Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 43-49 recibido 30-01-08, aceptado 03-03-08

From knowledge acquisition to information retrieval*

De la adquisición del conocimiento a la recuperación de información

M. Fernández Gavilanes S. Carrera Carrera M. Vilares Ferro
Computer Science Department, University of Vigo
Campus As Lagoas s/n, 32004 Ourense, Spain
{mfgavilanes,sccarrera,vilares}@uvigo.es

Resumen: Introducimos una propuesta en recuperación de información basada en la consideración de recursos sintácticos y semánticos complejos y automáticamente generados a partir de la propia colección documental. Se describe una estrategia donde el lenguaje y el dominio de documentos son independientes del proceso.
Palabras clave: adquisición del conocimiento, análisis sintáctico,extracción de términos, recuperación de información, representación del conocimiento

Abstract: We introduce a proposal on information recovery based on the consideration of complex syntactic and semantic resources which are automatically generated from the documentary collection itself. The paper describes a strategy where the language and the domain of documents are independent of the process.
Keywords: information retrieval, knowledge acquisition, knowledge representation, parsing, term extraction

1 Introduction

Efficiency in dealing with *information retrieval* (IR) tools is related to the consideration of relevant semantic data describing terms and concepts in the specific domain considered. This kind of resources are often taken from an external and generic module (Aussenac-Gilles and Mothe, 2004), which implies that we probably lose a number of interesting properties we would be able to recover if semantic processing was directly performed on the text collection we are dealing with.

In order to solve this and produce practical understandable results, we should allow easy integration of background knowledge from possible complex document representations, fully exploiting linguistic structures. So, we could compensate for missing domain-specific knowledge, which is a significant advantage for redeploying the system when no external resources are yet available. Also, access to a concept hierarchy so generated allows information to be structured into categories, fostering its search and reuse; as well as to integrate an interesting strategy to relate languages, using it as a semantic pipeline between them (Bourigault, Aussenac-Gilles, and Charlet, 2004; Aussenac-Gilles, Condamines, and Szulman, 2002).

In the state-of-the-art, methods to automatically derive a concept hierarchy from text can be grouped into *similarity-based* approaches and *set-theoretical* ones. The first type is characterized by the use of a distance in order to compute the pairwise similarity between vectors of two words in order to decide if they can be clustered (Faure and Nédellec, ; Grefenstette, 1994). Set-theoretical ones partially order the objects according to the existing inclusion relations between their attribute sets (Petersen, 2001). Both approaches adopt a vector-space model and represent a term as a vector of attributes derived from a corpus. Typically some syntactic features are used to identify which attributes are used for this purpose.

Our proposal aims to facilitate the knowledge acquisition task through a hybrid approach that combines *natural language processing* (NLP) strategies, such as shallow parsing and semantic markers, with statistical techniques and term extraction. A modular architecture allows for the addition of textual fonts on different topics and languages, providing the basis for dealing with multilingual IR. A collection of parallel texts on the

* Work partially supported by the Spanish Government from research projects TIN2004-07246-C03-01 and HUM2007-66607-C04-02, and by the Autonomous Goverment of Galicia from projects PGIDIT05PXIC30501PN, 07SIN005206PR and the Galician Network for NLP and IR.

