MNEMOSYNE

Δευ, 18/10/2010 - 17:23 — webmaster

MNEMOSYNE Document Collection Processing Environment

The **MNEMOSYNE** system constitutes a complete NLP system that incorporates advanced linguistic resources and computational tools aiming at the automatic extraction of structured information from unstructured electronic documents. It is mainly used for automatic processing of free text documents. It ensures:

1. processing of big volumes of information,
2. high precision in the recognition of named entities and events,
3. possibility of addition of new sources of information with low cost.

Main Characteristics

Input data: The system supports different formats of input documents (HTML, PDF, TXT) stored on various media (files, web pages, databases).

Language resources: The system uses a wide variety of language resources such as vocabularies of different languages, spelling and morphological dictionaries, domain-specific dictionaries, thesauri, etc.

Annotations: Incorporates semantic annotations of documents at different levels of abstraction. The various annotation levels are kept separately from the source document and this allows for greater flexibility as annotation levels can be linked together.

Analyzers and process flow: The text analyzers are mechanisms used to extract information from textual sources and they generate appropriate semantic annotations. The analyzers can be connected in pipelines using different process flows where the output of each analyzer is used as the input to the next. The process flows are executed in parallel and this achieves substantial improvement in the analyzer performance.

Semantic annotation rule language: The «Kanon» semantic annotation rule language is used for the definition of syntactic phenomena that are used to extract information from textual sources.

Filtering: The output of each analyzer can be filtered before it is used as input to another analyzer. Thus, the information is stored, is minimized and furthermore, the annotation rules used in the subsequent analyzers are simplified.

Fuzzy matching: The environment incorporates fuzzy matching between the extracted named entities (such as persons, organisations, companies, etc.) and the data stored in an existing corporate database system. Two kinds of fuzzy matching mechanisms are used: the lexicographical ones, which make use of spelling correction techniques (i.e. word distance), and the statistical ones, which count the degree of similarity based on the number and the weight of the trigrams (or qgrams) of each word compared with the words included in the database.

Dumpers: These are specialized analyzers that ensure the transfer of extracted information to specific destinations and formats (e.g. XML, Database tables, etc.).

Monitoring: A special mechanism that can monitor the whole process and can record the audit data for the process in specific destinations (e.g. files, database).

Extraction Audit: The semantic annotations can be viewed at any level of abstraction. This

mechanism is used for the debugging annotation process step by step.

Desktop application for validation, verification and correction of the extracted information. The application user can easily add, update and delete semantic annotations created by the automatic process in order to maximize the quality and accuracy of the extracted information.

Web application for presentation of the results in multiple ways such as coloured annotations on the text, tables for each category of the extracted information, filtering based on each of the attributes of the extracted information, aggregated results, comparisons with the information extracted by other methods, etc.

Source URL: http://www.neurolingo.gr/el/node/165