CHAPTER 3

RESEARCH METHODOLOGY

Every day, just about 5000 new articles appear as a result of biomedical research at an ever-increasing rate. Its topical growth replicates an interest in retrieval of information. Effective accessing of these data is a challenge in itself. This wealth of knowledge is used for demanding task for both computer scientists and biomedical experts.

Without automated means of retrieval and extraction, it is difficult for researchers to keep abreast of developments within biomedicine [99] [100]. Thus, the flare-up information in biotext corpora sets massive concentration in automated techniques for extraction. Text Mining and Information Retrieval at increasing interest within the biomedical field, enriches text via the addition of semantic metadata, and tasks as analyzing molecular pathways [74] and semantic searching [85].

Semantic searching on top of concept level requires pre-processing to recognize named entities and to categorize the relations between entities. Due to domain-specific features of texts and the types of events to be recognized, IE must be adapted to deal with specific domains.

The major key concepts of proposed system for biomedical research are defined in advance and then the text is searched for concrete evidence of the existence of such concepts. In order to perform this process, the research methodology is categorized into two major phases as follows as:

Phase I: Named Entity Recognition (NER)

Phase II: Named Entity Relationship or Relation Extraction (RE)
NER acts as a core step to access the higher level of information. It identifies the entity in shallow semantic categories and sluice ambiguity among gene, protein, disease category. RE does automatic detection of relevant types of entities and relations among entities with many definitions. Relationship as gene-disease relationship, protein-protein interaction and gene-gene relationship are primarily focused points for significant domain research efforts.

This chapter introduces the steps involved in the proposed system of Information Retrieval System (IRIS) and outlines the methodology proposed in each step.

3.1 Proposed Architecture of IRIS

The fundamental challenge is with dynamic updating and complex nature of textual sources from biomedical corpus [97]. Scripts in biomedical literature and inconsistent lexical variants are standstill challenges to access. It is tough for researchers and users to find out the measures which they want as output from biomed corpus. In this way, the current trend of research gets focused on the design and implementation of methods for retrieval over biomedical corpus.

Intended of these projections, IRIS formulates the knowledge basis by delimiting the grounded theory. The developed IRIS enriches and unlocks the knowledge present in natural language on biomed corpus. It simplifies technical texts to tune researchers. It fetches and conveys constructive information from amorphous textual data on exposure to the executive and desired patterns.

With the unprecedented amount of information available on web, the proposed methods of IRIS crack the retrieval and extraction problems which have gained more attention.
Figure 3.1 shows the architecture of the proposed Information Retrieval System (IRIS) for NER and RE on biotext corpus.
The information from the vast archives of literature shows potential to get tackled for consequential information discovery. Accordingly, it is concerned with computational model to develop automated tools for language processing and to gain a better understanding of research.

The foremost aim is to develop an integrated system for computational analysis of dynamical behaviour appealing in different aspects, in turn, to eliminate the ambiguity and to bring out semantic information among entities on biomed corpus. The tasks of first phase, NER algorithm robotically recognizes Named Entities (NE). Subsequently, the tasks of second phase, RE algorithms aim to detect relationships among named entities taking account of gene-disease relationship, gene-gene relationship, and protein-protein interaction.

The common and progressive tasks of both phases are Information Retrieval, Information Extraction, Information indexing, Text Classification and Text Clustering. These tasks get along the access to the community of biologists, and developers in the midst of platform prototype. These tasks give convenient access to a large, unstructured repository of text. As a consequence, unstructured repository is upshot into a structured one, thereby slackening the work for index, storage, and search.

3.2 Proposed Methodology for Named Entity Recognition

The rapidly increasing scientific discoveries, ever-growing list of relevant terms and resources have prolonged growth as long as scientific progress continues. The current research gets rid of complex biological systems and gives meaningful biological function. Since it is elusive and highly dependent on the specific underlying scientific question and overwhelming textual information, IRIS formulates the flexible and general approach for integrating heterogeneous data and knowledge sources for discovery. It locates boundary
of entity mentions in a text and tag them with their corresponding semantic types. Named Entity Recognition (NER) is a system to recognize and categorize the named entities automatically with a constraint on domain specific.

For biomedical domain, a NE is defined as a single word term or multi-words phrase that denotes a biomedical object, for instance a protein, gene, disease with which a semantic hierarchy is associated [105].

Enacting Named Entity Recognition is to

- Accumulate Entities on corpus
- Better indexing of biotext
- Construct dictionary and bring away semantic information
- Defined precise identification of relevant passages for curation
- Enhanced support for relation extraction
- Facilitate ambiguity removal

The initial surface work like Word Sense Disambiguator, Probabilistic Latent Semantic Analysis (PLSA) [92], Probabilistic Latent Semantic Indexing (PLSI) methods fail on ambiguity basis and the semantic information [90]. In tune to a crisis, a novel weighting and probabilistic scheme BM25 is integrated with PLSA named as PLSA-BM25++ for Named Entity Recognition is implemented. The pre-processed facts refer Disease Annotation UMLS Meta thesaurus [105] [116] and Gene Ontology [59] to build a dictionary for precise identification [88]. Sub-sequentially PLSA-BM25++ is computed to recognize Named Entities. Then the results are enriched by Information Indexing, Text Classification, and Text Clustering as a user’s prevalent choice.
The progressive tasks of phase I: Named Entity Recognition is as follows as:

**Phase- I: Named Entity Recognition Tasks**

- Information Retrieval - Vector Space Model.
- Pre-processing system.
- Information Extraction - PLSA-BM25++.  
- Information Indexing - Filtering based indexing.
- Text Classification - Multiple Kernel Learning - Support Vector Machine.
- Clustering - Entropy Agglomeration.

Overall progressive tasks get at once the keyword or query collected from user is propelled on biomed corpus. Sequentially, necessitated pre-processing process is done to corpus of tokenization, stemming, stop words removal, morphological analysis and word sense disambiguator. It is called as Query Evaluation Process for keyword selection, and keyword searching. Then the pre-processed units are fed into Information Retrieval (IR) for ‘Topic Tracking’ to do document classification. The output of IR i.e. identified documents is fed into Information Extraction (IE) for analyzing the documents. It undergoes ‘Text Simplification’ for discourse analysis and rapid adaptation to new relations using a representative collection and ‘Structured Search’ for performing document inference, document linking that fact the text to a knowledge base. Followed by that, analyzed documents of IE results i.e. recognized entities are given to Information Indexing for ranking and filtering the documents. The filtered documents are cultivated into Text Classification and Text Clustering for classification and to cluster the extracted and targeted output respectively. The extracted documents are stored in the knowledge base edifice.
In short, to regulate the phase NER, concerned algorithms get initiated and robotically recognize Named Entities (NE) as user query like entities gene, protein and disease in the interest of biomedical domain. The progressive tasks with their concerned algorithms to achieve NER are shown in figure 3.2.

**Figure 3.2 Progressive tasks of NER**

Information Retrieval

Information Retrieval (IR) signifies as a representation of documents/representation of unstructured information needed [55]. It is executed by Vector Space Model [72] on collections of unstructured or semi-structured data by initiating pre-processing. Consider categorizing a set of ‘m’ collected biotext corpus based upon the presence or absence of a list of ‘n’ given query / keyword terms [18]. The term-document matrix is measured by ‘cosine similarity’ from Singular Value Decomposition[69] where query vector ‘q’ in the matrix stands support to weighted frequency of occurrence of entity and a matrix column vector gives the frequency of occurrence of each entity terms on given biotext corpus. Such a construction immediately facilitates the application of matrix analysis between a query vector, and the document vector contained within the term-document matrix.

Pre-processing System

Text pre-processing episode as imperative and crucial work subsists the enduring tasks [97] [64]. It embraces and is implemented by the solutions of systematic functions as Tag Removal, Tokenization, Stemmer, Stop Words Removal, Morphological Analysis and Word-Sense Disambiguator (WSD) space. Successively, when once the query/keyword is collected and pre-processed keyword reaches the documents, pre-processing system gets functioned on biotext corpus which acts as an input for IR.

Tokenization: It tokenizes the retrieved information texts from source into meaningful units called tokens. The biomedical content encompasses the named entities that enclose exceptional characters like numerals, hyphens, slashes and possess various lexical
variants. Predetermined heuristics of tokenization perform well on information retrieval since heuristics are evaluated systematically by eliminating special characters [36].

Subsequently, a common strategy ‘CASE-FOLDING’ is processed to reduce all letters to lower case. Following that, ‘LEMMATİZATİON’ tries to map verb forms to the infinite tense and noun to the singular forms for protein and gene. It removes the inflectional endings and return to the base. Forwarding that, ‘STEMMİNG’ a heuristic process severs the words to fabricate the basic forms of words.

Stemming process: It removes the inflection variation of words. Semantically related words should map the same stem, base or root form, and this should compensate for data sparseness. A heuristic process severs the words to fabricate the basic forms of words, such as stripsthe plural ’s’ from nouns, the ‘ing’ from verbs and removes derivational affixes.

Stop Words Removal: ‘Stop words’ are the terms that often emerge with little significance in the bias of retrieved documents. Within definite dataset some of the standard stop words are shown in table 3.1 get eliminated immediately.

<table>
<thead>
<tr>
<th>A</th>
<th>Are</th>
<th>Around</th>
<th>Alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Also</td>
<td>Already</td>
<td>Along</td>
<td>Often</td>
</tr>
<tr>
<td>Hereupon</td>
<td>Become</td>
<td>However</td>
<td>Almost</td>
</tr>
<tr>
<td>i.e.</td>
<td>Just</td>
<td>Meanwhile</td>
<td>All</td>
</tr>
<tr>
<td>Normally</td>
<td>Again</td>
<td>After</td>
<td>Against</td>
</tr>
<tr>
<td>Only</td>
<td>Onto</td>
<td>Otherwise</td>
<td>Out</td>
</tr>
</tbody>
</table>
Morphological Analysis: Basically, several words are originated from identical base words. Such words are grouped under analysis and replaced to base word. For example, ‘bacteria’, there may exist one entry associating it with bactérie, one with bactério (as in bactério/lyse) and another one with bactéri (in myco/bactéri/ose), should be combined to form a single word, bacteria.

Word-Sense Disambiguator: It turns down the nominal words in accordance with their base and semantic meaning i.e. noun, adjective and verb using WordNet ontology [66]. By then, synonym terms are surfaced for forthcoming feature set.

**Information Extraction**

Information Extraction (IE) can be of great benefit for researchers allowing identification and extraction of relevant information on protein, gene and disease [20]. It focuses on dictionary-based text mining and its role in enabling practitioners to understand and analyze the large biotext datasets.

Named Entities (NE) are recognized to unlock the necessitate information hidden in articles / corpus [104]. Its effort is successively carried in the process of Information Extraction phase. The dictionary is constructed for retrieved terms with reference to GO Annotation for gene and protein and Disease Ontology annotation ‘UMLS’ for disease entities. They bring out the semantic relationship of the entities like gene, disease and protein. Herewith, PLSA-BM25++ named entity recognition algorithm robotically recognizes Named Entities (NE) as given user query which related to entities of biomedical domain on biotext corpus to remove irrelevant entity recognition and eliminate ambiguity.
**Information Indexing**

As extracted information related to Named Entity Recognition may hold different words, choice on apt and classification chore is becoming crucial. Decisive to reduce the complexity of the classification process, the information indexing method is obligatory for extracted information. In order to obtain a fixed and appropriate index terms over NER documents, a greedy strategy is applied.

**Text Classification**

Text classification applied to biological literature minimizes the effort by selecting only the relevant articles and terms to a given task. The classification method relies on indexed information and their indexing information results stored in databases. Classify the list of indexing terms obtained from analysis phase into several categories of gene, protein and disease. MKL-SVM functions on a gradient descent optimization algorithm to classify the indexing information results into classes [102].

**Text Clustering**

Find which classified user named entity recognition has common entities with similar user and rest the extracted information documents with the most words in common into the same groups [71]. The classified results give the detailed information about the classes of the gene, protein and disease information indexed results. But, if number of the biomedical users become larger, it is difficult to find similar user categories for the classified data for inter cluster. Hence, Entropy Agglomeration (EA) is exploited for clustering the classified results [3]. The classified classes based on indexed information are clustered into similar group of the entity. Chapter 4, design of the Named Entity Recognition provides the detailed description of above mentioned steps and its concern methodology.
3.3 Proposed Methodology for Named Entity Relationship

In broad-spectrum, most of existing work mere extracts the information rather than knowledge. Thus, the second subtask of IRIS extracts relevant knowledge on biomedical domain as ‘Relation Extraction’. It is performed to detect binary relationships include gene-disease relationship, gene-gene relationship and protein-protein interaction.

In order to execute the named entity relationship, phase II, forthcoming steps and methodologies are sorted out as follows as:

<table>
<thead>
<tr>
<th>Phase- II: Named Entity Relationship Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Information Retrieval - Vector Space Model</td>
</tr>
<tr>
<td>2. Pre-processing system</td>
</tr>
<tr>
<td>3. Information Extraction - Entity Extraction enhanced to Named Entity Relationship</td>
</tr>
<tr>
<td>4. Relation Extraction</td>
</tr>
<tr>
<td>• Reporting a direct interaction between proteins by proposing ‘Dictionary Based Text and Gene Ontology-PPI’ (DBTGO-PPI) methodology.</td>
</tr>
<tr>
<td>• Reporting a direct relationship between gene-disease by proposing ‘Logistic Model of Dictionary Construction using LocusLink’ (LL-GDD)</td>
</tr>
<tr>
<td>• Reporting a direct relationship between two genes by proposing ‘Logistic Model of Dictionary Construction using Locus Link’ (LL-GGD)</td>
</tr>
<tr>
<td>5. Information Indexing - Filtering based indexing</td>
</tr>
<tr>
<td>6. Text Classification - Multiple Kernel Learning-Support Vector Machine</td>
</tr>
<tr>
<td>7. Text Clustering - Entropy Agglomeration</td>
</tr>
</tbody>
</table>
The progressive conceded tasks for RE is explicated diagrammatically in figure 3.3.
**Information Retrieval**

Once the articles are received as input, a construction is made immediately that facilitate the application of matrix analysis for the sake of quantifying the degree of similarity between document vectors contained within the term-document matrix. It is executed by Vector Space Model [72] [73] on collections of unstructured or semi-structured data by initiating pre-processing. The term-document matrix is measured by ‘cosine similarity’ from Singular Value Decomposition [69].

**Pre-processing**

Text pre-processing event endures the tasks Tokenization, Stemmer, Stop Words Removal, Morphological Analysis and WSD space. Successively, when once the query/keyword is collected significantly and pre-processed keyword reaches the documents, pre-processing system gets functioned on biotext corpus which acts as an input for IR.

**Information Extraction**

The IE systems that extract relationship and interaction from the text are named as Named Entity Relationship or Relation Extraction (RE). Its parsing recovers the syntactic structure of the text, and domain analysis extracts the relationships among the named entities by means of the information from the other processing steps [70]. Conventional relationship extraction is concentrated on the study of biomedical relation extraction such as protein-protein interaction, gene-gene and gene-disease relationship from biomedical terms such as gene, protein, and disease. The proposed dictionary-based text mining approach is able to detect the specified semantic relationship between each pair of entities.
through dictionary construction by utilizing resources like GO, UMLS and LocusLink after entity recognition. These resources are likely to be perfect at any given moment, resulting in some synonymy relationships that may be captured exactly.

**Protein-Protein Interaction**

Protein-Protein Interaction (PPI) Interaction is a crucial process for biological processes to be handled in a great manner [42]. By accurate identification of the set of interacting proteins, it gets rid of new light on the functional role of various proteins in the complex surroundings of the cell. The ability to construct biologically consequential network and identification of the exact relationship in the network is critical for present day systems biology. Proteins, the working molecules of a cell, carry out many biological activities. PPIs help to decipher the molecular mechanisms underlying the biological functions. Earlier studies done by Random Forest get pitfalls [76].

Once the article is received as input, the pre-processing steps and retrieval process are enacted to retrieve the information. Sequentially, the lines retrieved refer to the constructed dictionary using of GO and bring out the semantic similarity after the recognition of entity. The incorporation of semantic similarity based on GO annotation with gene expression and protein-protein interaction data can greatly enhance the functional relevance of inferred modules. Gene Ontology (GO) Annotation Similarity Score in respect of all other proteins in the data is measured to determine the similarity of PPI pair \( S(p_i, p_j) \), by taking the gene protein weights into consideration from extracted entities. Here, cosine similarity measure has been considered for interaction induction as in Eqn. (3.1) as
where \( ce_i(p_i) \) and \( ce_i(p_j) \) are the weights of the \( i \)th common entity in the expanded query from user given for two different proteins. If the Gene Ontology Annotation Similarity \( S(p_i, p_j) \) between two proteins is less than 1.0, it is reflected and held as interacting terms to frame an interaction network of proteins. It is given in Eqn. (3.2)

\[
PPI \text{ Score} = \frac{\text{No. of interacting residues}}{\text{Probable number of interacting residues}}
\]

Gene-Gene Relationship Measurements

Gene pair relationships get to the bottom of inconsistencies and incomprehensiveness problems [32]. In regulation of computing the gene-gene relationship of relation extraction, the proposed Logistic Model of Dictionary Construction using LocusLink (LL-GGD) is utilized. A unique LocusLink identifier is assigned to each entry in the gene dictionary for consistent merge of gene information dispersed in different databases. Thus, dictionary is constructed with reference to LocusLink for gene description. With reference to dictionary, and from the recognized entities, each gene-gene pair is evaluated by computing the Gene Ontology Annotation Similarity. Taking the gene weights into consideration, standard similarity measure ‘cosine similarity’ calculates \( S(g_i, g_j) \) as in Eqn. (3.3)

\[
GO(S) = S(p_i, p_j) = \frac{\sum_{i=1}^{k} ce_i(p_i) * ce_i(p_j)}{\sqrt{\sum_{i=1}^{l} co_i^2(p_i) \sum_{i=1}^{l} co_i^2(p_j)}}
\]

\[
GO(S) = S(g_i, g_j) = \frac{\sum_{i=1}^{k} ce_i(g_i) * ce_i(g_j)}{\sqrt{\sum_{i=1}^{l} co_i^2(g_i) \sum_{i=1}^{l} co_i^2(g_j)}}
\]
where $ce_i(g_i)$ and $ce_i(g_j)$ are the weights of the $i^{th}$ common entity in the expanded query from user given for two different genes. If the Gene Ontology Annotation Similarity $S(g_i, g_j)$ between two genes is less than 1.0, then it is upshot to interacting phase, thus formulating a gene-gene relationship in network.

**Gene-Disease Relationship**

The proposed Logistic Model of Dictionary Construction using LocusLink (LL-GDD) extracts a relation between the disease and gene names in each sentence of biotext corpus. For each output entry linked to data sources; a gene dictionary is created using the unique LocusLink and disease dictionary is created by the UMLS Meta thesaurus for selected congenital disease with semantic types. By merging the recognized entity entries of multiple public biomedical databases, gene-disease relationship are found and computed by cosine similarity measure as Eqn. (3.4)

$$S(g_i, d_j) = \frac{\sum_{i=1}^{k} ce_i(g_i) * ce_i(d_j)}{\sqrt{\sum_{i=1}^{l} co^2_i(g_i)} \sqrt{\sum_{i=1}^{l} co^2_i(d_j)}}$$

(3.4)

where $ce_i(g_i)$ and $ce_i(d_j)$ are the weights of the $i^{th}$ common entity in the expanded query from user given for genes and diseases.

**Construction of the Gene and Disease Dictionaries**

IE identifies and extracts a range of specific types of information from texts of interest on constructing dictionary. For NER, dictionary is constructed with reference to GO and UMLS. For ‘PPI’, the dictionary is constructed significantly with reference to GO. For gene-gene relationship and gene-disease relationship, dictionary is constructed based on unique LocusLink identifier [30] for gene and UMLS for disease. In LocusLink
identifier, genetic loci are assigned to each entry in the gene dictionary, which in turn merge gene information dispersed in different databases. It holds all relevant literature information associated with a given gene.

To advancement of ontology, the dictionary has been constructed with full meta-data evidences. Gene Ontology (GO) is a collaborative effort to address the need for consistent descriptions of gene and protein in different databases. Disease Ontology (DO) annotation is human disease ontology provides consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary through collaborative efforts of researchers.

To build related concept dictionaries for annotating a collection of documents from a particular domain, dictionary $D = \text{Dict}(C,X)$ as a set of words that refers to or describes a semantic concept, as entity recognition and entity relationship concept ‘C’ in a pre-processed document ‘X’ is defined [88].

In this method, an online interactive framework is applied, where the user starts off with a small set of protein or gene or disease words, inspects the results, selects and rejects the entity words from the returned ranking, and iterates until get satisfied. With interactive supervision, the user provides positive and negative seed entities at each stage of iteration to the algorithm. This process gradually refines the seed sets and the ranking comes closer to the user’s preference as the iteration continues. Then the ranking returned by the system is inspected to create the adapted dictionary for the new document collection. Thus, re-using dictionaries can significantly make easier the task of specifying the semantic concept in the absence of semantic structure for dictionary construction for a concept.
The following algorithm explicates the steps to construct the dictionary.

Algorithm 1: Build Dictionary (Entities $g^e$ or $p^e$ or $d^e$, Int K, Double t)

1. Initialize candidate set $C_0$ to empty set
2. For all the entities ($G(g^e$ or $p^e$ or $d^e$))
3. If the $G(p^e, p') >$ thesholdp', $G(g^e, g') >$ thesholdg', $G(d^e, d') >$ thesholdd',
4. Add p' to candidate set $C_S$, add g' to candidate set $C_{S1}$, add d' to candidate set $C_{S2}$
5. For each protein terms p' in $C_S$, g' to candidate set $C_{S1}$, d' to candidate set $C_{S2}$
6. Compute similarity between two protein ($p^e, p'$) with $p^e$, similarity between gene terms ($g^e, g'$) with $g^e$ and two disease ($d^e, d'$)
7. Reject p' if the similarity below the threshold t', reject g' if the similarity below the threshold and reject d' if the similarity below the threshold t'
8. Sort remaining proteins, genes and diseases by the similarity and return the top most information that related to concept it is represented as K.

Information Indexing

Decisive to reduce the complexity of the classification process, the information indexing method is obligatory for extracted information. In order to obtain a fixed and appropriate index terms belonging to documents, a greedy strategy is applied.

Text Classification

To select the related or interacted entities that lay emphasis on user given document, text classification is applied to minimize the effort automatically. Classify the
list of indexing terms obtained from analysis phase into several categories of gene-gene relationship, protein-protein interaction and gene-disease relationship. Thus, MKL-SVM method functions on a gradient descent optimization algorithm to classify the indexing information results into classes [102].

Text Clustering

Entropy Agglomeration (EA) is exploited for clustering the classified results [3]. The classified classes based on indexed information are clustered into similar related group of the entities.

3.4 Summary

This chapter elucidates the methodology and algorithms that persist in the functions of NER and RE. Thus, the proposed named entity recognition and named entity relationship methodology extract the information from biotext corpus in efficient manner on improving the information extraction results and solves issues of relation extraction (RE), named entity recognition for biomedical researchers.