A Review on Mining Textual Information from Biomedical Data

K.Mary Sudha Rani[1], T.Krishnaveni[2]
[1]*Assistant Professor, Department of CSE, CBIT, Hyderabad*
[2]*M.Tech, Department of CSE, CBIT, Hyderabad*

***Abstract*- This paper presents a review on bio-medical text mining from scientific documents involving event extraction which is used to explore the data biomedical data in an easy and comprehensive manner. Information extraction is a challenging task from the biological texts as it considers many aspects from different areas such as natural language extraction, statistics and machine learning models. This event extraction involves a series of steps from preprocessing to post processing which includes many intermediate methods to represent in different patterns. This paper describes different resources available for biomedical text mining, information extraction and recent machine learning methods that are significantly used.**

1. INTRODUCTION

Biomedical text mining is a text mining which is applied on text and scientific documents. Biological systems consist of entities and the relationships between them in terms of how they interact and the downstream effects of such interactions. Automatic extraction of events have wide range of applications in molecular biology which support for the creation and annotation of pathways to improve the performances of databases. Event extraction systems can be trained to recognize a wide range of activities, including protein–protein interactions, pathway enrichment and construction, gene regulatory events, and metabolic or signaling reactions.

This survey begins with identification of resources for mining the biomedical literature ,event extraction and machine learning methods. This concludes with an expectation for further development of the domain.

2. BIOMEDICAL TEXT MINING IN DIFFERENT FIELDS

2.1 Protein-protein interactions based on graph kernel method[5]

Biomedical text mining is very essential in finding the extracting the complex interactions between proteins. These studies mainly involve the effects of training and testing on different resources for providing protein pairs. This graphical method is based on dependency scheme which gives the user a contextual and syntactic information which is needed to distinguish between interactions and non interactions. The dependency structures parser make the relationships between words which directly interacts with the parser. This process proposes three approach such as subsequence kernels , tree kernels and shortest path kernels for relation extraction. This is method that captures information in unrestricted dependency graphs to a format that kernel based learning algorithms can process. These interactions re not only binary but multiple complex structures will also be considered.

2. Bio-molecular event extraction[1]

Biomedical text mining used for recognizing the events in the text and also for identifying approaches to automatic event extraction. These events mainly includes genes, enzymes or transcription factors which play major role in biological processes. Event extraction mainly involves information retrieval, discovery and knowledge summarization. This describes the event extraction from preprocessing event extraction to post processing including entity recognition, trigger detection and edge detection.

3. Biological network extraction[8]

Biomedical text mining a way for network extraction to indicate the interactions between the bio-molecular events .this is done mainly to reduce the complexity and increase the functionality with gives completeness to the estimated network. This network

extraction provides research in finding associations between many complex structures and attributes.

3. RESOURCES

(i) CORPUS

MEDLINE was the first main resource in biomedical text mining. abstracts are available in different way

(a)GENIA corpus

The GENIA corpus a product of the GENIA project of which the objective is to develop information extraction (IE) and text mining (TM) systems for the specific subject domain of molecular biology and medical science. Currently, the GENIA corpus is made up of the titles and abstracts of journal articles which have been taken from the MEDLINE database. Whereas the MEDLINE indexes broad range of academic articles covering the general or specific domains of life sciences, GENIA is intended to cover a smaller subject domain: biological reactions concerning transcription factors in human blood cells. For example, to retrieve the corresponding records from MEDLINE, MeSH terms are used like "blood cells", "human", "transcription factors". The GENIA corpus has been being annotated by encoding human's interpretation of the text. The purpose of the annotation is : first, annotate the corpus to make the biomedical knowledge encoded in the text transparent. Second, also annotate the corpus to reveal the syntactic structure behind the text. Eventually, our objective is to establish the mapping between the knowledge pieces and the linguistic structures.

(b)OHSUMED

The OHSUMED test collection is a set of 348,566 references from MEDLINE, the on-line medical information database, consisting of titles and abstracts from 270 medical journals over a five-year period (1987-1991). The available fields are title, abstract, MeSH indexing terms, author, source, and publication type. The National Library of Medicine has agreed to make the MEDLINE references in the test database available for experimentation, restricted to the following conditions: The data will not be used in any non-experimental clinical, library, or other setting. Any human users of the data will explicitly be told that the data is incomplete and out-of-date. recent sets are more data from TREC genomes track from 1994-2003 which cross references PubMed, clinically oriented subset of MEDLINE.

(ii) BioNLP

Biomedical text mining (also known as BioNLP) refers to text mining applied to texts and literature of the biomedical and molecular biology domain. It is a rather recent research field on the edge of natural language processing, bioinformatics, medical informatics and computational linguistics. There is an increasing interest in text mining and information extraction strategies applied to the biomedical and molecular biology literature due to the increasing number of electronically available publications stored in databases such as PubMed. The main developments in this area have been related to the identification of biological entities (named entity recognition), such as protein and gene names as well as chemical compounds and drugs in free text, the association of gene clusters obtained by microarray experiments with the biological context provided by the corresponding literature, automatic extraction of protein interactions and associations of proteins to functional concepts (e.g. gene ontology terms). Even the extraction of kinetic parameters from text or the sub cellular location of proteins have been addressed by information extraction and text mining technology. Information extraction and text mining methods have been explored to extract information related to biological processes and diseases.

4.INFORMATION EXTRACTION

Information extraction is a process of deriving structured facts automatically from the unstructured or semi structured data. Biomedical text mining gave a rapid development in information extraction based on different shared tasks. It mainly has three main tasks.

4.1 Name entity recognition

It is the process of automatically identifying occurrences of biomedical terms in unstructured text. This faces the challenges of dynamicity of scientific discovery and synonymy.NER mainly involves three basic tasks.NER is typically capable of achieving favorable results.

4.2 Relation extraction

Relation entity is extracting named entities and finding the relationships among those entities. It considers only simple binary entities involving pair-

wise relations. This gives the recognition between protein-protein ,drug-drug interactions. Compared to name entity recognition annotation of relations is more complicated since relations are expressed or discontinuous texts. The main of relation extraction is to identify the medical problem –test, problem-treatment, problem-problem in clinical domains. This mainly resulted in the evolution of complex systems using syntaxes and dependencies.

4.3 Event extraction

Event extraction is the method to identify nested and complex structures from the recognized binary events. Events are based on characteristics of verbs or nominal verbs. This mainly involves proteins and other bio-molecules. Each event is identified based on trigger word and their arguments. Trigger detection is the method of identifying the trigger words which shows the effectiveness of strongly depending information. This involves the identification of chunk of text which serves as a predicate. There are various approaches for trigger detection which are categorized into: rule based, dictionary based, machine learning based.

5. MACHINE LEARNING MODELS

Machine learning models are used to train the events based on defining a sequence –labeling problem. These are mainly used for the feature extraction using different tools like BioCreAtivE , Metamap (MMTx) - NLM ,Negex, Context ,Ctakes etc..

5.1 Conditional Random Fields(CRF)

CRF is a discriminate model which uses conditional probability for interference. CRF is a [popular method for sequence-labeling problems which are justified by the fact to avoid label bias problem.CRF model is used for interference ,sequences based on input and output. These are mainly based on the strong independence assumptions that are required to learn the parameters of generative models.

5.2 Support vector machine(SVM)

SVM will be used for finding the hidden associations between the multiple associations. Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. In addition to performing linear classification, SVMs can efficiently perform a non- linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is often used in industrial applications either when data are not labeled or when only some data are labeled as a preprocessing for a classification pass.

5.3 Vector Space Model(VSM)

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System. Documents and queries are represented as vectors. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as (term) weights, have been developed. One of the best known schemes is TF-IDF weighting. The definition of term depends on the application. Typically terms are single words, keywords, or longer phrases. If words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary (the number of distinct words occurring in the corpus).Vector operations can be used to compare documents with queries.

5.4 TrigNER

TrigNER, a machine learning-based solution for biomedical event trigger recognition, which takes advantage of Conditional Random Fields (CRFs) with a high-end feature set, including linguistic-

based, orthographic, morphological, local context and dependency parsing features. Additionally, a completely configurable algorithm is used to automatically optimize the feature set and training parameters for each event type. Thus, it automatically selects the features that have a positive contribution and automatically optimizes the CRF model order, n-grams sizes, vertex information and maximum hops for dependency parsing features. The final output consists of various CRF models, each one optimized to the linguistic characteristics of each event type.

TrigNER was tested in the BioNLP 2009 shared task corpus, achieving a total F-measure of 62.7 and outperforming existing solutions on various event trigger types, namely gene expression, transcription, protein catabolism, phosphorylation and binding. The proposed solution allows researchers to easily apply complex and optimized techniques in the recognition of biomedical event triggers, making its application a simple routine task. This work is an important contribution to the biomedical text mining community, contributing to improved and faster event recognition on scientific articles, and consequent hypothesis generation and knowledge discovery.

6. CONCLUSION

This paper presents a review of biomedical text mining in recent years, the progress was made on (1) identifying the different resources for clinical text mining ;(2) different uses of biomedical text mining in different research activities;(3)information extraction using different entities, relations and bimolecular events from unstructured data;(4)machine learning methods for training for the obtained data from text mining. The efforts needed to make a system more capable for handling real-time workflow for further interactions and studies to find different associations and discoveries.

REFERENCES

[1] MindLab Research Laboratory, Universidad Nacional de Colombia, Bogotá, Colombia" An Overview of Biomolecular Event Extraction from Scientific Documents" DETI/IEETA, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, 18 August 2015

[2] D. Campos, S. Matos, and J. L. Oliveira, "A document processing pipeline for annotating chemical entities in scientific documents," Journal of Cheminformatics, vol. 7, supplement 1, article S7, 2015

[3] D. Campos, S. Matos, and J. L. Oliveira, "Current methodologies for biomedical named entity recognition," in Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data, pp. 839–868, John Wiley & Sons, 2013.

[4] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo, "Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)," in Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval '13), pp. 341–350, June 2013.

[5] Airola,A. et al. (2008) All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC bioinformatics, 9 (Suppl. 11), S2.

[6] Ananiadou,S. et al. (2010) Event extraction for systems biology by text mining the literature. Trends in Biotechnol., 28, 381–390.

[7] Björne,J. et al. (2009) Extracting complex biological events with rich graph-based feature sets. In Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, Association for Computational Linguistics, pp. 10–18.Bjorne,J. et al. (2010) Complex event extraction at PubMed scale. Bioinformatics, 26,i382–i390.