DOI: 10.21917/ijsc.2016.0179

NAMED ENTITY RECOGNITION FROM BIOMEDICAL TEXT -AN INFORMATION EXTRACTION TASK

N. Kanya[1] and T. Ravi[2]
Department of Computer Science and Engineering, Manonmanium Sundaranar University, India
E-mail: [1]kanyamtech@yahoo.co.in
Department of Computer Science and Engineering, Madanapalle Institute of Technology and Science, India
E-mail: [2]travi675@yahoo.com

Abstract

Biomedical Text Mining targets the Extraction of significant information from biomedical archives. Bio TM encompasses Information Retrieval (IR) and Information Extraction (IE). The Information Retrieval will retrieve the relevant Biomedical Literature documents from the various Repositories like PubMed, MedLine etc., based on a search query. The IR Process ends up with the generation of corpus with the relevant document retrieved from the Publication databases based on the query. The IE task includes the process of Preprocessing of the document, Named Entity Recognition (NER) from the documents and Relationship Extraction. This process includes Natural Language Processing, Data Mining techniques and machine Language algorithm. The preprocessing task includes tokenization, stop word Removal, shallow parsing, and Parts-Of-Speech tagging. NER phase involves recognition of well-defined objects such as genes, proteins or cell-lines etc. This process leads to the next phase that is extraction of relationships (IE). The work was based on machine learning algorithm Conditional Random Field (CRF).

Keywords:

Information Extraction, Information Retrieval, Text Mining, Named Entity Recognition, Data Mining

1. INTRODUCTION

The enormous amount of biomedical text provides a huge source of knowledge for biomedical scientists, researchers and doctors. Text mining facilitates to mine the knowledge and information from the massive resources. MEDLINE 2010 database contains over 12.00 million documents [1]. The quantity of biomedical literature is growing at such a speed that it is becoming hard to discover, without text mining it is hard to retrieve and handle the described information, which intend to extract information, discovering knowledge and hypotheses generation according to the user requirements. The aim of bio Text Mining is to allow scientist to recognize required information efficiently.

Text Mining comprise of:

- Information retrieval (IR)
- Information extraction (IE)

Information Retrieval's task is to collect and refine the relevant documents Whereas Information Extraction yields exact information about specified kind of entities and relationships of interest. Information Extraction applications range from Semantic Web technology to Bioinformatics. With recent digital trends in the Biomedical Industry, the need for automatically extracting the relevant information from biomedical documents is greater than ever before. In a nutshell, Information Extraction intends to develop textual data set into a form that promote Search and discover knowledge [1] [3]. The Text Mining Pipeline (Fig.2) consists of Information Retrieval (IR) and Information Extraction (IE). IR system aims to acquire preferred text documents on a particular topic; whereas, IE systems intend to extract predefined types of information such as Named Entity and Relations. In this paper, the section 2 will deal with the information Retrieval Module of the Text mining Pipeline and the section 3 of the paper deals with the Information Extraction Process of Text Mining task such as Named Entity Recognition. The methodology used is also described within each section.

This section includes detailed information about the dataset used. A detailed evaluation analysis of the obtained results was discussed in the section4. Future enhancement and conclusion are covered in section 5. Cited reference list added in the section 6 of this paper.

2. RELATED WORK

Named Entity Recognition is an essential component in Information Extraction. Identification of Named entities from biomedical Literature is a crucial process. Many researchers have experimented different methods to achieve the above tasks. Identification Named Entities from biomedical Literature has been analyzed automatically by many tools to generate the Entity classes [2] [3] [4] [5] [9] etc. Accuracy provided by many tools varies from one to other. Yoshimasa, Jun'ichi and Sophia [2] works on extracting entities by identifying the dynamic sentence and the Annotation process using two named entity corpora named as GENIA and CoNLL. This approach can reduce the number of sentences which need to be examined by the human annotator. Fei Zhu et al. proposed to identify the entities in the Biomedical Literature [3]. The work proposed machine learning approaches such as HMM, SVM, ME and used biomedical information as domain knowledge to recognize the protein, RNA, DNA etc. Kazama et al. [4] proposed a work to identify protein, lipid, and cell line and cell type using SVM. Lin et al. [5] presents an approach using ME to recognize 23 categories of biological terms. Saha S, Ekbal A, Sikdar UK [9] proposed a work based on Single Objective Optimization using classifier ensemble technique with search capability of Genetic Algorithm (GA) for Named Entity Recognition from biomedical texts. Z. Ju, J. Wang and F. Zhu presents a biomedical named entity identification system using Support Vector Machine (SVM), using Data from the GENIA corpus which is a collection of Medline Abstracts.

The above survey presents the extraction of entities from Biomedical Literatures from various Data sets, using various Machine Learning Algorithms such as Support Vector Machine (SVM), Conditional Random Field (CRF) and Hidden Markov

Model (HMM). The proposed work presents the extraction of entities from various Biomedical Literatures from the PubMed and Medline Databases with proper preprocessing, synthetic analysis and semantic analysis. The model generated with Feature Space Definition using CRF without manual intervention.

3. PROPOSED METHODOLOGY

The Identification of Entities from biomedical Literature involves sequence of diverse tasks. The Named Entity Recognition includes the task of Information Retrieval (IR) and Information Extraction (IE). The proposed work starts with Information Retrieval and lead to the task of Information Extraction. The Information Retrieval task retrieves the biomedical literature, which fulfill the criteria given by end user from the PubMed Database. From the retrieved document the corpus will be generated. The Information Extraction task includes various sub-tasks to extract the named Entities from the biomedical literature. The system architecture named as Workflow of NER is shown in Fig.1.

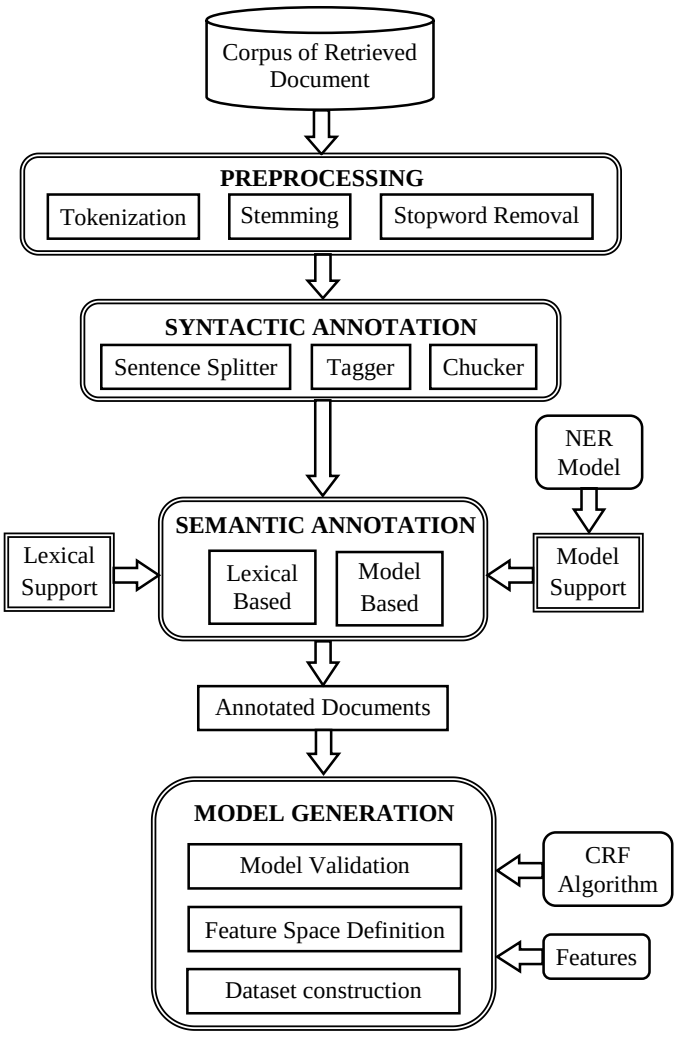

Fig.1. Workflow of NER

3.1 INFORMATION RETRIEVAL

IR is about identifying a subset of document from the corpus of document .The content of subset document is most relevant to the users need. The users' requirement will be specified in the query [2]. The Information Retrieval system retrieves the relevant document from the huge collection of document based on the users query.

The Information Retrieval (IR) system based on document consists of two sub modules.

- Document Acquisition Module.
- Document Structuring Module.

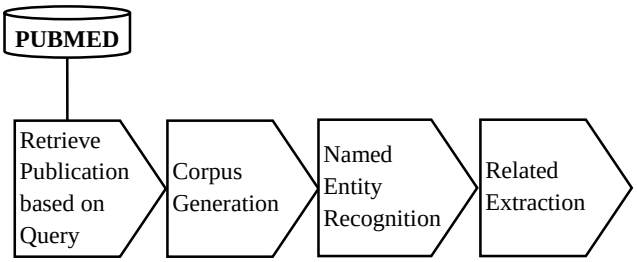

Fig.2. Text mining Pipeline

3.1.1 Document Acquisition Module:

A document repository consists of huge collection of documents of diverse topics of interest [4]. Domain specific documents can be retrieved from the various domain specific repositories. In the Biomedical domain at present PubMed is the voluminous bibliographic catalogue of biomedical domain [1]. It contributes publication facts such as Title of the article, affiliation of the authors, abstracts, full text information about the articles, statistical information about the article as well as journal. In an article full text provides more information than abstracts [5] [20].

Entrez Programming Utilities (eUtils) Web service is used to access the largest biomedical bibliographic catalogue [8] [21]. NCBI Entrez system is the broadly used interface for information retrieval from biomedical databases (Fig.3). The interface eUtils is used to access and navigate all the dimensions of database. The power of this system is that a single query can retrieve all information of the document.

The Document Acquisition Module, works based on the eUtils service. The users' requirement will be described in terms of keywords. The system will convert the keyword based user requirements in to query.

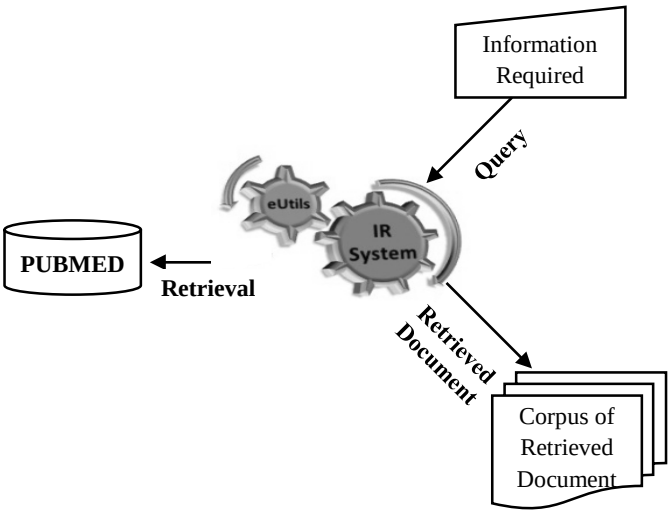

Fig.3. Information Retrieval

The IR system along with eUtils will explore the PubMed and retrieve the documents. The document collections contain documents of various domains. From this collection the system will retrieve the documents based on the interest of the user's information requirement. The key concern is to decide on the appropriate index terms.

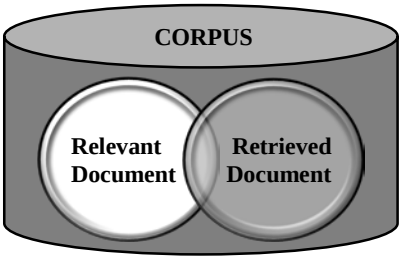

Fig.4. Corpus Generation

The major issue of Information Retrieval system is identifying and retrieving the relevant information needed by the user [6]. The retrieved documents may have relevant and non-relevant document to the user's need. The standards such as precision and recall have been used to measure the success of the retrieved document based on the user's need [6] [15] [10].

More than retrieving the document the Document Acquisition Module also helps in identifying the relevance of the retrieved document [2]. The query based on the keyword identification helps in extracting the key documents. Each document will be evaluated by the user manually. This is a time consuming and lengthy process. To automate the relevance assessment the component contain Machine Learning algorithms to obtain the domain specific document.

3.1.2 Document Structuring Module:

The Document Structuring Module is responsible for converting the document from PDF to Text. The documents need to be converted to a particular format which could be understood by natural language Processing Module [8] [12] [17] [13]. After considering various text format conversion software's that include PDFBox and Xpdf software in to our framework. XML based documents were converted using the standard rules.

		Retrieved	Not Retrieved
Irrelevant		Retrieved & Relevant	Not Retrieved & Relevant
Relevant		Retrieved & Irrelevant	Not Retrieved & Irrelevant

Fig.5. Corpus Document

3.2 INFORMATION EXTRACTION

The Information Extraction task is to recognize a predefined set of conceptions in a specific domain. It is to recognize the pre-specified class of entities, relationships or events of particular instances.

The Information Extraction Process includes the following classical task such as

- Named Entity Recognition
- Co-reference Resolution
- Relation Extraction
- Event Extraction

3.2.1 Named Entity Recognition:

The primary task of Information Extraction is Named Entity Recognition (NER). NER identifies and detects the named entities of the predefined classes [6] [9] [19]. The Named entities are phrases that holds the names of people, cites, positions ,etc. and particularly in biomedical text mining the entities for instance genes, proteins, cell line, DNA, RNA, drugs, disease or organisms.

In this work the NER processes consist of document preprocessing, syntactic and semantic annotations and a CRF machine learning algorithm based model generation for named entity recognition [5] [9] [16].

Further the model for NER will evaluate the various datasets and identifies the entities of various classes. The text preprocessing consists of various subtasks such as tokenization, sentence splitting and stopword removal [7].

3.2.1.1 NER Module:

The NER Module commences with the preprocessing of documents. The documents undergo tokenization, tagging and stop word removal in preprocessing phase. After preprocessing of the document automatic annotation occurs with the help of lexical resources such as dictionaries, lookup tables, lexical words and rules. The terms that are identified by the lexical resources are tagged for general recognitions.

Gene Ontology (GO) is used to identify the mentions. The technique identifies the following classes such as gene, protein, DNA, RNA, Cell line, Cell type [3] [14] [16].

The relational format will hold the retrieved bibliographic information and lexical resources. This work includes the MySQL database engine to store the information retrieved. For model creation and feature set generation, Rapidminer and Weka tool kit is used.

MALLET is used to deploy the conditional random Field algorithm. Feature vector (or) Feature weight efficiency are identified. The Quasi-Newton Method called L-BFGS is used to identify the feature vector and. Feature weight efficiency.

In the Information Retrieval Phase, based on the user query the system will retrieve the documents from the PubMed Database. From that document system will retrieve candidate documents and generate a corpus of retrieved.

In the Information Extraction phase, the primary task will be preprocessing. In this, the documents undergo the process of tokenization, Stemming and stopword removal. After preprocessing, the document undergoes the syntactic and semantic annotations [11] [18].

Syntactic annotation will have the phases of sentence splitting, tagging (Parts Of speech) and Noun Phrase chunking. Semantic Annotation is done using the lexical resources and Machine Learning based Natural Language Processing Model generation using the Condition Random Field (CRF) algorithm.

The Lexical resources includes Dictionaries, Lookup tables, rule sets written using Relational Expressions using Text: Rewriter tool [9] [12] [6] [20]. The model is generated by creating the feature space definition and data set construction. In Fig.2, the created model is evaluated by the CRF algorithm. Dataset is constructed using the features. The performance of the system is evaluated by measuring the precision and recall.

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

True positive (TP) = correctly identified
False positive (FP) = incorrectly identified
True negative (TN) = correctly rejected
False negative (FN) = incorrectly rejected

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F-measure} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. RESULTS AND DISCUSSION

The Primary process of Named Entity Recognition is Information Retrieval Process (IR). IR process works based on the Query given by the Text Miner [7] [14] [23]. In this work initially the query is set to retrieve the documents from the PubMed and Medline databases.

The keyword of the query is defined as “Breast Cancer”, organism value is defined as “human” and the publication duration of the documents were set to be in-between 2014 to 2015 [8] [22]. The Entrez Programming Utilities (eUtils) Web service based IR system retrieved about 1547 documents as per the specified constrains. The documents were stored as a corpus [10] [11] [19].

The Next Process of NER is to extract the entities from documents which are stored in the corpus. The document undergoes the process of Preprocessing, Syntactic Annotation based on sentence splitting, tagging and chunking. Semantic Annotation involves in Lexical Resources and Conditional Random Field Machine Learning Algorithm to generate the model [6] [8] [23]. The feature Space will be constructed, the Conditional Random Field Algorithm will work based on that. At the end six classes of entities were found from the retrieved.

Table.1. Number of Documents and Entities from the three Corpus

Corpus	Number of documents Retrieved	Number of Entities identified
Breast Cancer	1547	10339
Lung Cancer	1463	9738
Thyroid Cancer	1274	7638

The system retrieved three corpus based on the query, the corpus are Breast Cancer Corpus, Lung Cancer Corpus and Thyroid Cancer Corpus. From the corpus of document the Named Entity Recognition system (NER) identified six classes of entities. The classified classes are Protein, Gene, DNA, RNA, Cell Line and Cell Type. The Precision and Recall values are calculated for the Breast cancer corpus, Lung Cancer Corpus and Thyroid Cancer. The Information Retrieval system retrieved the documents from the PubMed Corpus. The system retrieved the documents based on the query given by the end user. Based on the query the Information Retrieval system retrieved documents from three corpus such as Breast Cancer Corpus, Lung Cancer Corpus and Thyroid Cancer Corpus. The retrieved 1547 documents from Breast cancer corpus is based on the condition on the given query. Similarly 1463 and 1274 documents are retrieved from lung cancer and Thyroid cancer corpuses respectively. The retrieved documents were stored in the corpus. The retrieved corpus is used to extract the named entities. The corpus will be the input for the Named Entity Reorganization system.

The Document in the corpus used to identify the entities and relationship. The study analyzed and compared using two machine learning algorithms such as Support Vector Machine (SVM) and Feature Space Based Conditional Random Field (CRF). On the study the system identified six classes of entities for all three Corpus. For the Breast cancer corpus Feature based Conditional Random Field outperforms the support Vector Machine in identifying the classes such as Protein, DNA, RNA, Cell Line and Cell type. SVM out performs the CRF on identifying Gene.

Similarly for Lung Cancer and Thyroid Cancer corpus CRF outperforms on identifying four classes out of six. And SVM outperforms on two classes out of six.

The frame work was analyzed with all the three corpuses with the two Machine Language Algorithms. On that Feature space based Conditional Random field outperforms the other existing system Support Vector Machine.

The data set consist of documents related to Breast Cancer, Lung Cancer and Thyroid Cancer. The comparative result shows that machine learning system out performs than another one. The corpus of documents experimented with machine learning algorithms Feature based Conditional Random Field and Support Vector Machine. The comparison results are represented in the Table.1, Table.2 and Table.3. The Fig.6, Fig.7 and Fig.8 represents the comparative chart between the three different corpus. In this work feature space based Conditional Random Field out performs Support Vector Machine for various classes. In some classes Support Vector Machine does better than the algorithm Feature space based CRF. In this work six classes of entities were identified by both of the learning algorithms.

The Feature vector is an n-dimensional vector of features. Features are equivalent to numerical representation of data. That represents some instance. Every feature can be thought of a dimension. Feature space will have N-Dimensions of features. The machine learning algorithms must be influencing that feature space in by some means in order to label new instances.

The Advanced features can be acquired from previously existing features and added to the feature vector. Adding new features from the set of existing features is known as feature construction.

In this work the feature vector is identified using Quasi-Newton Method called L-BFGS (Limited-memory BFGS (L-BFGS or LM-BFGS) method to identify the optimal feature weight effectively. Feature Space based CRF algorithm will validate the model using the feature vector generated by the feature space.

Table.2. Precision values for Breast Cancer, Lung Cancer and Thyroid Cancer Corpus

Precision						
Learning Algorithm	Breast Cancer Corpus		Lung Cancer Corpus		Thyroid Cancer Corpus	
Classes	SVM	Feature Space based CRF	SVM	Feature Space based CRF	SVM	Feature Space Based CRF
Protein	86.05	87.77	82.32	91.22	94.56	88.2
Gene	88.09	85.56	84.52	93.55	81.24	84.02
DNA	82.51	90.99	94.35	90.67	87.92	93.46
RNA	86.75	95.87	84.52	86.16	84.56	94.71
Cell Line	82.28	94.7	86.52	95.8	83.25	94.59
Cell Type	84.62	94.57	96.87	93.42	97.26	94.66

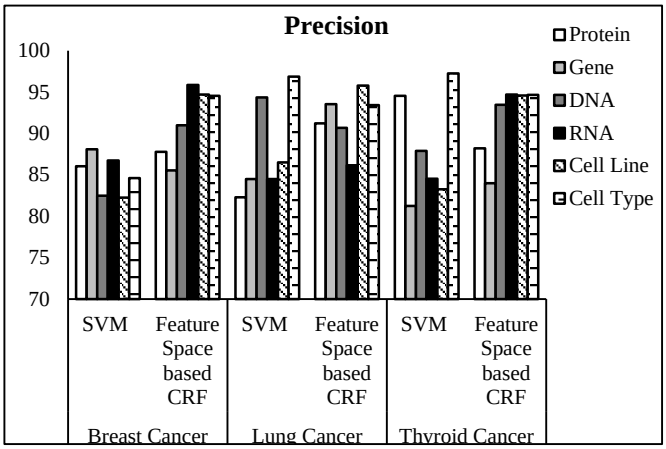

Fig.6. Precision values for Breast Cancer, Lung Cancer and Thyroid Cancer Corpus

Table.3. Recall values for Breast Cancer, Lung Cancer and Thyroid Cancer Corpus

Recall						
Learning Algorithm	Breast Cancer Corpus		Lung Cancer Corpus		Thyroid Cancer Corpus	
Classes	SVM	Feature Space Based CRF	SVM	Feature Space Based CRF	SVM	Feature Space Based CRF
Protein	86.52	94.02	84.21	94.3	82.54	90.36
Gene	82.35	95.07	82.14	96.67	86.35	97.54
DNA	96.35	93.13	85.32	91.77	98.53	94.53
RNA	89.36	98.97	81.62	97.61	97.79	97.79
Cell Line	99.63	96.76	89.36	97.89	96.35	94.59
Cell Type	92.62	96.69	83.25	96.6	84.53	96.68

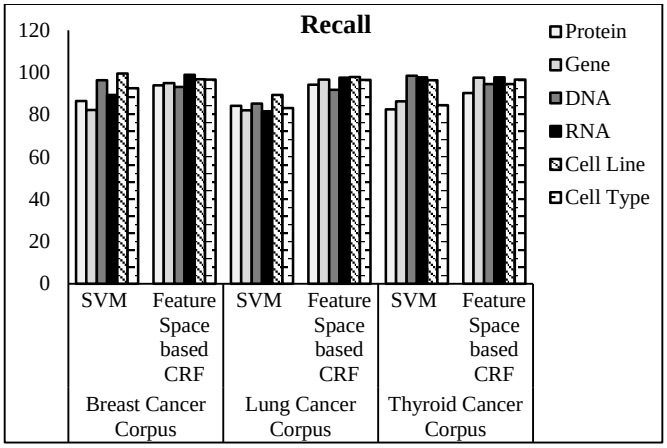

Fig.7. Recall values for Breast Cancer, Lung Cancer and Thyroid Cancer Corpus

Table.4. F-Measure values for Breast, lung, Thyroid Cancer Corpus

F-Measure						
Learning Algorithm	Breast Cancer Corpus		Lung Cancer Corpus		Thyroid Cancer Corpus	
Classes	SVM	Feature Space Based CRF	SVM	Feature Space Based CRF	SVM	Feature Space Based CRF
Protein	0.86	0.91	0.83	0.93	0.88	0.89
Gene	0.85	0.9	0.83	0.95	0.84	0.9
DNA	0.89	0.92	0.9	0.91	0.93	0.94
RNA	0.88	0.97	0.83	0.92	0.91	0.96
Cell Line	0.9	0.96	0.88	0.97	0.89	0.95
Cell Type	0.88	0.96	0.9	0.95	0.9	0.96

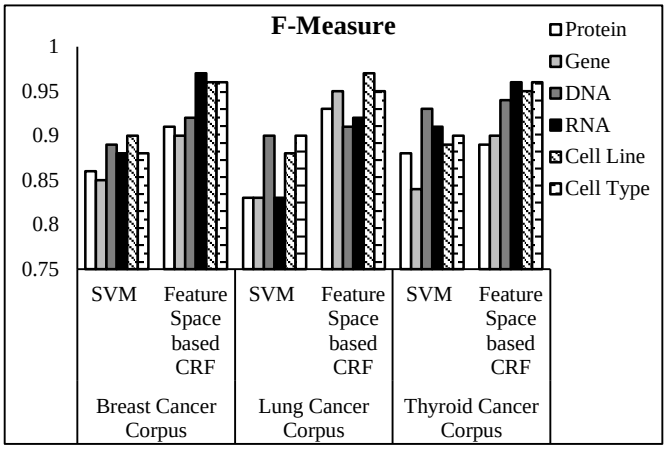

Fig.8. F-Measure values for Breast, lung, Thyroid Cancer Corpus

5. CONCLUSION

The proposed work to extract the entity from biomedical documents using the machine learning technique Feature space based Conditional Random Field works well within its limitations. The outcome of the work compared with the machine

learning algorithm Support Vector Machine. In many cases of identifying the entities of various classes, FS-CRF (Feature Space based Conditional Random Field) does better than Support Vector Machine. The result shows that the proposed work is better than the other techniques used for the same purpose. The feature work focus on the extraction of association (relationships) between the recognized entities and Event Extraction. Named Entity Recognition is a prerequisite for relation extraction systems.

REFERENCES

[1] Burr Settles, "ABNER: An Open Source Tool for Automatically Tagging Genes Proteins and other Entity Names in Text", *Bioinformatics*, Vol. 21, No. 14, pp. 3191-3192, 2005.

[2] Yoshimasa Tsuruoka, Junichi Tsujii and Sophia Ananiadou, "Accelerating the Annotation of Sparse Named Entities by Dynamic Sentence Selection", *BMC Bioinformatics*, Vol. 9, No. 11, pp. 1-10, 2008.

[3] Fei Zhu, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, Bairong Shen, "Biomedical Text Mining and its Applications in Cancer Research", *Journal of Biomedical Informatics*, Vol. 46, No. 2, pp. 200-211, 2013.

[4] Junichi Kazama,Takaki Makino,Yoshihiro Ohta and Junichi Tsujii, "Tuning Support Vector Machines for Biomedical Named Entity Recognition", *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, pp. 1-8, 2002.

[5] Yi Feng Lin, Tzong Han Tsai, Wen Chi Chou, Kuen Pin Wu, Ting Yi Sung and Wen Lian Hsu, "A Maximum Entropy Approach to Biomedical Named Entity Recognition", *Proceedings of 4th Workshop on Data Mining in Bioinformatics*, pp. 56-61, 2004.

[6] Lishuang Li, Wenting Fan and Degen Huang, "A Two-Phase Bio-NER System Based on Integrated Classifiers and Multiagent Strategy", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 10, No. 4, pp. 897-904, 2013,

[7] Lishuang Li, Rongpeng Zhou and Degen Huang, "Two-Phase Biomedical Named Entity Recognition using CRFs," *Computational Biology and Chemistry*, Vol. 33, No. 4, pp. 334-338, 2009.

[8] Z. Ju, J. Wang and F. Zhu, "Named Entity Recognition from Biomedical Text Using SVM", *Proceedings of 5th International Conference on Bioinformatics and Biomedical Engineering*, pp. 1-4, 2011.

[9] Sriparna Saha, Asif Ekbal and Utpal Kumar Sikdar, "Named Entity Recognition and Classification in Biomedical Text using Classifier Ensemble", *International Journal of Data Mining and Bioinformatics*, Vol. 11, No. 4, pp. 365-391, 2015.

[10] Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka and Yuka Tateisi, "Introduction to the Bio-Entity Recognition Task at JNLPBA", *Proceedings of International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 70-75, 2004.

[11] Phoebe M. Roberts, "Mining Literature for Systems Biology", *Briefings in Bioinformatics*, Vol. 7, No. 4, pp. 399-406, 2006.

[12] David P Hill, Barry Smith, Monica S McAndrews Hill and Judith A Blake, "Gene Ontology Annotations: What They Mean and Where They Come From", *BMC Bioinformatics*, Vol. 9, No. 5, pp. 1-9, 2008.

[13] S. Ananiadou, D.B. Kell, and J.I. Tsujii, "Text Mining and its Potential Applications in Systems Biology", *Trends in Biotechnology*, Vol. 24, No. 12, pp. 571-579, 2006.

[14] Manabu Torii, Zhangzhi Hu, Cathy H. Wu, and Hongfang Liu, "BioTagger-GM: A Gene/Protein Name Recognition System", *Journal of the American Medical Informatics Association*, Vol. 16, No. 2, pp. 247-255, 2009.

[15] Douglas E. Appelt, "Introduction to Information Extraction", *AI Communications*, Vol. 12, No. 3, pp. 161-172, 1999.

[16] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland and Mausam, "Open Information Extraction: The Second Generation", *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 3-10, 2011

[17] Rosie Jones, Rayid Ghani, Tom Mitchell and Ellen Riloff, "Active Learning for Information Extraction with Multiple View Feature Sets", *Proceedings of Workshop on Adaptive Text Extraction and Mining*, 2003

[18] Daniel M. Bikel, Richard Schwartz and Ralph M. Weischedel, "An Algorithm that Learns What's in a Name", *Machine Learning - Special Issue on Natural Language Learning*, Vol. 34, No. 1, pp. 211-231, 1999.

[19] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld and Alexander Yates, "Unsupervised Named-Entity Extraction from the Web: An Experimental Study", *Artificial Intelligence*, Vol. 165, No. 1, pp. 91-134, 2005.

[20] Fei Zhu, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, Bairong Shen, "Biomedical Text Mining and its Applications in Cancer Research", *Journal of Biomedical Informatics*, Vol. 46, No. 2, pp. 200-211, 2013.

[21] A.M. Cohen and W.R. Hersh, "A Survey of Current Work in Biomedical Text Mining", *Briefings in Bioinformatics*; Vol. 6, No. 1, pp. 57-71, 2005.

[22] Irena Spasic, Sophia Ananiadou, John McNaught and Anand Kumar, "Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text", *Briefings In Bioinformatics*, Vol. 6, No. 3, pp. 239-251, 2005

[23] Lishuang Li, Rongpeng Zhou, Degen Huang. "Two-Phase Biomedical Named Entity Recognition using CRFs", *Computational Biology and Chemistry*, Vol. 33, No. 4, pp. 334-338, 2009.