MBlab: Molecular Biodiversity Laboratory

Corrado Loglisci, Annalisa Appice, Michelangelo Ceci,
Donato Malerba, and Floriana Esposito

Department of Computer Science, University of Bari "Aldo Moro"
{loglisci,appice,ceci,malerba,esposito}@di.uniba.it

Abstract. Technologies in available biomedical repositories do not yet provide adequate mechanisms to support the understanding and analysis of the stored content. In this project we investigate this problem under different perspectives. Our contribution is the design of computational solutions for the analysis of biomedical documents and images. These integrate sophisticated technologies and innovative approaches of Information Extraction, Data Mining and Machine Learning to perform descriptive tasks of knowledge discovery from biomedical repositories.

1 Introduction and Motivation

The exponential increase in publication rate of new papers in biomedicine makes difficult for researchers to keep up with research progresses without the help of computational techniques. Over 16 million references to biomedical papers are currently contained in the Medline database, the main on-line repository of biomedical research literature. Textual data as Medline papers are generally unstructured and the available technologies do not provide adequate mechanisms for helping humans in deeply analyze very large amount of content. The need to analyze this volume of unstructured data has prompted the use of information extraction and data mining tools to automatically extract key biological information. Several methods have been presented so far. The approaches in information extraction have been more and more specialized and refined as much as to permit to identify and recognize facts of interest from text by considering both surface and deep information, such as keywords and syntactic structures. Data mining techniques are mainly used, as a layer of techniques on top of those of information extraction, to perform predictive (e.g., text categorization) and descriptive (e.g., association discovery) analysis.

Moreover, the rapid expansion of biomedical image repositories raised problems that go beyond simple acquisition issues, and cause the need to organize and classify the contents in order to improve the effectiveness of the retrieval procedure. To this aim, the integration of machine learning and image processing techniques represents a suitable approach to face the task.

2 Scientific Challenges

In the project "MBlab - Molecular Biodiversity Laboratory" we have designed three computational solutions to support the activity of the researchers

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 132–135, 2011.

in biomedicine when documents are considered. Typically, the researchers need to recognize relevant information present in the documents (e.g., named entities as SNPs, gene names) and then formulate new hypothesis or infer new findings (e.g., identification of SNPs or gene names involved in particular pathologies). The proposed approaches perform three distinct descriptive data mining tasks and, in particular, permit to extract facts of interest from biomedical literature in large repositories and mine regularities based on statistical evidence from these facts of interest: regularity can denote a particularly well-established process which therefore can be biologically relevant.

As for the biomedical image analysis, during the past years a considerable effort was spent in the definition of automatic tools for low level features extraction. However, they fail in dealing with one of the main characteristics of biomedical repositories represented by the existence of relations among the objects contained in the collection. Hence, the need of deriving relational high-level semantic annotations able to describe the objects in the collection in order to improve the indexing process.

3 Contribution by the Research Group

We explored the potentialities of the technologies and approaches of information extraction and data mining for three descriptive tasks of knowledge discovery in biomedical literature described in the following. Furthermore, we investigated the applicability of machine learning and image processing techniques for the extraction of meaningful annotations for effective indexing of biomedical images.

Discovering Frequent Syntactic Structures. We study the problem of how to mine an unstructured biomedical text corpus in order to identify any syntactic structures of named entities which frequently appear in the documents retrieved from biomedical repositories on a specific topic. We propose a knowledge discovery framework which first annotates the named entities (Semantic Text Processor in Fig. 1) in the retrieved corpus and integrates these entities in syntactic structures (Syntactic Text Processor in Fig. 1). Then, it uses a multi-relational data mining approach for frequent pattern discovery to identify frequent syntactic structures in the form of subject-object(s)-verb (Frequent Pattern Discovery in Fig. 1). Discovered frequent patterns play the role of informative syntactic knowledge shared by the papers under study and provide an indication of the existence of interesting verbal based dependencies among named entities.

Mining Temporal Links. Mining biomedical literature when considering the temporal dimension of the papers is another problem explored in the MBlab project. In particular, we focus on the identification of interesting and hidden relations (links) between seemingly unconnected entities when entities are cited or reported in scientific papers published in different periods of time. This provides us a means to unearth linkages which have not been discovered when observing the literature as static but which may have developed over time, when

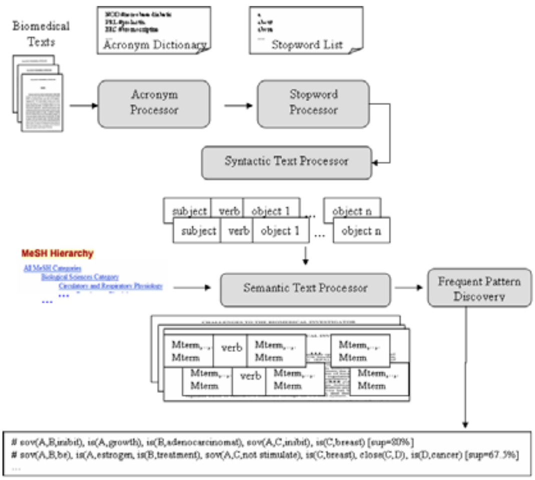
Biomedical Texts
Acronym Dictionary
Stopword List
Acronym Processor
Stopword Processor
Syntactic Text Processor
subject
verb
object 1
...
object n
MeSH Hierarchy
All MeSH Categories
Biological Sciences Category
Circulatory and Respiratory Physiology
...
Semantic Text Processor
Frequent Pattern Discovery
Mterm, ..., Mterm
verb
# sov(A,B,inibit), is(A,growth), is(B,adenocarcinoma), sov(A,C,inibit), is(C,breast) [sup=80%]
# sov(A,B,be), is(A,estrogen, is(B,treatment), sov(A,C,not stimulate), is(C,breast), close(C,D), is(D,cancer) [sup=67.5%]
...

Fig. 1. Discovering Frequent Syntactic Structures from Biomedical Literature

considering its dynamic nature. The adopted computational solution first partitions the corpus based on distinct and consecutive time-intervals, then identifies the biomedical named entities present in each partition (Named Entity Recognition, Feature Selection in Fig. 2). Multiple-level association rules are mined from each time-interval (Association Rules Mining in Fig. 2), and finally, a process of chaining of association rules is performed through all sets of association rules, namely through the time-intervals, in order to link over time two input entities (Link Discovery in Fig. 2). Therefore, temporal links are discovered as chains of association rules and denote hidden relations between the entities in the rules.

Pattern Discovery for Semantic Role Labeling. A bio-molecular event is a process that involves and transforms molecular entities. In the literature, these are reported and described as predicate-argument structures, where each argument corresponds to a single entity which plays a particular role in the described event. Our interest in the bio-molecular events is that of defining an approach able to label the roles of the entities present in the documents, or, in other words, to identify which entities are associated to the roles of bio-molecular events. The determination of the entities which more frequently play a specific role in a particular event may provide indications on the investigation to conduct on a specific biological process. The computational solution first identifies predicate-argument structures (Pas Extractor, Frame Filling in Fig. 2) by exploiting a knowledge base on the semantic structure of the events (Event FRAMEs in Fig. 2). Then, the arguments of these structures are recognized as named entities (Named Entity Recognition in Fig. 2). This permits us to preserve the relational information of each event. Finally, a relational frequent pattern mining approach is used to discover the associations roles-entities and the relationships of these associations with the events (Frequent Pattern Discovery in Fig. 2).

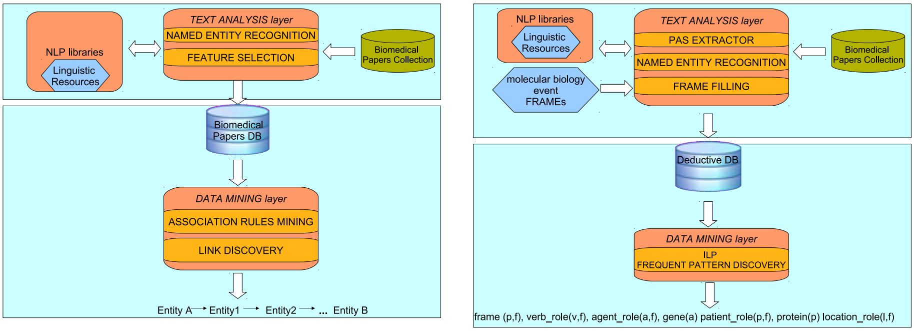

Fig. 2. Mining Temporal Links from Biomedical Literature (left) - Pattern Discovery for Semantic Role Labeling in Biomedical Literature (right)

Biomedical Image Annotation. In [3] we propose a suite of image processing and machine learning approaches to annotate biomedical microscope images with qualitative information rather then quantitative ones coming from low level features. In details, a set of morpho-structural features such as dimension, granularity and polarity of the cytoplasm of a cell are extracted. Generally, these morpho-structural features are manually analysed by the clinicians that use such information to assess the goodness of a cell for a specific therapeutic treatment. On this baseline process, aimed at extracting qualitative tags from an image, a first-order relational learning framework is designed to combine the qualitative information gathered from images with textual content about patients clinical data in order to learn relational models able to characterize the trend and the success of a therapeutic plan in indexing and retrieval processes.

Acknowledgment. This work is in partial fulfillment of the research objectives of the project MBlab DM19410 “Molecular Biodiversity Laboratory”.

References

1. Appice, A., Ceci, M., Loglisci, C.: Discovering Informative Syntactic Relationships between Named Entities in Biomedical Literature. In: Proc. of International Conference on Advances in Databases, Knowledge, and Data Applications, DBKDA 2010, pp. 120–125 (2010)
2. Loglisci, C., Ceci, M.: Discovering Temporal Bisociations for Linking Concepts over Time. In: Proceedings of European Conference on Machine Learning and Principle and Practices of Knowledge Discovery in Databases, Athens, Greece (2011)
3. Basile, T.M.A., Esposito, F., Caponetti, L.: A Multi-relational Learning Approach for Knowledge Extraction in in Vitro Fertilization Domain. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Chung, R., Hammoud, R., Hussain, M., Kar-Han, T., Crawfis, R., Thalmann, D., Kao, D., Avila, L. (eds.) ISVC 2010. LNCS, vol. 6453, pp. 571–581. Springer, Heidelberg (2010)