A data mining approach to retrieve mitochondrial variability data associated to clinical phenotypes

Berardi M. ,Attimonelli M.(1), Cascione I.(1), Santamaria M.(1), Accetturo M.(1), Lascaro D.(1), Berardi M.(2), Ceci M.(2), Loglisci C.(2), Malerba D.(2)

(1) Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Via E.Orabona 4, 70126 Bari – Italy
(2) Dipartimento di Informatica, Università degli Studi di Bari, Via E.Orabona 4, 70126 Bari – Italy

Motivation

The maintenance of biological databases is at present a problem of great interest since the progress made in many experimental procedures has led to an ever increasing amount of data. These data need to be structured and stored in databases and made accessible to the biological community in user-friendly ways. Although both the interest and the need of accessing biological databases are high, the mechanisms to fund their maintenance are unclear. Funding agencies cannot support data annotation in terms of labour costs and hence the development of new tools based on "data miming" technologies could greatly contribute to keep biological databases updated. Here we present a new approach aimed to contribute to the annotation in the HmtDB resource (http://www.hmdb.uniba.it/) of variability data associated to clinical phenotypes [1]. These data are prevalently available in literature where they are reported in a completely free style. Thus, we suggest the construction of a knowledge base derived from browsing papers on web and to be used in the retrieval phase. Nevertheless, problems in extracting data from literature come not only from the heterogeneity of presentation styles but mainly from the unstructured format (i.e. the natural language) in which they are represented. In this scenario, the goal is to feed a knowledge base by identifying occurrences of specific biological entities and their features as well as the particular method and experimental setting of the scientific study adopted in the publication. In this work, we describe some solutions to the problem of structuring information contained in scientific literature in digital (i.e., pdf) or paper format.

Methods

We present a document analysis framework to support biologists in the automatic extraction of relevant information from texts. Initially, the document layout structure is extracted from document images obtained by either scanning paper documents or converting pdf documents. The layout structure is then mapped into a semantic structure, which correspond to sections whose content is of interest for the application. While layout structure extraction is based on geometric features, the semantic structure extraction is based on textual content of layout components. The automation of both extraction processes is based on the application of machine learning techniques. In particular, the extraction of semantic structures demands for learning the distribution of relevant information among the semantic structures (which sections are the most informative for each data entity), as well as learning rules that are able to localize instances of predefined entities of interest (mutation type and position, data associated to a DNA sample such as tissue source, subject age, phenotype and geographic origin, etc.) in free text. Training examples for machine learning systems are generated by asking domain experts (biologists) to manually label the layout components of some training documents (HmtDB resources) and to then to annotate relevant textual information for the task at hand.

Results

The project is in progress. Initial results have been obtained by applying the proposed approach to a collection of scientific papers reporting studies on mitochondrial diseases associated to mtDNA mutations. Results on semantic structure extraction show that the framework is quite precise in classifying Abstract, Methods&Material and Results sections [2]. Lower accuracy on other components of the semantic structures, such as Introduction and Discussion, is less problematic, since those components are not considered very informative for subsequent annotation purposes. In addition, a pilot study has been conducted on information extraction from text. Results show that some annotations, as the mutation type and position, are easier to identify than others since they are characterized by a more homogeneous distribution in texts. Future work will focus on learning issues related to this last step.

Contact email: berardi@di.uniba.it

Supplementary Information: References: 1. M. Attimonelli, M. Accetturo, M. Santamaria, D. Lascaro, G. Scioscia, G. Pappadà, M. Tommaseo-Ponzetta, A. Torroni: "HMTDB, a Human Mitochondrial genomic resource based on Variability studies supporting population genetics and biomedical research". Submitted for publication. 2. M. Berardi, M. Lapi, and D. Malerba: "An integrated approach for automatic semantic structure extraction in document images". In: Proc. of the 6th IAPR International Workshop on Document Analysis Systems (DAS 2004), Florence, Italy, September 8-10, 2004, in A. Dengel & S. Marinai (Eds.).