International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 1, January 2015, PP 92-97
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Knowledge Discovery System (KDS) for Named Entity Recognition

K. Prabavathy
Research Scholar in Computer Science
Manonmanium Sundaranar University
Tirunelveli Town, TamilNadu, India
praba_bud@yahoo.co.in

Dr. P. Sumathi
Assistant Professor, Research Department of
Computer Science, Government Arts College
Coimbatore-18, TamilNadu, India
sumathirajes@hotmail.com

Abstract: *In recent years, the recognition of named entities in the biomedical scientific literature has become the spotlight of major research. Now, in this work, the focus is shifted towards uncovering of semantic hierarchy amongst entities like Gene, Disease and Proteins etc., on collection of Biomed Corpus. To regulate the proposed work KDS for Named Entity Recognition, 'NER', PLSA-BM25 (novel Weighting scheme & probabilistic background based on the modelling of relevant and irrelevant entity) algorithm is implemented robotically to recognize Named Entities in the interest of bio medical domain. Since the biomedical domain on the other hand, due to its complicated language, has consistently lagged behind, GO annotation and disease annotation are utilized to carry out the semantic information.*

Keywords: *Information Retrieval, Information Extraction, Entities, Clustering, Classification*

1. INTRODUCTION

In broad-spectrum, most of existing work mere extracts the information rather than knowledge.NE is defined as a single word term or multi-words phrase that denotes a biomedical object, for instance a protein, gene, disease, or drug where semantic hierarchy is associated. KDS locates boundary of entity mentions in a text and tag them with their corresponding semantic types. In this intend, it hold close the tasks Information Retrieval(IR), Information Extraction(IE), Indexing, Classification and Clustering to unlock the hidden knowledge on biomed corpus. Accurate extraction of these entities is central to the various text mining and knowledge discovery tasks that have now become essential due to the overwhelming amount of textual information being produced.

2. RELATED WORK

Recognizing the named entity is one of the most fundamental tasks in the biomedical knowledge discovery [1]. It is not sufficient for capturing biomedical phenomena in detail since there is a growing need for capturing more detailed and complex relations. Li Y, Lin H, Yang Z (2009) BIGNER developed for tagging gene and protein mentions using of feature coupling generalization of a Conditional Random Field (CRF) model achieving an F-score of 76%. Conrad Plake et al (2009) developed GoGene – annotation. It ranks according to novelty text mining extracting co-occurrences of genes and ontology terms from literature. Highest relationship between the gene and protein relationship are not performed. Tanabe et al (2002) developed for tagging Gene and Protein names in biomedical text using a combination of statistical and knowledge-based strategies but relies on transformation-based part-of-speech tagger, and manually generated rules. Yang & Zhou et al (2010) represents a two-phase approach based on semi-CRFs and novel feature sets with term boundary detection and semantic labelling.

3. PROPOSED WORK

The most basic obstacle is of dynamic nature of facts, rapidly increasing scientific discoveries, ever-growing list of relevant terms and resources can never be complete as long as scientific progress continues. In particular, most biomedical texts introduce specific notations, acronyms, and innovative names to represent new concepts, relations, processes, and automatic extraction of biomedical terminologies and mining of their diverse usage are major challenges in biomedical information processing system [2].

Information access is an iterative process, the goals of which shift and change as information is encountered. The key components and tasks grounding the structured information used in our framework are Corpora, Pre-processing - sentence splitting, tokenization and annotation encoding, IR, IE, Information Indexing, Text Classification, Text Clustering. It structured as following steps

Step-1: Query received from user gets pre-processed and fed into biotext corpus.

Step-2: IR using of Vector Space Model gathers relevant texts/represent document from unstructured text.

Step-3: General pre-processing system is acted.

Step-4: IE using of PLSA-BM25 with incorporation of gene and disease ontology are initiated for interactive supervision.

Step-5: Information Indexing using of greedy method verifies the correctness of the indexed results based on category next time.

Step-6: Text Classification using of Multiple Kernel Language – Support Vector Machine (MKL-SVM) on additional constraint of weights classifies the indexing information results into classes.

Step-7: Text Clustering using of Entropy Agglomeration (EA) find which classified user NER have common entities with similar user.

Step-8: Receive the desired output from clustered classes.

3.1. Steps Initiated

- Query Pre-processing - Receive the PolySearch query which support 50 different classes of queries against nearly a dozen different types of text, scientific abstract or bioinformatics databases
- IR - The systems rely on a combination of deep linguistic knowledge and richness of annotations obtained from biological resources. Applications of matrix analysis do quantify the degree of similarity between a query vector and the document vectors contained within the term-document matrix.
- IE - identifies and extracts a range of specific types of information from texts of interest. Formulate a rapid adaptation to new relations using of representative collection and capture temporal information using of ontologies. The user starts off with a small set of words, inspects the results, selects and rejects entity terms from the returned ranking, and iterates until get satisfied.

In this work, GO annotation and disease ontology annotation are utilized with the purpose of providing the biomedical community with consistent, reusable and sustainable descriptions of human entities [3]. This helps in precise identification about the gene, protein and disease entities.

Contributed Work

Initially a simple WSD method was performed for NER. It determines the boundaries of the NEs and classifying the entities into classes. But ambiguity basis, capitalization are under inconsistent and follows unreal naming principle and unconsider the semantic information of entities. In return, PLSI is used to depict the entities in a unified way for exposing similarity functions. But existence of Perplexing name is being encircling in biomed corpus is critical. Later, Probabilistic Latent Semantic Analysis (PLSA) method [4] is deemed to perform the NER utilizing and exploiting the semantic information of the entities. But raw frequencies counts of each entity have to be on each biotext are considered. Factors such as raw frequency counts and term rarity can lead to the favor of certain irrelevant entity recognition results.

In tune to overcome the above crisis such as dimensionality consideration and irrelevant entity recognition results problem, a novel Weighting schemes BM25 is integrated to PLSA have been developed [5] to remove such irrelevant entity recognition. Hence 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.

Consider the biotext corpus extracted information as being a bag filled with tokens. Each token has an associated terms related to gene, protein and diseases and semantically meaningful tag label attached from gene and disease ontology results. $P(d,t)$ is the probability take out a token with the document

tag label d results from gene and disease ontology and entity term t associated with it. Therefore if $f_{d,t}$ tokens are in the bag with labels d and t, implying that term t appeared $f_{d,t}$times in document d, obtain the sample probability:

$$\widehat{P}(d, t) = \frac{f_{d,t}}{\sum_{\delta \epsilon D} \sum_{t \epsilon T} f_{\delta,T}} \quad (1)$$

where D and T are the set of extracted information document and entity terms of genes ,protein and diseases respectively . PLSA attempts to model these sampled labeled tag–word probabilities as the sum of hidden topic) distributions:

$$P(d, t) = \sum_{z \in Z} P(d|z)P(t|z)P(z) \quad (2)$$

Where Z - set of hidden topics,P (d,t) - probability of word w being related to tags documents d, $P(d|z)$is the probability of document d given topic z, P(t|z)is the probability of named entity term t given topic z under gene, protein and diseases and P(z) is the probability of topic z. The BM25 weighting scheme has a probabilistic background based on the modelling of relevant and irrelevant document tags. The simplified document scoring equation can be shown as:

$$S(d, Q) = \sum_{t \in Q} we_{d,t} we_t$$

Where d - document, Q - set of query terms, $we_{d,t}$ and we_t are the document-entity term and entity terms weights respectively. The entity term weight is used to reflect the importance of the named entity of gene, protein and disease due to its rarity, therefore its weight should be high:

$$we_{d,t} = \frac{(k_1 + 1) f_{d,t}}{K + f_{d,t}}$$

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. Once the weights values are applied to every named entity terms of genes, protein and diseases based on frequency value, use the PLSA method to obtain the value of $P(d, t)$ and each of its components. Therefore our new PLSA-BM25 relationship becomes:

$$\widehat{P}(d, t) = \frac{\omega_{d,t}}{\sum_{\delta \epsilon D} \sum_{t \epsilon T} \omega_{\delta,t}}$$

1. Information Indexing - Decisive to reduce the complexity of the classification process, the information indexing method is obligatory for extracted information. The main functionalities of Information Indexing on Filtering based indexing is inclusive of 5N's,

- Nature to reduce the number of irrelevant words.
- Nature to reduce the complexity of the classification problem on other hand.
- 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.

2. Classification (MKL-SVM) - Text classification applied to biological literature can minimize this effort by automatically selecting only the relevant articles to a given task. MKL-SVM approach [6], gradient descent optimization algorithm helps in classifying the indexed information of the named entity recognition of proteins, genes and diseases as specific categories or types. Gradient descent on SVM objective value with weighted 2-norm regularization and formulation with an additional constraint on the weights improves the classification results.
3. Clustering(EA) - Find which classified user named entity recognition have common entities with similar user and rest the extracted information documents with the most words in common into the same groups. Entropy Agglomeration (EA) is exploited for clustering the classified results. This clustering technique must be able to construct "pure" clusters in order to have precise annotations. In that case, it is desirable to avoid fixing the number of clusters.

4. EXPERIMENTATION RESULTS

Analysis helps to uncover the detected NER results between different data sources which are not directly apparent. To compare the efficiency of named entity recognition results, the parameters such as precision vs. recall, F measure, Mean average precision (MAP), and Normalized discounted cumulative gain (NDCG) are considered and shown in figures. Hidden Markov Model (HMM)[7], Conditional Random Field (CRF)[8] and proposed PLSA-BM25 methods for NER task with above mentioned parameters are compared.

Precision = C/D, Recall = C/T, F= (2* P*R)/(P+R)

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \quad , \quad nDCG_k = \frac{DCG_k}{IDCG_k}$$

D - Number of pairs identified as interacting by the classifier, C - Number of pairs correctly identified as interacting, T -Number of pairs labeled as interacting, AveP(q)- average mean of precision values for user specified query,

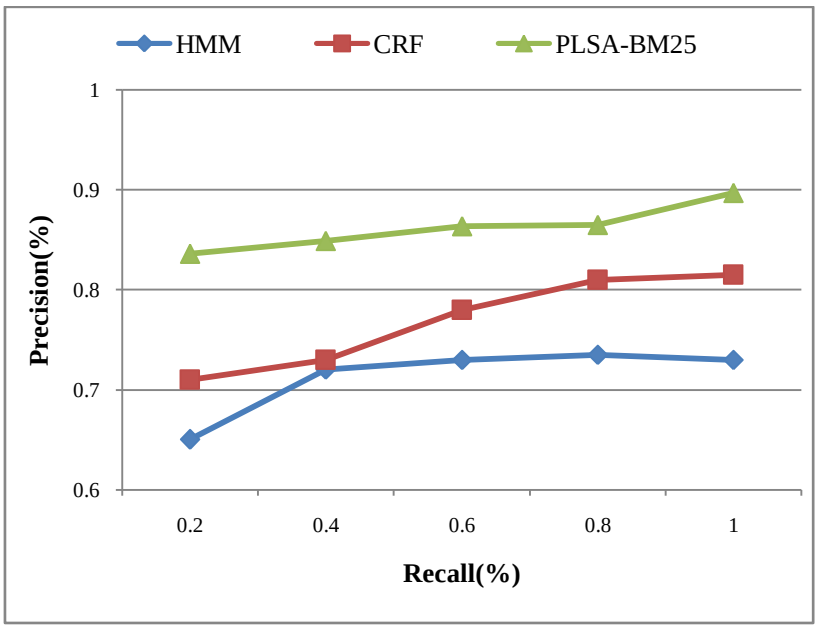

Fig1. *Precision vs. Recall curves*

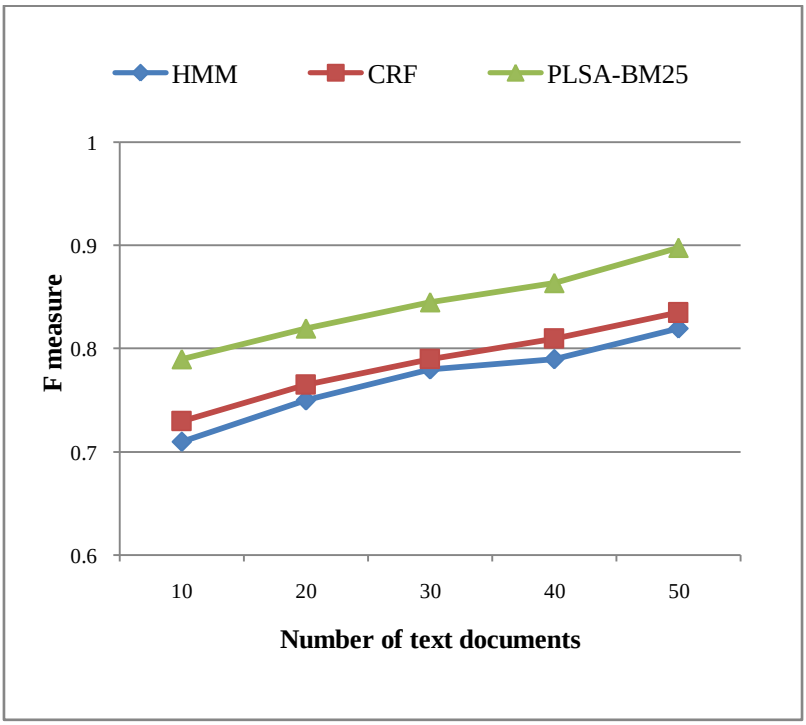

Fig2. *F-measure vs. methods*

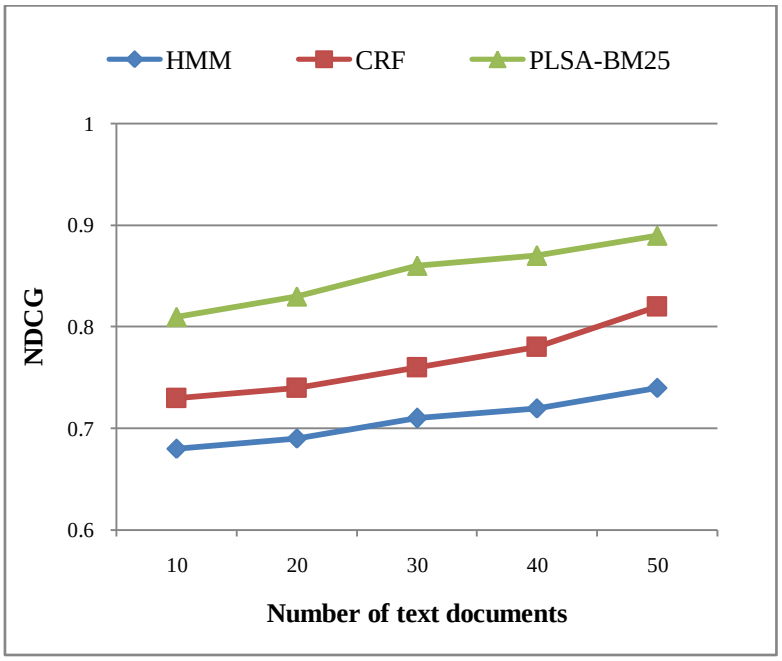

Fig3. *NDCG vs. methods*

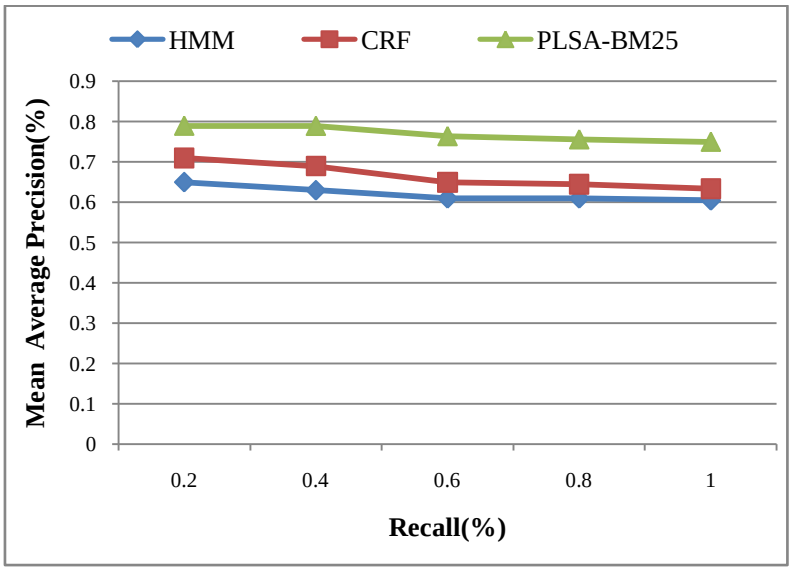

Fig4. *MAP vs. methods*

5. SUMMARY

Knowledge Discovery System is not an ephemeral problem. It addresses some of 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 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.The system incorporated various features of genes, protein, diseases and experimented with different strategies for combination of probabilistic methods. With dynamical behavior of entities it achieves higher results.

REFERENCES

[1] Haochang Wang, Tiejun Zhao, et al, "Biomedical Named Entity Recognition Based On Classifiers Ensemble", International Journal of Computer Science and Applications, Techno mathematics Research Foundation , 2006, Vol. 5, No. 2, pp 1- 11.

[2] Seonho Kim, Juntae Yoon, et al, "Two-Phase Biomedical Named Entity Recognition Using A Hybrid Method", Natural Language Processing – IJCNLP 2005 Lecture Notes in Computer Science , 2005,vol. 3651, pp 646-657.

[3] Bhattacharya I, Godbole S and Gupta A, "Building re-usable dictionary repositories for real-world text mining", CIKM'10, Toronto, Ontario, Canada, October 26–30, 2010.

[4] Ji S, Zhang W, Li R, "A probabilistic latent semantic analysis model for coclustering the mouse brain atlas", IEEE/ACM Trans Comput Biol Bioinform,2013 Nov-Dec;10(6):1460-8.

[5] S. Robertson and H. Zaragoza. "The probabilistic relevance framework: BM25 and beyond. Found",Trends Inf. Retr. , Apr. 2009, 3(4):333–389.

[6] Xiaoou Li, Xun Chen, et al "Classification of EEG Signals Using a Multiple Kernel Learning Support Vector Machine", Sensors (Basel). Jul 2014; 14(7): 12784–12802, 2014.

[7] Shaojun Zhao, "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 ,2004,pp. 84-87.

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

AUTHOR'S BIOGRAPHY

Smt. K.Prabavathy M.Sc,M.Phil.. , Doctoral Research scholar in department of Computer Science, Manonmanium Sundaranar University ,Tirunelveli, Tamil Nadu, India. She completed M.Phil in the area of Data Mining and received MCA degree through Bharathiar University, Coimbatore and M.Sc degree through Madurai Kamaraj University, Madurai. She has published number of papers in reputed journals and conferences. She has about five years experience of teaching and research experience. Her area of interest includes Data Mining, Bioinformatics and Computer Networks

Dr. P. Sumathi is working as an Assistant Professor and Doctoral Research Supervisior in PG & Research Department of Computer Science, Government Arts College, Coimbatore, Tamilnadu, India. She received her Ph.D., in the area of Grid Computing in Bharathiar University. She has done her M.Phil in the area of Software Engineering in Mother Teresa Women's University and received MCA degree at Kongu Engineering College, Perundurai. She has published a number of papers in reputed journals and conferences. She has about Seventeen years of teaching and research experience. Her research interests include Data Mining, Grid Computing and Software Engineering.