Towards A Language Infrastructure for the Semantic Web

Thierry Declerck[*], **Paul Buitelaar**[†],
Nicoletta Calzolari[‡], **Alessandro Lenci**[§]

[*]Department of Computational Linguistics, Saarland University
Postfach 15 11 50, D-66041 Saarbruecken, Germany
declerck@dfki.de
[†]DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrcken, Germany
paulb@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
§University of Pisa - Department of Linguistics
via Santa Maria 36 - 56100 Pisa
alessandro.lenci@ilc.cnr.it

Abstract

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 this conctext the semantic web allows for increasingly intelligent and therefore autonomous processing. This development brings new challenges for Human Language Technology (HLT), which require not only some adaptation of processes within the state of the art processing chain of HLT, but also changes at the infrastructure level of HLT resources.

1. Introduction

In recent years, the Internet evolved from a global medium for information exchange, that is directed mainly towards human users, into a global and virtual work environment, in which both human users and machines are involved. 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 machine processing.

In this context, it is important to realize that the upcoming semantic web (SW) will function more and more as the man-machine interface of this new global and 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. Human Language Technology (HLT) is thus confronted with new challenges, if it should both contribute to the success of the semantic web and benefit from the advances in the organization of knowledge, that will also concern linguistic knowledge, to be achieved within the SW. 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 not only in different languages, but in also in different cultures, thereby avoiding that the semantic web would force an unnecessary formalized semantic normalization in the human-machine interaction, leading to a reduction of linguistic and cultural diversity on the Web. 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.

But not only HLT tools should be adapted to the semantic web technologies: also the natural language resources should be organized in an infrastructure that allow them to be shared by semantic web applications.

2. Goal of the paper

The paper will presents some actual projects and initiatives geared towards a better synchronization and integrated development of HLT and SW technologies. Section 3 will sketch the role HLT can play in the development of the Semantic Web. Section 4 will address the possible impact of the general SW architecture on language technology, and briefly describe the kind of actions the HLT community should take in order to respond effectively to this new challenge.

3. 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 (e.g. Amilcare[1], GATE[2], OntoMat[3], Melita[4], MnM[5]) and facilities for ontology monitoring and adaptation (e.g. TextToOnto[6], OntoLearn[7], OntoLT[8].

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. This point has been partly stressed by major actors in the Semantic Web community, so for example Richard Benjamins et all. mention in their white paper Multilingualism as one of the 6 challenges of the Semantic Web (see (Benjamins et al., 2002)).

This aspect is also central in a FP5 EU project, Esperonto [9], that aims at bridging the actual Web towards the Semantic Web with the help of Semantic Annotation services that process actual web documents in order to provide them with semantic indices that can be interpreted by more "intelligent" search engines. Multilingualism in the project plays a role in connection with ontologies, NLP and language resources and should be realised at distinct levels:

- Content annotation for achieving Natural Language Understanding (see for example (Declerck, 2002)),

- Multilingual lexical semantics for supporting the mapping of ontologies in various natural languages,

- Ontology based natural language generation in the Semantic Web user interfaces (see for example (Wilcock, 2003)).

HLT will play an important role in knowledge markup, but can presumably also be used for supporting the automated extraction of ontologies, at least at a flat level, from free texts. Actual experiments on this are also done within the Esperonto project, also in collaboration with projects at DFKI (see (Buitelaar et al., 2004)).

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. (Asunción Gómez-Pérez and Corcho, 2003), dedicated on ontology engineering, presents a very good overview of what is needed in order to deal with the so-called life-cycle of ontologies. The authorsof this book have also identified elsewhere, besides time and place, the multilingual variation. We note here, that besides those aspects, quite importantly, includes adaptinontologies should also adapt to different cultures, thereby avoiding an unnecessary process of semantic normalization on the web.

4. Semantic Web Architecture for Language Technology

It is to be expected that semantic web methodologies (ontology-based 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. 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. At the

[1] http://nlp.shef.ac.uk/amilcare/

[2] http://gate.ac.uk/

[3] http://annotation.semanticweb.org/tools/ontomat

[4] http://www.aktors.org/technologies/melita/

[5] http://kmi.open.ac.uk/projects/akt/MnM/

[6] http://kaon.semanticweb.org/Members/rvo/Module.2002-08-22.4934

[7] http://www.dsi.uniroma1.it/ velardi/IEEE_C.pdf./

[8] http://www.dfki.de/ paulb/iswc03-demo.pdf. See also (Buitelaar et al., 2004)

[9] www.esperonto.net. See also (Benjamins et al., 2003).

lexical semantic level, important central contribution has been offered for example with EuroWordNet, mapping lexical senses to ontologies (see http://www.hum.uva.nl/ ewn or (Vossen, 1998)). In the domain of multilingual lexicon encoding, the ISLE-MILE[10] (Multilingual ISLE Lexical Entry) project has been very early taking into account ontologies. And beyond this aspect, there exist also in the meantime proposals for an RDF/S encoding of the lexical resources defined in the MILE framework, supporting thus accessibility of those in the context of the Semantic Web, as can be seen in (Ide et al., 2003).

At the infrastructural level, initiatives and projects have already started that can be immediately relevant for the Semantic Web architecture: e.g. ENABLER[11] (European National Activities for Basic Language Resources), which aims at improving cooperation among national activities established by national authorities for providing LRs for their languages, or ICWLR (International Committee for Written Language Resources), which has been estbalished very recently. At the metadata level, the IMDI[12] (ISLE Metadata Initiative) has been proposing a detailed structured list of Metadata descriptors supporting Web access to multimmodal and multimedia data. This Metadata set is complementary to OLAC[13] (Open Language Archives Community), which was at its beginning more closely related to the generic Dublin Core type of Metadata for a more global access to documents on the Web (see http://dublincore.org/). More details on IMDI , the differences and complementarities of IMDI and OLAC, as well as the relation to lexical resources are given in (Broeder et al., 2004).

The IMDI Set is being currently used, tested and extended in the context of a running European project, INTERA[14] (Integrated European Language Data Repository Area), that is defining a protocol for ensuring the Web interoperability of language resources and tools acting on those. The INTERA project can be seen as well as a first step to a web service for language resources and tools.

Some of the initiatives mentioned here are directly related to standardization efforts in the context ISO/TC37/SC4[15], where relation and links between linguistic resources and processes and knowledge sources and representation play a central role. This close connection to standard initiative is central if one wants to see the kind of infrastructure developed in INTERA being taken up in commercial scenario based on web services dealing with language resources. In this context the HLT community has to definitely take into account, as already shown in (Ide et al., 2003), emerging (semantic) web standards as specified within W3C or industry, e.g. RDF/S[16], OWL[17], TopicMaps[18], Web Services Choreography Group[19], DAML-S[20], JXTA[21] platform fro P2P technology.

To finalize this overview of relevant projects and initiatives, we wouls like to mention the information service provided by Language Technology World (LT-World). LT-world is a WWW-based virtual information center on the wide spectrum of technologies for dealing with human languages. It is a free service provided to the R&D community, potential users of language technologies, students and other interested parties by the German Research Center for Artificial Intelligence (DFKI). What make this information service particularily relevant in the scope of this paper, is the fact, that the whole knowledge about HLT has been encoded using Semantic Web tools and representation languages[22].

5. Conclusions

Effective acquisition, organization, processing, sharing, and use of the knowledge embedded in (textual and multimedia) web 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.

Given these considerations, we have been presenting on going initiatives and projects in the HLT domain that take the (multilingual and multicultural) Semantic Web as a challenge for the own field of ressearch and development and that address the changes to be achieved within the own community, not only at the level of resources and tools but also at the level of the language infrastructure in

[10] http://www.ilc.cnr.it/EAGLES96/isle/complex/clwg_home_page.htm

[11] http://www.enabler-network.org

[12] http://www.mpi.nl/IMDI

[13] http://www.language-archives.org/

[14] http://www.elda.fr/rubrique22.html

[15] http://www.tc37sc4.org/

[16] http://www.w3.org/RDF/

[17] http://www.w3.org/TR/owl-features/

[18] http://www.topicmaps.org/

[19] http://www.w3.org/2002/ws/chor

[20] http://www.daml.org/services/

[21] http://www.jxta.org

[22] See http://www.lt-world.org/ie_index.html

general. Those are encouraging steps towards a global research and development effort on establishing a distributed, standardized and semantically interoperable infrastructure of language resources and tools, which would enable a widespread integration of multilingual analysis tools into semantic web services and applications.

6. Acknowledgements

This research has in part been supported by EC grants IST-2000-29243 for the OntoWeb project, IST-2001-34373 for the ESPERONTO project and EDC-22076/27924 for the INTERA project.

7. References

Asunción Gómez-Pérez, Mariano Fernández-López and Oscar Corcho, 2003. *ONTOLOGICAL ENGINEERING*. Springer Verlag.

Benjamins, Richard, Jesús Contreras, Oscar Corcho, and Asunción Gómez Pérez, 2002. Six challenges for the semantic web. Http://www.isoco.com/isococom/whitepapers/files/ SemanticWeb-whitepaper-137.pdf.

Benjamins, Richard, Jesús Contreras, Thierry Declerck, Hans Uszkoreit, Ying Ding, Dieter Fensel, Asun Gomez Perez, Oscar Corcho, Michael Wooldridge, and Valentina Tamma, 2003. Esperonto: Application service provision of semantic annotation, aggregation, indexing and routing of textual, multimedia, and multilingual web content. In *Proceedings of WIAMIS 2003*.

Bontcheva, Kalina and Hamish Cunningham, 2003. The semantic web: A new opportunity and challenge for hlt. In *Workshop on Human Language Technology for the Semantic Web and Web Services at ISCW 2003*.

Broeder, Daan, Thierry Declerck, Laurent Romary, Markus Uneson, Sven Strmqvist, and Peter Wittenburg, 2004. A large metadata domain of language resources. In *Proceedings of LREC 2004*.

Buitelaar, Paul, Daniel Olejnik, Michaela Hutanu, Alexander Schutz, Thierry Declerck, and Michael Sintek, 2004. Towards ontology engineering based on linguistic analysis. In *Proceedings of LREC 2004*.

Capstick, J., T. Declerck, G. Erbach, A. Jameson, B. Joerg, R. Karger, H. Uszkoreit, W. Wahlster, and T. Wegst, 2002. Collate: Competence center in speech and language technology. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Declerck, Thierry, 2002. SCHUG: A platfom for integrating NLP and other sources of information for real worl d applications. In *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'02), Workshop Semantic Authoring, Annotation and Knowledge Markup (SAAKM)*. Lyon, France.

Ide, Nancy, Alesandro Lenci, and Nicoletta Calzolari, 2003. Rdf instantiation of isle/mile lexical entries. In *Proceedings of ACL'03 Workshop on Linguistic Annotation: Getting the Model Right*.

Uszkoreit, Hans and Brigitte Joerg, 2003. A virtual information center for language technology: Ontology, datastructure, realization. In *Nordic Language Technology Yearbook*.

Vossen, Piek (ed.), 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic.

Wilcock, Graham, 2003. Talking owls: Towards an ontology verbalizer. In *Workshop on Human Language Technology for the Semantic Web and Web Services at ISCW 2003*.