Ontology Lexicalisation: The *lemon* Perspective

Paul Buitelaar•, Philipp Cimiano⋆, John McCrae⋆, Elena Montiel-Ponsoda†, Thierry Declerck‡
•Unit for Natural Language Processing, DERI, National University of Ireland, Galway
⋆Semantic Computing Group, CITEC, University of Bielefeld, Germany,
†Ontology Engineering Group, Universidad Politécnica de Madrid, Spain,
‡Language Technology Lab, DFKI, Germany

1 Introduction

Ontologies (Guarino1998) capture knowledge but fail to capture the structure and use of terms in expressing and referring to this knowledge in natural language. The structure and use of terms is the concern of terminology as well as lexicology. In recent years, the relevance of terminology in knowledge representation has been recognized again (for example the advent of SKOS[1]) but less consideration has been given to lexical and linguistic issues in knowledge representation (Buitelaar2010).

2 Use Cases of Ontology Lexicalisation

Natural language is often the medium of choice for knowledge representation and transfer between humans. However, ambiguity is widespread in natural language. Words have multiple meanings and grammar can be ambiguous in structure and therefore in interpretation. However, such ambiguities appear to provide little issue to people, who can with little effort resolve these ambiguities in nearly all situations. Machines, on the other hand, have significant issues in resolving these ambiguities and this can lead to difficulties in defining precise interpretations in technical domains. To illustrate this we will now briefly explore some of the use cases of ontology lexicalisation, i.e. in knowledge acquisition from text and multilingual knowledge access.

2.1 Knowledge Acquisition from Text

In the case of knowledge acquisition from text we aim to identify relevant text segments and align these with formally defined knowledge structures, such as facts and axioms. Let us focus on ontology-based information extraction, that is, the extraction of facts from text relative to a given ontology. Consider for example an ontology on tourism with ontology labels (terms) in Spanish. The ontology defines concepts of relevance to tourism such as historical buildings, which will be defined by use of the Spanish term (ontology label) "edificio histórico". For instance, in the following sentence there is a specification of a set of facts concerning a historical building (*Universidad de Barcelona*), its architect (*Elies Rogent*), and building period (*1863-1882*):

- "El edificio histórico de la Universidad de Barcelona es obra de Elies Rogent, se inició su construcción en 1863, pero no se concluyó hasta 1882." (The historical building of the University of Barcelona is the work of Elies Rogent, its construction began in 1863, but was not completed until 1882.)

Observe that the match between ontology label and text is straightforward, as they are identical. However, this is not the case in the following example:

- "El Cabildo de Buenos Aires, ... El edificio, declarado Monumento Histórico Nacional desde el año 1933, fue objeto de sucesivas alteraciones, ... " (The Cabildo of Buenos Aires, ... The building, declared a National Historic Landmark in the year 1933, underwent successive alterations, ...)

In this case, the text segment again specifies a set of facts on a historical building (*El Cabildo*),

[1] http://www.w3.org/2009/08/skos-reference/skos.html

WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 33–36
Paris, 10 November 2011

its location (*Buenos Aires*), and dedication date (*1933*), but the match between ontology label and text is not straightforward and requires the representation of linguistic information to compute morphological and syntactic variants.

2.2 Multilingual Knowledge Access

Ontology lexicalisation can be extended to multiple languages, enabling applications such as multilingual ontology-based question answering. Consider the following question in English, Dutch, German and Spanish:

- "Who painted the Mona Lisa?"
- "Wie schilderde de Mona Lisa?"
- "Wer malte die Mona Lisa?"
- "¿Quién pintó la Mona Lisa?"

Intuitively, the answer to these questions should be the same and thus independent of the specific language the question is expressed in. According to our main hypothesis, we claim that these questions could be translated into a normalized language-independent representation that can be evaluated with respect to semantically structured data. For example, we could use a formal query in the SPARQL language to express these questions in a way that abstracts from the original language:

```
PREFIX rdf: .../22-rdf-syntax-ns#
select ?who  where {
<http://dbpedia.org/.../Mona_Lisa>
<http://dbpedia.org/.../artist>
?who
}
```

The strings enclosed in angle brackets represent URIs (Uniform Resource Identifiers) that uniquely identify a certain entity (*Mona Lisa*) and a property (*artist*). The fact that the label of the property artist is English should not mislead; the URI represents a real-world relation between paintings and their creators and just happens to be labeled with an English string for the sake of human readability. The existence of such a relation is however independent of a specific language. In any case, in order to map the above question into a normalized and language-independent representation, i.e. the SPARQL query above, we require knowledge about the fact that the verb "schilderen" in Dutch, "malen" in German, "pintar" in Spanish and "paint" in English all refer to the property artist.

3 A Lexicon Model for Ontologies

Given the motivations for ontology lexicalisation given by the use cases outlined above and the fact that a solution for this seems missing in current state of the art research and best practices, we propose a formal model for the proper representation of the continuum between: i) ontology semantics; ii) terminology that is used to convey this in natural language; and iii) linguistic information on these terms and their constituent lexical units. As this model in essence enables the creation of a lexicon for a given ontology, we call this a *lexicon model for ontologies*.

3.1 Requirements

The requirements for a lexicon model for ontologies address several different goals. In particular, the model should: i) represent linguistic information relative to the semantics given by the ontology, thereby avoiding the representation of unnecessary lexical features that may lead to overgeneration of term variants; ii) strict separation of 'world knowledge' (describing domain objects that are referenced by lexical objects) from 'word knowledge' (describing lexical objects); iii) enable easy uptake of the model by providing a simple core model, supplemented with a set of modules that can be used, extended or ignored upon need.

3.2 lemon: lexicon model for ontologies

The proposed lexicon model for ontologies ('lemon') is described in detail in the 'lemon cookbook'[2]. Here we provide a summary of its most prominent features, starting with the lemon core, which is organized around a *core path* as follows:

- **Ontology Entity**: URI of an ontology element to which a **Lexical Sense** points, providing a possible linguistic realisation for that **Ontology Entity**
- **Lexical Sense**: functional object that links a **Lexical Entry** to an **Ontology Entity**, providing a sense-disambiguated interpretation of that **Lexical Entry**

[2] http://lexinfo.net/lemon-cookbook.pdf

- **Lexical Entry**: morphosyntactic normalisation of one or more **Lexical Form**
- **Lexical Form**: morphosyntactic variant of a **Lexical Entry**, including inflection, declination and syntactic variation
- **Representation**: standard written or phonetic representation for a **Lexical Form**

In addition, lemon has a number of modules that allow for further modeling:

- The **linguistic description module** is concerned with the use of data categories such as ISOcat for describing lemon elements. Although lemon itself is a meta-model and therefore agnostic as regards the specific data category set used, specific data categories can be used in particular instances of the lemon model.
- The **morphology module** is concerned with the analysis and representation of inflectional and agglutinative morphology. The module allows the specification of regular inflections of words by use of Perl-like regular expressions.
- The **phrase structure module** is concerned with the modeling of lexical entries that are syntactically complex, such as phrases and clauses, to enable representation of the syntactic structure of such lexical entries.
- The **syntax and mapping module** is concerned with a description of lexical 'predicates' (sub-categorisation frames with syntactic arguments) and semantic predicates (properties with subject/object) on the ontology side and the mapping between them.
- The **variation module** is concerned with a description of the relationships between elements of a lemon lexicon: sense relations (e.g. translation) require a semantic context, lexical variations (e.g. plural) require a morphosyntactic context, form variations (e.g. homographs) include all other variations.

4 Conclusions

In this paper we presented a motivation for ontology lexicalisation that builds on use cases, among others, in knowledge acquisition from text and multilingual knowledge access. We argued that the representation of a lexical level in ontologies, beyond the semantic and terminological level, is needed for a proper use of ontologies in applications and also serves in integrating the terminology level with the ontology level. No previously available model (e.g. (Gangemi et al.2003), (Farrar and Langendoen2003), (Reymonet et al.2007)) fulfills all the requirements for an ontology lexicalisation model. We therefore developed a model (lemon) for this purpose, of which we discussed some of its main features and directions in which it is currently used. Full details of the model and details of its use are described in other papers to which we refer the interested reader (Buitelaar et al.2009), (McCrae et al.2011), (McCrae et al.forthcoming).

Acknowledgements

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion2) and the CITEC excellence initiative funded by the EU and the DFG.

References

Buitelaar P. (2010) **Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions** *In: Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Oltramari, Alessandro Lenci, Laurent Prevot (eds.) Ontology and the Lexicon: A Natural Language Processing Perspective* Cambridge Studies in Natural Language Processing, Cambridge University Press.

Buitelaar P., P. Cimiano, P. Haase, M. Sintek (2009) **Towards Linguistically Grounded Ontologies** *Proceedings of the 6th European Semantic Web Conference.* Lecture Notes in Computer Science, Springer.

Farrar S., D. Terence Langendoen (2003) **A linguistic ontology for the Semantic Web** *GLOT International. 7 (3)*, pp.97-100.

Gangemi A,, R. Navigli, P. Velardi (2003) **The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet** *Proceedings of ODBASE*, Springer.

Guarino, N. (1998). **Formal Ontology in Information Systems** *In: N. Guarino (ed.) Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998.* IOS Press, pp.3-15.

McCrae J., D. Spohr, P. Cimiano (2011) **Linking Lexical Resources and Ontologies on the Semantic Web with Lemon** *Proceedings of the 8th European Semantic Web Conference*, Lecture Notes in Computer Science, Springer, Volume 6643, pp.245-259.

McCrae J., G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner (forthcoming) **Interchanging lexical resources on the Semantic Web** Accepted for publication in *Language Resources and Evaluation*, Springer.

Reymonet A., J. Thomas, N. Aussenac-Gilles (2007) **Modelling ontological and terminological resources in OWL-DL** *Proceedings of the ISWC07 workshop From Text to Knowledge: The Lexicon/Ontology Interface (OntoLex '07).*