International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 5525-5535
© Research India Publications
http://www.ripublication.com

Interpretation of Knowledge Discovery System on formulation of Gene Disease Relationship

[1] Dr. P.Sumathi, [2] K.Prabavathy

[1]Assistant Professor, Government Arts College, Coimbatore-18, TamilNadu, India
sumathirajes@hotmail.com
[2]ResearchScholar in ComputerScience, Manonmaniam Sundaranar University,
Tirunelveli, TamilNadu, India
praba_bud@yahoo.co.in

Abstract

The flare up information is owing to high production of biomed corpus. By which, it detonates on increased availability solutions for storing, organizing, and retrieving the outsized text data. In this intend, Knowledge Discovery System (KDS) is designed to unlock the knowledge stored in natural language of biomed corpus to tune researchers. It extracts named entities like Genes and disease from unstructured textual data through the identification and exploration of interesting patterns and its relationship. It compares different documents , and relevance of the documents by incorporating Gene Ontology (GO) and disease annotation to hit upon patterns and trends across multiple documents. Named Entity Recognition (Gene& Disease) by Probabilistic Latent Semantic Indexing and Analysis with BM25 and stimulating Named Entity Relationship (formulation of Gene Disease relationship) by Logistic Model of dictionary construction method. Testing the GDD algorithm indicates that it perk up better identification of Gene Disease relationship result with all previously suggested methods .

Keywords: Entity, Entity Recognition, Entity Relationship, Information Retrieval, Information Extraction, MKL-SVM Text Classification, EA Text Clustering, Ontology.

Introduction

Information Extraction is the activity of obtaining information resources (Ontology) relevant to an information need from a collection of information resources. The proposed KDS recognizes the term entities and their relationships on free structural biomed text corpus without human intervention. The Challenges remains are dynamic nature of the domain i.e, inclusive of new terms of genes, proteins, disease, chemical compounds and constant updating of biomedical resources. This is a point why we in need to ground and develop a new application to triumph over structural information for formulation of Gene Disease relationship using Logistic Model of dictionary construction method

Analysis of Existing

Since the data are unstructured in nature and dynamic scenario, blind usage of knowledge discovery techniques to retrieve needed documents stand hard-hitting to researchers. In earlier days, subset of the corpus syntactic annotation is employed to compute the performance of the Link Grammar and Connexor Machinese Syntax dependency parsers in the biomedical area was studied[1] [2].

For the past years, biomedical systems have been developed using linguistic characteristics of the word, the orthographic features, the morphological features, and local context features [1][2]. The binary encoding of the feature set is used as input for the machine learning algorithm to train the NER model, along with the human annotation of NE mentions in the training dataset [1]. The gene expression can be employed in designing gene modules, even though these approaches could not contain wide knowledge about the presented gene annotation/function. Thus, these approaches would not be estimated to be maximally efficient in building modules with strong association to biological functions [3].

Proposed Work

Biotext corpus was originally annotated for disease and treatment mentions [4] is part of Biotext Project at UC Berkley. The motivation withheld facilitates the mining of literature for entity recognition and interactions between entities. It is capable to put together a knowledge base for entities. It search, sort, display and cites the wide associations from articles to bring out the functional information. In array of performing thorough investigation for source of information and knowledge discovery, the techniques are diluted in the current approach. To regulate the discovery of potential gene candidates for their role in specific disease-related behavior, further computational analysis and empirical validation are carried out using Logistic Model of dictionary construction method.

Knowledge Discovery System

The entire proposed model is illustrated in Figure 1. To regulate the Entity Recognition and Relation Extraction, the biotext corpus is taken as input. The input text gets hoard in local database. As an initial phase, Information Retrieval is swayed off using Vector Spacing model. Subsequently preprocess steps Tokenization, Stemming, Stop Word Removal, Morphological Analysis, Word Sense Disambiguation to remove the non-functional characters; stop words are performed.

Following, Information Extraction phase are conceded out for Entity Recognition and Relation Extraction with the help of Dictionary based methods.

Named Entity Recognition is taken to identify the set of entities from the extracted text as of biotext corpus, by Probabilistic Latent Semantic Indexing and Analysis method with BM25. Next to this Gene-Disease Relationship is progressed by Logistic Model of dictionary construction method.

The recognized and interrelated Gene and disease entities are indexed by information indexing technique. The indexed terms and entities are classified and clustered by SVM and Entropy clustering algorithm. The clustered entity is visualized and gets occupied as output relationship. The workflow has been as stride up as following steps

Step-1: Fed the biotext corpus/ article as input.

Step-2: Information Retrieval task by Vector Space Model.

Step-3: Preprocessing the retrieved input.

Step-4: Information Extraction task initiated.

Step-5: Subtask-1 Named Entity Recognition by Probabilistic Latent Semantic indexing which refer and compare the Gene Ontology and disease database.

Step-6: Subtask-2 Named Entity Relationship by Logistic Model of dictionary construction method is carried out.

Step-7: Construction of Gene and Disease dictionary.

Step-8: Indexing the entities.

Step-9: Classifying and clustering the entities and its relationship.

Step-10 Extracted output gets reposted.

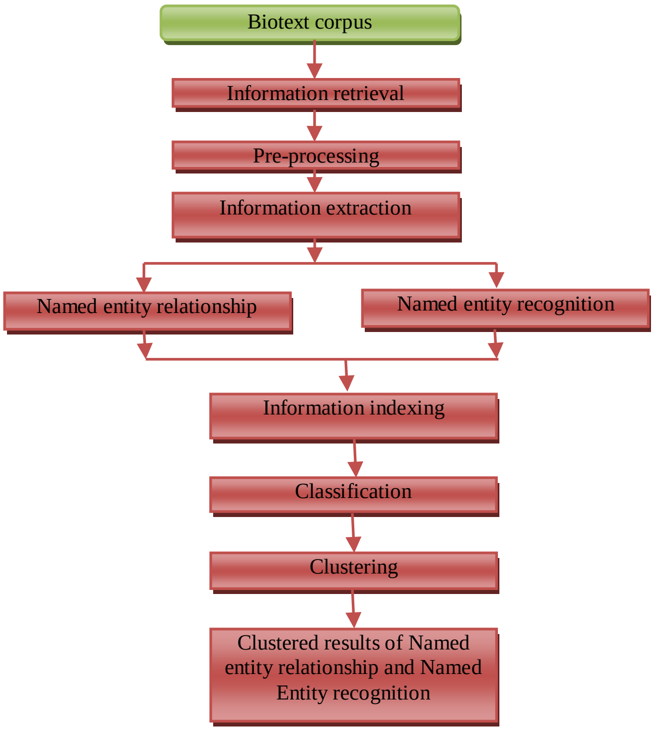


Figure 1 Architecture of KDS

Probabilistic Latent Semantic Analysis for Named Entity Recognition

Named Entity Recognition (NER) refers to the computational method to automatically recognize named entities in natural language documents [1]. It is evidenced by the facts of many aliases, different naming conventions , abbreviations, synonymy and variety of organism referring a same protein/gene with different terms, or a term may refer to different biologically different entities.

Initially a simple WSD method is to determine the boundaries of the NEs and classifies into classes such as genes, disease and proteins. But inconsistent ambiguity tunes to unreal naming principle and unconsiders the semantic information of the gene and disease. In return PLSI is used to depict the concepts verbally. But existence of Perplexing name is being encircling in biomed corpus like Drosophila gene names such as white and cycle as regular denotation. Later, a Probabilistic Latent Semantic Analysis (PLSA) method is deemed to perform NER by exploiting the semantic information of the gene and disease by raw frequency counts. But factors raw frequency counts and term rarity leads to the favour of certain irrelevant entity recognition results.

In tune to overcome the above crisis such as dimensionality consideration and irrelevant recognition results, a novel Weighting schemes BM25 probabilistic background using of Gene ontology and disease database is integrated with PLSA to take away irrelevant recognition. Hence it improves the overall quality of results from the retrieval system.

Logistic Model of Dictionary Construction Method for formation of Named Entity Relationship

The second subtask is relation extraction in midst of complex human disease by way of high transience rates (cancer, heart disease, obesity, diabetes, and common psychiatric and neurological conditions).Genetic and environment aspections are route cause of such defined disease but still proclivity relationships amongst these aspections and disease are unable to state explicitly. So it is tried by Logistic Model of dictionary construction method.

The system initially collects sentences that contain at least one pair of disease and gene names, using the dictionary-based longest matching technique [1]. The system then attempts to extract a binary relation between the disease and gene names in each sentence[1]. Tsuruoka and Tsujii (2003) proposed a dictionary-based longest matching approach for name recognition where they employed a Naive-Bayes classifier to filter out false positives [1]. However, their dictionary constructed is different from the real situation wherein in this work a Gene–Disease Dictionary (GDD) is constructed from biomedical databases [1].

Construction of the Gene and Disease Dictionaries

In order for each output entry to be linked to publicly available biomedical data sources, we created a human gene dictionary and a disease dictionary by merging the entries of multiple public biomedical databases[9][10].

The Gene Dictionary

A unique 'LocusLink identifier' for genetic loci is assigned to each entry in the gene dictionary making possible to merge gene information dispersed in different databases. Each entry in the merged gene dictionary holds all relevant literature information associated with a given gene. Each entry consisted of five items: gene name, gene symbol, gene product, chromosomal band, and PubMed ID tags.

The Disease Dictionary

Unified Medical Language System (UMLS) collect disease related vocabulary. From the 2003AC edition of the UMLS Meta thesaurus, selected 12 TUIs (unique identifiers of semantic types) that correspond to diseases names, types of abnormal phenomena, or their symptoms are denoted in Table 1.

Table 1 Selected TUIs (Unique identifiers of semantic type)

Diseases names	**Types of abnormal phenomena**
T019	Congenital Abnormality
T020	Acquired Abnormality
T033	Finding
T037	Injury or Poisoning
T046	Pathologic Function
T047	Disease or Syndrome
T048	Mental or Behavioral Dysfunction
T049	Cell or Molecular Dysfunction
T050	Experimental Model of Disease
T184	Sign or Symptom…..

Information Indexing

Decisive to reduce the complexity of the classification process, the information indexing method is obligatory for extracted information. To sort and to obtain a fixed number of index terms that appropriately cover all information about named entity relationship documents, a simple greedy strategy are applied.

Classification of Indexed Information Using Multiple Kernel Learning Support Vector Machines (MKL-SVM)

Choice on adapting the most informative entity relationship, Multiple Kernel Learning Support Vector Machine (MKL-SVM) approach [6] is considered for classification. This approach helps in classifying the indexed information of the named entity relationship of Gene-Disease as specific categories or types. MKL-SVM functions on a gradient descent optimization algorithm to classify the indexing information results into classes.

Clustering Using Entropy Agglomeration (EA)

If number of the biomedical users become larger, it is difficult to find similar user categories for the classified data for inter cluster. To perform this task, Entropy Agglomeration (EA) clustering is to group NEs with relationship which is mutually highly similar. This constructs "pure" clusters in order to have precise annotations.

Performance Metrics

The experimental results of the proposed KDS are performed against concerned algorithms with standard dataset. The performance evaluation of Information Extraction tasks such as Named Entity Recognition through PLSA-BM25 against existing Hidden Markov Model(HMM) [7] ,Conditional random field(CRF) [8] (for genes and disease based named entities) are shown in Figure 2(a),2(b),2(c),2(d) and tabulated in Table 2 and Named Entity Relationship through Logistic Model of dictionary construction method against existing Bayesian Networks(BN), Distributed Annotation System (DAS) are shown in Figure 3(a),3(b),3(c) and tabulated in Table 3. The evaluated measures are in terms of the precision, recall, F measure, MAP and accuracy. Simultaneously evaluated clustering and classification results are shown in Fig. 4 and Fig.5.

Recall (%)	**Precision (%)**		
	HMM	CRF	PLSA-BM25
0.2	0.65	0.71	0.836
0.4	0.72	0.73	0.849
0.6	0.73	0.78	0.864
0.8	0.735	0.81	0.865
0.2	0.65	0.71	0.836

Table 2 (a) Precision vs. Recall

Recall (%)	**Precision (%)**		
	HMM	CRF	PLSA-BM25
0.2	0.65	0.71	0.79
0.4	0.63	0.69	0.7897
0.6	0.61	0.65	0.764
0.8	0.61	0.645	0.756
1	0.605	0.634	0.75

Table 2(b) MAP vs. Recall

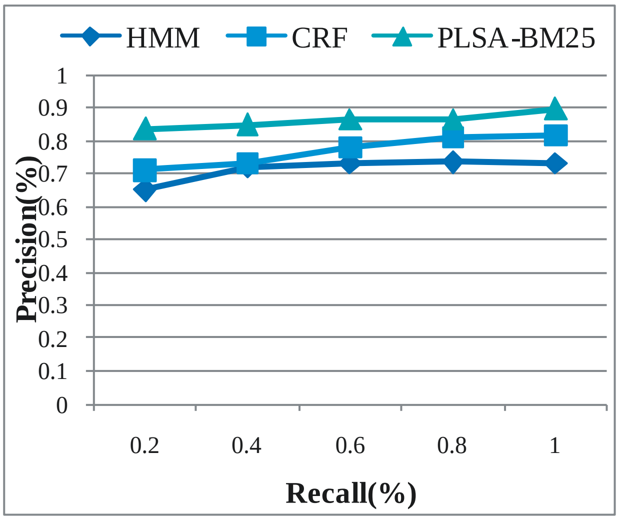


Figure 2(a) Precision vs. Recall curves

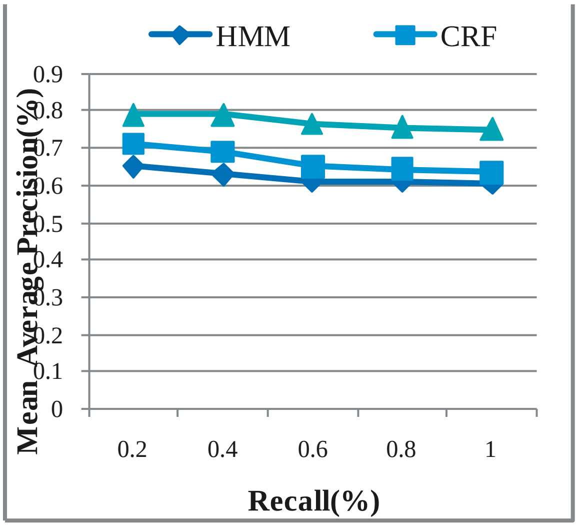


Figure 2(b) MAP vs. methods

Text doc	**F-measure (%)**			**NDCG(%)**		
	HMM	CRF	PLSABM25	HMM	CRF	PLSA-BM25
10	0.71	0.73	0.79	0.68	0.73	0.81
20	0.75	0.765	0.82	0.69	0.74	0.83
30	0.78	0.7898	0.845	0.71	0.76	0.86
40	0.79	0.81	0.864	0.72	0.78	0.87
50	0.82	0.835	0.898	0.74	0.82	0.89

Table 2(c) F measure and NDCG performance comparison

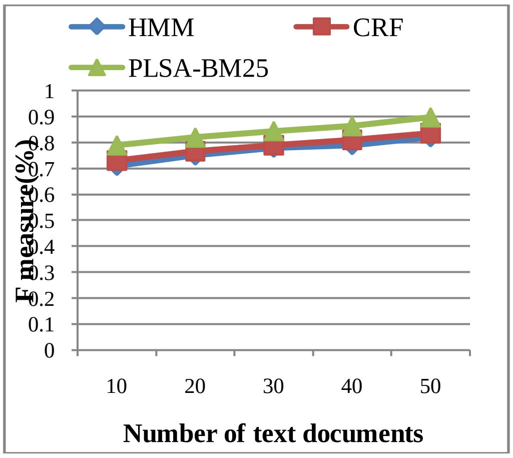


Figure 2(c) F--measure vs. methods

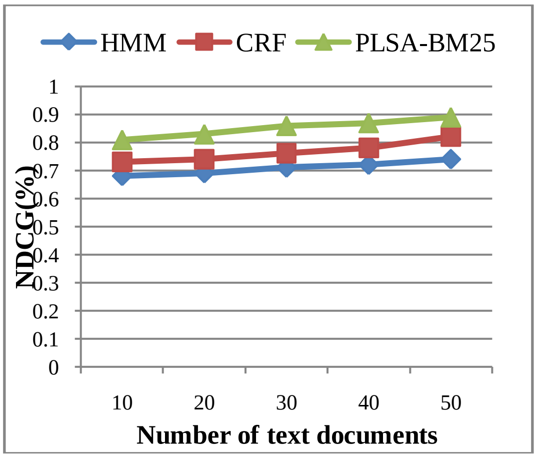


Figure 2(d) NDCG vs. methods

Recall (%)	**Precision (%)**		
	DAS	BN	GDD
0.2	0.685	0.752	0.83
0.4	0.698	0.748	0.834
0.6	0.702	0.751	0.84
0.8	0.71	0.76	0.85
1	0.73	0.774	0.86

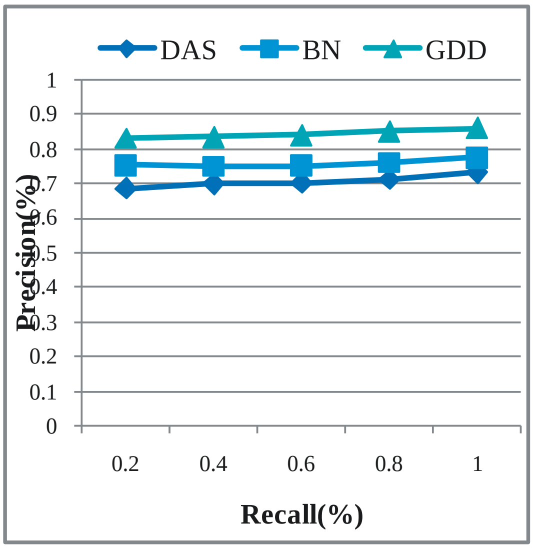
Table 3(a) Precision vs. Recall

Text doc	**F-measure (%)**			**NDCG(%)**		
	DAS	BN	GDD	DAS	BN	GDD
10	0.72	0.76	0.84	0.71	0.75	0.88
20	0.728	0.77	0.86	0.725	0.758	0.89
30	0.729	0.778	0.865	0.73	0.76	0.895
40	0.734	0.795	0.88	0.734	0.762	0.9
50	0.74	0.8	0.912	0.739	0.776	0.905

Table 3(b) F measure and NDCG

Recall (%)	Precision (%)		
	DAS	BN	GDD
0.2	0.68	0.72	0.82
0.4	0.698	0.736	0.835
0.6	0.7	0.7394	0.84
0.8	0.71	0.74	0.85
1	0.72	0.75	0.863

Table 3(c) MAP



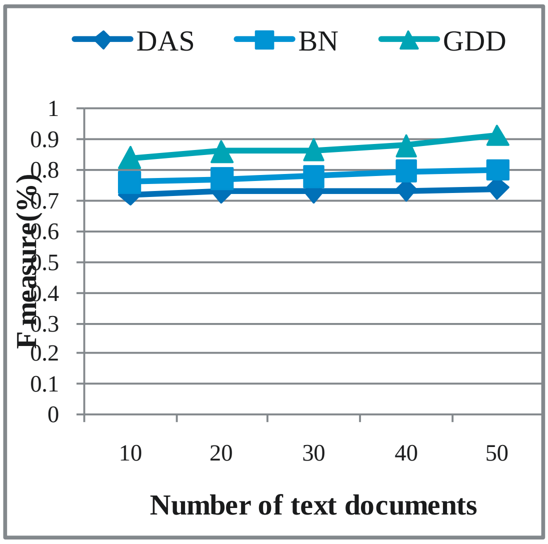
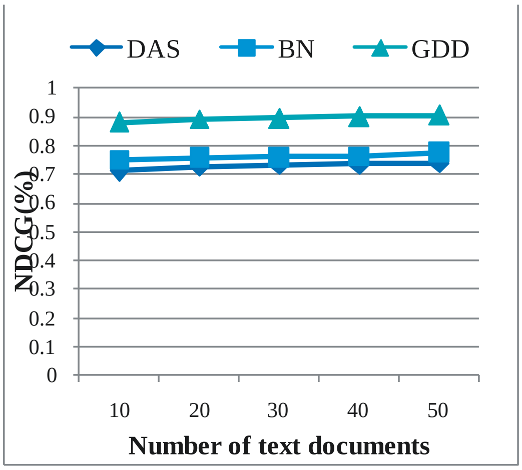


Figure 3(a) Precision vs. Recall curves **Figure 3(b)** F-measure vs. methods



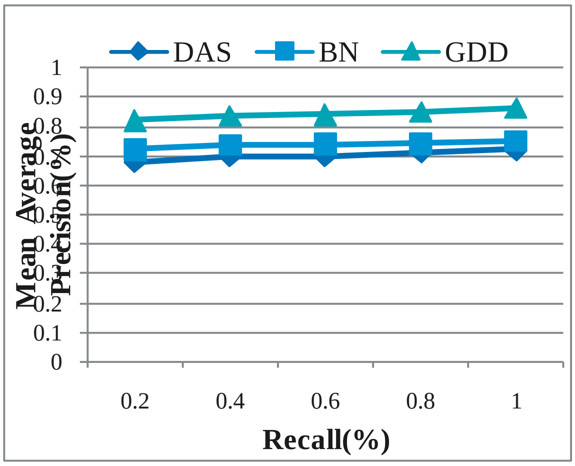


Figure 3(c) NDCG vs. methods **Figure 3 (d)** MAP vs. methods

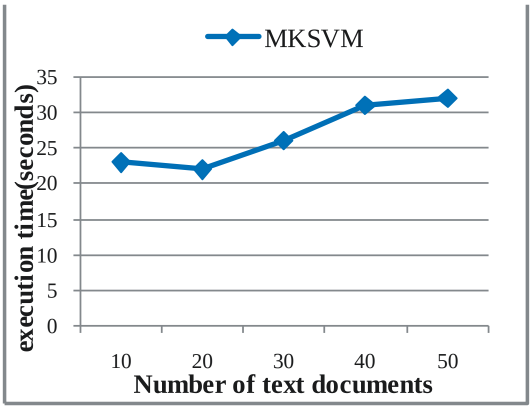


Figure 4 Time comparison of classification

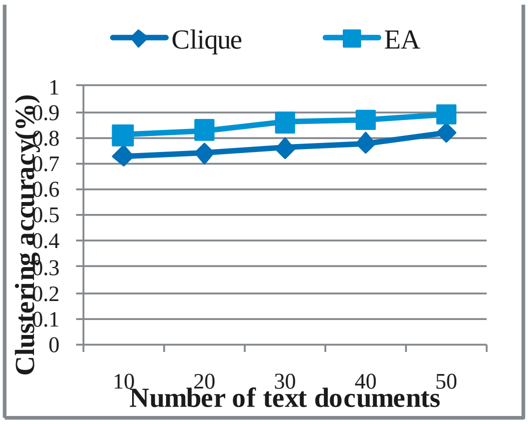


Figure 5 Clustering accuracy

Text doc	Clustering accuracy (%)		Time comparison(Seconds)	
	Clique	**EA**	**Clique**	**EA**
10	0.73	0.81	35	23
20	0.74	0.83	45	22
30	0.76	0.86	48	26
40	0.78	0.87	47	31
1.0	0.82	0.89	52	32

Table 5 : Clustering accuracy and time comparison

Conclusion

The foremost aim of Knowledge Discovery System for computational analysis of dynamical behavior of entities is achieved with higher results. The designed approaches have been demonstrated as the most robust method owing to capability. It handles high dimensional discriminative vector features in text processing and prediction of new terms and variations. It benefited greatly from enhanced access to services and tools for the community of biologists, bioinformaticians and developers in the midst of platform prototype for convenient access to a large, unstructured repository of text.

Future Enhancement

In further tasks, tool scalability and integration dependency on commonly accepted are supposed to be eliminated. Despite major initiatives towards seamless data exchange and interoperability, a pilot application can be included into workflows. The last direction is to focus on ongoing activity as development of text mining resources.

References

[1] Zhong Huang and Xiaohua Hu, December 2013,” Disease Named Entity Recognition by Machine Learning Using Semantic Type of Met thesaurus,” International Journal of Machine Learning and Computing, Vol. 3, No.6.

[2] Hong-woo chun, yoshimasa tsuruoka et al , October 2005, ”Extraction of Gene-Disease Relations from Medline using Domain Dictionaries and Machine Learning”, Proceedings of Pacific Symposium on Biocomputing.

[3] Sampo Pyysalo, Filip Ginter et al, 9 February 2007, “BioInfer: a Corpus for Information Extraction in the Biomedical Domain”, *BMC Bioinformatics.*

[4] Haochang Wang, Tiejun Zhao, Hongye Tan, Shu Zhang, 2006, “ Biomedical Named Entity Recognition Based On Classifiers Ensemble”, International Journal of Computer Science and Applications, Techno mathematics Research Foundation , Vol. 5, No. 2, pp 1- 11.

[5] Rosario, B., Hearst, M. ,2004,”Classifying Semantic relations in bioscience texts”, In Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL’04).

[6] Xiaoou Li, Xun Chen, Yuning Yan, Wenshi Wei,and Z. Jane Wang, 2014,“Classification of EEG Signals Using a Multiple Kernel Learning Support Vector Machine”, Sensors (Basel). Jul 2014; 14(7): 12784–12802.

[7] Shaojun Zhao,2004, “Named Entity Recognition in Biomedical Texts using an HMM Model”, Proceeding JNLPBA '04 Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications ,pp. 84-87.

[8] Zhong Huang and Xiaohua Hu,2013, "Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus," International Journal of Machine Learning and Computing vol.3, no. 6, pp. 494-498.

[9] Biology.stackexchange.com/questions/14088/standard-classification-of-Disease

[10] Bioinformatics.ca/links_directory/catego...nome/health-and-disease? Filter=databases