AI/NLP technologies applied to spacecraft mission design

Maria Teresa Pazienza [1], Marco Pennacchiotti [1],
Michele Vindigni[1], Fabio Massimo Zanzotto[2]

[1] Artificial Intelligence Research Group,
University of Roma Tor Vergata, Italy.
{pazienza,pennacchiotti@info.uniroma2.it}
[2] University of Milano Bicocca, Italy.
zanzotto@disco.unimib.it

Abstract. In this paper we propose the model of a prototypical NLP architecture of an information access system to support a team of experts in a scientific design task, in a shared and heterogeneous framework. Specifically, we believe AI/NLP can be helpful in several tasks, such as the extraction of implicit information needs enclosed in meeting minutes or other documents, analysis of explicit information needs expressed through Natural Language, processing and indexing of document collections, extraction of required information from documents, modeling of a common knowledge base, and, finally, identification of important concepts through the automatic extraction of terms. In particular, we envisioned this architecture in the specific and practical scenario of the Concurrent Design Facility (CDF) of the European Space Agency (ESA), in the framework of the SHUMI project (Support To HUman Machine Interaction) developed in collaboration with the ESA/ESTEC - ACT (Advanced Concept Team).

1 Introduction

An interesting field of application of information access technologies relates to scenarios in which several users work jointly to a common *project*, sharing their possibly different and specific knowledge, and providing their essential personal contribution to a common goal. Imagine, for instance, a *design process* in which a team of experts coming from different scientific disciplines, cooperates in a common task of designing and engineering a particular device, that requires their different competencies to be jointly used and intertwined. Moreover, they should be possibly supported during the process by a large repository of domain knowledge from which to extract information that can help in the design [1].

For instance, in designing a space missions (as it is the case of the SHUMI project [13]), the goal of the process is both to produce a spacecraft able to accomplish an envisioned mission and to plan the mission itself. The expert team, composed by engineers, physicians and other scientists, jointly works in the CDF. The planning

[1] This context is what specifically analyzed into SHUMI-ESA ESTEC funded study N.18149/04/NL/MV

activity needs a fast and effective interaction of involved disciplines and requires the access to several kinds of documentations, among which scientific papers, studies, internal reports, etc., produced by experts of related disciplines all over the World (*pre-existing knowledge*). Thus, during a design process a large quantity of knowledge is usually accessed in order to satisfy the team information need. Moreover, the design process produces itself a large amount of information, such as meeting minutes and deliverables (*on-going knowledge*). Tools for retrieving and coherently organizing documents are then necessary as complementary resources for a design environment (such as the ESA - CDF). We propose a model of an architecture whose aim is to provide the team of experts with such tools, in order to speed-up the design process and to improve the quality of the resulting project. The proposed system can be intended as a *virtual assistant* helping the team to use the *pre-existing* and *on-going knowledge* repositories.

In order to help the experts during the design process, the system should thus be able to interpret the information need of the team expressed implicitly in the on-going knowledge repositories or explicitly through direct queries by the experts. It should be then able to satisfy these needs extracting the required information from the pre-existing knowledge repositories. IR and NLP (such as syntactic parsing and information extraction) are the most promising technologies to carry out these activities. Moreover, the system could provide a way to model and express in a *design process ontology* the overall relevant knowledge shared by the experts . Such a formal ontological *conceptualization* has two main goals: to represent how the project contributed to the systematic representation of the knowledge about the specific domain of interest, and to support a useful indexing of the documentation produced and gathered during the design process. Finally, as an additional feature, the system could offer the possibility of understanding the common “jargon” and terminology used in the design process, fixing it in the *design process ontology*. Indeed, it is plausible that some new concepts arise during the design process and assume a status of shared concepts, expressed through their linguistic expressions, that is *terms*.

The technological scenario for the information access framework is a *virtual assistant* as depicted in Fig.1. In the overall architecture it is envisioned a *proactive* system, that “listens” at the dialogues going on among the project participants (through the minutes of the meetings, for example) and extracts information needs, later on used to query information access systems able to retrieve documents where they can be satisfied. Once selected as relevant by users, retrieved documents contribute to the definition of the *design process ontology*, that embodies the knowledge relevant for the design project.

The overall system could result in facilitating: the access to the project related documentation and external information, the definition of terminology and knowledge involved in the process (through the ontology of the mission), the creation of a central view of the knowledge stored in the project related documentation using the proposed terminology. Such a system could be realized with technologies ranging from Information Retrieval engines, to knowledge based systems using complex natural language models. Either generic linguistic (such as WordNet [10]) or specific domain semantic knowledge can be used to empower document clustering and to interpret ambiguous and unknown terms. In the framework of the SHUMI project, a modular architecture able to satisfy all users needs has been defined, while allowing to reach final results at different levels of automation. It is possible to set up several different architectures where more functionalities can be added, starting from a "core" system, composed by an Information Retrieval engine plus the Document Clusterer.

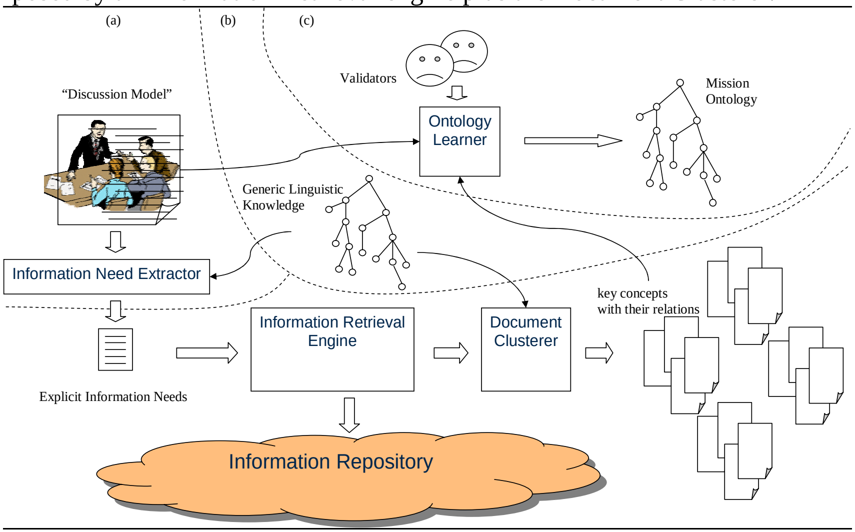

Fig.1 A complete solution for an Automatic Assistant

Additional capabilities define more complex systems (*(a)*,*(b)*,*(c)*):

- the system *(a)* behaves as an "active" information access system, "following" the conversation among experts and extracting *implicit* information needs (Sec. 2.2);
- the system *(b)* becomes more robust for lexical variations by using generic linguistic knowledge bases such as WordNet [10] (Sec. 2.3);
- the system *(c)* could acquire an explicit model of the knowledge embodied in processed documents as well as produced by the process. This explicit knowledge model is what has been called the design process ontology. It represents the memory the system has about the structure of the mission (Sec. 2.4).

All the linguistic processes carried out to implement the system are supported by an underlying modular syntactic parser (*Chaos*, [2]).

2. Architecture components

In the following sections we describe existing technologies that could be integrated to implement the different proposed architectures.

2.1 Information Retrieval and Clustering (IR&C)

The "core" of the proposed architecture, as described in the previous section, is based on both an IR engine and an automatic cluster components (IR&C).

Clustering results of a given query is often seen as a way to better publish documents retrieved by an information retrieval engine, driving users to the relevant documents by using indexing techniques.
Clustering algorithms as well as information retrieval methods are generally based on a vector space model, where documents are represented in the bag-of-word fashion. Nothing prohibits to use more relevant, i.e. more readable, features, such as the one we propose in [11], where features like terms and simple relations (verb-object and verb-subject pairs) are used to represent document content. Due to modular approach in architectural design, our technology for IR&C may be substitute by other tools accessible on the market (commercial information retrieval engine with clustering capabilities as Vivisimo ©, RealTerm and e-Knowledge Portal™). As it is a very active area in information retrieval research [17], several products have been produced as a follow-up.

2.2 "Active" Information Access System

Aim of the "active" system is to follow the conversation in a project session (through the use of a Speech Recognizer module) and to "extract" an *implicit information need*, that will be in turn used to query and enhance the information retrieval core system.
As carrying out directly all these activities using NLP state of the art technologies still is a challenge to be faced, the basic idea is not to produce a complex information need extractor but a simple model taking advantages from stable technologies. Meanwhile, instead of a speech recognition module to produce minutes of the meeting, we can start from a manually provided version. The meeting minute is then used to feed an *Information Need Extractor* module able to extract the *implicit information needs*. A criterion to model how an implicit information need is expressed may be to investigate and give information on things and ideas where the communication fails, i.e. a concept that is not understood by two or more people in the same way. Repeated *terms* may suggest that a disagreement exists as the underlying concept is not shared. This may be an easy way to decide a sort of list of candidates to be searched. The Information Need Extractor can be thus intended as a simple module that, relying on a terminological repository is able to find the most frequent terms in the minutes and to query the IR&C system. Moreover, it should be able to enrich the repository with new terminological expressions contained in the minutes, using ad hoc methodologies, as described in Sec. 3.
As an example, imagine that during the meeting the experts are discussing about different option in building a lunch vehicle. From the automatic minute produced by the speech recognition module frequent terminological expressions could then emerge, such as "launch vehicle" and "mechanical parts". The Information Need Extractor should recognize these frequent terms and query the IR&C system using them as keywords. At the end of the process, the experts could thus be provided with relevant documents that could support their decisions, organized in topical clusters, such as "Test design" (containing documents on previous design of vehicles) and "Reusable Launch Vehicle" (documents on designing general purpose vehicles).

2.3 The generic linguistic knowledge

Generic linguistic resources can be used to help the system in interpreting and disambiguating the content of both pre-existing and on-going knowledge repositories.
As natural language is rich of information and, as a consequence, very ambiguous, words may convey very different meaning while different words may be used to express the same concept. To tackle with this problem linguistic background knowledge resource such WordNet ([10]) can be used. These resources may be coupled with a graded activation of these relationships among words, that often take the form of probabilities [15], [6]. The use of the linguistic knowledge is particularly useful in the following phase of creating and enriching the domain ontology, as described in the next section.

2.4 The *design process* ontology

As a further relevant step, the architecture can be enriched with a domain specific ontology, able to represent the knowledge emerging from the design process through pre-existing knowledge repositories and document retrieved by the IR module.
A few approaches have been proposed to learn automatically or semi-automatically a domain ontology from textual material (e.g. [1],[9]). Here, we propose a novel methodology, able to fix in a single structured and harmonized knowledge base different types of information: an upper-level ontology of domain concepts (*domain concept hierarchy, DCH*), an set of semantic relations among concepts (*relation type system, RTS*), aterminology extracted from the knowledge repositories (*terms*), a set of verbal relations among terms (*relational patterns*), and a generic linguistic knowledge (*linguistic knowledge base, LKB*)
The *DCH* formalizes the knowledge of the design process in a conceptual hierarchy (e.g. in SHUMI, concepts like *spacecraft* and *orbit* are here represented). The *RTS* hierarchy stores important *semantic relations* among concepts in the DCH (e.g. the event of a *spacecraft reaching an orbit*). *Terms* are defined as “surface linguistic forms of relevant domain concepts” ([12]): the terminology, automatically extracted from the knowledge repositories, thus represents a synthetic linguistic representation of domain concepts as embodied in documents. Terms are then linked to their corresponding concepts in the DCH (for example the term *Earth’s_orbit* should be attached to the concept *orbit*). As terms are linguistic representation of concepts, in the same way *relational patterns* are (partially generalized) verbal relation prototypes that represent semantic relations in the RTS. For example the patterns *spacecraft gets close to Lunar orbit* (that can be a generalization of text fragments like *Shuttle gets close to Lunar orbit* and *Endeavour gets close to Lunar orbit*) should be associated to the semantic relation *spacecraft reaching an orbit.* As semantic relations are usually linguistically expressed through fragments governed by verbs, in our model they are supposed to be instantiated in text only by verbal patterns. The WordNet LKB represents a hierarchical linguistic repository of generic lexical knowledge: a link can be thus established between concepts in the DCH and synset in the LKB. For example the concept *spacecraft* can be associated with the synset *{spacecraft, ballistic capsule, space vehicle}.*

What we propose is an acquisition method that, starting from a pre-existing DCH and LKB , is able to derive the *linguistic interface* of the ontology (composed by the LKB, the relational patterns and the terms) suggesting linguistic patterns for known concepts and relations as well as to propose new concepts and new semantic relation. Knowledge textual repositories are the starting point of our analysis and are assumed to drive the discovery of new domain knowledge.

The overall learning process is organized as follows. Firstly terms and relational patterns are extracted from the corpus. Then, an analysis devoted to determine a concept hierarchy is applied to the more relevant concepts patterns extracted, making use of the pre-existing DCH. This activity generalizes the available evidence across the LKB and is called *Semantic Dictionary Building*. Domain concepts are also mapped into the general lexical database (we propose an automatic method, described in [5]). The resulting *concept hierarchy* can be successively used in the analysis and interpretation of relational patterns in the domain texts. This generalization allows to conceptually cluster the surface forms observed throughout the corpus. The derived generalizations can undergo the statistical processing during the *Domain Oriented Clustering* phase. The resulting generalized patterns can be organized according to their domain relevance score. The manual *Relation Type Definition* phase identifies a system of important domain concept relationships, which are in turn used for the manual or semi-supervised *Relational Pattern Classification* phase. The previously clustered relational patterns are thus mapped into the appropriate semantic relations. The result of this last activity is the set of linguistic rules for the matching and prediction of relations in RTS (*Linguistic Relation Interfaces*).

The ontological repository can be then used to support the design process, providing a central view of the overall knowledge. As a simple application, suppose for example that the team of experts is interested in finding all the textual material gathered so far (minutes and external documents previously queried via the IR&C) related to the modality of launch of spacecrafts. They could simply access the ontology to easily find the semantic relation "launching of spacecraft" navigating the RTS hierarchy. They would then retrieve all the relational patterns and the terms linked to the semantic relation and finally obtain the documents in which the patterns and the terms have been found.

3. Extracting terms and relational patterns

As stated above, one of the primary tasks in building the ontology is to extract terms and relational patterns. At the present we do the simplifying assumption that semantic relations are expressed in the text only through verbal fragments as it usually happens. *Terms*, defined as surface (linguistic) representations of domain key concepts, are automatically extracted from texts using NLP techniques supported by statistical measures. Many approaches to terminology extraction have been proposed in the literature, ranging from purely linguistic (e.g. [7]) to purely statistical (e.g. [16]). Usually, mixed approach are the most reliable and used (e.g. [8], [12]): *candidate terms* are extracted from text as noun phrases having particular syntactic structure (e.g. *adjective+noun*, *noun+noun*) and then ordered according to a specific statistical

measure that is supposed to capture the notion of *termhood* (the degree of reliability with which a text fragment is supposed to be a term). In our architecture a mixed approach has been chosen, mixing linguistic filters with a measure (frequency) that seems to capture the notion of termhood, according to different studies (e.g.[8],[14]) where a comparative analysis over different measures have been done.
Relational patterns are generalized forms of lexical knowledge that represent a sort of normalization of one or more actual textual *sentences*. In particular they are verb phrases, i.e., semantically generalized lexical fragments of text governed by a verb, representing the syntactic expressions of relational concepts. As for terms, also relational pattern extraction is carried out using a mix of linguistic and statistical methods [3]. In order to feed the ontology, once automatically extracted from the corpus, terms and relational patterns have to be validate by human experts.

3.1 Terminology and relational patterns extraction

The architecture of our *Term Extractor*, includes the modules hereafter described.
A **pre-processing module** takes as input the corpus documents in textual format, converting them into XML files readable by the syntactic parser, checking for possible corrections and adaptations. The **parsing module** invokes Chaos, a robust and modular parser architecture developed at the AI laboratory of Roma Tor Vergata University [2]. The **terminology extraction module** extracts *admissible surface forms* from the previously parsed text: specific syntactic rules are used to select candidates, identifying sequences of words with specific syntactic properties: for instance, syntactic sequences like *JJ NN* (an adjective followed by a singular common noun, as *"lunar mission"*) and *NN NNP* (singular common noun followed by a plural common noun, as *"spacecraft projects"*) are retained as possible surface forms. Finally, the **terminology sorting module** sorts by relevance the list of previously produced candidates. Relevance is evaluated as the frequency with which each form has been met in the corpus. In fact, while many statistical measures have been proposed in the literature to estimate term importance (Mutual Information, T-score, TfIdf, etc.), frequency has been demonstrated in several frameworks to be a good approximated measure to express term relevance, as underlined in [8] and [14]. The list of produced forms is the *candidate* terminology, as the set of candidate terms that still needs a manual validation by a human expert.
In our framework, each term can be a simple sequence of words (e.g. *"spacecraft_mission"*) or a semantically generalized form. In the latter case the candidate term is formed by words and *Named Entities* (NE) (semantic generalizations representing important entities of a specific domain, such as people or organizations). As an example the candidate term*"entity#ne#_mission"* indicates a mission of a generic *entity*, that is an organization, a person or a specific object (e.g. *"ESA mission"*).
Relational patterns are extracted from text using a strategy similar to the one adopted for terms. The *Relation Extraction* extracts surface forms by using as background knowledge the terms extracted by the Term Extractor, since relational patterns are intended as relations among terms. An architecture similar to the Terminology

Extractor is needed: corpus syntactic analysis is carried out to extract forms of interest.
The **relational pattern extraction module** analyses the parsed text produced by the parsing modules and extracts all verb phrases (text fragments): a list of *sentences* is thus produced, each of which is represented by the governing verb and its arguments. For each argument its lexical form and its syntactic role is indicated (for example *approach((SUBJ, the spacecraft), (OBJ,the orbit),(IN,ten minutes))*). The **relational pattern sorting module**, taking as input the corpus sentences, by first generalizes them into relational patterns, then ranks the patterns. The strategy we adopted for the generalization step is fully described in [3]. Once surface forms are produced, they are ranked accordingly to their frequency (calculated as the sum of the frequency of appearance of its corresponding sentences in the corpus). Candidate relational patterns are then validated by a human expert. An example of relational pattern that generalizes the above sentence, could be *approach((SUBJ, spacecraft), (OBJ,orbit))*
It must be noticed that, as in the case of terms, NE are used in the extraction of relational patterns, producing pattern like *approach((SUBJ, mission#ne#), (OBJ,orbit))*, where *mission#ne#* represents the entity class of spacecraft missions. The pattern thus generalizes all the sentences which have “approach” as verb, “orbit” as object and any spacecraft mission as subject (i.e. “Mariner”, “Voyager” etc.).

3.2 SHUMI case study: preliminary results

In order to estimate the validity of our term and relational pattern extraction methods, in the framework of the SHUMI project, we tested our architecture over a corpus of spacecraft design documents specifically provided by ESA, consisting in a collection of 32 ESA reports, tutorials and glossaries, forming 4,2 MB of textual material (about 673.000 words), fairly in line with other experiments in term extraction, such as [8] (240.000 words) and [7] (1.200.000 words). Extracted terms and relational patterns have been manually validated by a pool of ESA experts.
58.267 candidate terms have been extracted from the ESA corpus, among which 7821 (14%) have been retained as useful by the experts. Out of the 58.267 candidates, 4820 appear inside the corpus more than five times, with an accuracy of 38% (1814 terms retained). As the accuracy rises from 14% to 38%, a frequency of five can be thus empirically considered as a good threshold to automatically separate interesting term from spurious ones.
As outlined in [8] the most interesting and frequent terms are those composed by two *main items* (i.e., counting only meaningful words, such as noun, adjectives and adverbs): indeed, in our experiment roughly 60% of retained terms are 2-words. A list of the 10 *most relevant* terms (that is with highest frequency and retained by the experts) and a list of the 10 2-words most relevant terms is reported in Fig.2 (where entity#ne# is a generic NE standing for persons, companies and organizations), together with the list of 2-words non generalized most relevant terms (without NE). Terms as *“solar wind”* and *“magnetic field”* represent important concepts for an envisioned ontology for spacecraft design: those terms are in fact a useful hint both to identify concepts to insert into the ontology and to model the ontology itself.

For what concerns relational patterns, the system extracted 110.688 forms, among which the 21% has been retained by the experts (a quite good accuracy considering that the procedure of patterns extraction is affected by the problem of *overgeneration*,

Requirement	entity#ne#_system	application_datum
System	application_datum	magnetic_field
spacecraft	entity#ne#_packet	solar_wind
datum	entity#ne#_requirement	technical_requirement
test	entity#ne#_engineering	test_level
time	entity#ne#_state	source_packets
orbit	magnetic_field	source_datum
process	entity#ne#_model	launch_vehicle
operation	solar_wind	mechanical_part
design	entity#ne#_spacecraft	mission_phase

Fig.2 Ten most relevant terms (left), ten most relevant *2-words* terms (center) and most relevant not generalized *2-words* terms (right).

that is, each verb sentence met in the corpus creates several related surface forms, some of which can be sometimes too general to be considered interesting). Fig.3 shows the most relevant (i.e. frequent) patterns.

perform((SUBJ,test))
conform((TO,requirement))
meet((DIROBJ,requirement))
conform((SUBJ,null),(TO,requirement))
do((SUBJ,service))
conduct((SUBJ,test))
conform((DIROBJ,null),(TO,'space_organization#ne#'))
conform((TO,'space_organization#ne#'))
conform((DIROBJ,null),(DIROBJ2,null),(TO,'space_organization#ne#'))
perform((SUBJ,analysis))

Fig.3 Ten most relevant relational patterns validated by the experts.

As it can be inferred from previous table, most of the surface forms retained by the experts are governed by verbs whose driven semantic *meaning in phrases* usually directly refers to events regarding planning and design. That is, these verbs, used in specific context (i.e. spacecraft design) assume a particular meaning. For example, the verb “meet”, that in general can assume many senses and semantic values (10 according to *The Concise Oxford Dictionary*), in the analyzed spacecraft design context assumes a specific semantic value. This “sense restriction” has two important implications in the overall automatic process. From one side it underlies the importance of surface forms in order to build a correct DCH (it emerges how verbs behave either semantically or syntactically in specific domains). Moreover, verb senses a sort of *verb sense disambiguation* is automatically carried out.

4. Further improvements

At the moment we are focusing our major efforts in modeling and implementing the ontology building process. We are trying to develop a framework in which semi-automatic techniques cooperate in learning the domain ontology using linguistic and semantic approaches (see [5]). The relational pattern semantic clustering activity is also a challenging issue we are still exploring, using Machine Learning techniques based on linguistic and semantic features ([4]). Techniques to cut down the need for human support is also an important point: so far, domain experts are in fact requested to validated terms and relational patters and to help in building at least the top levels of the DCH and RTS hierarchies. While the latter task is an unavoidable and "one time" step, the former is highly time consuming, as it involves a vast amount of data. We are thus developing interactive tools able to support and speed up validation.

References

1. Agirre, E., Ansa, O., Hovy, E., and Martinez, D. Enriching very large ontologies using the WWW. In: Proceedings of the Workshop on Ontology Construction of ECAI-00 (2000)
2. Basili, R., Pazienza, M.T., Zanzotto, F.M.: Customizable modular lexicalized parsing. In: Proc. of the 6th International Workshop on Parsing Technology (2000)
3. Basili, R., Pazienza, M.T., Zanzotto, F.M.: Learning IE patterns: a terminology extraction perspective. In: Workshop of Event Modelling for Multilingual Document Linking at LREC 2002, Canary Islands, Spain (2002)
4. Basili, R., Pazienza, M.T., Zanzotto, F.M.: Exploiting the feature vector model for learning linguistic representations of relational concepts. In: Workshop on Adaptive Text Extraction and Mining (ATEM 2003). Cavtat, Croatia (2003)
5. Basili,R., Vindigni, M., Zanzotto, F.M.: Integrating ontological and linguistic knowledge for Conceptual Information Extraction Web Intelligence (WI 2003) Halifax, Canada (2003)
6. Basili, R., Cammisa, M., Zanzotto, F.M.: A semantic similarity measure for unsupervised semantic disambiguation. In: Proceedings of the Language, Resources and Evaluation LREC 2004 Conference, Lisbon, Portugal (2004)
7. Bourigault, D.: Surface grammatical analysis for the extraction of terminological noun phrases. In: Proceedings of the Fifteenth International Conference on Computational Linguistics (1992) 977-981
8. Daille, B. : Approach mixte pour l'extraction de termilogie: statistique lexicale et filters linguistiques. PhD Thesis, C2V, TALANA, Universitè Paris VII (1994)
9. Hahn, U., Schnattinger, K.: Towards text knowledge engineering. In Proceedings of AAAI '98 / IAAI '98, Madison, Wisco (1998)
10. Miller, G.A.: WordNet: A lexical Database for English. In: Communication of the ACM, 38(11) (1995) 39-41
11. Moschitti, A., Zanzotto, F.M.: A robust summarization system to explain document categorization. In: Proceedings of ROMAND2002, Frascati, Italy July (2002)
12. Pazienza, M.T.: A domain specific terminology extraction system. In: International Journal of Terminology. Benjamin Ed., Vol.5.2 (1999) 183-201
13. Pazienza, M.T., Pennacchiotti, M., Vindigni, M., Zanzotto, F.M.: Shumi, Support To Human Machine Interaction. Technical Report. ESA-ESTEC cont.18149/04/NL/MV (2004)
14. Pazienza, M.T., Pennacchiotti, M., Zanzotto F.M.: Terminology extraction: an analysis of linguistic and statistical approaches. In Knowledge Mining, Springer Verlag, 2005
15. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (1995)
16. Salton, G., Yang, C.S., Yu, C.T.: A Theory of term importance in automatic text analysis. In: Journal of the American Society for Information Science 26(1) (1972) 33-44
17. Wu,W., Xiong, H., Shekhar, S.: Clustering and Information Retrieval. Kluwer Academic Publishers, Boston (2003)