LREC 2012 Tutorial: Enriching the Web with Ontology-lexica

John McCrae, Brian Davis and Jorge Gracia

Monday 21st May

This tutorial will introduce, motivate and detail models for representing linguistic information relative to ontologies and sharing it on the web with the aim of improving the quality and scope of applications that work with natural language. In particular, we will focus on a model called *lemon* (Lexicon Model for Ontologies), which has been proposed for the representation of lexica and machine-readable dictionaries as lexical linked data.

Languages Resources (LRs) such as lexica and terminological databases play a vital role in many NLP applications, however such resources are often published in applications-specific formats, resulting in linguistic data silos which are not directly reusable by other NLP applications.

In contrast, the use of ontologies to describe the semantics of the Web endeavors to bring machine processable meaning to the content of web pages and envisions the Web as a universal medium for data, information and knowledge exchange. It follows that this could allow for new paradigms for the life cycle of language resources as the data is often already open and can easily be converted into linked data. This makes the data available on the web in such a direct manner it is easier to create applications that can access these resources. Secondly, the use of web standard formats such as XML and RDF and the ecosystem of applications around them, such as query languages, structured storage and alignment techniques, should ease the effort in interact with diverse resources. Finally, the existence of LRs on the web allows for links to be created to and from existing resources.

lemon uses ontologies as the basis for its semantic modelling for two reasons, as ontologies in web standard formats, such as RDFS and OWL, form the majority of the existing linked data cloud. Furthmore, ontologies are playing an increasing role in state of the NLP systems either to model the internal domain knowledge of such systems or to drive the semantic annotation or ontology based information extraction. Multilinguality is also a concern, since multilingual ontology building is a very expensive and time-consuming undertaking, methods and tools to aid users in the localization of ontologies, becomes crucial.

Ontology lexicalization, as developed by the Monnet project, drives a new standard for sharing lexical information that can help to enhance existing language resources by linked data principles. Furthermore, lexicalized ontologies can improve the performance of applications as varied as ontology localization, machine translation, information extraction, multilingual access and presentation and natural language generation. As such, we hope to equip the audience with an understanding of how to enrich ontologies with lexical information by conforming to current practice in the fields of the Computational Linguistics and Language Resources and methods for sharing such data as Lexical Linked Data.

This tutorial is aimed at PhD students, post-docs, or industry people working with NLP applications and using ontologies for their applications. It would be advantageous for participants to have some experience with an ontology language, preferably OWL, and has some familiarity with Web standards such as XML and ideally RDF.

Outline of talks

- **Current models and trends *(Brian Davis)***
- **Modelling lexica on the web (*Jorge Gracia)***
- **Hands-on: Creating lexica with *lemon***
- **Syntax, structure and semantics for ontology-lexica *(John McCrae)***
- **Hands-on: Applications of ontology-lexica**

Presenters

Brian Davis (Digital Enterprise Research Institute (DERI), National University of Ireland Galway)
Brian Davis graduated in 2001 with an Honours B.Sc in Computational Linguistics and German from Dublin City University. After receiving an M.Sc. in Computer Science by Research from Trinity College Dublin in 2003, he worked as a Software Engineer at Sun Microsystems Ireland for 2 years. He joined DERI Galway in 2006 as a Language Engineer attached to the Lion/Nepomuk projects from 2006-2009. He is currently a PhD student in the Unit for Semantic Collaborative Systems (USCS) under the supervision of Dr. Siegfried Handschuh and Prof. Hamish Cunningham (Sheffield NLP group). His research interests are Controlled Natural Languages, semantic annotation and Natural Language Generation(NLG). He has a certified Text Analyst (GURU level) in GATE(General Architecture for Text Engineering) .

Jorge Gracia (Ontology Engineering Group (OEG), Universidad Politécnica de Madrid (UPM))
Jorge Gracia obtained his degree in Physics and his PhD in Computer Science at University of Zaragoza. Currently, he is a postdoctoral researcher in the Artificial Intelligence Department of Universidad Politécnica de Madrid, Spain. His research interests include ontology matching, semantic disambiguation, semantic measures, and multilingualism in the field of the Semantic Web.

John McCrae (Center of Excellence Cognitive Interaction Technology, University of Bielefeld)
John McCrae received an MSci in Mathematics and Computer Science from Imperial College London, UK in 2006 and a PhD from the National Institute of Informatics, Japan in 2009. In 2009 he joined the Semantic Computing Group at CITEC in the University of Bielefeld. Currently he is working on multlingual ontology representation and localisation in the Monnet project.