

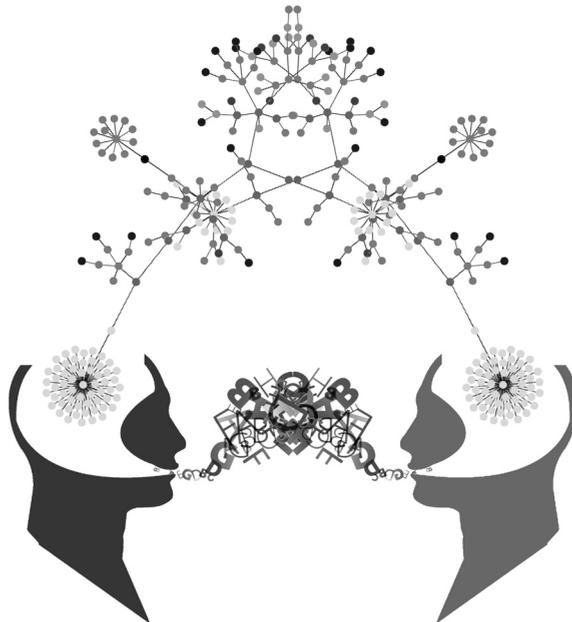


UNIVERSITÀ DEGLI STUDI DI SALERNO

Dottorato in Scienze della Comunicazione

Exploring Formal Models of Linguistic Data Structuring

Enhanced Solutions for Knowledge Management Systems
Based on NLP Applications



Federica Marano

Supervisor
Prof. Annibale Elia

Coordinator
Prof. Alessandro Laudanna

X Ciclo – Nuova Serie
2008-2011

ABSTRACT

The principal aim of this research is describing to which extent formal models for linguistic data structuring are crucial in Natural Language Processing (NLP) applications. In this sense, we will pay particular attention to those Knowledge Management Systems (KMS) which are designed for the Internet, and also to the enhanced solutions they may require. In order to appropriately deal with this topics, we will describe how to achieve computational linguistics applications helpful to humans in establishing and maintaining an advantageous relationship with technologies, especially with those technologies which are based on or produce man-machine interactions in natural language.

We will explore the positive relationship which may exist between well-structured Linguistic Resources (LR) and KMS, in order to state that if the information architecture of a KMS is based on the formalization of linguistic data, then the system works better and is more consistent.

As for the topics we want to deal with, first of all it is indispensable to state that in order to structure efficient and effective Information Retrieval (IR) tools, understanding and formalizing natural language combinatory mechanisms seems to be the first operation to achieve, also because any piece of information produced by humans on the Internet is necessarily a linguistic act. Therefore, in this research work we will also discuss the NLP structuring of a linguistic formalization Hybrid Model, which we hope will prove to be a useful tool to support, improve and refine KMSs.

More specifically, in section 1 we will describe how to structure language resources implementable inside KMSs, to what extent they can improve the performance of these systems and how the problem of linguistic data structuring is dealt with by natural language formalization methods.

In section 2 we will proceed with a brief review of computational linguistics, paying particular attention to specific software packages such as Intex, Unitex, NooJ, and Cataloga, which are developed according to Lexicon-Grammar (LG) method, a linguistic theory established during the 60's by Maurice Gross.

In section 3 we will describe some specific works useful to monitor the state of the art in Linguistic Data Structuring Models, Enhanced Solutions for KMSs, and NLP Applications for KMSs.

In section 4 we will cope with problems related to natural language formalization methods, describing mainly Transformational-Generative Grammar (TGG) and LG, plus other methods based on statistical approaches and ontologies.

In section 5 we will propose a Hybrid Model usable in NLP applications in order to create effective enhanced solutions for KMSs. Specific features and elements of our hybrid model will be shown through some results on experimental research work. The case study we will present is a very complex NLP problem yet little explored in recent years, i.e. Multi Word Units (MWUs) treatment.

In section 6 we will close our research evaluating its results and presenting possible future work perspectives.

Keywords

Knowledge Management System, Natural Language Processing, Linguistic Formal Model, Hybrid Formal Model.