MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-412

A Unified Medical Language System (UMLS) Based System for Literature-Based Discovery in Medicine

Matteo Gabetta[a], Cristiana Larizza[a], Riccardo Bellazzi[a]

[a]*Department of Industrial and Information Engineering, University of Pavia, Italy*

Abstract

Literature-Based Discovery (LBD) is a technique that can be used in translational research to connect the very sparse and huge information available in scientific publications in order to extract new knowledge. This paper presents an LBD system based on the open discovery paradigm exploiting NLP techniques and UMLS medical concepts mapping, to provide a set of tools useful to discover unknown relationships. The system has been evaluated on the problem of discovering new candidate genes potentially related to dilated cardiomyopathies (DCM), and can be used in any medical context to connect different type of concepts. The validation of the system involves reproducing the discovery of genes currently associated to DCM. Validation showed that the system is able to discover many gene-disease associations by using the literature available before their first publication in a scientific article.

Keywords:

Literature-Based Discovery, UMLS, NLP.

Introduction

Translational research projects aim to combine the science of -omics, structural and functional studies, with clinical investigation results to translate basic knowledge of diseases into routine clinical practice. Biomedical informatics can fruitfully support this research by implementing information technology solutions to support the researcher carrying out this task, thus improving the diagnostic and therapeutic process. A task in translational research is combining the available knowledge on a domain with clinical data to try obtaining new knowledge or test new hypotheses. However, nowadays the available scientific knowledge is very huge, increases very rapidly, and is usually located on very restricted domains. These attributes make it difficult to consider the complete literature on a domain or, more ambitiously, to link efficiently the existing knowledge related to different domains and finding new relations. Literature-Based Discovery (LBD) helps the researcher discover unknown relationships among scientific knowledge and applies text-mining techniques applied to the scientific literature [1]. The goal is to generate new hypotheses representing potential new scientific discoveries. In this paper, an LBD system, based on the UMLS, searches for conditions potentially involved in disease aetiology. The system can be used for every medical subject, but has been validated in the context of dilated cardiomyopathies (DCM) where the problem is to find gene mutations causing the disease. The paper is structured as follows: after a short description of the LBD technique and an overview of the available systems, the architecture of the implemented system and the LBD workflow is described. Finally, the system validation process and the results are discussed.

Materials and Methods

Literature-Based Discovery

LBD is a technique introduced by Swanson [1] that automatically searches a large set of documents and reveals the connections that can be inferred between relevant concepts, but not explicitly reported in the literature. The knowledge discovery process results from the combination of existing knowledge and observations in a way capable of obtaining evidence of new hypotheses. The LBD process is based on two distinct ideas: the *concepts* relevant to the research domain and the available *literature*, that is the set of documents related to the domain, which potentially refer to these concepts. Usually the documents are public scientific papers, and the concepts are medical terms mentioned in the documents.

Definition: Two literatures L_A and L_C referring to the set of concepts *A* and *C*, respectively, are disjointed if there is no overlapping between the two sets *A* and *C*.

The theory for new knowledge discovery introduced by Swanson, called *ABC model*, is illustrated in Figure 1. If *A* and *B* are related, and *B* and *C* are related, it follows that *A* and *C*, even if $L_A \cap L_C = \emptyset$, might be indirectly related through *B*. The process can be carried out in two different ways, called *Open Discovery* and *Closed Discovery*: the first one is used to discover new connections and the second one is used to confirm potential new hypotheses.

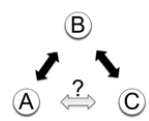

Figure 1- The Swanson ABC model

The search starts at *A*, for instance a disease, and results in *C*, possibly a drug. The intermediate *B* steps may represent, for instance, (patho)physiological mechanisms.

In particular, the Open discovery process performs three sequential actions: 1) given a literature L_A related to a concept *A* (e.g. a disease), search in L_A the set of concepts *B* related to *A*; 2) creation of the literature L_B comprising the documents related to a subset of *B* concepts, properly filtered to restrict the analysis only to the very relevant ones (e.g. disease effects); 3) extraction from L_B of the interesting concepts *C*, excluding the ones already known as related to *A*. Concepts *C* (e.g. substances used to treat such effects) are potentially related to *A* through *B*, therefore these relations can represent new knowledge. The Closed Discovery process searches in the literatures L_A and L_C for the concepts *B* related both to the concepts *A* and *C*. The validation of a new hypothesis consists

in finding among the concepts *B*, those justifying the relation between *A* and *C*.

Overview of LBD systems

After the Swanson work, many systems for automating the LBD process have been proposed. They can be distinguished on the basis of the discovery paradigm adopted. Arrowsmith [2] is a tool for LBD based on the Closed Discovery model. It identifies the common words or phrases (B-terms) in the titles of two disjointed literatures. These terms are then ranked on an estimated relevance probability to filter out the not interesting concepts. Although Arrowsmith has been the first effort to automate the discovery process, the program has many steps that require a lot of manual intervention. A recent version of the tool exploits MeSH terms to rank B-terms also on the basis of the domain context of the articles.

The system proposed by Gordon & Lindsay [3] analyses complete Medline records to compute several lexical statistics, such as word frequency counts and record counts. Such statistics are used to rank the documents and assess their relevance in the discovery of hidden connections. Their system allows to manually filter out less relevant concepts and manage synonyms and generalizations.

Weeber et al. have developed a system called DAD (Disease-Adverse Reaction Drug-Drug) [4], based on the Open Discovery paradigm, that can extract from the literature interesting concepts –instead of simple words– mapped onto the Unified Medical Language System (UMLS) [5]. The filtering of *B* concepts can be based on the semantic types defined in UMLS.

One of the most interesting systems is the one developed by Pratt and Yetisgen, called LitLinker [6]. LitLinker couples the Open Discovery model to a data-mining algorithm to automatically extract an ordered list of intermediate concepts *B* to be filtered out on the basis of different criteria (too much general, too much cited, not belonging to a specified semantic type). Finally, the system clusters the similar concepts and, using the Apriori algorithm [7], finds association rules between the starting concept and the clusters (*B* concepts). After proper filtering of the rules based on their support, the same process is applied to the intermediate concepts, as starting concepts, to find out the final concepts *C*.

One system that is more similar to the one presented in this paper is Bitola [8], developed by Hristovski. In Bitola each record is represented by the MeSH terms indexing the article and the gene names and symbols found in the title and in the abstract. The system uses three different reference databases to map gene names: LocusLink [9], OMIM [10] and HUGO [11]. The association between two concepts *A* and *B* is scored on the basis of its support (number of articles containing both *A* and *B*) and confidence (percentage of articles containing both *A* and *B* over the total number of articles containing *A*).

Considering the efforts above in the field of LBD, the most critical aspects of the discovery process are: a) the choice of the knowledge sources (article titles, abstracts, MeSH terms, etc.) that, in some cases, can be affected by human errors (e.g., in manually indexing articles) or cannot be publicly available or can generate high volumes of spurious associations; b) the concepts representation problem (in many contexts it could require specific procedures for acronyms disambiguation, synonyms management); and c) the design of a robust validation process (for new knowledge discoveries, a lot of time could be needed to confirm their validity).

The LBD system presented in this paper proposes a system based on the open discovery paradigm that allows good flexibility in setting up the different steps toward the knowledge discovery process, depending on the specific domain. Also it can manage concept generalization/specialization through medical concept mapping based on UMLS. The concept ranking is based on the support and confidence of the relations found in the literature.

This LBD system proposes a discovery algorithm that combines a set of new solutions with features already adopted by LBD systems presented above; in particular, we chose to implement the features that proved to be effective for the specific task of gene-disease associations. The final goal of our approach is to harmonize these solutions and achieve an open discovery system able to be effective without either disregarding the available information sources or loosing in generality.

Similarly to the system presented in Weeber [4], our system considers the whole abstract of the article and extracts the UMLS concepts contained. Like all the evaluated systems, also ours considers as linked two co-cited concepts. Like Weeber [4], Pratt [6] and Hristovski [8], our system also allows filtering the concepts on the basis of their UMLS Semantic Types. Differently from other systems we have chosen to implement a persistence layer in order to save the results of the concept extraction process from the available literature, that is not limited to a set of pre-calculated MeSH-based association rules as is in Hristovski [8]. Differently from Smalheiser [2], Gordon [3] and Weeber [4] where the results of the discovery process are evaluated in terms of frequency of paths joining *A* and *C*, our system, like Pratt [6] and Hristovski [8], adopts an approach based on the association rule theory. It combines the confidence scores of the links $A \rightarrow B$ and $B \rightarrow C$ to produce a score measuring the strength of the indirect link $A \rightarrow C$ and, moreover, it allows the user to evaluate the confidence scores of the $A \rightarrow B$ and $B \rightarrow C$ steps.

System Workflow

The LBD system we have developed has been created entirely with Java-based technologies integrated with freely accessible datasets and web services. The concepts, core of data representation inside the LBD system, are codified using UMLS and their persistence is entrusted to a MySQL DBMS called Literature Mining Database (LM-DB). The literature access is performed with the Entrez Programming Utilities (EUtils), a set of server-side programs providing an interface into the Entrez query and database system at the NCBI [12]; for the purpose of this work we have used the Java APIs provided to search and fetch the PubMed database via web service.

Once downloaded, the abstract of PubMed articles pass through a text mining process aimed at extracting the contained UMLS concepts; this part of the system has been implemented on top of the General Architecture for Text Engineering (GATE) [13], which libraries have been used to define a standard text mining pipeline and to develop an additional plug-in able to extract UMLS concepts from free text and to store them in the LM-DB. The choice of deploying an additional persistence layer in order to store the concepts extracted from literature is due to the high computational costs of this operation. These costs would have made the LBD system practically unusable if the literature mining process was made in a hurry for each execution.

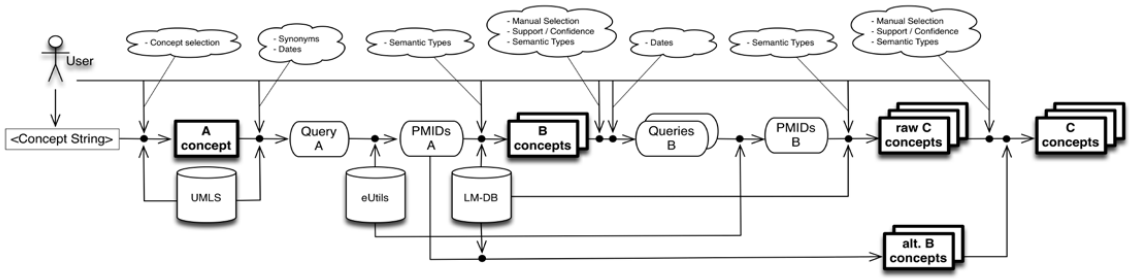

Figure 2 - Complete workflow of the discovery process implemented by the LBD system.

Finally, the Graphical User Interface of our LBD system has been created with the Google Web Toolkit in order to easily expose the Java-based system on the Internet.

The overall workflow of the discovery process underlying the LBD system is represented in Figure 2. The approach is based on the open discovery paradigm that, starting from a source concept, tries to discover related knowledge (i.e., other concepts) that have never been directly associated with the starting one (i.e., co-cited inside an abstract). but prove to have a strong relationship with it in terms of intermediate concepts directly associated with both.

The first step the user performs is to choose the starting concept to query (*A concept*); in practice, the user provides the name of the concept and the system tries to match it in the UMLS. If the matching succeeds, the system asks the user to choose, if necessary, one or more synonyms of the A concept that will be included in the PubMed query. Furthermore, the system asks to choose a temporal interval for the publication dates in order to reduce the queried literature; then the resulting query is built up and sent to PubMed. Once the EUtils web service returns the list of articles matching the query (in term of their unique identification number - PMID), the user is asked to identify the Semantic Types of interest which the concepts extracted from these literature will belong to; afterwards, for each article, the system queries the LM-DB to generate the list of intermediate concepts (*B concepts*) belonging to the Semantic Types selected by the user that are co-cited in literature with the starting concept, within the time span selected.

For every B_i concept shown, the system calculates support and confidence of the $A \rightarrow B_i$ relationship; these are standard measures in the association rule theory that have been adopted, for instance in Pratt [6] and Hristovski [8]. Support is defined as the number of articles where A and B_i are co-cited:

$$Support(A, B_i) = \|L_A \cap L_{B_i}\|,$$

where L_A and L_{B_i} are the literatures that contain A and a B_i concept, respectively; confidence is defined as the relative number of articles where A and B_i are co-cited on the whole literature relative to A:

$$Confidence(A, B_i) = \frac{\|L_A \cap L_{B_i}\|}{\|L_A\|}.$$

Now, the user can manually select one or more B concepts, set up a threshold for support and confidence or apply a new filter on the Semantic Types in order to reduce the set of B concepts that will be used in the final step of the discovery process.

For each intermediate concept selected, the system performs the same steps that have been described before the generation of the set of B concepts from the starting concept A. It is important to notice that the user is once again asked to define the filtering criteria (publication time interval and Semantic Types) that will be applied in the generation process of the final concepts (*C concepts*) and that these criteria may be different from the ones used in the previous step; this operation may lead to an ambiguity in the discovery process that will be solved in the very final step performed by the system.

Therefore, each B concept generates a single query (without the manual addition of the UMLS synonyms) that is sent to PubMed to obtain a list of PMIDs. All these lists are then joined together and used to query the LM-DB in order to achieve the set of UMLS concepts that are cited in the whole literature relative to the intermediate concepts (given the filtering criteria defined by the user).

Similarly to what happened with the B concepts, also for these concepts the user can apply some filters such as manual selection of single concepts or Semantic Types and set up of a threshold for support and confidence. Since for the $B \rightarrow C$ step there is not a single concept to start from (B is a set of intermediate concepts), the definition of support and confidence has to be adapted to the case:

$$Support(\hat{B}, C_k) = \left\| \bigcup_i (L_{B_i} \cap L_{C_k}) \right\|,$$

$$Confidence(\hat{B}, C_k) = \frac{\left\| \bigcup_i (L_{B_i} \cap L_{C_k}) \right\|}{\|L_{C_k}\|},$$

where $\hat{B}$ is the whole set of B concepts, B_i and C_k are respectively single B and C concepts and $\bigcup_i(\cdot)$ is the union operator.

After this filtering process the system returns to the user a set of C concepts called *raw C concepts*; in fact this set could still contain some concepts that are directly associated with A and therefore they don't represent new associations. In order to delete these concepts, the system can use the list of B concepts if and only if the filtering criteria used in the $B \rightarrow C$ step (publication time interval and Semantic Types) are the same used for the $A \rightarrow B$ step. Therefore the system performs a further step: it searches for concepts directly cited with A using the filters of the $B \rightarrow C$ step (*alternative B concepts*) and removes them from the set of raw C concepts. Let us explain this aspect with an example: assume our A concept to be a disease and to search for intermediate concepts (B) that still represent diseases (*Semantic Type T047 - Disease or Syndrome*). Now, for the $B \rightarrow C$ step, let assume we want the LBD system to search for genes (*Semantic Type T028 - Gene or Genome*). The raw C concepts set will then be composed of genes and we have no assurance that these genes have been never co-cited with A; the last step removes from C the genes co-cited with A.

The final set of C concepts, that represent potential new knowledge, is then shown to the user and each concept is associated with its support, confidence and an additional heuristic score that takes into account the quality of the intermediate concepts that link it with A. This score is defined as:

$$Score(C_k) = \sum_{i=1}^{N} (Support(A, B_i) * Support(B_i, C_k)),$$

where N is the number of intermediate concepts that link A whit the single final concept C_k. The structure of this heuristic score function guarantees that C concepts that are linked with A by a larger number of links and, in particular, by links with a high support, will achieve a better ranking.

Results

As previously underlined the validation process of an LBD system is complex and can require much time, so we chose to validate our system on already known relations. In particular, we applied our approach in the context of DCM where the goal is to discover new gene-disease associations. At this time, literature reports a list of several genes mutations currently recognised as responsible for DCM [14], so our validation process consisted in trying to discover for each gene its association to DCM from the scientific literature before its first publication. The test has been done on a subset of 16 genes and for each of them we followed several steps:

Publication date filter - To limit the analysis to the literature available before the gene-disease association it is necessary to correctly identify its first publication date. To this aim we build a query that refers not only to the complete gene name, but also to the most diffused synonyms of the gene.

A→ B step - *Given the literature on DCM (concept A) available before the publication of the first gene-DCM association, the goal of the first step of the discovery process is to identify the concepts B, related to DCM. In order to reduce the number of documents considered, the literature relative to DCM is obtained exclusively from those articles that are indexed with the MeSH term "Cardiovascular Disease" or those terms that are hierarchically dependent from it; this operation allows considering about 1.100.000 articles instead of the entire PubMed corpus. From the overall set of B concepts cited in this literature, we remove those which are over-cited in order to exclude the concepts that are not enough specific and which informative content is therefore poor [15]; this operation consisted in removing the concepts that are cited more than a specific number of times; currently, this threshold has been empirically set to 100.000, in order to limit the number of B concepts and keep the system usable, anyway further analyses are needed to identify an optimal value. Afterwards we take into account exclusively those B concepts that belong to the list of Semantic Types reported in Table 1; we have defined this list with the scope of including all the concept types that could be most likely the intermediate concept between DCM and a gene and, on the other side, to exclude those concepts that, despite being part of UMLS, are too distant from this scope.*

Table 1 – Semantic types of B concepts (TUI is the Type Unique Identifier)

Semantic Type (TUI)	
• Gene or Genome (T028)	• Neoplastic Process (T191)
• Nucleic Acid, Nucleoside, or Nucleotide (T114)	• Anatomical Abnormality (T190)
• Congenital Abnormality (T019)	• Biologic Function (T038)
• Amino Acid, Peptide, or Protein (T116)	• Cell or Molecular Dysfunction (T049)
• Mental or Behavioral Dysfunction (T048)	• Laboratory or Test Result (T034)
• Disease or Syndrome (T047)	• Sign or Symptom (T184)
• Finding (T033)	• Acquired Abnormality (T020)
• Pathologic Function (T046)	• Amino Acid Sequence (T087)
• Molecular Sequence (T085)	• Clinical Drug (T200)
• Cell Function (T043)	• Carbohydrate Sequence (T088)
	• Nucleotide Sequence (T086)

Furthermore support and confidence of each A→B association have been evaluated; since we had not at our disposal any reference value for this indexes, we chose to set as a threshold, for both support and confidence, the average value of these indexes and to exclude from the next phase those concepts that were under these thresholds for at least one of the scores.

B→ C step - *Starting from the set of selected B concepts, keeping the same filter on publication dates used in the A→B step, we identify, from the set of 1.100.000 articles used for the A→B step, the whole literature related to this set and then the system extracts C concepts; in order to identify genes, this time only concepts belonging to the "Gene or Genome" Semantic Type have been used. At the end of these steps the system returns the list of the genes that, despite never being co-cited with DCM in the considered time span, could be connected with it. Each element of this list is characterized by the three described indexes: support, confidence relative to the B→C association and the heuristic score relative to the complete A→C association. In particular, for the sake of evaluating the potential new knowledge, we sorted the C concepts on the base of their heuristic score. The results are shown in Tables 2 and 3.*

Table 2 – Time spans valid for the discovery and number of papers and concepts found

Gene	*First date*	*First date w/ DMC*	*B concepts*	*# Papers*
TNNT2	1994 May	2000 Jan	Not Found	5
TTN	1975 Jan	1994 Oct	64	546
MYBPC3	1993 Feb	1997 Mar	Not Found	17
ACTC	1977 Feb	1998 May	98	1313
TPM1	1974 Jan	2000 Jan	Not Found	51
MYH7	1989 Feb	2000 Jan	Not Found	35
ABCC9	2001 Apr	2004 Apr	Not Found	9
CLP	1991 Sep	1997 Feb	Not Found	11
DES	1976 Dec	1990 Jan	82	943
DMD	1978 May	1990 Feb	35	290
DSP	1982 Jan	2000 Oct	189	313
LDB3	1993 Jan	2003 Dec	Not Found	14
LMNA	1983 Jan	1999 Dec	166	214
MVCL	1985 Jan	1997 Jan	Not Found	30
PLN	1975 Jan	1990 May	45	203
SGCD	1999 Aug	1999 Aug	Not Available	2

Conclusion

Recent trends in translational research go towards multidisciplinary approaches where bioinformatics play a relevant role in providing methodologies to support the investigation. LBD systems provide the researchers with a set of tools useful for analysing the scientific literature and extracting potential new knowledge.

Table 3 - Summary results of the system validation for the genes the system is able to discover as associated to DCM. The scores ("Rank Sup" and "Rank Score") are obtained by comparing the relative measure with the ones of all the C concepts found in the same run of the system.

Gene	*Score*	*Support*	*Rank Sup*	*Rank Score*
TTN	26832	92	68/542	41/542
ACTC	203577	1025	7/662	6/662
DES	21598	150	11/349	8/349
DMD	15268	300	2/349	21/349
DSP	256598	1115	5/887	8/887
LMNA	252739	752	9/822	5/822
PLN	7906	47	69/380	75/380

The system presented in this article implements an LBD model based on the Open Discovery approach used in the context of DCM to discover new genes potentially related to the disease. The results obtained in the validation process confirmed that the algorithm implemented, resorting to association rules and several (semantic and statistical) filters, is effective both in selecting the most relevant concepts and in removing less interesting ones (very critical aspects that could strongly affect the research results). The validation of the system demonstrates its efficacy, as it is able to replicate many known connections between genes and DCM. Moreover, the results show that the heuristic function implemented (score) is a valid measure of the concept relevance, better than other types (e.g. support), since it permits the user to verify which associations between the starting (A) and final concepts (C) are stronger on the basis of the intermediate concepts (B). In particular it is clear that, when it is possible to evaluate the discovery results (Table 3), the genes associated with DCM achieve relatively high scores (in particular "Rank Score").
The validation results show also that the system cannot discover new connections if the time span between the first appearance of the concept in literature and the discovery is too short and the number of articles describing the concept is small. Such shortcoming is probably due to the reduced set of articles currently available in LDB that could ignore some articles relevant for the discovery. Indeed, the definition of a gold standard to validate LBD systems should be advantageous to be able to easily compare and evaluate different LBD systems. A possible improvement of the system is embedding more advanced text mining techniques capable of extracting from literature not only the co-occurrence of two concepts, but also specific types of relations (e.g., cause-effect and negative relations).

Acknowledgments

This work was supported by the INHERITANCE project, funded by the European Commission.

References

[1] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine 1986: 30(1): 7-18.

[2] Smalheiser NR and Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed 1998: 57(3): 149-53.

[3] Gordon MD and Lindsay RK. Toward discovery support systems: A replication, reexamination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science 1996: 47(2): 116-128.

[4] Weeber M, Klein H, de Jong-van den Berg LTW and Vos R. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology 2001: 52(7): 548-557.

[5] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004: 32(Database issue): D267-70.

[6] Pratt W and Yetisgen-Yildiz M. LitLinker: Capturing Connections across the Biomedical Literature. Proceedings of the International Conference on Knowledge 2003.

[7] Agrawal R, Imielinski T and Swami A. Mining associations between sets of items in massive databases. Proceedings of the ACM-SIGMOD 1993.

[8] Hristovski D, Stare J, Peterlin B and Dzeroski S. Supporting discovery in medicine by association rule mining in MEDLINE and UMLS. Medinformation 2001: 10(2): 1344-1348.

[9] Pruit KD and Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001: 29: 137–140.

[10] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {December 9th 2012}. World Wide Web URL: http://omim.org/

[11] Seal RL, Gordon SM, Lush MJ, Wright MW and Bruford EA. genenames.org: the HGNC resources in 2011. Nucleic Acids Res. 2001: 39(Database issue): D519-9.

[12] Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25501/

[13] Cunningham H, Maynard D, Bontcheva K, Tablan V, Ian Roberts I, Gorrell G, Funk A, Roberts A, Damljanovic D, Thomas Heitz T, Greenwood MA, Saggion H, Petrak J, Li Y and Peters W. Developing Language Processing Components with GATE Version 7, The University of Sheffield, Department of Computer Science, 8 February 2012.

[14] Elliott P, Andersson B, Arbustini E, Bilinska Z, Cecchi F, Charron P, Dubourg O, Kühl U, Maisch B, McKenna WJ, Monserrat L, Pankuweit S, Rapezzi C, Seferovic P, Tavazzi L, Keren A. Classification of the cardiomyopathies: a position statement from the European Society Of Cardiology Working Group on Myocardial and Pericardial Diseases. Eur Heart J. 2008: 29: 270-6.

[15] Spärck Jones K. A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation 1972: 28:11-21.

Address for correspondence

Matteo Gabetta, Via Ferrata 3, 27100 Pavia, Italy, tel +39 0382985981, fax +39 0382985060, email matteo.gabetta@unipv.it