Language Resources and the Semantic Web

Paul Buitelaar[♦], Thierry Declerck[♦], Nicoletta Calzolari[◊], Alessandro Lenci[*]

[♦]DFKI Language Technology, Stuhlsatzenhausweg 3,
D-66123 Saarbrücken, Germany
{paulb,declerck}@dfki.de

[◊]Istituto di Linguistica Computazionale (ILC) - CNR
Area della Ricerca CNR, Via Alfieri 1 (San Cataldo)
I-56010 PISA, Italy
glottolo@ilc.cnr.it

[*]Dipartimento di Linguistica, Università degli Studi di Pisa
Pisa, Italy
alessandro.lenci@ilc.cnr.it

1 Introduction

In recent years, the Internet evolved from a global medium for information exchange (directed mainly towards human users) into a "global, virtual work environment" (for both human users and machines). Building on the world-wide-web, developments such as *grid technology*, *web services* and the *semantic web* contributed to this transformation, the implications of which are now slowly but clearly being integrated into all areas of the new digital society (e-business, e-government, e-science, etc.) In particular, grid technology allows for distributed computing, web services for a distributed workflow, and the semantic web for increasingly intelligent and therefore autonomous processing.

In this, it is important to realize that the semantic web will function more and more as the man-machine interface of this "global, virtual work environment". The underlying semantic web infrastructure of shared knowledge (ontologies) and markup of resources and services with such knowledge (ontology-based metadata) ensures that a common understanding will exist between the human user and the machine-based processes. However, as much of human knowledge is and will be encoded in language, multilingual and multicultural aspects (culture as specific to countries, regions and nations, connected with language) will play an important role in establishing and maintaining such common understanding. Given these considerations, we emphasize the following two important issues in future semantic web development:

- **Making the semantic web accessible in many languages:** Authoring support for automatic knowledge markup should be available for many languages thereby avoiding that only documents in some languages will become part of the semantic web

- **Allowing the semantic web to represent many different cultures:** Ontologies should express concepts as used in different cultures, thereby avoiding that the semantic web would force an unnecessary semantic standardization. Therefore, tools for ontology adaptation and for mapping different ontologies should be an integral part of the semantic web infrastructure.

In both cases, there will be an important role for a combination of language technology, ontology engineering and machine learning, in order to provide text analysis for knowledge markup and text mining facilities for ontology mapping and learning. A growing integration of language technology tools into semantic web applications is therefore to be expected with the following characteristics:

- **Language Technology for the Semantic Web:** Language technology tools will be used for efficient, (semi-)automatic knowledge markup (based on information extraction) and ontology development (based on text mining), allowing web documents in many languages and from different cultural backgrounds to be integrated on a large scale within the semantic web.

- **The Semantic Web for Language Technology:** Semantic web methodologies (metadata, web services) and standards (RDF/S, OWL) will be used in the specification of web-based, standardized language resources – data (corpora, lexicons, grammars) and tools – allowing for a distributed and widespread use of these resources in semantic web applications.

2 Language Technology for the Semantic Web

As human language is a primary mode of knowledge transfer, a growing integration of language technology tools into semantic web applications is to be expected. Language technology tools will be essential in scaling up the semantic web by providing automatic knowledge markup support and facilities for ontology monitoring and adaptation.

Turning the web into a Semantic Web implies widespread annotation of documents with ontology based knowledge markup. Many of these documents consist of free text in different languages, which can only be marked up in an efficient way by use of automatic, language technology tools. Obviously, it will then be of political and cultural importance that such authoring support for automatic knowledge markup will be available for many languages, thereby avoiding that only documents in some languages will become part of the semantic web.

Ontologies, as used in knowledge markup, are views of the world that tend to evolve rapidly over time and between different applications. Currently, ontologies are often developed in a specific context with a specific goal in mind. However, it is ineffective and costly to build ontologies for each new purpose each time from scratch, which may cause a major barrier for their large-scale use in knowledge markup for the Semantic Web. Creating ambitious semantic web applications based on ontological knowledge implies the development of new, highly adaptive and distributed ways of handling and using knowledge that enable existing ontologies to be adaptable to new environments. Besides time and place this also, quite importantly, includes adapting to different cultures, thereby avoiding an unnecessary process of semantic standardization.

In all of this, there will be an important role for a combination of technologies (language technology, ontology engineering and machine learning) to provide linguistic analysis and text mining facilities for knowledge markup, ontology mapping (between cultures and applications) and ontology learning (for adaptation over time and between applications).

There are a number of ongoing projects working on these issues, e.g.: AKT[1] - Advanced Knowledge Technologies, ContentWeb, Dot.kom[2] - Designing infOrmation extracTion for KnOwledge Management, Esperonto[3], OntoBasis[4]

Tools for the integration of language technology within semantic development that are being developed (further) within these projects include: Amilcare[5], GATE[6], KAON[7], Melita[8], MnM[9], MuchMore[10], TERMINAE, WebODE[11]

---

1 http://www.aktors.org, http://www.dcs.shef.ac.uk/research/groups/nlp/akt/
2 http://nlp.shef.ac.uk/dot.kom/
3 http://www.esperonto.net/
4 http://cnts.uia.ac.be/cnts/projects/2002ontobasis.html
5 http://nlp.shef.ac.uk/amilcare/
6 http://gate.ac.uk/
7 http://kaon.semanticweb.org/
8 http://www.aktors.org/technologies/melita/
9 http://kmi.open.ac.uk/projects/akt/MnM/

3 Semantic Web Architecture for Language Technology

It is to be expected that semantic web methodologies (ontology-based metadata, web services) and standards (RDF, OWL) will be used in the specification of web-based, standardized language resources – data (corpora, lexicons, grammars) and tools – allowing for a distributed and widespread use of these resources in semantic web applications. Therefore, platforms will be needed for the discussion, implementation and dissemination of semantic web standards and protocols for the syntactic and semantic interoperability of language tools and resources across languages, cultures and applications.

This work should build on and reinforce previous and ongoing national, European and world-wide projects and initiatives in this area within language technology (e.g. ENABLER[12] - European National Activities for Basic Language Resources, ICWLR - International Committee for Written Language Resources, IMDI - ISLE[13] Metadata Initiative, INTERA[14] - Integrated European Language Data Repository Area, ISLE[15] - MILE: Multilingual ISLE Lexical Entry, ISO/TC37/SC4[16], LT-World[17], OLAC[18] - Open Language Archives Community, OLIF[19]), while taking into account emerging (semantic) web standards as specified within W3C or industry (e.g. RDF[20], RDF(S)[21], OWL[22], TopicMaps[23], Web Services Choreography Group[24], DAML-S[25], jxta[26] platform for P2P technology).

---

10 http://muchmore.dfki.de/demos.htm
11 http://delicias.dia.fi.upm.es/webODE/
12 http://www.enabler-network.org
13 http://www.mpi.nl/IMDI
14 http://www.ilsp.gr/intera_eng.html
15 http://www.mpi.nl/ISLE
16 http://tc37sc4.org
17 http://www.lt-world.org
18 http://www.language-archives.org/
19 http://www.olif.net
20 http://www.w3.org/RDF/
21 http://www.w3.org/TR/rdf-schema/
22 http://www.w3.org/TR/owl-features/
23 http://www.topicmaps.org/
24 http://www.w3.org/2002/ws/chor
25 http://www.daml.org/services/
26 http://www.jxta.org

4 Language Infrastructure: Some Issues

The integration of heterogeneous and distributed language resources into a unified, semantic web based language infrastructure touches on a number of fundamental issues in the nature, construction and use of such resources. Here we are highlighting:

- **Language and Knowledge:** Integration and progressive equation of language resources (e.g. lexicons and grammars for a domain-specific sublanguage) and knowledge resources (e.g. ontologies describing a specific domain) will lead to a unified framework for the representation and use of concepts and their linguistic realizations. Such a unified framework will be enabled by the availability of an open semantic web architecture based on common representation formats (RDF(S), OWL, Topic Maps) that will be used to describe any possible knowledge resource, which may therefore become more easily merged.

- **Language, Knowledge and Culture:** An important aspect of the emerging unification of domain knowledge and linguistic knowledge will be the possibility of designing ontologies in a more effective way. As mentioned before, ontologies change over time, between applications and between cultures. Connecting concepts as represented in ontologies with their multilingual realizations allows for an automatic mapping of ontologies on the basis of the use of concepts in language (i.e. the linguistic context of the words corresponding to these concepts).

- **Language, Knowledge and Multimedia:** There is a need for an integration of different types of content analysis as provided by image and video processing, mature multilingual language technology (including speech) and semantic web methods to enhance access to distributed multimedia content. One of the main goals in this will be to investigate and discuss how the so-called 'semantic gap' in the analysis of video data can be reduced with the help of semantic annotations delivered by advanced language technology tools applied on textual documents related to the video or image data.

- **Static and Dynamic Language and Knowledge Resources:** In order to make the use of language and knowledge resources more effective, a continuing cycle needs to be established between static (existing lexicons, grammars, ontologies, etc.) and dynamic resources (semi/automatically acquired or adapted lexicons, grammars, ontologies, etc.). In fact, such a

cycle implies a growing integration of language and knowledge (and multimedia) resources as described above, because adapting a knowledge resource (i.e. an ontology) will involve the analysis of linguistic contexts of concepts in text (or multimedia) mining. In turn, language resources will need to be adapted to specific domains by mapping linguistic objects (e.g. words, phrases, terms) onto concepts in a given ontology for a domain. Importantly, such automatically acquired or adapted resources are to be validated, e.g. by defining common standards and protocols for assigning "quality labels". The definition of such protocols should be well connected with a strong user base that will provide evaluation of language and knowledge resources through their use in different application scenarios.

5 Conclusions

Effective acquisition, organization, processing, sharing, and use of the knowledge embedded in multimedia content as well as in information- and knowledge-based work processes plays a major role for competitiveness in the modern information society and for the emerging knowledge economy. However, this wealth of knowledge implicitly conveyed in the vast amount of available digital content is nowadays only accessible provided that considerable manual effort has been invested into its interpretation and semantic annotation, which is possible only for a small fraction of the available content. Therefore the major part of the implicit semantic knowledge is not taken into account by state-of-the-art information access technologies like search engines, which restrict their indexing activities to superficial levels, mostly the keyword level.

Multilinguality and multicultural expression are important aspects of human society. Texts and documents are - and will be - written in various native languages, but these documents are relevant even to non-native speakers. We could imagine bypassing the multilingual problem by focusing directly onto knowledge itself, rather than on language, but in fact, human knowledge is and will be encoded in language, and multilingual and multicultural aspects (culture as specific to countries, regions and nations, connected with language) will play an important role in establishing and maintaining such common understanding. The Semantic Web must represent and structure concepts in multilingual and multicultural ontologies, which can be obtained only by linking conceptual nodes with the various language specific lexical realizations.