Natural Language Processing for the Semantic Web

Diana Maynard[1], Johanna Völker[2], Wim Peters[1],

[1] University of Sheffield, 211 Portobello, Sheffield, S1 4DP, United Kingdom
email: [diana,wim]@dcs.shef.ac.uk
phone: +44 114 222 1800
fax: +44 114 222 1810
[2] University of Mannheim, 68159 Mannheim, Germany
email: voelker@informatik.uni-mannheim.de
phone: +49 621 1812661
fax: +49 621 1812682

1 Overview

The proposed tutorial takes a detailed view of several human language technical aspects and current approaches that use NLP for automating Semantic Web-specific knowledge acquisition tasks. After a short introduction to the foundations of ontologies and the Semantic Web, the tutorial will focus on NLP technologies and their important role within applications such as ontology-based information extraction, ontology learning and population, or semantic metadata generation. Each of these technologies will be exemplified with descriptions of some key open-source NLP tools, thus enabling the participants to put their newly learned skills into practice. The tutorial will conclude by demonstrating several Semantic Web applications and the underlying data sets.

The tutorial will last **half a day** with an expected audience of 30-50 people. There are no specific **technical requirements**, although an internet connection would be useful.

2 Motivation

The Semantic Web aims at complementing the current text-based web with machine interpretable semantics to facilitate automated processing and integration of the vast amount of available information. The provision of such semantics, materialised in the tasks of ontology building and population, is a basic requirement for establishing the Semantic Web. However, as the manual construction of ontologies is very tedious and time-consuming, many people are still reluctant to add formal semantics to their web-sites. Automatic approaches such as methods for ontology learning and population can help to overcome this knowledge acquisition bottleneck by extracting relevant information from the Web and transforming it into a machine-processable representation. Given the enormous amount of textual data that is available online, it seems natural that these methods rely largely on the use of natural language processing techniques. The field of NLP has matured over the last decade to a point at which robust and scalable

applications are possible in a variety of areas, and current Semantic Web projects are now poised to exploit this development.

However, while NLP specialists have begun to realise and exploit the potential of their technology for the provision of semantics, the growth of the Semantic Web requires that experts from different domains (e.g. biology, law, tourism) build their own domain-specific tools for quick acquisition of semantics for their particular domain. The real challenges therefore are to bring NLP closer to ontology engineers from any background, and to extend current NLP techniques to ontology-based applications. This involves providing toolkits of basic NLP modules and, most importantly, guidelines for how these toolkits can be used for ontology enrichment activities and adapted to different domains. This tutorial will address these needs by

- providing an introduction to ontologies and the Semantic Web for NLP specialists,
- describing methodologies for manual or automatic ontology construction, and
- illustrating the techniques with examples of applications such as semantic annotation, ontology-based information extraction and ontology learning.

3 Outline of the tutorial

The tutorial will be divided into 3 sections, as follows:

Introduction to ontologies and the Semantic Web. This section will cover the foundations of the tutorial, giving an introduction to the Semantic Web, and describing different types of ontologies and their role for natural language engineers.

NLP and ontology engineering. This section will explore further the concept of ontology engineering and describe ways in which NLP techniques can be used. It will cover the use of controlled natural language, the representation of linguistic information in ontologies, and some techniques for ontology learning from unstructured text. It will also touch briefly on some of the challenges for NLP, such as the use of more expressive ontologies, the handling of logical inconsistencies, and ontology evaluation techniques.

Applications. Finally we shall illustrate the techniques introduced with examples of real research applications. These will cover topics such as ontology-based information extraction, semantic annotation and ontology learning.

4 Benefits for the attendees

This tutorial will introduce the attendees to the Semantic Web and how it benefits from NLP technologies. It will cover state-of-the-art research as well as established methods and tools for important tasks such as ontology learning or semantic annotation. Since all of the NLP tools to be presented are open source, the tutorial will provide the attendees with skills which are easy to apply and do not require special software or licenses. The **target audience** will consist primarily of researchers in the area of NLP looking to extend their work to Semantic Web applications, and to learn about the role of ontologies and how they may be used, e.g., for connecting ontologies to textual data. No previous knowledge of the Semantic Web or of ontologies is necessary.

5 Tutorial speakers

Speaker 1 (Main Contact): Dr Diana Maynard (University of Sheffield) has been a Research Associate in the Natural Language Processing Group at the University of Sheffield for the last 10 years. She holds a PhD in Automatic Term Recognition from Manchester Metropolitan University (UK) and has over 15 years of experience working in the field. Her main interests are in Information Extraction, robust and adaptable tools for language engineering, terminology, evaluation and accessibility of technology. Over the past 10 years she has led the development of Sheffield's open-source multilingual IE tools, and has led research teams on a number of projects including the EU NoE KnowledgeWeb and the EU projects NeOn and Musing. She has published over 50 scientific papers in conferences, journals and books, has reviewed numerous conference and journal papers, organised several workshops and will chair the Semantic Web Challenge at ISWC'10, one of the most prestigious annual Semantic Web events. She has given lectures on Text Mining, NLP and the Semantic Web, including a keynote speech on NLP and terminology at TIA'09[3] as well as tutorials at international NLP and Semantic Web conferences.
Website: http://www.dcs.shef.ac.uk/~diana

Speaker 2: Dr Johanna Völker (University of Mannheim) is a postdoctoral researcher at the University of Mannheim, where she is currently lecturing on Semantic Technologies. She has extensive experience in ontology-based knowledge representation and published numerous papers at international conferences and workshops. Her areas of expertise include ontology learning, formal concept analysis and description logics. She holds a diploma in Computer Science with special focus on Computational Linguistics from the University of Saarbrücken and a PhD in Applied Informatics from the University of Karlsruhe. Her PhD thesis focuses on methods for the semi-automatic refinement and evaluation of expressive ontologies. Before joining the Artificial Intelligence group in Mannheim, she had been affiliated with the Institute AIFB at Karlsruhe University for more than 5 years, working in the EU projects SEKT and NeOn. Her teaching experience includes tutorials at ESWC'05, EKAW'06 and ISWC'08.
Website: http://ki.informatik.uni-mannheim.de/people/johanna_voelker.html

Speaker 3: Dr Wim Peters (University of Sheffield) is a postdoctoral research fellow at the University of Sheffield. He has been active in the field of computational linguistics for 19 years, and has participated in many computational linguistic projects covering lexicon and thesaurus construction, corpus building and annotation, and (ontology-based) information extraction. As a linguist, he has worked with resources in various formats for linguistic and terminological representation, such as relational, XML and RDF/OWL. Amongst others, he has been involved with the EuroWordNet (multilingual thesaurus building) and NeOn projects. His main interests within the ontological sphere are knowledge acquisition and modelling issues in various application domains, and the interface between (multilingual) lexicons and ontologies. He is an active member of the ISO committee on the standardization of terminological and lexical resources (TC37/SC4), and has presented a number of seminars and tutorials.
Website: http://www.dcs.shef.ac.uk/~wim

[3] http://www.irit.fr/TIA09/ProgrammeEN.html#keynote2