Basic Techniques for the Extraction and Annotation of Machine Understandable Information

Manuela Kunze, Dietmar Rösner
Otto-von-Guericke-Universiät Magdeburg
Institut für Wissens- und Sprachverarbeitung
P.O.box 4120,
39106 Magdeburg, Germany
{makunze, roesner}@iws.cs.uni-magdeburg.de

June 28, 2002

Project Description

The core of the semantic web are machine understandable information about resources (mostly these are documents). Currently the documents available in the www are mostly without any explicit information about their content. In our project we want to support the automatic extraction of information and annotation of documents with machine understandable information. Our document suite XDOC provides a collection of tools for the flexible and robust processing of documents in German. The analyses of documents are based on linguistic methods and techniques of knowledge extraction. Presently we deal with the document content for the following application fields:

- automatic annotation of documents with metadata e.g. in a kind of RDF(S) and
- extraction of company profiles from webpages[1].

The document suite contains tools for preprocessing, structure and POS tagging, syntactical parsing, and semantic analysis[2]. The latter includes methods like the semantic tagging of unique words (tokens), a case frame analysis and the mapping from syntactic structures into semantic relations. As resources several lexica for linguistic and semantic analysis are used. Most of these lexica are manually recorded but we also implemented techniques for automatic acquisition of entries for the semantical lexicon resp. for possible relations inside an ontology[3]. The XDOC collection of tools is based on the use of XML as unifying formalism for encoding input and output data, resources (e.g. lexica) as well as process information. Through XSL transformations different views of the results are generated, in particular for the annotation with metadata. We can present these information as a Topic Map or in RDF(S) or in another KR language. Through the use of XML as internal format we are independent of the finally required target markup language.

References

[1] S. Krötzsch and D. Rösner, "Towards extraction of company profiles from webpages," in *2nd International Workshop on Databases, Documents, and Information Fusion*, (Karlsruhe, Germany), July 2002. to appear.

[2] D. Rösner and M. Kunze, "Die Document Suite - XML-basierte Sprachverarbeitung als Basistechnologien für das Semantic Web," in *XML Technologien für das Semantic Web - XSW 2002* (R. Tolksdorf and R. Eckstein(Hrsg.), eds.), no. P-14 in Lecture Notes in Informatics (LNI) - Proceedings, (Berlin, Germany), pp. 119–133, GI-Edition, Köllen Druck + Verlag GmbH, Bonn, June 2002.

[3] D. Rösner and M. Kunze, "Exploiting sublanguage and domain characteristics in a bootstrapping approach to lexicon and ontology creation," in *Proceedings of the OntoLex 2002 - Ontologies and Lexical Knowledge Bases at the LREC 2002*, (Las Palmas, Canary Islands), May 2002.