Terminology Extraction and Knowledge Management

Roberto Guarasci[1], Anna Rovella[1], Stefano Vuono[1] and Paolo De Gasperis[2]

[1]Dip.Linguistica - Università della Calabria – Cosenza - Italy ; [2]SeGID CNR – Roma - Italy

Abstract

The reform of the National Research Council (C.N.R.) has introduced a new typology for the management of research planning activities, which are now seen as operating transversely as opposed to their being organised by single institutes. The project which is now underway, which is the result of a public/private partnership, hypothesises the use of the term clustering for the organisational analysis and functional reunification of the research policies of the Council. Experts in computer studies, documentation and terminology, working in synergy, have begun to produce the first significant results, which have already taken concrete form in the creation of the C.N.R.'s documentary management system, and in the activation of an Integrated System for the Management of Research Policy (SIGLA).

1.Introduction

The growing national production of legislation with regard to the necessity and imperativeness of the production of documents which are not in paper form and a parallel, though more tranquil, European trend in the same sense have, in the last decade, caused many disciplinary certainties to waver, damaged many timeworn training curricula and brought to light the necessity of convergence between disciplines which previously had never built ties in terms of organic collaboration. This requirement, which is particularly evident when one comes to specific aspects of textual analysis, has found concrete form, for example in France, with the establishment of TIA, a research group dedicated to Terminology and Artificial Intelligence, fruit of the convergence of terminologists, linguists, documentalists and ICT specialists (Aussenac Gilles and Condamines, 2004).

The multidisciplinary approach to knowledge classification, which is substantially new on the Italian scene, allows for the presentation of a global view of the problem, by postulating the collaboration of diverse forms of expertise in relation to subsequent moments in the construction of the knowledge environment.

Furthermore, such an approach allows for the humanistic re-appropriation of ambits which, for too much time and through our fault, have been the exclusive prerogative of other disciplines which were not always endowed with the necessary and specific competencies, yet surely have been holders to an all-inclusive claim which is difficult to justify on a theoretical level. It is in this sense and in this direction the idea of using methods proper to documentation, computational linguistics and terminological analysis for the construction of an adaptable semantic system aimed at the classification of the research of the National Research Council (CNR), acquires particular significance and value in terms of planning (the project group is coordinated by Prof.Giovanni Adamo of Iliesi-CNR in Roma and other participants are the Laboratory of Documentation, the University of Calabria, the University of Catania, the ISUFI at the University of Lecce and Bologna, and the Scuola Superiore Interpreti e Traduttori of Forlì).

The Legislative decree n. 127 (namely, d.lgs. 4 giugno 2003 n.127, published on G.U.R.I. Serie Generale n.129 on June 6th 2003) aimed to provide a radical change of the National Research Council with a view to "promoting and linking up operational centres of excellence, to avoid the duplication of their objectives, and to ensure the maximum level of flexibility, autonomy and efficiency, as well as a more favourable drawing up of understandings, programme agreements and consortiums….." (article 1). Consequently, this has brought about a substantial re-think of the entire scientific network of the Council, with the constitution of 11 departments, corresponding to a certain number of macro areas of research (i.e., Agricultural and Food, Life Sciences, Medicine, Materials and Devices, Molecular Planning, Land and Environment, ICT, Production Systems, Cultural Identity, Cultural Heritage, Energy and Transport), the elimination of sections, and the organisation of projects and job commissions for research work, with the idea of constructing a matrix system "in which institutes intersect – with the functions of running research activities and the dynamic management of competencies – and Departments – with functions related to planning – which commission research both on the part of the relevant Institutes, and other Institutes in the CNR, as well as externally to these" (AA.VV., 2004). Projects and work orders are, however, almost always transversal and multi-disciplinary and involve, therefore, the creation of research groups which do not necessarily take account of the partitioning which today characterises the geographical articulation of the Council's scientific network. As above mentioned, it is worth noting that the CNR organisation regulations, as resulting from the reordering Decree, does not foresee the existence of "sections which are territorially distinguished" from the institutes while, according to earlier rules, they were articulations of the institutes themselves, also constituting "autonomous centres of expenditure" in terms of activity management. From June 1st 2005, according to CNR presidential act n.35/2005, sections have formally ceased to exist and an institutes has become a single structure whose unique responsible is the director himself. In the same time, a job commission is defined as the segment of activity which an executive organ (for instance, an Institute) agrees to carry out for an external costumer – the project head – whose reference point is a Department.

Once a system with this configuration is established, some form of equilibrium and coherence should entail in terms of research demands, expressed by Departments and the research offered, proposed by the Institutes, thus initiating a functional re-organisation which it will be possible to define, after the system has been allowed to function for a preliminary period.

To support this organisational model, SIGLA (Information System for the Management of Activity Policy) (Guarasci et al., 2005) has been created, with the task not only of managing the finance and accounting of research activities, but also of creating a more complex system for the representation and recording of events, as well as monitoring and evaluating results.

2.Technical aspects and developments

In detail, SIGLA is and integrated information system offering support to:

1) the process of triennial planning;

2) the process of proposal and negotiation of work commissions;

3) the production of budgets (estimates and final balances) and relative management;

4) the production of balances (estimates and final balance) and relative management;

5) the production of the Economic Balance and of the Financial Standing;

6) the verification of the achievement of planning objectives and verification of their results.

Functionally, this is interconnected with the Council's system for document management which links up the entire research structure for the country. In this last system, with a view also to putting into practice the existing regulations with regard to the de-materialisation and management of not paper documentation, classification of these is foreseen, apart from the obligatory protocol of actions, including research documentation, by means of the use of a classification scheme and optical scansion of those documents which are in paper form. The recuperation of information, apart from by means of the obligatory numerical indicators, is carried out by tokening and stemming, limited to the text strings typed into the database fields.

From a formal standpoint, with the resolution n.57/2004 (December 23rd, 2004), the CNR Administrative Council has defined the structuring of the Homogeneous Organisational Areas with characteristics and specifications in agreement with art.50 of DPR 445/2000; with the resolution n.21/2005 (February 9th, 2005) it has approved two collaborative agreements with the University of Calabria, Prisma Engineering and FileNet Italia for the realisation of a computerised protocol system and the management of document flows, to be integrated with the Information System for the management of activity policy (SIGLA); with the resolution n.39/2005 (April 6th, 2005) the classification scheme has been approved for the classification, relating to the service for the computerised management of documents, document flow and archives, according to art.61 of DPR 445/2000. By the President's provision n.44/2005 (June 21st, 2005), the Office for the computerised management of document, document flow and archives has been instituted and its Director nominated; from September 8th 2005, the system has been activated limiting to the AOO of the Central Administration and of the Presidency; on February 28th 2006, the link with the institutes has been activated, including the entire connection of 107 institutes on March 30th 2006.

This is also in consideration of the fact that documentary production is for the most part in paper form and the obligation – as stated in the management manual – to attach the file to the scanned documents, even in the absence of digital signature – has not, so far, produced significant effects, making other retrieval systems difficult to apply. The necessity of circumscribing the research within a chain composed of a limited number of syntagms, the typing of which is certainly not free of a variable degree of arbitrariness depending on the human operator, apart from the known limits characteristic of the typology of retrieval used, has led to the prefiguration of more performing hypotheses with regard to the logics of interrogation. An early attempt was that of asking for the insertion of three keywords –freely chosen – by each of the document compilers. This artifice, the limits of which are known, was not necessarily foreseen so much as an optimal solution, as, rather, an instrument to evaluate whether the specialist context of use, though multi-domain, managed either wholly or in part, to provide a remedy to the generic nature of the terms normally introduced in similar contexts. Results were substantially ambivalent. On the one hand, the response level of the keywords introduced with respect to the original texts was satisfactory, yet relative solely to the principle theme of the research, without taking any account of its internal articulation. This last information emerged, however, as often being more significant than the first in terms of the comprehension of activity policy and the industrial spin-off of the results, also in consideration of the fact that it was precisely the already cited logic of transversal aggregations for projects and work commissions that imposed a high degree of internal complexity. A further step was the creation of predefined term lists with controlled implementation. Starting with the supposition that research activities are generally triennial and that, therefore, the corpus of texts for work commissions for the year 2005 is super-imposable by at least 80% compared to that of the current year, an attempt was made to extract from that corpus – composed of around 500 CNR work commissions in 2005 – term lists through a process of conceptual analysis and morphological insertion of the subject strings obtained at the end of the process.

This produced 998 index terms the significance of which was certainly greater compared to the keywords directly typed by users, but which – removed from the context of reference – were certainly not exhaustive compared to the results expected. Even the hypothesis of considering the extracted terms as segments of a thesaurus, implementing, therefore, for each of these, the set of relations contained in a pre-coordinated indexing dictionary did not produce a practical application due to the already noted multi-domain typology of the documents treated and the consequent unavailability of the relative taxonomies. Furthermore, in the meantime, a further request arrived from Council management to organise indexing terms in a spatial and three-dimensional scenario in which the axes X Y Z would have represented, respectively, the *subject*, the *objective* and the methodological *approach* of the work commission. This with a view to allowing for the immediate verification of cases of superimposition/duplication of activities and the transformation of those verified into synergic activities through the eventual modification of one of the variables.

A work commission: *Chemical-physical study of dumps for land analysis*…..would thus be segmented as, study of dumps (objective) chemical-physical (approach) land analysis (subject).

In a Euclidean vision, every work commission should therefore be traceable to a point (x,y,z) in the three-dimensional space. When several work commissions occur in a given interval, they constitute a *cluster* (Lapalut, 1995). Thematic clusters thus created represent the future and hypothetical organisational aggregations within the Council.

The response to this requirement encountered a limitation in the "…current statistical methods and instruments in data mining which extract recurrent patterns from data and information, considering only their attributes and not knowledge of the domain. Such systems are therefore not able to recall part of the knowledge assimilated previously, to disambiguate the content of words and sentences, assigning new meanings to the text which are not directly obtainable from its content" (Lella, 2006). Further to this, the creation of three-dimensional clusters postulated the assignation of weights to the links between terms, weights which, with respect to the original corpora, had necessarily to be updated automatically with respect to the variation of the content of the documents used or the addition of new texts.

A solution to the problem could be constituted by the use of associative networks (Mc Clelland and Rumelhart, 1986). The propagation of an activation signal which spreads across the semantic nodes of a network, nodes constituted of the meanings which a term possesses within a text or a portion of it, until its value becomes stable after being amplified and/or accumulated according to the weights of the links crossed, allowing the identification of the nodes most activated and therefore defining the context of use for the terms in the document in question.

3.Conclusions

Thus we have the textual analysis of a work commission: Materials and processes for energy, identify the nodes correspondent to the terms: "materials", "processes" and "energy" and by means of the diffusion of the activation signal, it can activate the nodes "combustion" and "high temperatures" to which the node "materials" is linked, while the level of the activation signal in the presence of the node "anemometry" would be low, though it could be activated more in the case of the commission "materials for wind power". The process of assignation of domain meaning is to be understood, however, as semi automatic, in that it is possible to verify the co-presence – at an equal signal level – in two semantic sub domains also, though not necessarily, alternative. The advantage of the system in the presence of texts which are multi-domain texts and not completely structured, even though partly pre-formatted, is its independence from syntactic parsing which gives it a high degree of adaptability to the various specialist languages used by writers.

The identification of univocal nodes with determined weights is not only functional to the creation of clusters, but, as we set about prefiguring not only the domain ontology as a sum of the relevant nodes, but also the single pseudo-ontologies for each single text, it allows for the construction not only of the profile of the writer on the basis of the query, but also of a profile of the document and its relationship with the three axes identified.

The really new things – as we said at the beginning – are in the multi-disciplinary approach, of which the technological solutions are natural emanations, in the sphere of application – the only real multi-domain sphere in the national research panorama and, in conclusion, in internal organisational goals, certainly not commonly foreseen as the product of the activities of document management or terminology extraction.

Since dealing with documents homogeneous from a contents standpoint and formally pre-formatted (as XML-based data outputs), the problem of the classification of classifications for textual typologies has been only theoretically treated, but not practically implemented as, in such an approach, the differentiation level does not appear to be particularly meaningful. In the case of added documentation, a classification for textual features has been introduced to recognize the two following typologies: i) administrative documentation and ii) research reports based on various compilation dtd.

4. References

Aussenac Gilles N. and Condamines A. (2004). Documents èlectroniques et constitution de resources terminologiques ou ontologiques. *Information-Interaction-Intelligence*, vol.4:75-93

AA.VV.(2004). La gestione per commesse delle attività di ricerca e sviluppo del CNR. *Documento CDA*, n.16/04

Guarasci R., Lancia M, Mascari J.F., Tuzi F., Puccinelli R., Rovella A. (2005). SIGLA: a flexible portal and reusable ERP system for Research Management. *Proceedings of the 3rd International Conference on Politics and Information Systems:Technologies and Applications.* Orlando (USA).

Lapalut S. (1995). Text Clustering to support Knowledge Acquisition from Documents. Rapport de Recherche INRIA n.2639.

Lella L.(2006). Oltre il Datamining: verso l'estrazione della conoscenza tacita. *www.itconsult.it*

Mc Clelland J and Rumelhart D.E. (1986). *Parallel distributed processing*. MIT Press.