Journal of Biomedical Informatics 43 (2010) 1020–1035

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora

Jahiruddin [a], Muhammad Abulaish [a,*], Lipika Dey [b]

[a] Department of Computer Science, Jamia Millia Islamia (A Central University), New Delhi, India
[b] Innovation Labs, Tata Consultancy Services, New Delhi, India

ARTICLE INFO

Article history:
Received 18 May 2010
Available online 24 September 2010

Keywords:
Biological text mining
Biological relation extraction
Biomedical knowledge extraction and visualization
Semantic network
Biomedical query answering

ABSTRACT

A number of techniques such as information extraction, document classification, document clustering and information visualization have been developed to ease extraction and understanding of information embedded within text documents. However, knowledge that is embedded in natural language texts is difficult to extract using simple pattern matching techniques and most of these methods do not help users directly understand key concepts and their semantic relationships in document corpora, which are critical for capturing their conceptual structures. The problem arises due to the fact that most of the information is embedded within unstructured or semi-structured texts that computers can not interpret very easily. In this paper, we have presented a novel Biomedical Knowledge Extraction and Visualization framework, `BioKEVis` to identify key information components from biomedical text documents. The information components are centered on key concepts. `BioKEVis` applies linguistic analysis and Latent Semantic Analysis (LSA) to identify key concepts. The information component extraction principle is based on natural language processing techniques and semantic-based analysis. The system is also integrated with a biomedical named entity recognizer, ABNER, to tag genes, proteins and other entity names in the text. We have also presented a method for collating information extracted from multiple sources to generate semantic network. The network provides distinct user perspectives and allows navigation over documents with similar information components and is also used to provide a comprehensive view of the collection. The system stores the extracted information components in a structured repository which is integrated with a query-processing module to handle biomedical queries over text documents. We have also proposed a document ranking mechanism to present retrieved documents in order of their relevance to the user query.

1. Introduction

The number of text documents disseminating knowledge in biomedical field has gone up many folds as scientific publications and other forms of text-based data are produced at an unprecedented rate due to growing research activities in the recent past. Most scientific knowledge is registered in publications and other unstructured representations that make it difficult to use and to integrate the information with other biological data sources. Given that almost all current biomedical knowledge is published in scientific articles, researchers try to make use of this information. Consequently there is an increasing demand for automatic curation schemes to extract knowledge from scientific documents and store them in a structured form without which the assimilation of knowledge from this vast repository is becoming practically impossible. Knowledge discovery could be of major help in the discovery of indirect relationships, which might imply new scientific discoveries. Such new discoveries might provide hints for experts working on specific biological processes. While search engines provide an efficient way of accessing relevant information, the sheer volume of the information repository on the Web makes assimilation of this information a potential bottleneck in the way its consumption. One approach to overcome this difficulty could be to use intelligent techniques to collate the information extracted from various sources into a semantically related structure which can aid the user for visualization of the content at multiple levels of complexity. Such a visualizer provides a semantically integrated view of the underlying text repository in the form of a consolidated view of the concepts that are present in the collection, and their inter-relationships as derived from the collection along with their sources. The semantic net thus built can be presented to users at arbitrary levels of depth as desired.

Several disciplines including *information extraction, document classification, document clustering, and information visualization* have

* Corresponding author.
E-mail addresses: jahir.jmi@gmail.com (Jahiruddin), abulaish@ieee.org (M. Abulaish), lipika.dey@tcs.com (L. Dey).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.09.008

been developed to ease extraction and understanding of information embedded in unstructured text documents [3–7]. However, knowledge that is embedded in natural language texts is difficult to extract using simple pattern matching. Although, techniques such as simple pattern matching can highlight relevant text passages from large abstract collection, generating new insights to future research is far more complex. Text mining has emerged as a hybrid discipline on the edges of the fields of information science, bioinformatics and computational linguistics which attempts to find hidden knowledge in the literature by exploring the structure of the knowledge network created using textual information [1,2,8].

In this paper, we have proposed the design of a novel biomedical knowledge extraction and visualization framework, BioKEVis, for conceptualization of document corpora and biomedical query answering. Conceptualization of document corpora here means representation and visualization of document corpora with a set of concepts and their relationships which can provide distinct user perspectives and allows navigation over documents with similar information components. BioKEVis applies Latent Semantic Analysis (LSA) to identify key concepts. Relationships among key concepts are extracted using natural language processing and semantic-based analysis. The information components are centered on key concepts and their relationships, and stored in structured form. The process of extracting relevant information components from text documents and automatic construction of structured knowledge bases is termed as curation which is very effective in managing online journal collections [15]. Schutz and Buitelaar [14] state that verbs play an important role in defining the context of concepts in a document. BioKEVis is designed to locate and characterize verbs within the vicinity of biological entities in a text, since these can represent biological relations that can help in establishing query context better. The verbs thus mined from documents are subjected to feasibility analysis and then characterized at concept level. We have shown that relation mining can yield significant information components from text whose information content is much more than entities.

Besides mining relational verbs and associated entities, the novelty of the system lies in extracting *validatory entities* whose presence or absence validates a particular biological interaction. For example, in the following PubMed sentence, "*regulates*" is identified as relational verb relating the biological entities "*Rac1*" and "*transcription of the APP gene*" while "*primary hippocampal neurons*" is identified as *validatory entity*.

''... Racl regulates transcription of the APP gene in primary hippocampal neurons (PMID: l9267423).''

We have also presented a scheme for semantic integration of information extracted from text documents using semantic net. The semantic net highlights the role of a single entity in various contexts, which is useful both for a researcher as well as a layman. The network provides distinct user perspectives and allows navigation over documents with similar information components and is also used to provide a comprehensive view of the collection. It is possible to slice and dice or aggregate to get more detailed or more consolidated view as desired.

The system is also integrated with a biomedical named entity recognizer, ABNER [13], to identify a subset of GENIA ontology concepts (*DNA*, *RNA*, *protein*, *cell-line*, and *cell type*) and tag them accordingly. This helps in answering biological queries formulated at different levels of specificity. Given a query, BioKEVis aims at retrieving all relevant sentences that contain a set of biological concepts stated in a query, in the same context as specified in the query, from the curated database. We have also proposed a document ranking mechanism to present retrieved documents in order of their relevance to user query. The efficacy of BioKEVis is established through experiments on GENIA corpus [28].

The remaining paper is structured as follows: Section 2 presents a review of related works on biomedical text mining. The architectural detail of BioKEVis is discussed in Section 3. Section 4 presents the experimental detail and evaluation of various modules. Section 5 presents a critical discussion to highlight the novelties of the proposed system over existing ones. Finally, Section 6 concludes the paper and provides direction for possible enhancements to the proposed system.

2. Related works

In this section, we present an overview of some of the recent research efforts that have been directed towards the problems of biological relation extraction from text documents. A brief review of the existing biomedical knowledge visualization and query answering systems will be also a part of this section.

2.1. Biological relation extraction

Though, named-entity recognition from biological text documents has gained reasonable success, reasoning about contents of a text document however needs more than identification of the entities present in it. Context of the entities in a document can be inferred from an analysis of the inter-entity relations present in the document. Hence, it is important that the relationships among the biological entities present in a text are also extracted and interpreted correctly. Related works in biological relation extraction can be classified into the following three categories:

Co-occurrence based approach: In this approach, relations between biological entities are inferred based on the assumption that two entities in the same sentence or abstract are related. Negation in the text is not taken into account. Jenssen et al. [21] collected a set of almost 14,000 gene names from publicly available databases and used them to search MEDLINE abstracts. Two genes were assumed to be linked if they appeared in the same abstract; the relation received a higher weight if the gene pair appeared in multiple abstracts. For the pairs with high weights, i.e. with five or more occurrences of the pair, it was reported that 71% of the gene pairs were indeed related. However, the primary focus of the work is to extract related gene pairs rather than studying the nature of these relations. In [32], an ontology-based biological information extraction and query answering (BIEQA) System is proposed which extracts biological relations from MEDLINE abstracts using NLP techniques and co-occurrence based analysis from tagged documents. Each mined relation is associated to a fuzzy membership value, which is proportional to its frequency of occurrence in the corpus and is termed a fuzzy biological relation. The fuzzy biological relations along with other relevant information components like biological entities occurring within a relation, are stored in a database which is integrated with a query-processing module. The query processing module has an interface, which guides users to formulate biological queries at different levels of specificity. The recall values ranged from 84.68% to 86.23% and precision from 94.73% to 98.87%.

Linguistics-based approach: In this approach, usually shallow parsing techniques are employed to locate a set of handpicked verbs or nouns. Rules are specifically developed to extract the surrounding words of these predefined terms and to format them as relations. As with the co-occurrence based approach, negation in sentences is usually ignored. Sekimizu et al. [9] collected the most frequently occurring verbs in a collection of abstracts and developed partial and shallow parsing techniques to find the verb's subject and object. The estimated precision of inferring relations is about 71%. Thomas et al. [10] modified a pre-existing parser based on cascaded finite state machines to fill templates with