Semantic networks for improved access to biomedical databases

Sassolini Eva, Cucurullo Sebastiana, Picchi Eugenio

Organization:	Istituto di Linguistica Computazionale "Antonio Zampoli"
Address:	Via Moruzzi, 1, Area della Ricerca di Pisa
Postal Code/City/Country	56124, Pisa, Italia
Topics:	Repurposing Grey literature
	Adapting New Technologies
Telephone:	050 315-2853/2759
Fax:	0503152839
Email:	{eva.sassolini|nella.cucurullo|picchi}@ilc.cnr.it

1. Introduction

The development of strategies and tools to access and analyze large amounts of data, so to discover correlations between seemingly unrelated data, capture associations and draw conclusions, is a research area of recent development in the acquisition of knowledge.
It is a fact that the study of innovative technologies has enabled an exponential growth of knowledge in the biomedical field, but the large amount of information available and the heterogeneity of information sources are a severe constraint to the full exploitation of such knowledge. The availability of systems for collecting and aggregating data as well as of analysis systems has therefore become a priority, mainly in fields which public health.

2. State of the art

Some research groups are developing tools and methods for summarizing of medical documents; others use specific Information Extraction (IE) techniques for building medical and biomedical ontologies, for example the infrastructure AMBIT (Acquiring Medical and Biological Information from Text), developed within the CLEF [1] and myGrid [2] projects. AMBIT aims to provide a intelligent access to large and unstructured biomedical data resources [1].
Recently, many efforts have been directed to the creation of large-scale terminological resources that merge information contained in various smaller resources: large thesauri based on a normalized nomenclature[2], extensible lexical and terminological databases like TERMINO[3] and the specialized Lexicon (e. g. BioLexicon[4], its peculiarity is to combine features of both terminologies and lexicons, within the project BootStrep[3]).

[1] The Clinical e-Science Framework (CLEF) project provides a repository of structured and well-organized clinical information which can be queried and summarized for biomedical research and clinical care.
[2] The project myGrid presents research biologists with a single unified workbench through which component bioinformatics services can be accessed using a workflow model.
[3] BootStrep (Bootstrapping Of Ontologies and Terminologies STrategic Project) is a STREP project of the FP6 IST (call 4), that involves six partners from four European countries (Germany, U.K.,Italy, France) and one Asian partner from Singapore.

3. SUBITO project

SUBITO (Unique Social Network for Innovation in Biomedical Tuscany) is a "POR Creo" project, promoted by the Tuscany region and funded by the European Community (FESR). The project goal is creating an archive and a website to collect all specific domain information, regarding institutional or private players in the field of life science. The project inherits and improves some previous experiences carried out in Tuscany, like ORBIT, THRAIN, Net-TLS. These projects have already identified some synergies existing in the territory, regarding technical knowledge, activities, skills and potential scientific-technological.
The project involved the Institute of Computational Linguistics “Antonio Zampolli” (hereafter ILC), the Institute of Clinical Physiology (IFC) and a consortium of private companies. Particularly, ILC has developed tools and resources for the extraction and classification of textual data in order to enable a more efficient browsing.

4. Textual database

We created the knowledge base with the retrieval of abstracts and other information from three main websites: PubMed.gov, Espacenet.com, ClinicalTrials.gov.
PubMed is known as the most reliable and used repository for the publishing of biomedical articles, since it is a service of the U.S. National Library of Medicine and the National Institutes of Health that comprises more than 22 million citations for biomedical papers from MEDLINE and journals with content related to life sciences. PubMed has become currently the standard of reference for the scientific papers in biomedical domain.
This type of text is available in the Web but its consultation is difficult, especially if we want a selection of documents related to a specific sub-domain: typical problems that a information retrieval system has when a user wants to retrieve information from any knowledge base. It is important to improve and innovate the quality of services offered to users. The above-mentioned text material can be identified as "grey literature" because internet has transformed the electronic publishing. The Web offers new tools and channels for producing, disseminating and assessing scientific literature. Author/producer and reader/consumer changed their roles. The transformation of the research environment and the birth of new channels of scientific communication show clear that grey literature needs a new conceptual framework[4].

5. Terminological resources

In a more general context if is have systems for extraction, management and browsing of semantic relevant information, it is also possible to experiment new approaches for the automatic selection of those terms that are able to identify a specific domain of interest, even if these not are ontological nodes known. For example if we want to identify all the articles dealing with “rare diseases” or if we want to study what the “emergent diseases” in Tuscany, we need to identify a different domain.

[4] D. J. Farace & J. Schöpfel (eds.) (2010). Grey Literature in Library and Information Studies. De Gruyter Saur

Our research team works in the ILC and builds upon experiences in NLP techniques (text mining, text analysis). As recent research pointed out, the knowledge, intended as relationships and dependencies among the various relevant information contained in a text, can be extracted by means of text mining techniques and particular linguistic-statistical algorithms.
Typical text mining tasks do include text categorization, text clustering, concept/entity extraction; for our purpose, however, this is not sufficient. As a matter of fact, analyzing the collected texts material through linguistic tools (morphology and tagger) and resources (terminological dictionaries, lists of proper names, last names, geographical places, etc.) is fundamental for a productive application of the statistical functions of extraction, that would not by themselves offer guarantees to ensure the validity of the extracted data. Our aim is to develop not only tools for the analysis and synthesis of linguistic evidences, but also terminological data bases and specialized linguistic resources for textual analysis and named entities recognition.
Correct identification of terms and of all the semantic information in a text is essential for building a system of textual analysis, but also for classification and browsing of the text. Such a system is able to create relationships among semantically relevant information and also suggest synergies among private companies and public institutions, such it is required in SUBITO. The goal is to build a network of “knowledge” useful for an intelligent browsing, which is the real richness of the web services we offer to the project.
For this reason we developed a specific browsing system, text classification tools and semantic knowledge extraction systems.
The extracted features are mostly proper names, names of institutions, names of places and other relevant terms that characterize the specific domain.

6. Reference (Text) Corpus

In a first phase, we applied our attention to the creation of a reference (text) corpus of the biomedical domain. This training corpus was made up of a set of documents (in particular abstract of scientific articles) extracted from the PubMed website, where all texts are in English. All the resources and tools that constitute our background are in Italian, so it was necessary to adapt them in order to work in English. The same strategy will be adopted for the three other types of text documents considered: descriptions of projects, patents (EP, US e WO categories) and clinical trials (extracted by clinicalTrials.gov), in case the text size him will permit.
The creation of a specific reference corpus is a really important task and constitutes the basis of the whole process of creation of the specialized resources; the adaptation of the procedures to the project requirements, as well as the final editorial phase, are quite important since they can suggest new adaptations and improvements to the whole process.

7. Multi-word term extraction

After the creation of a specific reference corpus we extracted the relevant terms that will constitute the semantic information of the specific domain.

The creation of a biomedical ontology remains a valuable starting point for the extraction of knowledge and semantic associations by means of our statistical and linguistic tools. In order to meet the project requirements, we started the documents categorization using the tree MeSh[5] as knowledge base, like in PubMed. On the basis of MeSh tree we have then enriched the terminology by using our classification tools.
The extracted terms are not only those existing in online thesauri and dictionaries and belonging to different categories such as genes, proteins, drugs and molecules, etc. but are also those retrieved by semantic analysis procedures.
For example, through the automatic extraction of ontological trees:

- ✓ Acid (.. acid, etc.);
- ✓ Agent (.. immunosuppressant, etc.);
- ✓ Alcohol (methyl .. etc..);
- ✓ Rare diseases.

The extraction of terms (simple and compound words) linked to a domain terminology can be another example:

- ✓ immuno-suppressant agent, chromosome

It is also possible to extract the events:

- ✓ tumor growth, low blood sugar, cardiovascular collapse.

Once the reference corpus is built, the elaborate terminology can be extracted and used for the creation of a knowledge network.

8. Semantic filtering

From the same set of features, we extracted the terms for the creation of domain dictionaries, which, in our case, coincided with the main MeSh sub-tree, for example, starting from the category “Diseases [C]”, vocabularies have been created for the 23 subcategories: “Bacterial Infections and Mycoses [C01]”, “Virus Diseases [C02]”, etc.

Figure 1: example of MeSh tree structures

[5] The Medical Subject Headings (MeSH) is a huge vocabulary created by the National Library of Medicine (NLM) of the United States, with the goal of indexing scientific literature in the biomedical field. The 2008 version of MeSH contains a total of 24,767 subject headings, also known as descriptors. Because of these synonym lists, MeSH can also be viewed as a thesaurus.

Each terminological lexicon was created with statistical procedures that measure the relevance of a term to the domain, in order to create the semantic filters or “topics”.
The term "topic" identifies an area of interest chosen according to the project requirements. In fact the whole MeSh tree contains sub-trees that, after assessment, were deemed inadequate for the construction of a specific domain lexicon.
In general, the creation of a domain lexicon begins with the selection of pivot terms that have a high semantic value for the same domain, in this case specifically are the MeSh nodes. The lexicon also includes those terms having a higher co-occurrence value with the pivot terms, but that are not necessarily MeSh nodes.
Hence, the decision to acquire all nodes as basic terminology, but to use only some of these to create the semantic filter. In the light of the above considerations, we made a targeted decrease of nodes and sub-nodes, aimed at selecting the categories of greatest relevance to the areas of new technologies and research in the biomedical field.

9. Text bowsing system

The browsing system “DBT-Faccette” provides primitives to be integrated in the project website using the terminological basis identified and allows the automatic re-organization of content, based on the salient concepts.
This approach allows the user to dynamically discover the concepts semantically relevant for the domain, and to carry out search refinements through the interrelated concepts.
An alternative access to content is the search for topics, which is as important as a traditional browsing of textual content. This research modality provides the user with a selection of crucial documents, ordered by their relevance to the topic. In this way it is allowed to measure the ranking of an document with respect to the topic.

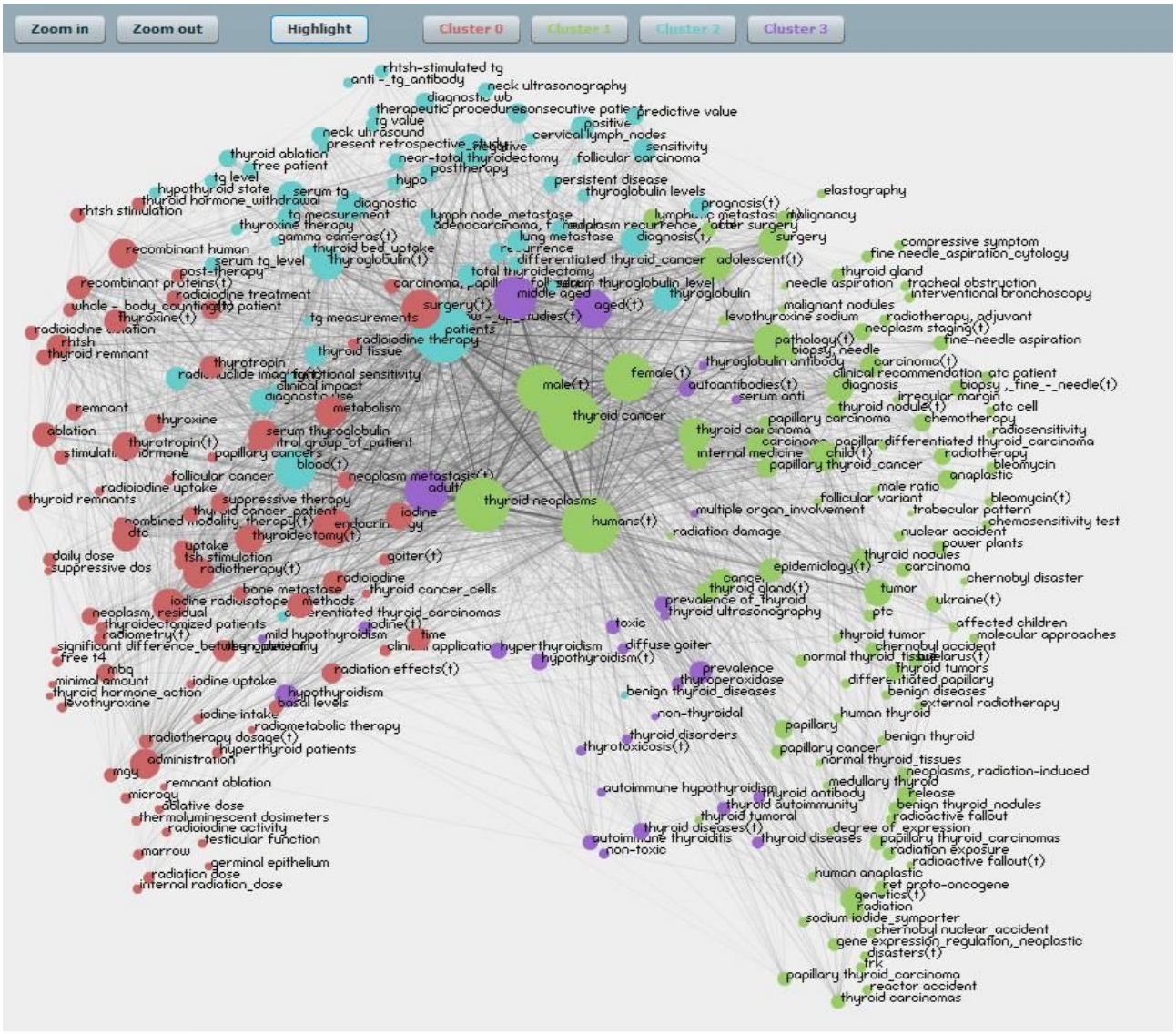

Figure 2: graph of "thyroid cancer" query

10. Conclusion

Through the semantic browsing tools, SUBITO can allow to researchers of the Tuscany region to benefit from a set of information that will facilitate the development of new synergies with consequent positive effects on employment, economic growth and citizens' welfare.

Furthermore, these initiatives will allow to an audience of European researchers to use such information through the pages of the portal CORDIS, that provides to the regions a space for the dissemination of research activities on its territory.

11. References

1. Harkema H., et al.: Information Extraction from Clinical Records. In S.J. Cox (ed.), Proceedings of the 4th UK e-Science All Hands Meeting, Nottingham. UK (2005)
2. Kors, J.A., et al.: Combination of Genetic Databases for Improving Identification of Genes and Proteins in Text. In: Proceedings of the BioLINK 2005. ACL (2005)

3. Harkema, H., et al.: A Large Scale Terminology Resource for Biomedical Text Processing. In: Proceedings of the BioLINK 2004, pp. 53–60. ACL (2001)
4. Quochi V., et al.: A Standard Lexical-Terminological Resource for the Bio Domain. In: Lecture Notes in Artificial Intelligence, vol. 5603 pp. 325 - 335. Human Language Technology - Challenges of the Information Society. Z. Vetulani and H. Uszkoreit (eds.). Springer Berlin / Heidelberg. (2009)
5. Picchi E., et al.: The "Micro semantics" for intelligent browsing. In: CHC 2011 - 4-th Intl. Congr. Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin (Istanbul, 22-25-11 2011). In: Proceedings of Congress, pp. 286. Valmar, Roma (2011)