LINGUISTIC INFORMATION FOR MULTILINGUALITY IN THE SEMBYSEM PROJECT

Ingrid Falk, Samuel Cruz-Lara, Nadia Bellalem and Lotfi Bellalem
INRIA Nancy Grand-Est, Nancy Université, Nancy, France

Keywords: Knowledge representation, Ontology engineering, Multilingual linguistic representation, Ontology and lexicon.

Abstract: In this paper we discuss ways to handle multilingual linguistic information within the framework of the SEMbySEM project[a]. The SEMbySEM project aims at defining tools and standards for the supervision and management of complex and dynamic systems by using a semantic abstract representation of the system to be supervised or managed. As we want our system to conform to an end-user's point of view, the conceptual information must be available and presentable in the end-user's language. On the other hand, lately the need for and benefits of more accurate linguistic information associated to ontological knowledge representations have become more evident and there emerged models of how this articulation could be achieved. Two of these models are LexInfo (Buitelaar et al., 2009) and LIR, the Linguistic Information Repository (Montiel-Ponsoda et al., 2008). In this paper we explore these models under the prospect of putting one or both in praxis in the setting of the SEMbySEM project.

[a]SEMbySEM (http://www.sembysem.org) is a research project within the European ITEA2 programme (http://www.itea2.org/). It started June 2008 and will end December 2010.

1 INTRODUCTION

The SEMbySEM project aims at providing a framework for universal sensors management using semantic representations. A detailed description can be found in (Brunner et al., 2009), here we give a brief overview and concentrate on the aspects related to language and linguistic information. A sensor system supervises and manages the data coming from various sensors with varying technical specifications and placed on various objects. The sensors collect and transmit data and a sensor management system must make sense of and visualise this data. To achieve this the SEMbySEM system will be organised in a three layered architecture as shown in Figure 1. The interaction with the sensors (registering and processing events from the sensors) is done in the basic layer, the *Faade Layer*. The information from the sensors is unified and processed and may then trigger an update of the semantic model of the system. The semantic model together with a rule system make up the middle layer, the *Core Layer*. End-users connect to the system through the top layer, the *Visualisation Layer*. They have access to tailored view points designed by expert users and HMI experts through which the data from the semantic model is displayed. From the linguistic point of view the relevant modules are the *Core* and the *Visualisation Layer*. It was decided against using OWL and Description Logic which are habitually employed to represent semantic information in this setting (Brunner et al., 2009). The main reasons were that they were difficult to handle by business users and that OWL fails to express specific business needs in some use cases. The business-

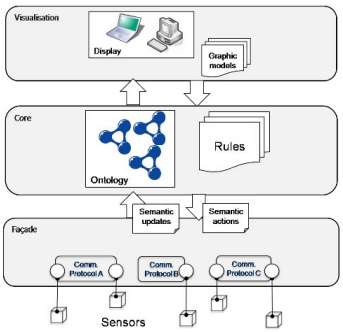

Figure 1: SEMbySEM architecture.

Falk I., Cruz-Lara S., Bellalem N. and Bellalem L. (2009).
LINGUISTIC INFORMATION FOR MULTILINGUALITY IN THE SEMBYSEM PROJECT.
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 447-451
DOI: 10.5220/0002304404470451

oriented model developed instead (**MicroConcept**) is designed to meet the additional requirements and use existing standards wherever possible in order to leverage existing design tools eg. the lexicalisation tools to be discussed later in this paper. The ontology design is intended to be done at design time, as well as the viewpoints and HMI modeling. The latter will not change at runtime.

2 LINGUISTIC NEEDS IN SEMBYSEM

SEMbySEM needs (multilingual) linguistic information on the conceptual level (cf. the *Core Layer*, Fig. 1) and on the GUI or visualisation level. It only needs to be provided at design time and does not need to be updated at runtime. At the conceptual level it is needed to help at the design and maintenance of the semantic model.With regard to the visualisation level the needed (conceptual) information comes from two sources: from the domain ontology on one hand and from elements pertaining to the HMI or GUI. The latter objects can also be modeled by an ontology thus allowing for a uniform data representation. These two ontologies could then be *lexicalised* using one of the two representation models we will describe in Section 3 and 4. The result of the lexicalisation will be one or more lexical repositories, which can then be used in the visualisation process, ie. at the design of the viewpoints and HMIs which are employed in the *Visualisation Layer* (cf. Figure 1).

3 RELATED WORK

Linguistic Information for Ontologies. The need to ground ontology elements in natural language has become evident on the basis of some or all of the following reasons: Ontologies are developed by several people from different communities, over a longer period of time and are meant to be reused – linguistic information is needed to establish and assure the consistence of the human linguistic and cognitive systems with the ontological machine-readable conceptualisation system; Precise linguistic information allows for automatic procedures for ontology-based information extraction from text which in turn help at semi-automatic ontology population; Richer linguistic models capture how concepts and relations are realised in language and therefore help to verbalise and explain ontology elements. We identified two emerging best practices which we could use to attach linguistic and lexical information to ontology elements: LexInfo (Buitelaar et al., 2009) and LIR, the Linguistic Information Repository (Montiel-Ponsoda et al., 2008). In both representations domain knowledge and linguistic information are clearly separated while the articulation between language and ontological systems remains flexible. They both explicitly take into account multilingual aspects. The linguistic information is represented as lexical ontologies. Both systems explicitly adhere to the integration of other existing standards as eg. LMF (Francopoulo et al., 2007), LingInfo (Buitelaar et al., 2006) and LexOnto (Cimiano et al., 2007) but the implementations as ontologies differ mainly in that LIR focuses on representing aspects related to the meaning of lexical entries (eg. synonymy, antonymy, semantic relatedness, variations in meaning conditioned by cultural or regional differences) whereas LexInfo allows for accurate and elaborate representations of morphological relations between terms, morphological and syntactical decomposition of terms and complex linguistic patterns (eg. the mapping between subcategorisation frames[1] and predicate-argument structures[2]). We will further illustrate the two linguistic models and their possible integration with SEMbySEM in Section 4.

4 LINGUISTIC MODELS AND THEIR APPLICATION TO SEMBYSEM

Here we briefly describe the linguistic models LexInfo and LIR and explore ways to use them to associate linguistic information to the semantic model of SEMbySEM. In both frameworks the procedure is the following: (1) Starting from the domain ontology . . . (2) the system builds an empty or default lexical ontology, the main building blocks of which are the *lexical entries* (LEs) which roughly correspond to the domain ontology components i.e. the classes and properties. The domain ontology elements are linked to at this stage possibly empty LEs which in turn are associated to the corresponding ontology elements. LEs are constructed based on linguistic analysis of the domain ontology labels, comments and/or identifiers; (3) The lexicon is enriched (semi-)automatically using domain relevant texts and/or external lexical re-

[1] A subcategorization frame of a word describes a syntactic construction this word may be used in.

[2] The predicate-argument structure of a sentence represents the meaning of this sentence as a combination of a predicate and its arguments

sources eg. WordNet[3] or Wikipedia to search for further information (eg. definitions, translations); (4) The lexical ontology is further completed manually.

LexInfo models linguistic information in an ontology combining three previously proposed approaches: LingInfo (Buitelaar et al., 2006), LexOnto (Cimiano et al., 2007) and LMF (Francopoulo et al., 2007). At the time of writing, it mainly handles ontological properties which are in most cases realised in language as verbs and relational nouns[4]. For example starting from the following property definition from a SEMbySEM use-case:

```
<owl:ObjectProperty rdf:about="ontologyNS#engineOf">
  <rdfs:domain rdf:resource="ontologyNS#Engine"/>
  <rdfs:range rdf:resource="ontologyNS#Train"/>
</owl:ObjectProperty>
```

the tool would generate a lexicon for English which would link to 3 types of elements of the lexical ontology: a lexical entry for the *engine* component of the property identifier *#engineOf*, a predicate entry representing the predicate *engineOf* and to one or more subcategorization frames representing possible linguistic realisations of this predicate.The predicate-argument structure of the *engineOf* property ⟨ *engineOf*, Domain, Range ⟩ may be inferred from its ontological representation and is represented in LexInfo as shown in Figure 2, the rectangle with gray background. A possible linguistic realisation is represented by the subcategorisation frame *NounPP* which is shown in Figure 2 as the rectangle with filled nodes. The mapping between the semantic predicate argument structure and the syntactic frame is modeled as can be seen again in Figure 2. Finally, the link between lexicon entries and ontology elements is provided through the *hasSense* property (not shown in the figure). In this example we started from the representation of a property in a domain ontology from which **LexInfo** inferred a (linguistic) semantic representation (the predicate argument structure) which then was mapped to a syntactic construction (the subcategorisation frame) representing a possible realisation in language which, in a subsequent step, can help detect linguistic realisations of the involved ontological elements in relevant texts.

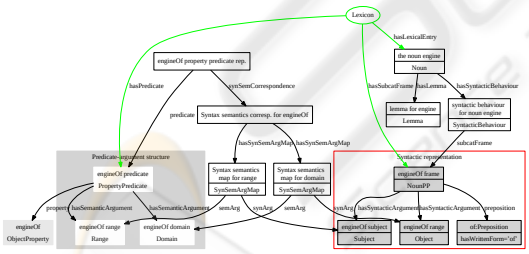

Figure 2: Syntax ↔ semantics mapping in LexInfo.

Localisation. Once the English lexicon is constructed, similar lexica must be built for each other required language. Within **LexInfo** lexical information pertaining to a given language is grouped in one *Lexicon*, so there would be one lexicon for each language.

LIR. (Linguistic Information Repository) is a model for associationg lexical information to OWL ontologies proposed within the NeOn project. The linguistic information pertaining to a given domain ontology is modeled as an OWL ontology. In contrast to LexInfo, LIR concentrates mainly on ontology classes, which are in general represented as plain noun or noun phrase class labels. As for LexInfo we start from an ontology element (in this case a class):

```
<owl:Class rdf:about="ontologyNS#Train">
  <rdfs:label xml:lang="fr">Train</rdfs:label>
  <rdfs:comment xml:lang="en">Train</rdfs:comment>
</owl:Class>
```

Figure 3 schematically shows the way it is represented linguistically within the LIR framework.

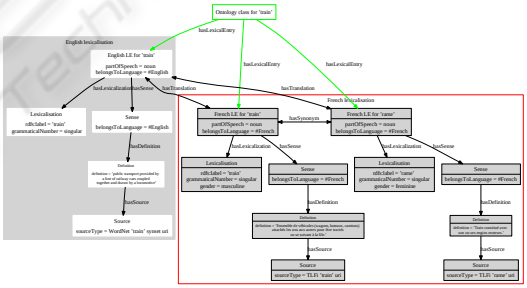

Figure 3: Linguistic representation with LIR.

Localisation is performed at the lexical entry level. As Fig. 3 shows, the ontology class is connected via the *hasLexicalEntry* relation to English and French lexical entries (LEs). LEs in different languages may be linked through the *hasTranslation* relation and within the same language through synonymy or antonymy relations. Their sense is typically represented through a definition from an external resource, which may also come from a different language. In LIR it is possible to express the fact that two words are synonym in most cases by linking their sense definitions through the *isRelatedTo* relation. Finally, within LIR all lexical entries, irrespective of the language they belong to, are members of the same lexical ontology.

[3] http://wordnet.princeton.edu/

[4] Most nouns as eg. *train* refer to an ontology class. The nouns we call *relational*, eg. *speed*, may be used to express an ontology property.

Discussion. LexInfo and LIR are both conceptualisations of lexicalising a domain ontology in the form of an ontology structure and they both aim at integrating the same existing standards (LMF). Still the resulting ontological structures are quite different. Firstly they are different from a syntactic point of view. Second, they also differ semantically: LexInfo rather emphasises the representation of properties and in particular the syntax ↔ semantics interface. For example with LexInfo it would be possible to express that the sentences *The train speeds at 100 km/h.* and *The speed of the train is 100km/h.* are linguistic realisations of the same meaning. LIR adopts a more traditional lexicographic position: one could express for example that *train* and *rame* are both possible French translations of the English *train*, that they may be used as synonyms in most cases but are not entirely synonymous. LexInfo is arguably more suitable at verbalising ontologies or translating natural language queries to database or web search queries, whereas LIR would be more useful to human knowledge engineers or lexicographers for building and maintaining a domain and lexicon ontology. It is in principle possible to translate one lexicon format to the other, but the syntax – semantics mapping information of LexInfo can not be extracted from LIR and conversely, the lexicographic aspects represented in LIR can not be generated from LexInfo. On this note the two models are complementary. This difference in perception also implies different (semi-)automatic acquisition methods from texts: to enrich a LexInfo lexicon one needs deep syntactic and semantic linguistic analysis whereas for LIR one would make use of statistic semantic similarity measures and high-quality linguistic semantic resources as (Euro)WordNet or Wikipedia. From the SEMbySEM position a point in favour of LexInfo is that it generates a separate lexicon for each language whereas the fact that SEMbySEM needs lie more on the side of ontology building (by humans) speeks in favour of LIR. These differences also entail different APIs and differences in the manipulation of the lexical ontologies.

5 CONCLUSIONS

In this paper we presented the SEMbySEM semantic model and explored ways of associating multilingual information to its elements. Two state-of-the-art techniques for representing and associating linguistic information to ontological structures are LexInfo and LIR. We presented and compared the principles of these models and investigated whether and how they could be integrated with the SEMbySEM semantic model. LexInfo and LIR both are designed as lexical ontologies building on a domain ontology represented in OWL. Although the SEMbySEM semantic model will not be represented as an OWL ontology, it will be designed following the same basic principles and it will therefore be possible to represent the linguistic information using LIR or LexInfo. We showed that LIR and LexInfo take up different positions mainly with respect to the kind of linguistic information they focus on: the syntax – semantics interface for LexInfo and more traditional lexicographic aspects for LIR. These differences entail different acquisition methods, different APIs and differences in the handling of the obtained lexical ontologies. Both present advantages compared to the current way of representing linguistic information by the `rdfs:label` and `rdfs:comment` elements, in that they clearly separate domain knowledge from lexical knowledge. Thus domain and lexicon ontology can be developed separately by domain and linguistic experts and can be more easily maintained and reused. However, at this stage of the project we have not yet decided which of them to use.

ACKNOWLEDGEMENTS

This work is carried out by the EUREKA ITEA2 project SEMbySEM partly funded by French, Spanish, Finnish and Turkish governments. We would also like to thank Elena Montiel-Ponsoda and Philipp Cimiano for providing very helpful insights regarding the LIR and LexInfo model respectively.

REFERENCES

Brunner, J.-S., Goudou, J.-F., Gatellier, P., Beck, J., and Laporte, C.-E. (2009). SEMbySEM: a Framework for Sensors Management. In *1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009)*.

Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards Linguistically Grounded Ontologies. In *The 6th Annual European Semantic Web Conference (ESWC2009)*, Heraklion, Greece.

Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., Porzel, R., and Cimiano, P. (2006). Linginfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of the OntoLex Workshop at LREC*, pages 28–32. ELRA.

Cimiano, P., Haase, P., Herold, M., Mantel, M., and Buitelaar, P. (2007). LexOnto: A Model for Ontology Lexicons for Ontology-based NLP. In *Proceedings*

of the OntoLex07 Workshop held in conjunction with ISWC'07.

Francopoulo, G., Bel, N., Georg, M., Calzolari, N., Monachini, M., Pet, M., and Soria, N. (2007). Lexical Markup Framework: ISO standard for semantic information in NLP lexicons. In *Proceedings of the Workshop of the GLDV Working Group on Lexicography at the Biennial Spring Conference of the GLDV*.

Montiel-Ponsoda, E., Peters, W., auguado de Cea, G., Espinoza, M., Pérez, A. G., and Sini, M. (2008). Multilingual and lozalization support for ontologies. Technical report, D2.4.2 NeOn Project Deliverable.