CHAPTER 4

FORMULATION OF NAMED ENTITY RECOGNITION

4.1 Introduction

The most essential is recognition of biological entities. There is an ever growing demand for entity recognition methods among biomed corpus owing to rapid publications of documents in open databases. In parallel, the increasing variation of biological terms needs syntactical extraction of information by a great effort. It imposes on imperative function on extraction. Named Entity Recognition (NER) is particularly to mine the textual data and identify concepts of interest in the text by mapping all relevant words and phrases to a set of predefined categories. Thus, Information Retrieval System (IRIS) has been developed for successive NER in interpretation of the knowledge discovery from biomed corpus.

Successive NER methods necessitate syntactic structure among entities and get rid of ambiguity within text. Its evolution derives knowledge from source by initiating tasks and proclivity. It adopts the input text from user and assesses the output within the structured, unstructured and semi-structured data.

It aims at identifying special kind of data from domain-specific document collections. Thus, automatic IRIS access is in demand on robust to sort out applicable and significant information which in turn, renewal of terms and information across biomed corpus are done. Knowledge acquisition work has been carried out to assists knowledge navigation in essence acquirement from biomedical literature for obtaining desired facts.
4.1.1 Challenges of Named Entity Recognition

NER are being standstill challenges to the users due to forthcoming facts. Handling of unknown terms, long compounding of terms are major concerns. On the whole, biological entity recognition is exigent on typical heterogeneity names and invariable changes in molecular biology. NE does not perceive identical text of entities but discovers source on scientific queries. The most basic obstacle results from the dynamic nature of scientific discovery. They are unlikely to be complete at any given moment, resulting in some synonymy relationships that may not be captured.

In the biomedical domain [28], there exists a vast amount of semantically relevant entities that is constantly and rapidly increasing as new scientific discoveries. In biomedical literature, the same concept may be expressed using different words, despite being expressed differently. When many synonyms for a particular concept are in use, it becomes difficult to integrate knowledge from multiple sources without a comprehensive synonymy resource such as the UMLS Meta-thesaurus or Gene Ontology. Finally, the abundant use of acronyms and abbreviations among entities, cascaded and embedded entities has to handle carefully while merging all text documents in collection.

4.1.2 Categories of Named Entity Recognition

The biomedical domain on its complicated nature stands in the wake of fact variations. Pertinent entity extraction paves the intention of discovery chores stand forth of crucial on vast crushing textual information. Much of these efforts are geared based on extraction of gene, protein and disease entity terms. This stacking category of entity framework helps in knowledge discovery on biomed corpus.
The objectives of categorization of the named entities such as gene, protein or disease names are shown in figure 4.1.

**Figure 4.1 Categories of Entities**

### 4.1.3 Key Components of Named Entity Recognition

Information Retrieval System framework holds close and includes the tasks as Information Retrieval, Information Extraction, Indexing, Classification and Clustering. Text Mining in the biomedical field spotlights on two tasks [77]: a) Named Entity Recognition, the task to recognize various types of biological entities and b) Relation Extraction, the task to discover relationship among named entities.

IRIS induces a new hypothesis not only in literature but by combining culled text fragments. User information need is of static and information seeking process is of iterative process and therefore the goals reallocate on dilemma source. Biomedical information search passes on to slanting techniques which inquire advance access on information archives.
The key components and tasks grounding the structured information used in this framework are as:

- **Corpora**: annotated abstracts and articles are the most used corpus type
- **Pre-processing**: sentence splitting, tokenization and annotation encoding are fundamental for input data processing
- **Information Retrieval (IR)**
- **Information Extraction (IE)**
- **Information Indexing**
- **Text Classification**
- **Text Clustering**

With the explosion of specific notations on biological provinces, stipulations of automated extraction techniques are being necessitated. In these processes, biomedical Named Entity Recognition is the core step to access the higher level of information [88].

### 4.2 Internal Representation of Named Entity Recognition

Information seeking process in NER is an iterative process that does refining a query until it receives all and only those documents relevant to the original information need [67]. The current research gets rid of complex biological systems and gives edges off the association with meaningful biological function. It is elusive and highly dependent on the specific underlying scientific question and overwhelming textual information. Hence, NER formulates the flexible and general approach for integrating heterogeneous entities and knowledge sources for discovery. It locates boundary of entity mentioned in a text and tags them with their corresponding semantic types.
Internal representation of NER is shown in figure 4.2.
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 [75]. Named Entity Recognition module is enacted to accumulate entities on corpus with semantic meaning by constructing up dictionary. It gives off precise identification of relevant passages for curation and facilitates ambiguity removal. It enhances support for relation extraction among entities.

Overall progressive tasks get initiated 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 for ‘Topic Tracking’ to do document classification.

4.2.1 Topic Tracking by Information Retrieval

IR system gets transpired at rapid growth on huge collection of textual data where traditional techniques are feeble to cope. Escalation of technologies is required for enlarging textual data on comprehensive access to hit upon relevant items from biotext databases. IR systems become ubiquitous to overcome the frequent occurrences and lexical variants of the entities. Thus, Vector Space Model is induced for supporting IR [18].

Consider categorizing a set of ‘m’ collected biotext corpus based upon the presence or absence of a list of ‘n’ given query / keyword terms. Under this, an ‘n x m’ term-text corpus matrix is constructed, in which in the matrix; all query vectors ‘q’ represent the weighted frequent entity term and column vector represents the frequency of occurrence of each named entity terms on given corpus. This matrix analysis results in
the resemblance between a query vector and the document vectors contained within the term-document matrix.

Given the Vector Space Model (VSM), a natural measure of similarity arises from the inner product. If both ‘q’ and the columns of ‘A’ have been normalized to unit magnitude, then the inner product between ‘q’ and the $j^{th}$ column vector of $A$ becomes as Eqn. (4.1)

$$q^T a_j = ||q|| ||a_j|| \cos \theta_j = \cos \theta_j$$

(4.1)

where the ‘T’ superscript denotes the transpose. Since all components of $q$ and $A$ are of non-negative, all inner products will evaluate to a value such that $0 \leq \cos \theta_j \leq 1$. If the resulted value is equal to value ‘one’ then it implies, there is a resemblance and indicates a small angle between ‘the query and column vector’ and if resulted value is equal to ‘zero’ then it implies, it is dissimilar. Hence above found result is measured and defined as ‘cosine similarity’ and given in Eqn. (4.2),

$$\cos \theta = q^T A$$

(4.2)

where $\cos \theta$ denotes a row vector and implies the relevance between the biotext query and each column vector of $A$. Then term-text corpus is represented by Singular Value Decomposition (SVD) [69] as given in Eqn. (4.3)

$$A = U \Sigma V^T$$

(4.3)

where $U$ denotes ‘$n \times n$’ and $V$ denotes ‘$m \times m$’ orthogonal matrix (i.e. $V^{-1} = V^T$).
Furthermore, $\Sigma$ is an $n \times m$ diagonal matrix of singular values as given in Eqn. (4.4),

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$  

(4.4)

where and $\sigma_i \equiv \Sigma_{ii}$. Thus, it implies linear dependencies between query terms and biotext corpus them and represented as Eqn.(4.5)

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T$$  

(4.5)

where $u_j$ and $v_j$ denotes the $j^{th}$ columns of $U$ and $V$ and associated singular value $\sigma_j$ weighing each product $u_j v_j^T$ with initiation of low rank approximation and truncation. The non-negligible contribution is given in Eqn. (4.6)

$$A \approx \sum_{j=1}^{L} \sigma_j u_j v_j^T$$  

(4.6)

Moreover, the subspace signified as by the first ‘L’ entity query terms columns of $U$ represents inferring linear dependencies among query terms in the biotext corpus. Hence Information Retrieval brings out the relevant documents by measuring low rank approximation for the given search query of biotext corpus.

### 4.2.2 Word Summarization by Pre-processing

It does parsing on corpus of text as on the query received and the retrieved document in order to identify certain aspects. Pre-processing is necessitated to support the textual revolution and tremendous change in the availability of information [75]. Thus, a structure can significantly simplify the access to a document collection for a user. The first and foremost crucial step is to collect the biomed corpus for data analysis. As such illustration is done with a set of documents and articles from the various
biological domains; they are fetched to deposit in specified information retrieval. The tasks commence for processing and analyzing the multivariate datasets and article consideration before the initiating the discovery of data and errands are carried out. Text pre-processing episode as imperative and crucial work subsists the remaining and enduring tasks.

It can be implemented by the solutions of systematic functions. It does tokenization on the retrieved information texts with a sentence which is in raw text into meaningful units called tokens. The biomedical content is inclusive out of ordinary entities seeing the gene, protein and disease names. Such an entity terms hold close specific typeset as numerals, hyphens, slashes and brackets, and the same entity may cover different lexical variants. It works on tokens to eliminate the symbols and wild card characters by tokenization generalized heuristics. Significance of stop words removal plays a vital role on discrimination of retrieved documents that is specific to datasets. Hence they are eliminated by using heuristics.

Subsequently, a common strategy ‘Case-Folding’ is processed to reduce all letters to lower case. Following that, ‘Lemmatization’ tries to map verb form to the infinite tense and noun to the singular forms for entities. It removes the inflectional endings and returns to the base. Forwarding that, ‘Stemming’ a heuristic process severs the words to fabricate the basic forms of words. Sequentially, the retained words are then analyzed to combine different words that having the same basis to a single one and Word-Sense Disambiguator resolves the meaning of affective words (i.e. nouns, adjectives and verbs) according to their context on using a semantic similarity measure.
4.2.3 Document Inference by Information Extraction

Mere extraction of entities without semantic information fails to give away the investigation research for researchers. Moreover, the development of tools for automatically extracting protein, gene and disease is essential for improving and updating the knowledge databases. Hence IE identifies and extracts a range of specific types of information from texts of interest. Automatic extraction of structured information such as entities and attributes from unstructured sources are incited. The systems rely on a combination of deep linguistic knowledge and richness of annotations obtained from biological resources.

Here IRIS, as a ‘Text Simplification process’, executes local text analysis, discourse analysis and formulate a rapid adaptation to new relations using of representative collection and capture temporal information using of ontology and as a ‘Structured Search process’ it proceeds Document Inference, Terminology Extraction to link the facts to a knowledge base and promote research in discovering facts about entities and promotes in expanding a knowledge source automatically by implementing PLSA-BM25++.

Dictionary Construction

Role of dictionary construction developed by Bhattacharya et al [88] enables practitioners to understand and analyze the large biotext datasets. To build concept dictionaries for annotating a collection of documents from a particular domain, a dictionary ‘D=Dict(C,X)’ is defined as a set of words that describes a concept of genes, diseases and proteins ‘C’ in a pre-processed document ‘X’.
In this method, an online interactive framework is applied, where the user starts with a small set of entity term, which scrutinizes the results, and finally selects or rejects the entity terms from the returned ranking. This iteration process continues until it gets satisfied. With interactive supervision, the user provides positive and negative seed gene or protein 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. The steps of algorithm for dictionary construction are given in earlier section of proposed methodology.

**Ontology Convention**

Recognizing named entities semantically, a crucial task unlocks the information stored in unstructured text into particular categories of domain. The grip focus on dictionary construction is done by referring to gene ontology and disease annotation. The topical growth of progress is due to fact variants which are in idiosyncratic nature like alias name, distinctive naming traditions, abbreviations, assortment of organisms may allude to a same entity with diverse terms, or a term may allude to distinctive naturally diverse entities [105]. The most basic obstacles result from the dynamic scenario of scientific discovery. This ever-growing list of relevant terms is problematic since it relies only on a dictionary of known terms or other curated resources to identify named entities. The growth of such resources can never be complete as long as scientific progress continues. As such in this case, the synonymy resources like UMLS Meta thesaurus and Gene Ontology integrate knowledge from multiple sources.

Gene Ontology (GO) is a collaborative effort to address the need for consistent descriptions of gene and protein in different databases. The Gene Ontology [114] is a
controlled vocabulary for standardizing the description of gene and its attributes, protein and its attributes across species and databases. To advancement of ontology, the dictionary has been constructed with full meta-data evidences.

Disease Ontology (DO) annotation [105], a human disease ontology provides consistent, reusable and sustainable descriptions of human disease terms, phenotype characteristics and related medical vocabulary through collaborative efforts of researchers. The DO is successfully used to build a detailed disease phenotype from data available in the electronic medical record.

In this work, GO annotation and DO annotation are utilized in a precise identification of gene, protein and disease entities.

**Document Linkage by PLSA-BM25++**

PLSA-BM25++ method is of assistance in recognizing entity on corpus. It passes to dictionary constructed for particular entry term on retrieved documents. This in turn, the articles fetched as result are with highlighted entities in around of terms with fact variants. Thus, PLSA-BM25++ gives document linkage to entity terms over biomed corpus.

Initially a simple WSD errand [27] was performed for the Named Entity Recognition task. It resolves the limits of the NEs and categorizing the entities into classes such as gene class, protein class and disease class. In it ambiguity basis is inconsistent and follows unreal naming principle. Further, capitalization and other surface clues are not solid markers of elements in the field. Also considerations of the semantic information of the entities are not performed.
Later a concept proposed is PLSI, which is depicted to recognize the entity conceptually. It represents the biotext in a unified way for exposing similarity functions and solving the ambiguity problems while retrieving. But existence of perplexing name encircling in biomed corpus is unachieved.

Later, Probabilistic Latent Semantic Analysis [68][35] method is deemed to perform the Named Entity Recognition utilizing and exploiting the semantic information of the gene, protein and disease. But PLSA term relationships have only been calculated using the raw frequency counts of each entity. Factors such as raw frequency count and term rarity unfavours the recognition results by firming irrelevant entity.

In tune to overcome the above crisis of dimensionality consideration and irrelevant entity recognition results problem, a novel weighting scheme BM25 is integrated to PLSA is developed and named as PLSA-BM25++. It improves the overall quality of results from the retrieval system. The BM25 weighting scheme has a probabilistic background based on the modelling of relevant and irrelevant entity.

Thus, the proposed PLSA-BM25++, a NER algorithm robotically recognizes Named Entities (NE) as given user query related to entities gene, protein and disease in the interest of biomedical domain on biotext corpus from IR step. Consider the retrieved documents containing extracted information as a bag filled with tokens; one token for each event of a words related to gene, protein and disease in the extracted information set. Each token has a related terms identified with gene, protein and disease and semantically important tag label appended from gene and disease ontology results. \( P(d,t) \) is the probability taken out of a token with the document tag label ‘d’ results from gene and disease ontology and entity term ‘t’ associated with it. In this way, if \( f_{d,t} \) tokens are
clinched with names d and t, inferring that term "t" showed up \( f_{d,t} \) times in archive d, then abstain the sample probability as Eqn. (4.6),

\[
P(d, t) = \frac{f_{d,t}}{\sum_{s \in D} \sum_{t \in T} f_{s,T}}
\]

where ‘D’ and ‘T’ are the set of extracted information document and entity terms of gene, protein and disease respectively.

The BM25 weighting scheme [82] has a probabilistic background based on the modelling of relevant and irrelevant document tags. The simplified document scoring is given in Eqn. (4.7)

\[
P(d, Q) = \sum_{t \in Q} w_{d,t} w_{e_t}
\]

where d is document, Q is the set of query terms, \( w_{d,t} \) and \( w_{e_t} \) are the document-entity term and entity terms weights respectively. The named entity terms of the gene, protein and disease weight is calculated as either the log odds of the terms that related to gene, protein and diseases appearing in a tag document as Eqn. (4.8) and its negative log of the probability of the terms appearing in a document is given in Eqn. (4.9),

\[
w_{e_t} = \log \left( \frac{N - f_t + 0.5}{f_t + 0.5} \right)
\]

\[
w_{e_{t+}} = \log \left( \frac{N}{f_t} \right)
\]
where $N$ is the number of tagged documents and $f_t$ is the number of tagged documents containing entity term $t$ related to gene, protein and disease. The entity term weight is utilized to mirror the significance of gene, protein and disease due to its rarity, therefore its weight ought to be high as given in Eqn. (4.10),

$$\text{we}_{d,t} = \frac{(k_1 + 1)f_{d,t}}{K + f_{d,t}}$$  \hspace{1cm} (4.10)

where $f_{d,t}$ is the frequency of named entity term $t$ in tagged name document $d$, $k_1$ is a positive constant, and $K$ is the pivoted document normalization value.

This function has two purposes. The principal is to decrease the impact of vast $f_{d,t}$ values. The second is to standardize the weight of the named entity terms of the gene, protein and disease due to document length.

To utilize the weighted named entity term frequencies, simply substitute the weighted values to raw entity term frequencies of the gene, protein and disease that are found. Once the weight values are connected to every named entity terms of gene, protein and disease based on frequency value, use the PLSA method to obtain the value of $P(d, t)$ and each of its components. Therefore new PLSA relationship becomes as Eqn. (4.11)

$$P(d, t) = \frac{\omega_{d,t}}{\sum_{\delta \in D} \sum_{t \in T} \omega_{\delta,t}}$$  \hspace{1cm} (4.11)
The precise methodology comprises the following steps as:

**Methodology Steps: Document Linkage Utilizing Dictionaries**

1. The retrieved documents are given as input to Extraction step.
2. Ontology reference with GO and UMLS find out the fact variants for a query term.
3. Dictionary construction gives out the entities with fact variants.
4. PLSA-BM25++ method brings out the documents in prior and preference with semantic information and eliminates the ambiguity.

**4.2.4 Information Indexing**

As extracted information related to Named Entity Recognition is of more different words, choice on opt and most informative entity one for a specific classification chore becomes crucial. The main functionalities of Information indexing is inclusive of 5N’s,

- Nature to reduce the number of irrelevant words.
- Nature to reduce the complexity of the classification problem.
- Nature to identify the top relationship among entities.
- Nature to verify and identify the category of the disease for a particular user.
- Nature to ease the biomedical researchers to verify the correctness of the information indexed results based on category next time.

Hence for extracted information, Information Indexing works on Filtering based indexing which removes words from the dictionary and form the document (in frequents). To sort out and to obtain a fixed number of index terms that appropriately cover all information about Named Entity Recognition documents, a simple greedy strategy is
applied [56]. It performs on highest relative entropy for each word $k$ in the vocabulary of the entropy and is computed as Eqn. (4.12) and in Eqn. (4.13),

$$W_k = 1 + \frac{1}{\log_2 n} \sum_{j=1}^{n} p_{jk} \log_2 p_{jk}$$  \hspace{1cm} (4.12)

$$p_{jk} = \frac{t f_{jk}}{\sum_{l=1}^{n} t f_{lk}}$$  \hspace{1cm} (4.13)

where $t f_{jk}$ is the frequency of named entity related word $k$ to gene, protein and disease in extracted Named Entity Recognition document $j$, and $n$ is the number of documents in the collection related to NER. Here the entropy gives a measure of how well a word is suited to separate NER related documents into several indexed names. For instance, words that occur in many documents will have low entropy.

Hence it is not considered as index terms to separate the NER based gene, protein and disease documents into several ways. Index words choose a number of words that have high entropy relative to their overall frequency. Decisive to reduce the complexity of the classification process, the Information Indexing method is much obligatory for extracted information.

This procedure has empirically been found to yield a set of relevant words that are suited as index terms. Its functionalities are evaluated and support classification and clustering progressive tasks.
The step by step methodology is given below as:

**Methodology Steps: Information Indexing**

1. To obtain a fixed number of index terms as categories of entities, a greedy strategy was applied.
2. From the first NER related documents in terms of entities, select the term with the highest relative entropy as an index term.
3. Then mark this selected entity related document with indexed term.
4. From the first document of the remaining unmarked documents, select again the index term with the highest relative entropy as an index term.
5. The process can be terminated when the desired number of index terms has been selected for NER related documents.
6. Consequently indexed information results are fed into classifier.

### 4.2.5 Text Classification

The classification of biological literature becomes crucial on substantial corpus for progressive upgrading by curators. Hence, Text Classification using MKL-SVM approach [102] is implemented to get optimized biomed corpus on selecting just the applicable articles to a given undertaking classes as protein, gene and disease. The arrangement technique of classification depends on filed data and their indexing data results put away in databases. MKL-SVM approach brings rapid convergence and favourable efficiency results.

Given a indexed information of ‘N’ independent and identical training instances \(\{(x_i, y_i)\}_{i=1}^{N}\) from NER samples, SVM hit upon the direct discriminant induced by the
mapping function $\Phi: \mathbb{R}^D \rightarrow \mathbb{R}^S$, where $x_i$ is the D-dimensional input of NER sample vectors and $y_i \in \{-1, +1\}$ is its class label of the gene, protein and disease either yes($+1$) or no class ($-1$).

Thus, the resulting discriminant function is given in Eqn. (4.14),

$$f(x) = \langle w, \Phi(x) \rangle + b$$

(4.14)

Then classifier is given in Eqn. (4.15)

$$\text{minimize} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i$$

(4.15)

With respect to $w \in \mathbb{R}^S, \xi \in \mathbb{R}_+^N, b \in \mathbb{R}$ subject to function as Eqn. (4.16)

$$y_i((w, \Phi(x_i)) + b) \geq 1 - \xi_i \forall i$$

(4.16)

where ‘$w$’, vector of weight coefficients for indexed information of NER, $C \in \{-1, +1\}$ and $b$ is the bias term. Instead of solving the above optimization problem directly, the Lagrangian dual function enables to obtain the following dual formulation in Eqn. (4.17). With respect to $\alpha \in \mathbb{R}_+^N$ subject to $\sum_{i=1}^{N} \alpha_i y_i = 0, C \geq \alpha_i \geq 0 \forall i$

Minimize $\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\Phi(x_i), \Phi(x_j)), k(x_i, x_j) = \Phi(x_i), \Phi(x_j)$

(4.17)

where $k: \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ is named the kernel function and ‘$a$’ is the vector of dual variables corresponding to each separation constraint. Solving Eqn. (4.17) the result becomes as Eqn. (4.18) as

$$w = \sum_{i=1}^{N} \alpha_i y_i \Phi(x_i)$$

(4.18)
Thus, the discriminant function is given in Eqn. (4.19)

\[
f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b
\]  

(4.19)

Here in this developed IRIS system, MKL approach is utilized by using multiple kernels instead of one specific kernel function and so its corresponding parameters becomes as Eqn. (4.20)

\[
k_\eta(x_i, x_j) = \{k_m(x_i^m, x_j^m)\}_{m=1}^{n}
\]  

(4.20)

Kernel function \(\{k_m : \mathbb{R}^{D_m} \times \mathbb{R}^{D_m} \rightarrow \mathbb{R}\}_{m=1}^{n}\).

Considering ‘n’ indexed information (4.23) of NER sample \(x_i = \{x_i^m\}_{m=1}^{p}\) and its dimensionality factor ‘\(D_m\)’ where \(x_i^m \in \mathbb{R}^{D_m}\) in a kernel function, \(H\) parameterizes the combination function to Eqn. (4.21) and Eqn(4.22).

\[
k_\eta(x_i, x_j) = \{k_m(x_i^m, x_j^m)\}_{m=1}^{n}
\]  

(4.21)

\[
f(x) = \sum_{i=1}^{N} \alpha_i y_i k_\eta(x_i, x_i) + b
\]  

(4.22)

where ‘\(\eta\)’ is the weight of kernel function. By applying MKL SVM approach, the retrieval of classified results on a particular class entity is done that is shown in an example.

### 4.2.5 Text Clustering

Clustering of entities is being an imperative research. It is persuaded by the vast number of natural articles that users need to peruse or essentially to know about advancement in a particular region. Clustering discovers the classified user with common
entities and by resting the extracted information documents with the most common words into the same groups.

Even though the classified results give the extricated indexed information results, it is difficult to find similar user categories for the classified data for inter clustering the huge data users. Hence, Entropy Agglomeration (EA) is exploited for clustering the classified results.

Figure 4.3: Text Representation by a Feature Allocation

Entropy Agglomeration clustering technique clusters the commonly profoundly comparable NEs. The consisted NE hits upon the diverse conceivable annotations of NE and formulates the cluster with exact annotations. In that case, it is alluring to abstain from fixing the number of clusters as shown in figure 4.3.

This data of characterized gene, protein, disease related data in the content records as graphically shown in bipartite diagram, where word elements below are linked to the classes’ information above them. Text corresponding to entities with ‘n’ distinct terms is denoted by its aspects of entity class.
A feature allocation of elements (gene, protein, and disease class) is given by \([n] = \{1, 2, \ldots, n\}\), and its multi-set blocks are given by \(F = \{B_1, \ldots, B_{|F|}\}\), where \(B_i \subseteq [n]\) and \(B_i \neq 0\) for all \(i \in \{1, \ldots, n\}\). These blocks \(\{B_1, \ldots, B_{|F|}\}\) denote the classified results information of gene, protein and disease of the NER text results. Entropy measures the components regarding the blocks of \(F\). It gets to be most extreme at the blocks that incorporate about portion of the components. Spotlighting on the specific subset, Projection of gene, protein and disease features onto ‘S’ limits the extent of ‘F’. Projection Entropy (PE) works out the segmentedness for specific components of gene, protein and disease subset, by anticipating \(F\) onto it. In it, low PE subset implies components comprise entropic connection. Number of subsets computes the projection size \(|\text{PROJ}(F, \{S\})|\) of an element ‘S’.

Hence, Entropy is given as in Eqn. (4.23), Eqn. (4.24) and Eqn. (4.25) as

\[
\text{Entropy} = H(F) = \sum_{i=1}^{|F|} \frac{|B_i|}{n} \log \frac{n}{|B_i|}
\]  

\[
\text{PROJ}(F, S) = \{B \cap S\}_{B \in F \{\emptyset\}}
\]

\[
\text{Projection Entropy : } H(\text{PROJ}(F, S))
\]

The above stride of the calculation still contains an excess of words. At that point, the categories of gene, protein and disease content with ‘n’ distinct words are anticipated onto each of these word sets, and EA keeps running on each of this anticipated gene, protein and disease content with \(n\) particular words of features allocations. The sequential steps included in EA algorithm are as follows as:
Methodology Steps: Entropy Agglomeration

1. Initialize $\Psi \leftarrow \{[1], [2], \ldots, [n]\}$

2. Finds the subset of ‘gene, protein and disease’ user files pairs $\{S_a, S_b\} \in \Psi$

3. Update $\Psi \leftarrow (\Psi \{S_a, S_b\} \cup \{S_a \cup S_b\})$

4. if $|\Psi| > 1$ then go to 2

5. Generate the dendrogram of chosen pairs by plotting minimum entropies for every bifurcation

Thus, EA generates a dendrogram for each word set to show the entropic connections among its entities as clusters. Therefore, the clustered results among entities are detailed in example.

4.3 Exemplification of Named Entity Recognition

For an example, input term given by the user is entity gene, “FOXH1”. By parallel select the category under which class to extricate its semantic information. At once, the keyword reaches the collected biomed corpus; it instigates to do pre-processing steps ‘tokenization, stop words removal, morphological analysis and WSD’.

Here few paraphrases of four articles such as Article - 1 (d1), Article - 2 (d2), Article - 3 (d3) and Article - 4 (d4) from collected biomed corpus are selected randomly and executed the progressive tasks of IRIS and exhibited for sample output.

Article – 1 (d1)

The Forkhead Box H1 (FoxH1) protein is a co-transcription factor recruited by phosphorylated Smad2 downstream of several TGFbetas, including Nodal-related
proteins. We have reassessed the function of zebrafish FoxH1 using antisense morpholino oligonucleotides (MOs). MOs targeting translation of FoxH1 disrupt embryonic epiboly movements during gastrulation and cause death on the first day of development. The FoxH1 morphant phenotype is much more severe than that of zebrafish carrying foxh1/schmalspur (sur) DNA-binding domain mutations, FoxH1 splice-blocking morphants or other Nodal pathway mutants, and it cannot be altered by concomitant perturbations in Nodal signaling. Apart from disrupting epiboly, FoxH1 MO treatment disrupts convergence and internalization movements.

Late gastrula stage FoxH1 morphants exhibit delayed mesoderm and endoderm marker gene expression and failed patterning of the central nervous system. Probing FoxH1 morphant RNA by microarray, we identified a cohort of five keratin genes--cyt1, cyt2, krt4, krt8 and krt18--that are normally transcribed in the embryo's enveloping layer (EVL) and which have significantly and reduced expression in gene FoxH1-depleted embryos. Simultaneously disrupting these keratins with a mixture of MOs reproduces the FoxH1 morphant phenotype. Our studies thus point to an essential role for maternal FoxH1 and downstream keratins during gastrulation that is epistatic to Nodal signalling.

**Tokenization**

It splits the text of a document into a sequence of tokens. There are several options to define the splitting points. Depending on the mode, split points are chosen differently. The range chosen are as non letters; specifying characters, regular expression and setting the default value as non letters. For the above defined steps, some of the tokens depending on the constraints implied to paraphrase are given in table 4.1.
### Table 4.1 Result - Token Representation

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>The, disrupting, these, keratins, with, a, mixture and etc</td>
<td>// Word</td>
</tr>
<tr>
<td><strong>FoxH1, Smad2</strong></td>
<td>// protein name or gene name</td>
</tr>
<tr>
<td>cyt1, cyt2, krt4, krt8 and krt18</td>
<td>// gene types</td>
</tr>
<tr>
<td>TGFbetas,</td>
<td>// protein type</td>
</tr>
<tr>
<td>MOs, DNA, EVL, RNA</td>
<td>// ABBR</td>
</tr>
<tr>
<td>Gastrulation, morphant phenotype, zebrafish, TGFbetas</td>
<td>// BIOTERMS</td>
</tr>
<tr>
<td>--</td>
<td>// SYMBOLS</td>
</tr>
</tbody>
</table>

The outputs of the tokenization for first two lines are as in table 4.2.

### Table 4.2 Tokenized Paraphrases

<table>
<thead>
<tr>
<th>Input</th>
<th>Tokenization results</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Forkhead Box H1 (<strong>FoxH1</strong>) protein is a co-transcription factor recruited by phosphorylated <strong>Smad2</strong> downstream of several TGFbetas, including Nodal-related proteins.</td>
<td>The // Word Forkhead Box H1 (<strong>FoxH1</strong>) protein // gene name is // word a // word co-transcription // word factor // word recruited // word by // word phosphorylated // word <strong>Smad2</strong> // gene name downstream // word of // word several // word TGFbetas // BIOTERMS, including // word Nodal related // word proteins // protein type.</td>
</tr>
<tr>
<td>We have reassessed the function of zebrafish <strong>FoxH1</strong> using antisense morpholino oligonucleotides (MOs). MOs targeting translation of <strong>foxH1</strong> disrupt embryonic epiboly movements during gastrulation and cause death on the first day of development.</td>
<td>We // word have // word reassessed // word the // word function // word of // word zebrafish // word <strong>FoxH1</strong> // gene name using // word antisense // BIOTERMS morpholino // BIOTERMS oligonucleotides // BIOTERMS (MOs) // ABBR. MOs // ABBR targeting // word translation // word of // word <strong>foxH1</strong> // gene name disrupt embryonic // BIOTERMS epiboly // BIOTERMS movements // word during // word gastrulation // BIOTERMS and // word cause // word death // word on // word the // word first day // word of // word development // word</td>
</tr>
</tbody>
</table>
**Stemming**

Stemming in information retrieval task describes the process for reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form. The stem need not be an identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Sample terms are given in table 4.3.

<table>
<thead>
<tr>
<th>Word</th>
<th>Stemming</th>
<th>Word</th>
<th>Stemming</th>
</tr>
</thead>
<tbody>
<tr>
<td>co-transcription</td>
<td>Cotranscription</td>
<td>Disrupting, disrupts</td>
<td>Disrupt</td>
</tr>
<tr>
<td>Recruited</td>
<td>Recruit</td>
<td>Movements</td>
<td>Movement</td>
</tr>
<tr>
<td>Phosphorylated</td>
<td>Phosphorylate</td>
<td>Morphants</td>
<td>Morphant</td>
</tr>
<tr>
<td>Reassessed</td>
<td>Reassess</td>
<td>Delayed</td>
<td>Delay</td>
</tr>
<tr>
<td>Including</td>
<td>Include</td>
<td>Failed</td>
<td>Fail</td>
</tr>
<tr>
<td>Altered</td>
<td>Alter</td>
<td>Patterning</td>
<td>Pattern</td>
</tr>
<tr>
<td>Targeting</td>
<td>Target</td>
<td>Identified</td>
<td>Identify</td>
</tr>
<tr>
<td>Perturbations</td>
<td>Perturbation</td>
<td>Transcribed</td>
<td>Transcribe</td>
</tr>
<tr>
<td>Signalling</td>
<td>Signal</td>
<td>Reduced</td>
<td>Reduce</td>
</tr>
</tbody>
</table>

Reducing variant word forms to a single "stem" forms “-’s, -ing, -ed, -s; in-, ad-, pre-, sub- etc.” Therefore, fancy suffix removal and affix removal heuristics are applied to perform stemming.
The above mentioned steps are applied and observed result is shown in table 4.4.

**Table 4.4 Stemming Representation - 2**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Stemming results</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Forkhead Box H1 (FoxH1) protein is a cotranscription factor recruited by phosphorylated Smad2 downstream of several TGFbetas, including Nodal-related proteins.</td>
<td>The Forkhead Box H1 (FoxH1) protein is a cotranscription factor recruit by phosphorylate Smad2 downstream of several TGFbeta include Nodal proteins.</td>
</tr>
<tr>
<td>We have reassessed the function of zebrafish FoxH1 using antisense morpholino oligonucleotides (MOs).</td>
<td>We have reassess the function of zebrafish FoxH1 using antisense morpholino oligonucleotide (MOs).</td>
</tr>
</tbody>
</table>

**Stop words removal**

Stop words are words which are filtered out before or after processing of natural language data (text). The result with eliminated stop words is given in table 4.5.

**Table 4.5 Eliminated Stop Words Representation**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Stop words results</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Forkhead Box H1 (FoxH1) protein is a cotranscription factor recruit by phosphorylate Smad2 downstream of several TGFbeta include Nodal proteins.</td>
<td>Forkhead Box H1 (FoxH1) phosphorylate Smad2 TGFbeta.</td>
</tr>
<tr>
<td>We have reassessed the function of zebrafish FoxH1 using antisense morpholino oligonucleotides (MOs).</td>
<td>zebrafish FoxH1 antisense morpholino oligonucleotide (MOs)</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 words.

Word-Sense Disambiguator

WSD turns down the nominal words in accordance with their base and semantic meaning such as noun, adjective and verb using of WordNet ontology. By then, synonym terms are surfaced for forthcoming feature set.

Article-2 (d2)

Bistability in developmental pathways refers to the generation of binary outputs from graded or noisy inputs. Signaling thresholds are critical for bistability. Specification of the left/right (LR) axis in vertebrate embryos involves bistable expression of transforming growth factor beta (TGFbeta) member NODAL in the left lateral plate mesoderm (LPM) controlled by feed-forward and feedback loops. Here we provide evidence that bone morphogenetic protein (BMP)/SMAD1 signaling sets a repressive threshold in the LPM essential for the integrity of LR signaling. Conditional deletion of Smad1 in the LPM led to precocious and bilateral pathway activation. NODAL expression from both the left and right sides of the node contributed to bilateral activation, indicating sensitivity of mutant LPM to noisy input from the LR system. In vitro, BMP signaling inhibited NODAL pathway activation and formation of its downstream SMAD2/4-FOXH1 transcriptional complex. Activity was restored by overexpression of SMAD4 and in embryos, elevated SMAD4 in the right LPM robustly activated LR gene expression, an effect reversed by superactivated BMP signaling.
We conclude that BMP/SMAD1 signaling sets a bilateral, repressive threshold for NODAL-dependent Nodal activation in LPM, limiting availability of SMAD4. This repressive threshold is essential for bistable output of the LR system.

**Article-3 (d3)**

Abnormalities of embryonic patterning are hypothesized to underlie many common congenital malformations in humans including congenital heart defects (CHDs), left-right disturbances (L-R) or laterality, and holoprosencephaly (HPE). Studies in model organisms suggest that Nodal-like factors provide instructions for key aspects of body axis and germ layer patterning; however, the complex genetics of pathogenic gene variant(s) in humans are poorly understood. Here we report our studies of FOXH1, CFC1, and SMAD2 and summarize our mutational analysis of three additional components in the human NODAL-signaling pathway: NODAL, GDF1, and TDGF1. We identify functionally abnormal gene products throughout the pathway that are clearly associated with CHD, laterality, and HPE. Abnormal gene products are most commonly detected in patients within a narrow spectrum of isolated conotruncal heart defects (minimum 5%-10% of subjects), and far less commonly in isolated laterality or HPE patients (approximately 1% for each). The difference in the mutation incidence between these groups is highly significant. We show that apparent gene dosage discrepancies between humans and model organisms can be reconciled by considering a broader combination of sequence variants. Our studies confirm that (1) the genetic vulnerabilities inferred from model organisms with defects in Nodal signaling are indeed analogous to humans; (2) the molecular analysis of an entire signaling pathway is more complete and robust than that of individual genes and presages future studies by whole-genome
analysis; and (3) a functional genomics approach is essential to fully appreciate the complex genetic interactions necessary to produce these effects in humans.

**Article- 4 (d4)**

FoxH1, a member of the Forkhead-box (FOX) gene family of transcription factors, takes part in mediating transforming growth factor-beta/activin signaling through its interaction with the Smad2, Smad4 complex. Using a series of experiments, we found that FoxH1 repressed both ligand-dependent and -independent transactivation of the AR on androgen-induced promoters. This action of FoxH1 was independent of its transactivation capacity and activin A but relieved by Smad2,Smad4. In addition, the repression of the AR by FoxH1 did not require deacetylase activity. A protein-protein interaction was identified between the AR and FoxH1 independently of dihydrotestosterone. Furthermore, a confocal microscopic analysis of LNCaP cells revealed that the interaction between the AR and FoxH1 occurred in the nucleus and that FoxH1 specifically blocked the foci formation of dihydrotestosterone-activated AR, which has been shown to be correlated with the AR transactivation potential.

Taken together, our results indicate that FoxH1 functions as a new corepressor of the AR. Our observations not only strengthen the role of FoxH1 in AR-mediated transactivation but also suggest that therapeutic interventions based on AR-coregulator interactions could be designed to block both androgen-dependent and independent growth of prostate cancer.

The results of above articles d2, d3, d4 after pre-processed steps are shown in table 4.6, 4.7 and 4.8 respectively.
### Table 4.6 Pre-processed Representation of d2

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Pre-processed results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormalities of embryonic patterning are hypothesized to underlie many common congenital malformations in humans including congenital heart defects (CHDs), left-right disturbances (L-R) or laterality, and holoprosencephaly (HPE).</td>
<td>Abnormal embryo pattern hypothesize congenital malformation human congenital heart defect CHD leftright disturbance LR lateral holoprosencephaly HPE</td>
</tr>
<tr>
<td>Here we report our studies of FOXH1, CFC1, and SMAD2 and summarize our mutational analysis of three additional components in the human NODAL-signalling pathway: NODAL, GDF1, and TDGF1.</td>
<td>report study FOXH1 CFC1 SMAD2 summarize mutation analysis components human NODAL signalling pathway NODAL GDF1 TDGF1</td>
</tr>
</tbody>
</table>

### Article-3 (d3)

### Table 4.7 Pre-processed Representation of d3

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Pre-processed results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Here we provide evidence that bone morphogenetic protein (BMP)/SMAD1 signaling sets a repressive threshold in the LPM essential for the integrity of LR signaling. Conditional deletion of Smad1 in the LPM led to precocious and bilateral pathway.</td>
<td>provide evidence bone morphogenetic protein BMP SMAD1 signal set repressive threshold LPM essential integrity LR signal Condition delete Smad1 LPM led precocious bilateral pathway</td>
</tr>
<tr>
<td>NODAL expression from both the left and right sides of the node contributed to bilateral activation, indicating sensitivity of mutant LPM to noisy input from the LR system.</td>
<td>NODAL expression left right side node contribute bilateral activate indicate sense mutant LPM noise input LR system</td>
</tr>
</tbody>
</table>
**Table 4.8 Pre-processed Representation of d4**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Pre-processed results</th>
</tr>
</thead>
<tbody>
<tr>
<td>FoxH1, a member of the Forkhead-box (FOX) gene family of transcription</td>
<td>FoxH1 member Forkheadbox FOX gene family transcript factor part</td>
</tr>
<tr>
<td>factors, takes part in mediating transforming growth factor-beta/activin</td>
<td>mediate transform growth factor beta activin signal interact</td>
</tr>
<tr>
<td>signaling through its interaction with the Smad2.Smad4 complex</td>
<td>Smad2 Smad4 complex</td>
</tr>
<tr>
<td>Using a series of experiments, we found that FoxH1</td>
<td>series experiment FoxH1 repress ligand transactivate AR androgen</td>
</tr>
<tr>
<td>repressed both ligand-dependent and independent transactivation of the</td>
<td>induce promoter</td>
</tr>
<tr>
<td>AR on androgen-induced promoters.</td>
<td></td>
</tr>
</tbody>
</table>

**Information Retrieval**

Later on, IR system gets transpired on pre-processed documents by Vector Space Model. It gives comprehensive access to hit upon relevant items from biotext databases. IR systems became ubiquitous to overcome the frequent occurrences and lexical variants of the entities Support Vector Machine.

In the above example, the stemmed results are converted into the matrix form as

\[
\begin{bmatrix}
\text{documents} & \text{FoxH1} & \text{SMAD1} & \text{holoprosencephaly} & \text{SMAD2} \\
\hline
d1 & 1 & 0 & 1 & 1 \\
d2 & 1 & 0 & 1 & 0 \\
d3 & 1 & 1 & 1 & 0 \\
d4 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\]

As from the mentioned documents d1, d2, d3, d4, d5 etc., it retrieves all four documents from the corpus since it contains most relevant to user input. Followed by that
moment, information extraction phase gets generated to fetch the documents with semantic information.

**Information Extraction**

In Information Extraction phase, PLSA-BM25++ method is deemed to perform the Named Entity Recognition utilizing dictionary constructed based on ontology reference. It exploits documents with the semantic information without ambiguity. Here term recognition have been calculated using the frequency counts of each entity from dictionary unlike factors using raw frequency count and term rarity that wavering the recognition results by firming irrelevant entity.

Sample output document with high priority is given in table 4.9.

**Table 4.9 Extracted paraphrase**

FoxH1 (also known as FAST) is a forkhead or winged-helix DNA-binding protein that was initially identified by its ability to bind to an activin response element in the promoter region of the Xenopus Mix.2gene (Chen et al. 1996). One FoxH1 homolog has been identified in human (also known as FAST2 or FoxH1a), as well as in human (FoxH1) and zebrafish (schmalspur) (Labbé et al. 1998; Zhou et al. 1998; Pogoda et al. 2000; Sirotkin et al. 2000). FoxH1 forms complexes with heteromers of Smad4 and either phosphorylated Smad2 or Smad3 (Labbé et al. 1998; Hoodless et al. 1999). Binding of this complex to DNA is stabilized by Smad4 contact with DNA at Smad-binding sites that lie adjacent to the FoxH1-binding site. Interestingly, FoxH1 does not contain a transcriptional activation domain and requires Smad interaction for transcriptional regulation, probably through the recruitment of transcriptional cofactors (Attisano and Wrana 2000). In the human, FoxH1 is expressed throughout the epiblast prior to and during gastrulation (E6.0–E7.5), with low levels detected in the extraembryonic endoderm (Weisberg et al. 1998). At early somite stages, FoxH1 is expressed bilaterally in the lateral plate mesoderm, and expression is subsequently restricted to the heart (Weisberg et al. 1998; Saijoh et al. 2000). These expression patterns suggest that FoxH1 functions during early embryonic patterning, and in zebrafish loss of FoxH1 causes variable defects in axial structures, indicative of potential functions in modulating nodal signaling in the organizer (Pogoda et al. 2000; Sirotkin et al. 2000). Consistent with this, FoxH1 regulates

**Information Indexing**

An inverted index is an index structure that maintains two hash indexed tables: document table and term table, where document table consists of a set of document records which belongs to the FoxH1, each containing two fields: doc id (FoxH1) and
posting list, where posting list is a list of genes names (or pointers to terms) that occur in the FoxH1 document, sorted according to some entropy measure. It performs on highest relative entropy for each term t in the vocabulary the entropy is computed. Once the entropy find all of the documents associated with a set of genes, first find a list of document identifiers in gene table for each gene, and then intersect them to obtain the set of relevant FoxH1 documents. Forkhead Box H1 (FoxH1)- FoxH1 morphant - antisense morpholino oligonucleotides (MOs)- Probing FoxH1 morphant RNA- five keratin genes--cyt1, cyt2, krt4, krt8 and krt18– Nodal-related proteins- TGFbetas- zebrafish - embryo's enveloping layer (EVL).

Sample indexed paraphrase are shown in table 4.10.

**Table 4.10 Indexed Paraphrases**

<table>
<thead>
<tr>
<th>doc id</th>
<th>posting list</th>
</tr>
</thead>
<tbody>
<tr>
<td>FoxH1</td>
<td>genes names</td>
</tr>
</tbody>
</table>

**Classification**

Once the inverted index is formed, then classification is done depending on the set of document records with two fields: doc id (FoxH1) and posting list. In gene documents list, if the specific document belongs to the FOXH1 with highest entropy values, then support vectors are converted in to those documents. Then those highest entropy values in the FOXH1 is classified as the yes class and other documents with gene list is named as no class. FoxH1 documents., Forkhead Box H1 (FoxH1) FoxH1 morphant - antisense morpholino
oligonucleotides (MOs) Probing FoxH1 morphant RNA - five keratin genes--cyt1, cyt2, krt4, krt8 and krt18– Nodal-related proteins- TGFbetas- zebrafish - embryo's enveloping layer (EVL). In classification task it named into two classes such as FoxH1 - ‘YES’ and Alias names of FoxH1 - ‘NO’

**Clustering**

Once the classification is done with yes and no classes, then inverted indexed in the gene and documents list refers to term list. The term list specifies the most number of details of the gene categories under same category then clustering is performed to same gene. Further, FoxH1 documents with many alias names information are verified and grouped into same classes with different alias names. In clustering process the FoxH1 and alias names related data are grouped into the same class.

**4.4 Summary**

This chapter is with suite of methodologies brings the result with semantic information and eliminates the ambiguities among entities on taken biomed corpus. Thus, the proposed IRIS for NER addresses and resolves the technical issues in representing, analyzing primitive features and variation in the use of natural language expressions. The process of scanning text for information relevant to some interest, including extracting entities, and events are tuned to researchers over here by using of PLSA-BM25++ algorithm. The system incorporated various features of gene, protein, disease and experimented with different strategies for combination of probabilistic methods. To further increase the efficiency of the work, information indexing, classification and clustering tasks are carried out. Information indexing is done by using
Formulation of Named Entity Relationship / Relation Extraction
greedy strategy. Classification is done by MKL-SVM that classifies the NER results into several numbers of classes for named entities such as gene, protein and disease. The classified NER results are grouped by Entropy Agglomeration clustering methods which groups similar user named entity.