., Aguado-de-Cea, G., McCrae, J.: Representing term variation in *lemon*. Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011,47-50, Paris (November 2011)
10. ISO 16642 (2003): Computer applications in terminology - Terminological markup framework, Geneva, ISO.
11. McCrae, J., Espinoza, M., Montiel-Ponsoda, E., Aguado-de-Cea, G. and Cimiano, P.: Combining statistical and semantic approaches to the translation of ontologies and taxonomies. Proceedings of the 5th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5) (2011).
12. ISO 12620 (2009): Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources, Geneva, ISO.