A Model for Representing Diachronic Semantic Information in Lexico-Semantic Resources on the Semantic Web

Fahad Khan, Francesca Frontini, Monica Monachini

Introduction

The Semantic Web offers a way of publishing structured data online that facilitates the interlinking of different datasets stored at different online locations; indeed one of the main aims of the Semantic Web movement is to actively encourage this enrichment of online datasets with information from other resources, in order to avoid the problem of so called 'data islands'. In contrast to conventional hyperlinks however the links between different resources on the Semantic Web can be given semantic types and classified hierarchically. Data published on the Semantic Web is referred to as Linked Data; if, in addition, this data is available with an open license then it can be referred to as Linked Open Data (Heath 2011).

The benefits of Linked Open Data for cultural resources, in particular for heritage resources in languages like Ancient Greek, Latin, and Classical Arabic are manifold (Oomen 2012). Therefore the development and/or appropriation of models for representing cultural heritage resources on the Semantic Web is an important area of research. In this article we will focus on representing lexical resources as linked data, and in particular the representation of diachronic semantic information.

Lemon

In order to motivate what follows we will give a little background on the model used to represent data on the Semantic Web, namely the Resource Description Framework (RDF). In the RDF model, facts or statements are represented by subject-predicate-object triples. Each member of a subject-predicate-object triple is a so-called resource with a unique identifier which is referred to, unsurprisingly, as its Unique Resource Identifier (URI)[1]. Given a statement like "Scott is the author of Waverley", then, we can write this as a triple in the following format:

<http://dbpedia.org/resource/Walter_Scott><http://dbpedia.org/ontology/author><http://dbpedia.org/resource/Waverley_(novel)>.

Here the text between each pair of angular brackets represents a URI. In this instance all three URIs belong to the dbpedia dataset, a linked data version of Wikipedia which is currently the largest and, arguably one of the most important nodes on the semantic web. Each of the three URI's above represents the entity which it refers to, and each can be 'dereferenced' in the sense that its URI will give us access to a description of the entity in question with further links to other resources referred to by URIs.

When it comes to representing lexical resources on the Semantic Web the linked data paradigm enables the linking together of lots of different kinds of information using categories and relations

[1] Objects in Subject-Predicate-Object triples can also be literals, e.g., strings or integers.

from diverse datasets. Take for example a Latin RDF lexicon with an entry puella representing the the Latin word *puella*. We can attach various kinds of information to this entry respecting the morphosyntactic properties associated with the word, for example its part of speech, its declensions, etc, using concepts and relations from various other online vocabularies. In addition, we can represent the meaning of the entry by linking the entry itself a concept in an ontology; this ontology will then allow us to relate the concept associated with the word to a network of other concepts and if the ontology is written in a language like OWL we can make use of a number of pre existing reasoners to carry out inferences.

The *lemon* model (McCrae et al. 2011), following up on previous work in LMF and lexinfo, provides a model for describing lexico-semantic resources in RDF. *lemon* uses sense objects to represent the meaning of words, these sense objects are reified pairings of a word and its extension where the extension is represented by an ontological concept: the sense object itself can be viewed as the intension of the word. A lexical entry is linked to a sense by the *sense* relation and this sense object in turn links to an item in an ontology using a *reference* relation. So for example the entry puella links to a sense object which has a reference in an ontology. A word may be linked to two or more sense objects if it is polysemous.

LemonDia

lemon allows the addition of temporal information via the use of the usedSince property[2]. This may not be sufficient, however, for describing information relating to the temporal validity of the different senses of a word or for tracking how different word senses evolve one into another. It is useful to have a more detailed representation of the evolution of word senses when it comes to constructing lexica for classical languages (or even explicitly diachronic lexica for modern languages like English or French) in which we want to represent information about a language at different stages in its evolution.

Unfortunately it turns out that adding a temporal dimension to RDF triples is a notoriously difficult problem since we are in effect confined to binary relations and breaking up n-ary relations into binary relations raises a number of other issues. The best solution seems to be a more high level one, namely, to work with perdurants, that is, entities with an associated temporal span in the course of which various (essential) properties may or may not hold at different intervals (Welty 2006).

In (Khan 2014) we describe an extension of *lemon* in which the sense objects linked to words are modelled as perdurants, namely as an entities with an associated temporal span for which various (essential) properties may or may not hold at different intervals of time. In addition these different sense entities can be combined in one meaning shift entity. This allow us to explicitly track and to query the meaning shifts of words.[3] With this explicit encoding of word senses as

[2] See the lemon cookbook for more details: http://lemon-model.net/lemon-cookbook.pdf

[3] The lemonDIA model is available here:
http://www.languagelibrary.eu/lemonDia/lemonDia.owl
Two example lexica written using the lemonDIA model are available here:

temporal entities we can more easily represent the dynamic semantic aspects of the lexicon. We believe that this work can contribute towards developing new appropriate models for the representation of cultural heritage resources as linked data.

Bibliography

Tom Heath and Christian Bizer (2011) Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

Fahad Khan, Federico Boschetti and Francesca Frontini (2014), Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources .LDL-2014 3rd Workshop on Linked Data in Linguistics.

J McCrae, D Spohr, P Cimiano (2011). Linking lexical resources and ontologies on the semantic web with lemon. The Semantic Web: Research and Applications.

J Oomen, LB Baltussen, M van Erp (2012). Sharing cultural heritage the linked open data way: Why you should sign up. Museums and the Web.

Christopher Welty , Richard Fikes , Selene Makarios (2006). A reusable ontology for fluents in OWL. In Proceedings of FOIS.

---

http://www.languagelibrary.eu/lemonDia/example_lexicon_1.rdf
http://www.languagelibrary.eu/lemonDia/example_lexicon_2.rdf