SARA GOGGI,
GABRIELLA PARDELLI,
MANUELA SASSI
CNR, Institute of Computational Linguistic, "Antonio Zampolli", Pisa, Italy
[Sara.Goggi, Gabriella.Pardelli, Manuela.Sassi]@Ilc.cnr.It
SILVIA GIANNINI,
STEFANIA BIAGIONI
CNR, Institute of Science and Information Technologies "A. Faedo", Pisa, Italy
[Silvia.Giannini, Stefania.Biagioni]@Isti.cnr.It

A Terminological Survey on the Titles of the Seventh Framework Programme (FP7)

Abstract

This paper focuses on the automatic extraction of domain-specific knowledge from the European Commission projects of the 7th Framework Programme, hereinafter referred as FP7.
The study is divided in three parts: the first part introduces the work starting from the building up of a corpus containing the titles of European Projects of the whole FP7 in order to obtain a relevant terminological sample for the different domains; the second describes software and methods while the third part focuses on the evaluation of results. Finally, we conclude by suggesting possible directions for further development of a comparison between terminological extraction from FP7 and FP5/FP6.
Keywords: 7th Framework Programme (FP7), Natural Language Processing, Terminology, knowledge extraction

1. Introduction

The creation of a corpus of titles of projects extracted from various European Framework Programmes would constitute a terminological niche, a sort of "cluster map" which would offer an overall vision on the terms used and the links between them. Within this scenario, the minimal purpose was to build a corpus of European project titles belonging to FP7 which would allow to obtain a terminological mapping of relevant words in the various research areas: particularly significant will be those terms spread across different domains. A term could actually be found in many fields and being able to acknowledge and retrieve this cross-presence means being able to linking those different domains by means of a terminological mapping.
The analysis of the projects titles has been carried out by using the new version of a software named "Text-to-Knowledge2 **(**T2K[2]**)"** [2] developed by the ItaliaNLP Laboratory of the Institute of Computational Linguistics "Antonio Zampolli" of CNR[1], a hybrid system for knowledge extraction and document indexing that combines natural language processing (NLP) technologies and statistical techniques.
The 7th Framework Programme (FP7) focuses its attention on important research areas such as health, information and communication technologies as well as spatial ones. The FP7 lasted 7 years (2007-2013) and this was an innovation with respect to the previous framework programmes that were on a five-year basis: the aim was maintaining continuity and thus guarantee a greater coherence of action. Within the Seventh Framework Programme, European research projects under the theme of Socio-economic Sciences and Humanities were funded since 2007: these projects concern the Lisbon strategy[2], sustainable development and regional cohesion, major societal trends, Europe in the world, the citizen in the EU, indicators, foresight and infrastructure.
The analysis of data provided by the extraction tool focused in particular on the mapping of some relevant terms in the domain of Information and Communication Technology (ICT), that is the set made up of technologies (software programmes, components and systems) which allow the processing and exchange of any kind of information (digital, textual, visual, audio or multimedia). ICT also represents one of the macro research areas (Departments) of the Italian National Research Council (CNR), to which the two research Institutes (ILC and ISTI) of the authors belong: the ICT Department monitors the growth of the above-mentioned technologies and studies their impact on society, with the goal of developing innovative products and know-how for responding to the current technological, strategical, societal and economical needs.
Since the second half of the XXth century, the gradual and pervading dissemination of information technologies caused a rapid spread of terms of the ICT domain to other fields: a sort of "lingua franca" [3] comprehensible both to common citizens and experts of the field has originated and the computer science lexicon has permeated other domain-specific lexica thus facilitating the exchange of information across multidiciplinary environments. This study centers on interpreting the terminological distribution on the titles which constitutes the FP7 corpus in order to verify to which extent the exacteness of a scenario where some domain-specific (i.e. information technology) terms spread to different disciplines – and those which are the most affected ones - can be confirmed.

---

[1] www.italianlp.it
[2] Cfr. Wikipedia "… The Lisbon Strategy, also known as the Lisbon Agenda or Lisbon Process, was an action and development plan devised in 2000, for the economy of the European Union between 2000 and 2010…
http://en.wikipedia.org/wiki/Lisbon_Strategy
[3] From Wikipedia: "A lingua franca also called a bridge language, or vehicular language, is a language systematically (as opposed to occasionally, or casually) used to make communication possible between persons not sharing a mother tongue, in particular when it is a third language, distinct from both mother tongues".
http://en.wikipedia.org/wiki/Lingua_franca

2. Materials and methods

The starting point of our analysis was the creation of the corpus of FP7 project titles. Data was collected between January and March 2013 and the information was extracted from CORDIS[4]. We gathered Grant number, Title, Specific program, Dates (start and end of project). The dataset was "fed" to the software T2K^2, a suit of tools for automatically extracting domain–specific knowledge from collections of Italian and English texts. T2K^2 relies on a battery of tools for NLP, statistical text analysis and machine learning which are dynamically integrated to provide an accurate and incremental representation of the content of vast repositories of unstructured documents.
Extracted knowledge ranges from domain–specific entities and named entities to the relations connecting them and can be used for indexing document collections with respect to different information types. T2K^2 also includes "linguistic profiling" functionalities aimed at supporting the user in constructing the acquisition corpus, e.g. in selecting texts belonging to the same genre or characterized by the same degree of specialization or in monitoring the "added value" of newly inserted documents. T2K^2 is a web application which can be accessed from any browser through a personal account which has been tested in a wide range of domains (Dell'Orletta, 2014). In our case study, T2K^2 elaborated the data contained in the field "Title" and performed three levels of automatic extraction:

1. list of single and multi-word terms ordered by relevance with respect to the context;
2. creation of fragments of taxonomical chains;
3. creation of clusters of related terms.

Thanks to the analysis performed by the extraction tool it is indeed possible to statistically group the terms as to minimize the "logical distance" within each group of words and to maximize the distance among the different groups. Those groups - called clusters – become the single unit of the analysis.
This work specifically analyzes the cluster "Network/Networks": this is a term with various lexical "flavours" which are very well-connected with the contents of our FP7 corpus: the resulted terminological occurrences and associations have then been investigated for proving their use in different fields.

3. Case studies: *Network* mapping

A network is a structure made up of a circuit of connected information originating from several entities (such as academic communities, professional associations or just common citizens) which choose to exchange and share any kind of resource by means of a web connection. By now the Internet has become a platform for exchanging information, alongside with tools and services, playing the role of distributing knowledge. The funding of the European Commission allows the building up of several networks of excellence which successfully carry on research programmes providing technologies and methods useful for developing and sharing resources amongst research groups and institutions at an international level: sharing of knowledge is therefore more efficient and transparent.
In the Dictionary Reference[5] the word Network, within the ICT context, is defined as follows: "A system of interconnected computer systems, terminals, and other equipment allowing information to be exchanged". As a result of the processing of the FP7, 888 occurrences of the term "Network/Networks" are found; in Table 1 the first 28 terms, ordered by frequency (from 1097 to 290) and corresponding to the first level of the automatic extraction, are presented: the first five prototypical forms are *Development, Research, Systems, Network, Technology*.

Prototypical_Form	Frequency	Prototypical_Form	Frequency
Development	1097	applications	387
Research	1045	role	367
Systems	1010	Design	365
Network	888	dynamics	361
Technology	629	Quantum	361
cell	609	model	352
Europe	551	cancer	351
Energy	532	Science	340
Analysis	520	health	322
study	491	materials	322
Control	423	approach	318
disease	410	production	309
mechanisms	392	Novel	292
Management	388	change	290

Table 1

Some couples and triples of terms belonging to the cluster "Network" are analyzed and shown below in Figure 2: they are significant keys for identifying the scientific area of reference. In the next table (Table 2) the terminological associations of 2 or 3 words associated with the word *Network* – that is the taxonomical chains automatically obtained from the second level of extraction - are represented.

[4] Community Research and Development Information Service, is an information space devoted to European research and development (R&D) activities and technology transfer, http://cordis.europa.eu/guidance/helpdesk/faq_en.html
[5] http://dictionary.reference.com/browse/network?s=t.

--> molecular networks --> Network in Europe --> brain network --> Network for Earth --> regulatory networks --> Access Network --> Initial Training Network --> Future Networks --> European Network	--> Training Network --> Wireless Networks --> Social Networks --> Communication Networks --> Network of Excellence --> Sensor Networks --> Wireless Sensor Networks	--> European Training Network --> Optical Networks --> Research Network --> neuronal networks --> Transport Networks --> Network of National --> International Network --> gene regulatory network --> electricity networks --> neural networks --> Next Generation Networks

Table 2

In Figure 1 the third level of extraction – where clusters of terms related to "Network/Networks" are created - is visualized. The single and multi-word terms are ordered by relevance.

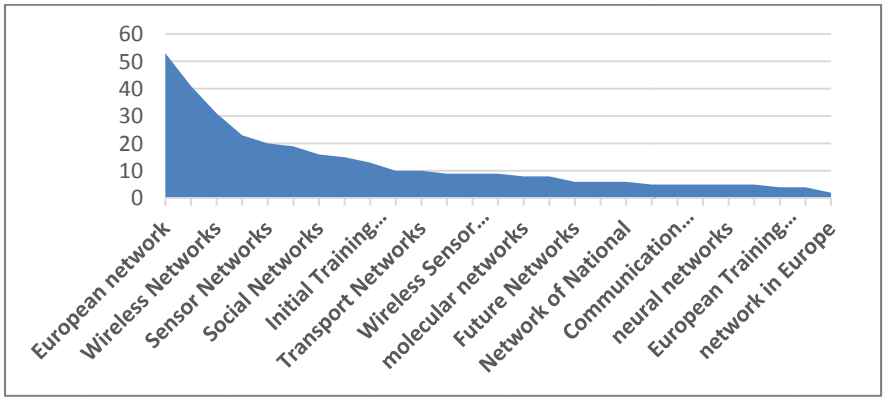
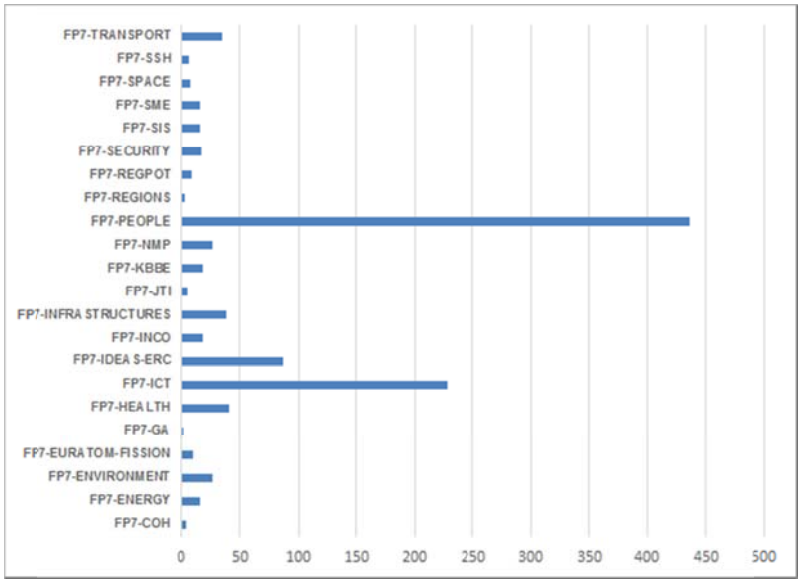


Figure 1 - 53 to 2 occurrences

Graph 1 points out the FP7 Programmes where the term "Network/Networks" appears: it occurs in mostly the 92% of them but with a very different effect, being higher in the PEOPLE Programme than in the ICT itself: a significant relevance can be noticed in the IDEAS, HEALTH, INFRASTRUCTURES and TRANSPORT Programmes as well.



Graph 1

The analysis of occurrences helps us in verifying the numerous terminological associations of "Network/Networks": the relevant presence of terms such as *European*, *Europe* and *EU* is in line with the objectives of research Programmes like PEOPLE and IDEAS that are to explicitly potentiate the "European" features of the research activity.
In this last Framework Programme, the 'Marie Curie Actions' had been regrouped and reinforced in the PEOPLE Specific Programme: entirely dedicated to human resources in research, it had a significant overall budget of more than 4,7 billion euros until 2013, which represented a 50% average annual increase over FP6.

The objectives of PEOPLE were to strengthen, both quantitatively and qualitatively, the human potential in research and technology in Europe, by stimulating people to invest energies in research. Efforts were also made to increase the participation of female researchers.
IDEAS deals with the basic "frontier" research and is devoted to foster wealth, social progress and promotion of a shared and competitive European area: communication and dissemination of results are a milestone of this Programme.

Figure 2 shows the first 10 more frequent lexical couples and/or triples of the cluster "Network" as visualized in Table 2: hereinafter the taxonomic chains will be briefly described and their participation with respect to the specific programmes will be surveyed.

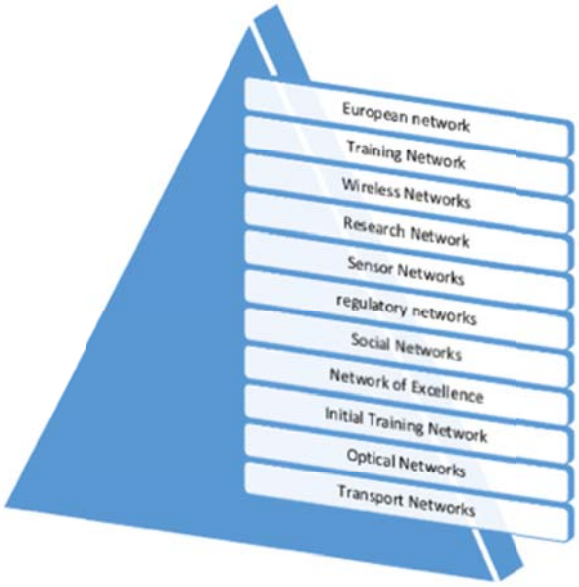


Figure 2

- Amongst the titles of our FP7 corpus, reference to Europe can be found in the acronyms as well both using the adjective *European* or shortened forms such as *euro* and *eu*. The pair of terms *European Network* counts 53 occurrences and derives from projects totally hinged on the constitution of scientific groupings spread all over the various fields of knowledge. The acronyms of these projects sometimes show the abbreviation *NET* (instead of Network).
- *Training network* is one of the most frequent pairs occurring within the corpus; it counts 43 occurrences and it is in second position after *European network*. The triple *Initial training network* appears 10 times and can be considered a sub-set of *Training network*.
- *Wireless network* impacts with many Programmes: ICT, PEOPLE, IDEAS; ERC; SME since the wireless technology turned upside down the worlds of telecommunications and in particular mobile computing as well as multimedia applications (radio and television broadcasts) and computer networking.
- Adequate *Research networks* for sharing and managing the flux of research information in Europe are distributed all over the areas of the FP7 Programmes: SECURITY, INFRASTRUCTURES, REGPOT, HEALTH, PEOPLE, etc.
- *Sensor networks* can be used in many applications: one of the most common applications is the monitoring of physical environments such as traffic in a big city or data collected from an area devastated by an earthquake.
- In FP7 there are 10 project titles containing the pair *Regulatory networks* (the word *Regulatory* means an authoritative rule or a condition that ensures something does not go beyond a given capacity) and are connected to PEOPLE and IDEAS Programmes; also the triple *Gene regulatory network* – related to the field of Bio-science - can be retrieved from our corpus.
- *Social networks* has 14 occurrences and it is a well-fitted example of a terminological use of wide range: the pair appears in titles of projects belonging to different fields and PEOPLE, ICT and IDEAS Programmes and there is a wide range of issues tackled (economical and financial, societal, technological, etc.).
- The phrasing *Network of excellence* identifies the so-called scientific "cradles" which witness the most-advanced European research and involve several domains such as ICT, security, public health.
- *Transport networks* can be found in the Programmes TRANSPORT (one of the most important in FP7), SECURITY, ICT and PEOPLE.
- *Optical networks* occurs 10 times in our corpus and is connected to ICT and PEOPLE Programmes only.

Conclusions

The trend towards a deeper specialization in the various fields alongside with the widespread inter-disciplinarity of research leads to the mixture of various domains (such as bio-physics, bio-informatics, bio-chemistry, psycholinguistics and so on) and contributes to the tendency of tackling issues from different perspectives: for example, transportation can be analyzed with respect to security of citizens, technical development and demolition of geographic barriers.

The $T2K^2$ software strengthens this vision by extracting groups of terms which deal with strategic European research areas and disclose the inter-disciplinarity amongst the several topics dealt with in the 7^{th} Framework Programme.

Our terminological case study on *Network* has driven us through different branches of knowledge by associating various uses of the same word: for example, the association of the term *Network* with the term *Wireless* leads to the implications and the *I*benefits that this technology brought in the field of telecommunications while the association with the term *Regulatory* faces the interactions between proteins and genes. In the digital era the literary cafés of the $XVIII^{th}$ and XIX^{th} centuries have been substituted by the *Networks of Excellence* and – at another level – by the *Social Networks*: new strategies of human communication for conveying high-level scientific knowledge as well as just moods and requests of friendship.

Other relevant examples are represented by the pairs *Future networks* and *Neural/Neuronal networks*: in this case as well terminology shifts from issues closely connected with technological strengthening to its application in different environments, ranging from the economic and social development of rural areas to the study of the brain passing through geo-morphological studies for the exploration of archaeological sites.

Also the terminological associations which are not included amongst the first 10 lexical pairs and/or triples provide numerous information and cover extensive fields: a good example is given by the pair Communication network which appears only 5 times within the corpus but is a sort of preamble to the several research perspectives in this domain represented by pairs and triples such as *Access Network*, *Future Networks*, *Wireless Networks*, *Social Networks*, *Communications Networks*, *Sensor Networks*, *Wireless Sensor Networks*, *Next Generation Networks*.

Future work

In our opinion, the evolution of terminology is a very rich field and for this reason we would like to further investigate the use of terms in FP4, FP5, FP6, FP7 corpora. As future work we are evaluating the possibility of extracting topic information from all the 80,000 FP titles for designing an efficient recognition process of obsolete and new words in order to identify old and most recent fields of European research in the time-span 1994-2013.

Essential Bibliography

[1] DELL'ORLETTA F., LENCI A., MARCHI S., MONTEMAGNI S., PIRRELLI V., VENTURI G. (2008). Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio. In: AIDAinformazioni: Rivista di Scienze dell'informazione, vol. 26 (1-2) pp. 197-218. Associazione Italiana per la Documentazione Avanzata.

[2] DELL'ORLETTA F., VENTURI G., CIMINO A., MONTEMAGNI S. (2014). T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts". In: LREC'14 - Ninth International Conference on Language Resources and Evaluation (Reykjavik, Iceland, 26-31 May 2014), edited by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis. Paris, European Language Resources Association (ELRA), pp. 2062-2070.

[3] PAZIENZA M.P., VIRDIGNI M.(2003). Agent based ontological mediation in IE systems. In: Information Extraction in the WEB Era. Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents, edited by M.T. Pazienza (Lecture Notes in Artificial Intelligence, v. 2700), Heildelberg, Springer.

[4] DAGAN I., MARCUS S., MARKOVITCH S. (1993). Contextual word similarity and estimation from sparse data. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, (22-26 June 1993, State University Columbus, Ohio, USA). New York, ACL, pp. 31-37.

[5] ZHANG M. et alii., (2010). Molecular network analysis and applications. In: Knowledge-Based Bioinformatics: from Analysis to Interpretation, edited by Gil Alterovitz and Marco Ramoni. Chichester, John Wiley & Sons.

[6] Community Research and Development Information Service, http://cordis.europa.eu/home_en.html.

Acknowledgments

The authors would like to thank Simonetta Montemagni, Director of the Institute of Computational Linguistics "Antonio Zampolli" (ILC) of the National Research Council and Felice Dell'Orletta, researcher at ILC for their support in the data curation process of the $T2K^2$ system.