Grey Literature and Computational Linguistics: From Paper To Net

Claudia Marzi, Gabriella Pardelli, and Manuela Sassi
Istituto di Linguistica Computazionale; Consiglio Nazionale delle Ricerche, CNR, Italy

Abstract
The advent and exponential development of the World Wide Web has led to an increasing availability of unstructured knowledge and distributed information sources, meeting general public requirements that are hardly addressed by other more traditional information channels. This trend has concurrently raised a considerable interest in the application of Computational Linguistics (CL) methodologies to document access and retrieval, as they offer the unprecedented opportunity to make the subjective, user-centred information demands of Net citizens meet the ever changing and heterogeneous information flow of the web.
Over the last five years, more and more Italian Universities have introduced CL courses into their Humanities curricula, making available on-line teaching materials, tutorials and language engineering software that appear to supply the lack of offer from traditional Italian publishing houses. In this paper, we consider in some detail the role played by this type of Grey Literature in bringing up a wider and increasingly more aware community of web users in Italy.

Keywords: Computational Linguistics, Grey Literature, Web-based information

1. Introduction

1.1 Computational Linguistics and Language understanding

Computational Linguistics (CL) and Natural Language Processing (NLP) have profoundly changed the way we look at human language as a subject of scientific inquiry, shifting emphasis from abstract knowledge to real usage (Manning & Schütze 1999).
Understanding language requires the ability to master a heterogeneous system of manifold skills, based on the processing of complex information structures in context (Bybee & Hopper 2001, Jackendoff 2002). In turn, these structures may vary, depending on the speaker's communicative intentions and purposes (Barsalou 1999). In this specific sense, knowledge of language can no longer be decoupled from "doing things with words" (e.g. reading, learning, recalling, guessing, judging etc.). The application of computer technologies to issues of language understanding and production epitomizes such a profound change of perspective in the most exemplary way. CL concerns itself with the empirical testing of language models and has contributed, over the last 15 years, to shedding considerable light on the interplay between linguistic abilities and general cognitive functions such as inference, classification and learning (Mitchell 1993, Guarino 1998). Awareness of these issues can have a tremendous impact on the general-public daily demands for real-time, goal-oriented and personalized information paving the way to a new generation of web users.
To date, the vast majority of information available on the web is conveyed through a flood of largely unstructured and uninterpreted text material, ranging from large digital archives to blogs, electronic newswires and dedicated web pages. Flexible, intelligent access to such a digital haystack is a logical precondition to its very existence: information needles are non existing if the haystack is not structured, disseminated and made available to respond, in real time, to the personalized needs, queries and ever changing goals of daily web users (Lawrence & Lee 1999).

1.2 Language, ontology and web contents

Gaining intelligent access to web-based contents presupposes, among other things i) a formal representation (ontology) of the knowledge areas of interest, ii) an algorithm for classifying a new text item according to known ontological categories, iii) an automated procedure to relate terms and complex relations among terms to knowledge categories, iv) ways for automatically indexing contents on the basis of the terms and relations they contain, v) ways for querying a document repository by terms, relations and concepts (Lancaster 2003, Buitelaar, Cimiano & Magnini 2005).
Clearly, non professional web users can profit from ontology-based and language-based technologies for document access and management. Although considerable progress has been made in this area, however, the search for relevant web-based information is still either a) channelled through the strictures of largely pre-defined, context-free, fully-interpreted document ontologies, or b) limited to simple text queries, containing key words and relatively unstructured word patterns connected through few logical operators. Both modes of access have serious limitations.