UNL as a Text Content Representation Language for Information Extraction*

Jesús Cardeñosa, Carolina Gallardo, and Luis Iraola

Dpto. Inteligencia Artificial Universidad Politécnica de Madrid
{carde, carolina, luis}@opera.dia.fi.upm.es

Abstract. This paper describes a new approach for describing contents through the use of interlinguas in order to facilitate the extraction of specific pieces of information. The authors highlight the different dimensions of a document and how these dimensions define the capacities of their respective contents to be found in the scalable process of finding information. A specific interlingua, UNL, will be described. This approach is illustrated both with rich examples of the followed model and with actual applications, that includes the description of some running projects based on the interlingual representation of contents.

Keywords: Textual contents representation, Interlinguas, UNL.

1 Textual Contents Representation

Information extraction and retrieval greatly benefits from intelligent text content representations. Most of the techniques employed for representing the knowledge contained in documents rely on some kind of linguistic analysis: the bottom line is that a content representation can be more easily achieved if we start from a representation of the linguistic meaning.

The overabundance of information accessible electronically and the standards currently employed for publishing it in the web force us to consider the semantic content as one of the many **dimensions** a document has, so a more holistic view of documents can be thought of if we consider as dimensions the distinctive sets of features that disjointly characterize a document. A plain text representation (one with no format or mark-up) may be viewed as not having any of these features. Layout, formatting and hyper-linking constitute a first dimension of the document. This first dimension may provide cues about specific information pieces contained in a document and can facilitate searching and extraction tasks. Thus, **format,** as it is typically encoded using HTML, can be considered as a first document dimension.

The recent emergence of the semantic web is the result of the application of new mark-up standards that are progressively enriching the document with what can be considered as a new, second dimension of a document: its **structure**. Specific information pieces can be extracted provided that they have been previously marked with

* This paper has been sponsored by the Spanish Research Council through the project PATRILEX-HUM2005-0726.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 507 – 518, 2006.

some meaningful tags. The most commonly used mark-up standard is XML. When employed in text processing applications, it allows any degree of analysis and consequently the information so tagged can be easily extracted by any XML-aware application.

Undoubtedly, XML has revolutionized the way we process textual contents, providing us with powerful and flexible mark-up languages for expressing document structure. However, finding specific information pieces requires a previous analysis of the document that contains it, and for that human intervention is still required. Question answering is in this respect quite different from information retrieval; a deeper analysis of the document is needed and this requires a new, extra dimension present in a document (the third one): its semantic **content**. This new dimension demands a powerful formalism for content representation and also suitable for deduction and inference, both required in question answering tasks. The general solution to this problem, currently unsolved, could take as its basis, in a credible and reliable way, the idea of content representation by means of an interlingua accompanied by a knowledge base that could support the tasks of finding and inferring information. This paper will describe an interlingua able to support these representational and deductive requisites.

2 Interlinguas as Textual Content Representation

The issue of representing (and extracting) the knowledge contained in texts written in a natural language dates back to pioneering works in knowledge representation in the AI field [1], [2], which will be referred to as "conceptual representations". A Conceptual Representation can be defined as a data-structure that represents the meaning of natural language expressions in an unequivocal way. Early conceptual representations can be characterized as very precise, domain dependent formalisms oriented towards inference but quite restricted in their expressivity. Thus Artificial Intelligence sought other ways for representing linguistically expressed knowledge while overcoming the narrowness of earlier conceptual representations, which resulted in the Interlinguas.

Interlinguas are mainly defined by the following characteristics:

- They deal with the representation of meaning, the most abstract and the deepest level of linguistic analysis. The interlingual approach attempts to find a meaning representation common to many (ideally to all) natural languages.
- An interlingua is another language and its vocabulary (usually concepts or semantic primitives), syntax (thematic and functional relations, formalism) and semantics (a subjacent ontology or knowledge base) need prior specification.

These facts support the idea of using an interlingua for representing the knowledge expressed in natural language, and thus becoming a candidate for the third semantic dimension of the representation of textual contents.

However, there are some obstacles in the design and further use of an interlingua, so that it has been proved almost unfeasible to find a suitable way to represent word meanings that is at the same time a) able to accommodate a wide variety of natural languages, b) easy to grasp and use, c) precise and unambiguous and d) expressive enough to capture the subtleties of word meanings expressed in natural languages.