Abstract. The ever-increasing number of publicly available biomedical articles calls for automatic information extraction from digitized publications. We have implemented a pipeline which, by exploiting text mining and semantic technologies, helps researchers easily access semantic content of thousands of abstracts and full text articles. The text mining component analyzes the articles content and extracts relations between a wide variety of concepts, extending the scope from proteins, chemicals and pathologies to biological processes and molecular functions. Moreover, the relations are extracted along with the context which specifies localization of the detected events, preconditions, temporal and logic order, mutual dependency and/or exclusion. Extracted knowledge is stored in a knowledge base publicly available for both, human and machine access, via web interface and SPARQL endpoint. To address the data accessibility, reusability and interoperability, all the extracted relations are standardized using unique resource identifiers (URIs) and a custom ontology based on Genia ontology.

1 Introduction

Information extraction from biomedical literature is becoming a common practice due to the huge amounts of available textual data, and technological maturity which allows to gain insight into scientific content. Text analysis evolved from spotting relevant concepts in the text [18, 19] to co-occurrence statistics [20–22] and, finally, extraction of complex events which seek to reveal cause-effect relation between various entities involved in the biomedical processes [23–25]. Some approaches use textual data as the only source for the analysis [26, 27] while some other combine it with experimental data available from dedicated databases [28, 29]. It turned out though that all the wealth of the results obtained from heterogeneous sources was hardly beneficial for anyone beyond the producer due to lack of a common language: each system came up with its own nomenclature if any [30]. It is where semantic technologies come into play to become an integral part of the information extraction process. To increase data
reusability and interoperability several solutions have been proposed. PubAnnotation [31], micro [32] - and nanopublications [33] are important examples of how to represent extracted knowledge in a standardized format as to be accessible and shared between machines and humans.

Knowledge discovery systems and platforms vary in scope. Many of them are focused on specific sub-domains. For example, EVEX [23] targets directed interactions between proteins; DisGeNET [34] explores genetic mechanisms of diseases, while LimTox [35] searches for toxicity associations of compounds, drugs and genes with the special interest in liver. Other systems adopt less centered strategies, trying to cover more aspects involved in biomedical processes. One such system is PolySearch [36], which searches associations between more than twenty entity types, exploiting data from medical literature, Wikipedia articles and 14 databases, among which are UniProt, DrugBank and HMDB. While leading in scope, Polysearch does not specify association types or directionality, leaving these important pieces of knowledge to be completed by the user. The BioKB platform we introduce here aims to discover cause-effect relations between multiple entity types and deliver standardized representation of knowledge relevant for various research scenarios, such as finding entities which demonstrate specific behaviour with respect to a pathological process; localization of biomolecular processes within cellular compartments or tissues; clarification of mechanism underlying a process.

The paper is organized as follows. Section 2 gives an overview of the BioKB platform. In Section 3 we describe the text mining component. Section 4 focuses on semantic technologies employed by BioKB. Description of the web interface follows in 5. Section 6 offers a discussion, while conclusions are presented in Section 7.

2 System Overview

![Fig. 1. BioKB platform. The publications, initially stored in Solr, are processed by the text-mining module. The results are then stored in Virtuoso for dissemination through a web interface and a SPARQL endpoint.](image-url)
We developed the BioKB (for bio-medical knowledge base) platform, which aims at processing scientific publications to extract bio-medical events, mapping those events to a dedicated ontology and storing them in a triple store. A web interface allows users to browse this knowledge base and a SPARQL endpoint offers a machine readable access. In this context, a bio-medical event consist of a labelled relationship connecting two bio-medical entities, i.e. an RDF triple. As illustrated in Figure 1, the publications (abstracts and full texts retrieved from PubMed and PubMed Central) are initially indexed by a Solr instance. The second step consists of executing our text-mining module (c.f. Section 3) for each publication. As this process is relatively slow (about one minute per publication), a high performance computing facility is used to process the publications in parallel. By using 160 cores, the processing capacity results in about 230,000 publications per day. The results of the text-mining component are converted to RDF and then stored in a triple store, the community edition of Virtuoso, as explained with more details in Section 4. To allow both human and machine access to the knowledge base, we offer two dissemination channels. Virtuoso’s built-in SPARQL endpoint provides machine readable access while a web application allows users to browse the content of the knowledge base. We developed this web application in Python 3 using the Flask framework and the SPARQL-Wrapper library to query the triple store. We use the vis.js [17] library to render the bio-medical events as a graph.

3 Text Mining Component

Fig. 2. Main stages of the text mining processing. The sentence detailed in Steps II and III is PGC-1 mediates this increased GLUT4 expression, in large part, by binding to and coactivating the muscle-selective transcription factor MEF2C. Triggers are marked in green, entities in red.
The main steps executed by the text mining component of the BioKB are: i) named entity recognition; ii) syntactic parsing; iii) semantic interpretation (see Figure 2). In the following subsections we briefly describe how we get them into our system.

3.1 Named Entity Recognition

During the Named Entity Recognition (NER) stage, biomedical concepts are identified in the text. Our choice of a NER engine was driven by two major requirements: a) capability to identify multiple concept types (bio-entities). This would have saved the effort of using and synchronization of multiple NER tools within one pipeline. b) ability of the engine to map entity name to its unique identifier in a dedicated database (process known as normalization). Normalization is indispensable in order to avoid multiple instantiations of the same concept in a database or graph which would otherwise occur due to multiple lexical realizations of the same concept. One of the systems which meet our criteria is Reflect [1]. Reflect recognizes proteins, chemicals, diseases, tissues, cell types, GO processes. It can be invoked on sentence and document levels and is very fast: the complete MEDLINE/PubMed dataset can be processed in about one hour. In Step I of the Figure 2 entities identified by Reflect are marked in turquoise and grey.

3.2 Trigger Generation

Availability of the trigger dictionary is another prerequisite for semantic analysis. Our trigger dictionary is derived from Genia annotated corpus [2] which is a collection of PubMed abstracts with the detected biomolecular events of various types: gene expression, (positive/negative) regulation, binding, cell process etc. Genia corpus is used also to learn so-called 'knowledge cues' which express negative statements and author attitude toward facts being described, such as hypothesis, uncertainty, etc. Each entry in the trigger/knowledge cue dictionary is assigned a relative weight calculated based on positive and negative examples learned from the corpus. During the text analysis, triggers and knowledge cues are detected as dictionary match; those which satisfy a pre-set threshold are retained. Although very useful, Genia corpus is limited to 2000 abstracts which might result in somewhat limited vocabulary. To increase potential coverage of the text mining component, we expanded triggers and knowledge cues with synonyms using the WordNet [4], which we access using NLTK [3].

3.3 Syntactic Analysis

With the entities and triggers being identified in the publication, we can proceed toward syntactic analysis. In order to reduce amount of text that has to be processed, we focus only on the sentences which contain at least two entities and one trigger. This requirement maximizes the probability to identify
a "subject-predicate-object" triple (e.g., "RFLAT-1 activates RATES"). For syntactic analysis we use Stanford parser [5] with Stanford dependencies [6]. Step II in Figure 2 shows dependency graph into which the surface structure of the sentence has been transformed by the parser. A proven benefit of using dependency parsing in information extraction task is the ability to map syntactic dependencies onto semantic roles [7, 24, 8] (see Subsection 3.4 for more details).

3.4 Semantic Interpretation

In order to ensure transfer between syntax to semantics, we opt for the rule-based approach. It consists of assigning semantic roles to entities which are syntactic arguments of a trigger. As a result, relations are typed (mostly, they inherit type from the type of their trigger) and, whenever applicable, directed. For example, direction of a regulatory event is from semantic subject (cause) toward semantic object (theme). On the contrary, relations of type binding and correlation are naturally not directed. We collect syntactic arguments of the triggers via the depth-first search (DFS) of the sentence graph. The rules are applied on the ensemble of trigger and its dependencies. For example, syntactic subject of mediates, PGC-1, is the semantic subject of the regulatory event whose predicate is mediates. Sometimes syntactic connections may lead to nodes merge in favor of a more straightforward semantic interpretation. This is the case of expression and its adjectival modifier (amod), increased. Taken individually, they trigger events of types Gene expression and Positive regulation, respectively. Jointly they are interpreted as Positive regulation whose semantic object is GLUT4.

Biomedical processes are subject to rich variety of conditions under which they could take place. We attempt to account for these by processing information conveyed by certain prepositions, complements, relative and adverbial clauses. Our working example illustrates one such case in which the main event, PCG-1 mediates positive regulation of GLUT4, is communicated along with the description of its mechanism. The latter is introduced by the adverbial clause (advcl) headed by trigger verb binding. By taking this bit of information into account we can logically order the events described in the sentence: (1) PCG-1 binds and coactivates MEF2C; (2) GLUT4’s expression is increased. This sequence is summarized in Step III of the Figure 2.

4 Semantic Web Technologies

The choice of using semantic web technologies for this project was dictated by two main reasons. First, using an ontology to represent the hierarchy of relationships permits, thanks to the reasoning capabilities of Virtuoso, to offer different level of query granularity. For instance, one can ask if two entities are connected by a property ”regulates” and the query engine will be able to return also results for the property ”increases” because the ontology specifies that those two properties are linked by a sub-property relationship (rdfs:subPropertyOf). Additionally, the ontology and thus the hierarchy of properties can be updated
without having to re-process the publications. Besides this reasoning capability, using semantic web technologies offers full machine readable access to the complete knowledge base. Not only can the knowledge base then be used by third parties directly but it becomes possible to combine BioKB data with external sources using federated queries.

4.1 BioKB Ontology

![Classes hierarchy of the BioKB ontology](image1)

![Properties hierarchy of the BioKB ontology](image2)

Fig. 3. Classes hierarchy of the BioKB ontology

Fig. 4. Properties hierarchy of the BioKB ontology

We created a simple ontology, see Figure 3 and Figure 4, to represent the hierarchy of classes and properties that are used to categorize entities and relationships identified by the text-mining component. This ontology is heavily inspired by the GENIA ontology. Our decision to allow inferences on sub-relationships resulted in the need to create a custom ontology. Indeed, in the GENIA ontology, relationships are represented by classes rather than properties. In the proposed ontology, a relationship between two bio-medical entities can be directly translated to a single triple, \( s \ p \ o \) where \( s \) and \( o \) are the entities and \( p \) is a sub-property of \texttt{biokb:bioRelation}, the top level property in the BioKB model. We then use the named graphs feature of Virtuoso to add metadata about this relationship. This includes information such as creation date, provenance and confidence score.

For each type of entity (BiologicalProcess, CellularComponent, Chemical, Disease, MolecularFunction, Protein and Tissue) a well established ontology is used. The Gene Ontology (GO) \cite{9} is used for biological processes and molecular functions. The BRENDA Tissue Ontology \cite{10} has been chosen for tissues and cellular components. PubChem \cite{11} represents all the chemicals identified by
the text-mining module while the Disease Ontology [12] is used for all diseases. Proteins are referred to using their Ensembl [13] identifiers.

4.2 Triple Store

In the current deployment of the platform, a single instance of the open source edition of Virtuoso 7 hosts the knowledge base and provides the SPARQL endpoint. The server hosting the Virtuoso instance has the following characteristics: 128GB Ram, 8 cores, Hard Drive 500GB 10000 RPM. At the date of this publication, the database weights 22GB for 215 million triples. On top of the content generated by the text-mining module, the different ontologies mentioned in Section 4.1 have also been loaded into the triple store. The actual number of triples constituting the BioKB specific content is about 156 million triples. Those triples are the result of the processing of more than 800 000 publications. About 10 million events were extracted from approximately 6.5 million sentences.

5 BioKB Web Interface

Fig. 5. Graph visualization of Asthma. Central node is Asthma. Other nodes with corresponding edges represent relationships identified by the text-mining component. Each color correspond to a different entity type (Disease, Genes, etc).

Besides the SPARQL endpoint, we created a web interface to access the BioKB content. This web application is publicly and freely available at http://biokb.lcsb.uni.lu. The home page displays a unique search field providing auto-complete functionality for all supported bio-medical entities. Once the user clicks on an entity, the entity page will be displayed. This page shows a textual description of the entity, the list of most common co-occurrences for this entity (as a tag cloud) and two tables with the list of relationships involving this entity as extracted by the text-mining module. Those incoming and outgoing relationships are also represented visually as an interactive graph (Figure 5). On this graph, the central node is the entity corresponding to the current page and all other nodes and edges represent the most common relationships involving this
entity. For each edge, on mouse over, the label and the number of occurrences of this relationship are displayed. Each node is clickable and leads to the corresponding entity page. Each edge is also clickable and results in the display of the relationship details page (Figure 6). This page displays the list of publications where this relationship was found and the specific sentences. On the entity page, a download button proposes an export of the result of the SPARQL DESCRIBE command in RDF/XML and in CSV.

Fig. 6. Specific bio-medical event. The page shows a list of publications containing this specific event.

6 Discussion

6.1 Use Cases

The primary goal of our information extraction system and knowledge base is to help researchers focusing on various types of biomedical data analysis. We illustrate its functionality with two use cases related to disease network construction and enrichment.

Chronic obstructive pulmonary disease: the network verification challenge. Gathering disease-related factors into a large-scale network became a common practice. Such networks provide a comprehensive model which helps to elucidate mechanisms involved in pathological processes. We used our system to participate in the network verification challenge dedicated to the Chronic Obstructive Pulmonary Disease (COPD). Using the system ability to provide typed, directed (if applicable) relations between various concepts, we have scanned the literature and extracted candidate relations which have been verified by a human expert and made part of the collaborative community curated network yielded by the Challenge [15]. Figure 7 shows part of the graph that we built. Nodes around the disease correspond to genes/proteins and contain information about their involvement in the disease condition: up- or under-regulated, related (“involved”), possibly related (“susceptible”). Some states are marked as hypothetical (e.g., TSLP is probably up-regulated).
Parkinson’s Disease map: integration and visualization of disease related data

Similar in flavour, our system is used to extract supporting evidence and/or suggest new candidates for inclusion to disease maps which is another instance of disease modeling networks. Parkinson’s Disease map is one such example [16]. Step IV in the Figure 2 shows how GLUT4 was approved and appropriately integrated in the PD map.

6.2 System Strengths, Limitations and Future Work

Our system is constructed with the goal of detailed knowledge extraction from textual data, its availability to human and machine. Its strength is the ability to process abstracts as well as full texts; extract semantic relations between various concept types and contextualize them in terms of location, conditions, logic and temporary order. A web interface offers public and free access to the knowledge base while a SPARQL endpoint offers a machine readable access.

Some aspects of the system will be further developed and there remains room for change and improvement. From the text mining perspective, it operates on the sentence level which limits its recall. Although extracted knowledge is normalized with respect to concepts and relations, various nomenclatures are used. To increase knowledge interoperability we plan to adopt Unified Medical Language System (UMLS) which aims at facilitating communication between various systems processing biomedical and health related data. Currently triple store covers main attributes of the extracted relations, such as subject-predicate-object while contextual aspects need to be incorporated. Future work will include extending the current web application by adding, among other developments, an advanced search feature, a personalized notification system, a REST web service and some bibliographic management system to easily cite the publications. BioKB will also have to be continuously extended by processing more publications.

7 Conclusions

In this paper we described an information extraction system along with the storage database and web interface in the field of biomedicine. The system employs text mining and semantic technologies to help discovery and accessibility of biomedical knowledge. As a proof of concept, we have shown its applicability to disease network construction and enrichment. Along with the strengths, we have pointed out the system’s limitations and outlined future work directions.

8 Acknowledgements

This work was partially conducted in the scope of the eTRIKS project that received funding from the European Union and from the European Federation of Pharmaceutical Industries and Associations as an IMI JU funded project.
Fig. 7. Genes/proteins related to the COPD. The figure shows the network of genes/proteins involved in the Chronic Obstructive Pulmonary Disease (COPD) and figure legend provides information about their involvement in the disease condition.

(no. 115446). The Reproducible Research Results (R3) team of the Luxembourg Centre for Systems Biomedicine is acknowledged for support of the project and for promoting reproducible research.

References


14. Genia Tagger http://www.nactem.ac.uk/GENIA/tagger/


17. Vis.js http://visjs.org


29. STRING: Protein-Protein interaction networks. [https://string-db.org](https://string-db.org)


