An XML-UNL Model for Knowledge-Based Annotation

Jesús Cardeñosa, Carolina Gallardo and Luis Iraola

Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, 28660 Madrid, Spain
{carde,carolina,luis}@opera.dia.fi.upm.es

Abstract. Efficient document search and description has radically changed with the widespread availability of electronic documents through Internet. Nowadays, efficient information search systems require to go beyond HTML-annotated documents. Complex information extraction tasks require to enrich text with semantic annotations that allow deeper and more detailed content analysis. For that purpose, new labels or annotations need to be defined. In this paper we propose to use UNL, an interlingua defined by the United Nations University, as a language neutral standard content representation in Internet. The use of UNL would open documents to a new dimension of semantic analysis, thus overcoming the limitations of current text-based analysis techniques.

1 Introduction

XML [1] is an standardized annotation language currently employed for a variety of purposes. For any given domain, the set of tags defined in its DTD attempts to capture the logical content structure of typical documents of the domain. So annotated, documents can be exploited by sophisticated document management systems that provide precise answers to users' queries. One the most promising uses of XML is the possibility of replacing textual document bases by their XML counterparts for document management purposes as well as for content management.

The capability of the XML standard to define the different information items present in a given document facilitates subsequent information extraction operations. This capability makes XML an ideal choice for annotating text corpora.

Annotated corpora have been one of the most useful resources in the last years for the study of linguistic phenomena. This orientation towards linguistic analysis has frequently associated corpus annotation with tasks such as part of speech tagging, chunking and parsing.. The Brown Corpus [2] or the British National Corpus [3] are examples of such annotated corpora. This sort of annotation is useful for many purposes but may be insufficient for information management tasks and for the location of very specific information items.

Corpus annotation poses significant difficulties when the goal is the representation and classification of information expressed in text form. While one could say that lexical and syntactic annotation of textual corpora is a more or less solved problem, semantic tagging is still a challenging goal currently aimed by several research lines.

© J. Cardeñosa, A. Gelbukh, E. Tovar (Eds.)
Universal Network Language: Advances in Theory and Applications.
Research on Computing Science 12, 2005, pp. 300–308.