THE "MICRO SEMANTICS" FOR INTELLIGENT BROWSING

Picchi Eugenio[1], Sassolini Eva[1],
[1]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR), Pisa, Italy,
{picchi|eva.sassolini}@ilc.cnr.it

Keywords: Computational Linguistics, Semantic Analysis, Information Retrieval.

ABSTRACT

Study and development of methodologies to improve systems of "information retrieval". Our approach is based on the integration of techniques, originally created to disciplines such as philology, lexicography, literature, with linguistic and statistical tools for the extraction and analysis of information in the text. Also we experimented a special methodology, for the creation of specific semantic metadata for text materials. In this paper, we describe "SmartCity", a project in which we applied these strategies. The project aims at designing and developing multimedia content (audio-guide for the new generation of interactive media and off-line and on-line) for the use of custom-cultural tourist routes, both physical (in the context of museums and cities) and virtual.

INTRODUCTION

SmartCity[1] is a project "POR Creo" of Tuscany region, funded by the European Community (FESR). The purpose is to allow a quantum leap in terms of productivity, versatility and adaptability of digital content, both textual and multimedia, through the use of knowledge engineering techniques. Several research activities are preparatory to the prototyping of innovative solutions for:

- ✓ the creation and delivery of interactive audio guide for new systems, both web and mobile;
- ✓ implementation of a prototype of "authoring system" for the production of guides and tourist-cultural paths, even virtual.

Our contribution to the project [1] is oriented to the development of an intelligent browsing system to be made available first and foremost of the "authoring system" and then of the creation of multimedia tourist guides.

The intelligent text browsing system is based on "TextPower" technology [2], developed by our research group. We are dealing with enrich the text with semantic annotations (micro-semantics) that we able to identify. The enrichment process allows the identification of relationships between concepts (proper names, geographical locations, acronyms, works of art, monuments, institutions, etc..), even when are potentially unknown.

This enrichment process makes a system of "information retrieval", much more performant and allows to specialize in different areas of research. In our particular case, is used to create an authoring system. This should be able to provide the author / publisher of the guide / tourist-cultural route a complete overview of available information.

The current technologies for information retrieval, as well as those for the extraction of knowledge, use computational tools of large capacity, but their potential it is still not fully expressed. There is a gap between the knowledge contained and recovered, which in fact prevents a further evolution step in the algorithms, in term of results and performance. The problem is that the knowledge, intended as terminology, concepts and semantic relationships between entities is not explicitly expressed, then not exploited by procedures for processing information. Many studies, designs, prototypes are trying to remedy this problem.

This particular text enrichment technique can be defined "generalized" and is obtained through application of statistical and linguistic analysis technologies, without using assumptions or structures and ontologies predefined. All information extracted, are then associated at the text in

1 Project title: "SmartCity: new solutions for content engineering and ambient intelligence as support of cultural tourism"

a paratextual formalism, enriching the text of all lexical information, semantic and factual (e. g. named entities, terminologies composed by a single word and multiword). This wealth becomes a source of information exploitable by search engines and by classification and summarization systems.

RESULTS

The intelligent text browsing system, named "DBTFaccette", is a customization of the categorization system used in librarianship. Its uniqueness lies in the possibility of exploiting the "micro semantics" detected in the text. therefore it is important to have an annotated corpus.
The text corpus is originally a set of text files which can be added annotations of various kinds, that is, additional information related to words, phrases and piece of text. Both texts and annotations must be organized in a coherent and orderly so as to be easily stored, managed and expandable over time. The textual corpus becomes so a collection of digital sources, and metadata, which, besides containing the precise annotations must contain all information necessary for the management and description of textual sources.
A useful way to imagine the "authoring system" is to assume the final usage scenarios. It is possible to create a history of use of the complete system on the part of one or more specific users. Assume usage scenarios has a double meaning:
- Is useful in design of a "authoring system";
- Allows us to hypothesize an end user and is a very effective way to evaluate the product, that yet to be designed, in its possible contexts of use real. It often tends to adopt "reference systems" implicit, done of individual experiences. Since the reference systems of users are not necessarily identical to ours, it's easy to fall into misunderstandings harmful to the design of a complex product.
The browsing system is able to exploit the knowledge extracted from textual materials in several ways:

- ✓ to expand the search by exploiting the ability to search for words in multi-words and offer suggestions to research incomplete or vague. For example, by proposing the query "madonna" are also returned the occurrences of "Madonna del Pozzo", "Madonna del Latte" and "Santuario della Madonna del Pozzo" (Fig. 1), (Fig. 2);
- ✓ to improve the correlation process among documents identified, using specific linguistic resources;
- ✓ to improve the ranking of results in the case of classification of responses (Fig. 3);
- ✓ to better organize the display of responses. In fact, the system responds to queries with a variety of contexts where all relevant entities are highlighted or "micro-semantics". In this way, the next refinement of the research is facilitated.

In our approach, the creation of linguistic resources must be designed to the development of navigation and information retrieval in a position to exploit them, or tools that capture the information, organizing, classifying and distributing it according to the desired objectives. Rarely can the information be classified only with hierarchical criteria, in open systems, as in the case of Cultural Heritage. It is a more effective to use an approach based on principles of "semantic similarity" that allow you to link information crosswise, apparently belonging to different categories, but that lead back to the same informative need. One important role played within the project by ILC has been the development of a system capable of suggesting to the user /author, keys to refine your search and bring out more content that meet his need for information.
Initially all textual materials acquired, were indexed with DBT[2] procedures. By using a specific tool, developed by ILC, named PiTagger[3], we then applied a tagging phase, so to identify all

[2] DBT (Data Base Testuale, Textual Data Base) is specific module for the treatment and analysis of textual and lexical material.

lemmas and relative POS of each occurrence. The ambiguities producted of tagging phases are solved by following a statistical approach on the basis of a training corpus statistically analyzed and summarized. By exploiting pattern matching techniques we extracted the "micro-semantic" more complex. Typically, in the Italian syntactic construction “Npreposition-N” and “ADj-N/N-ADj” are the most productive linguistic patterns. Once candidate terms obtained, we applied statistical algorithms to analyze the frequency distribution of each pattern identified. On the basis of results we extracted a set of semantically relevant terms and concepts for specific domain (in this case “Empoli e dintorni”).
For the aims of the project, we had to characterize also some parts of Corpus. These parts not necessarily correspond to the original text units, but identified parts of the Corpus of particular interest to the user.
For example, it is important to know how to define a Points of Interest (POI). Typically, descriptions of POI can be traced to known categories such as: works of art (painting, sculpture), museum objects or archaeological finds and excavations and we can identify only, macro areas, on which we can ensure a correct attribution. In fact the concept of "Work of Art" is too broad and contains within it, objects of different nature.

CONCLUSIONS

At the end of the project we has created a corpus annotated for the specific domain, with all the relevant information (micro semantics) identified. Were also extracted information of the high level (POI) to classify, where possible, the textual materials. In this way the intelligent text browsing system that works on these data it is able to offer more information and better performance.

REFERENCES

Example:

[1] Spadoni F., Tariffi F., Sassolini E., SMARTCITY: Innovative Technologies for customized and dynamic multimedia content production for Tourism applications. In: EVA 2011 Florence Electronic Imaging and the Visual Arts. (Firenze, 4-5-6 may 2011). Proceedings, Cappellini Vito (ed.). Pitagora Editrice Bologna, 130 - 135.
[2] Picchi E., Sassolini E., "Text power": Tools for the cultural heritage. In: CHC 2010 - 4-th Intl. Congr. Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin (Il Cairo, 6-7-8 december 2009). Proceedings, vol. 1, Fondazione Roma Mediterraneo, (2010), 435 - 439.
[3] Sassolini E., Cinini A., Cultural Heritage: Knowledge Extraction from Web Documents. In: LREC 2010 - Seventh International Conference on Language Resources and Evaluation (Valletta, Malta, 17-23 May 2010). Proceedings, Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, Daniel Tapias (eds.). European Language Resources Association (ELRA), (2010), 3363 - 3368.
[4] Picchi E., Ceccotti M. L., Cucurullo S., Sassi M., Sassolini E., Linguistic Miner. An Italian Linguistic Knowledge System. In: LREC Fourth International Conference on Language Resources and Evaluation (Lisboa-Portugal, 26-27-28 May 2004). Proceedings, M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silvia (eds.), (2004), 1811 – 1814.

[3] PiTagger is an important component for text lemmatization and tagging and constitutes a software module of PiSystem:
integrated system for processing of textual and lexical materials.

FIGURES

Fig. 1 – results of search “madonna”

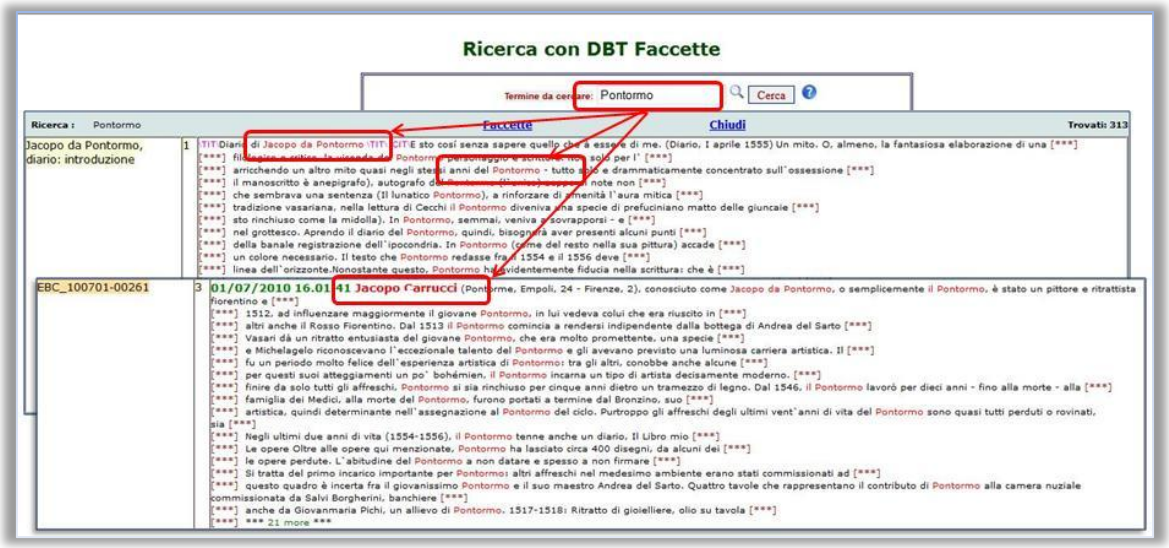

Fig. 2– results search “Pontormo”

Topiche
archeologia-Empoli
pitturaEmpoli
restauro
restauropittorico
vetro-Empoli

CD_terre_rinascimento\percorsi \archeologia\b00.htm	1	Archeologia antica Empoli e il territorio circostante nell`antichità La Storia "figlia" della Geografia La salubrità, la vicinanza di corsi d`acqua navigabili, la fertilità del suolo, la [***] [***] e culturale di stabili insediamenti umani sin dalla preistoria. Così è stato per gli attuali territori comunali di [***] [***] zona. Di fronte a tale situazione solo la ricerca archeologica sistematica poteva fare chiarezza e dare conferma a [***] [***] di volontariato archeologico (su tutti l`Associazione Archeologica di Volontariato del Medio Valdarno di Empoli e il [***] [***] . L`inizio del processo di Antropizzazione: la Preistoria e i primi insediamenti Protostorici Nella preistoria, agli albori della presenza dell`uomo, [***] [***] la materia prima per la lavorazione dei primi strumenti in pietra. I siti preistorici individuati sono oltre un centinaio [***] [***] umana accertata, finora, risale al Paleolitico Inferiore e più precisamente alla fase finale dell`Acheuleano ([***] [***] e Poggio Carbone (recentemente anche alcuni ritrovamenti in località Corniola presso Empoli), con ricca [***] [***] e alcuni bifacciali.E` con il Paleolitico Medio, corrispondente alla fase più antica della glaciazione di Wurm [***] [***] Montalbano, Poggio alla Malva). Il Mesolitico (circa 10.000 anni fa, postglaciale) [***] [***] attenta fase di studio per i recenti ritrovamenti sempre presso Montelupo. Il Neolitico è, per ora, scarsamente documentato. Con [***] [***]	100%
EBC_100703-00166	2	**04/07/2010 11.22.31** Comune di Empoli Soprintendenza Archeologica per la Toscana EMPOLI E IL TERRITORIO CIRCOSTANTE La presente relazione intende portare solo un modesto contributo allo studio della storia medievale del territorio di Empoli, senza pretendere [***] [***] delle ricerche infatti, nessun documento scritto di epoca romana parla di Empoli e dei suoi dintorni, [***] [***] delle notizie indirette, ovverosia tramite una ricerca archeologica seria, così che lo storico possa colmare le [***] [***] nel corso della relazione anche ricerche e scoperte archeologiche riferite a fasi storiche che a prima vista sembrerebbero estranee [***] [***] tra gli altri, anche da uno scavo in via di completamento nel centro storico di Empoli che il [***] [***] ,sta eseguendo con l`appoggio dell`Associazione Archeologica Volontariato Medio Valdarno e sotto la direzione scientifica della dott.ssa Anna Rastrelli della Soprintendenza Archeologica per la Toscana. EMPOLI E DINTORNI: [***] [***] fonti. E` impensabile progettare una qualsiasi indagine archeologica seria senza prima aver letto o, meglio, [***] [***] comunque ricordato che i risultati di una ricerca archeologica scientifica diventano anch`essi fonti storiche attendibili perché l` [***] [***] ,Frati) anche una ricerca di archeologia dell`architettura. Fonti importanti,pur essendo indirette, possono essere considerate: la Tabula Peutingeriana dove è riportata, fra le altre, una [***] [***] in corso) sulla storia e l`archeologia tardo romana e medievale locale, sintomo di [***] [***] arrivata,la pubblicazione periodica Milliarium, dell'Associazione Archeologica, che intende perseguire la strada del coordinamento fra documentazione storica classicae ricerca archeologica. RICERCHE PRELIMINARI: LE NOTIZIE E LE [***] [***] , dalla Britannia all`India, denominata Tabula Peutingeriana. Questo nome le deriva da Konrad Peutinger [***] [***]	100%
SourceHTML_16	3	**15/02/2011** EMPOLI ARCHEOLOGICA EMPOLI E IL TERRITORIO CIRCOSTANTE A cura di Leonardo Terreni La presente relazione intende portare solo un modesto contributo allo studio della storia medievale del territorio di Empoli, senza pretendere di rappresentarne un` [***] [***] delle ricerche infatti, nessun documento scritto di epoca romana parla di Empoli e dei suoi dintorni, [***] [***] delle notizie indirette, ovverosia tramite una ricerca archeologica seria, così che lo storico possa colmare le [***] [***] nel corso della relazione anche ricerche e scoperte archeologiche riferite a fasi storiche che a prima vista sembrerebbero estranee [***] [***] tra gli altri, anche da uno scavo in via di completamento nel centro storico di Empoli che il [***] [***] ,sta eseguendo con l`appoggio dell'Associazione Archeologica Volontariato Medio Valdarno e sotto la direzione scientifica della dott.ssa Anna Rastrelli della Soprintendenza Archeologica per la Toscana. EMPOLI E DINTORNI: [***] [***] fonti. E` impensabile progettare una qualsiasi indagine archeologica seria senza prima aver letto o, meglio, [***] [***] comunque ricordato che i risultati di una ricerca archeologica scientifica diventano anch`essi fonti storiche attendibili perché l` [***] [***] Frati) anche una ricerca di "archeologia dell`architettura". Fonti importanti,pur essendo indirette, possono essere considerate: la Tabula Peutingeriana dove è riportata, fra le altre, una [***] [***] in corso) sulla storia e l`archeologia tardo romana e medievale locale, sintomo di [***] [***] arrivata,la pubblicazione periodica Milliarium, dell'Associazione Archeologica, che intende perseguire la strada del coordinamento fra documentazione storica "classica" e ricerca archeologica. RICERCHE PRELIMINARI: LE NOTIZIE E LE [***] [***] , dalla Britannia all`India, denominata Tabula Peutingeriana. Questo nome le deriva da Konrad Peutinger [***] [***] l`unico "itinerarium pictum" di età romana giunto fino a noi. Non trattandosi di una [***] [***]	100%

Fig. 3– ranking of result