International Journal of Biomedical Research
ISSN: 0976-9633 (Online); 2455-0566 (Print)
Journal DOI:10.7439/ijbr
CODEN: IJBRFA

Short Communication

A machine learning approach for Biomedical Named Entity Recognition

Kanimozhi U[*1] and Manjula D[2]

[1]*Research Scholar, Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu 600025 India*
[2]*Professor, Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu 600025 India*

***Correspondence Info:**
Kanimozhi U
Research Scholar,
Department of Computer Science and Engineering,
College of Engineering Guindy, Anna University,
Chennai, Tamil Nadu 600025 India
E-mail: kanimozhiu.03@gmail.com

Abstract

This paper aims to develop a Named Entity Recognition (NER) that creates new tags that facilitate fast query processing, information retrieval and data preprocessing of Biomedical Domain. We have used a machine learning approach that uses domain specific knowledge to train the data and label the entities with appropriate tags. Conditional Random Fields (CRF) is implemented and used to train the input domain specific data that yields good performance. Experimental and evaluation result shows that the learned model yields a Biomedical domain specific NER recall of 82%, precision of 85%, accuracy of 82% and F-measure of 83%. Thus learned CRF model builds a domain specific NER for Biomedical Domain and tags the domain keywords with appropriate tags.

Keywords: Named Entity Recognition; Statistical Model; Machine Learning; Conditional Random Field Classifier; Bio medical Data analytics.

1.Introduction

There is an explosion of biomedical literature in the past decades. It is clear that the overwhelming amount of biomedical literature could only be managed efficiently with the aid of automated text information extraction methods. Our aim is to automate the transfer of unstructured textual information into a structured form. As reported by Fog Computing World [3], huge volumes of zettabytes of data are generated worldwide through internet and other sources which are evolving in digital and non-digital form. Ge *et al* (2014) pointed out that due to the huge volume of date, during the current years there is a shift to data driven marketing. These data consists of very useful knowledge and hidden insights. Enormous data are generated in Biomedical Domain through articles, web etc., The necessity to retrieve the required information from internet that is domain specific solves problems and can be used for fast query processing, efficient understanding of context of data, understanding, improving and for personalized healthcare recommendation.

In order to retrieve information from this data, the data has to be cleaned and pre-processed. One of the pre-processing steps is the extraction of named entities from text Named Entity Recognition (NER) in data analysis. According to Wikipedia, "NER is a subtask of information extraction that explore to discover and categorize elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.". It is also called as entity identification or entity chunking or entity extraction. Named entity recognition of well-defined objects, such as genes or proteins, diseases has achieved a sufficient level of maturity such that it can form the basis for the next step: the extraction of relations that exist between the recognized entities. The best standard example for NER is Stanford's Named Entity Tagger. This Stanford NER developed by Finkel *et al* (2005) tags the given sequence of words into name of person, organization, percent, money, data, location and time.NER is basically used for defining entities in generic domain. The generic domain NER is not suitable for Closed or Domain Specific NER as the characteristic changes between domains. The generic domain NER can only retrieve common information such as name of the people,

places and organizations. The predefined tags and entities and tags should be modified and tuned to suit the Domain Specific NER. There has been information retrieval in all fields like biological, chemical, agriculture and medical industry. As basic standard NER are not helpful to retrieving domain related entities and keywords it resulted in the need for domain specific NER. Lot of research work has been carried out in the medical domain and few researches in the agricultural domain.

In the current scenario, all applications related to Natural language processing rely on a domain independent NER for their needs with development of domain dependent NER is on the increase in the areas of Bioinformatics, Agriculture, etc. Due to the high source of biomedical literature there is a need to gain high knowledge and intelligence that provides real-time, incremental and right decision making and for the provision of personalized treatment recommendations. In order to provide significant advancement in the NER for Biomedical Industry, we have decided to go ahead with the CRF based machine learning approach for domain specific NER as the rule based approach cannot handle all possible scenarios and a HMM based approach would require a huge data corpus. As the application for Biomedical NER is multifold, we proposed a set of tags as named entities over and above the generic domain NER. These tags are very useful in information retrieval and question answering for domain specific data and very important useful preprocessing tool for biomedical data analytics.

The rest of the paper is organized as follows: Section 2 provides the related works on NER. Section 3 explains the materials and methods used in the proposed framework of the work done. Section 4 gives an insight about the various evaluation measures of NER and results of our NER. Section 5 discusses the results of our Biomedical NER. Section 6 is concluded with summary of the research work done along with the future directions.

1.1 Related Work

In the Biomedical Domain, NER is the subtask of information extraction that automatically labels the entities such as genes, proteins, diseases etc., The preliminary works that had been done in the area of NER are explained below.

There are several methods to develop a NER such as statistical analysis, semantics analysis, knowledge base, domain specific, rule based, machine learning approach such as supervised, semi-supervised and unsupervised learning and Hybrid. The research paper on NER was initially based on handcrafted rules and heuristic methods published in the year 1991 by Rau *et al* (1991). MUC-6 triggered the use of NER which introduced many other NERs not only in English language but in many other languages such as Chinese, French, Greek, Italian, Hindi, Punjabi and Malayalam. The Named Entity Recognition and Classification was introduced by Coates *et al* (1992) which further classified the entities into subcategories. Thielen *et al* (1995) also introduced the machine learning technique of classification into NER for classifying the entities.

Rule based Named Entity Recognition system enables extraction of real life task through a set of rules which are either hand written or learnt through various real life examples. A rule based NER system basically consists of set of rules and set of policies governing the rules. There are various approaches of NER used in the literature like multiple entity recognition, boundary recognition and whole entity recognition. Cunningham *et al* (2002) used the whole entity recognition, which is the classic approach that is used in rule based NER. This classic NER ensures that there is no dependency between various entities and the rule is modeled based on left and right content of the keyword to be named which was proposed by Cunningham *et al* (2002) and Mary *et al* (2003). Fabio *et al* (2001) and Dayne *et al* (2000) proved that the rules can be applied to find out the boundaries using the left and the right context and various rules are identified to independently identity the tags using pre and post words. Rules for Multiple Entities can be used for modeling the dependency that exists between entities Soderland (1999).

Machine Learning based NER system, the identifying suitable tags are considered as a classification problem and statistical method is used to solve it. There are many machine learning supervised approaches that had been used for Named Entity recognition such as Hidden Markov Model (HMM), Decision Trees, Maximum Entropy Models (ME) and Support Vector Machines. These supervised models learn rules based on discriminative features. HMM model is the first machine learning model that is used in English language for solving NER. IdentiFinder designed by Daniel *et al* (1999) used HMM model for solving NER problem. Maximum Entropy based machine learning model is used for classification of NER by directly learning the weightage for discriminative features. The MENE system proposed by Andrew (1999) and Curran *et al*'s [1] ME Tagger applied the Maximum Entropy algorithm for tagging entities. SVM was used by Paul *et al*[10] to tackle NER as a binary decision problem.

The semi supervised machine learning algorithms that are used for NER are CRF and Bootstrapping based methods. This algorithm solves the problem of unavailability of golden standard data and sparse availability of data. CRF algorithm was later used in which learning is done based on input which is sequence of words by John *et al*[4]. Finding NER tags using Adaboost

algorithm which is a boot strapping based machine learning method was proposed in the year 2002 by Xavier *et al*[16], which uses BIO labeling scheme. BIO is a very popular model which is used in NER where B indicates beginning word of NE, I point the inside or intermediate word in a NE and O is the word outside the NE.

Unsupervised machine learning methods ruled out the need for large annotated corpus and it build the representation from data. KNOWITALL is a NER which uses domain independent system which extracts information from the web proposed by Oren *et al* (2005). Unsupervised NER across various languages was also developed by Robert *et al*[12]. The Hybrid NER is the combination of both rule based and machine learning based approaches. It uses the strength of both the approaches while eliminating its weakness. A Hybrid NER called LTH system was proposed by Mikheev *et al* (1998) which uses a document centered approach. A NER that used the combination of hand-crafted rules, HMM and Maximum Entropy model was also proposed by Srihari *et al*[14].

NER has been designed for scientific, religious text and also emails by Einat *et al* (2005). Lot of research work has been done in NER for biomedical domain by Kazama *et al*[5], Hwang *et al* (2003) and Saha *et al*[13]. NER for agriculture domain called AGNER has been proposed to identify the various crops and pesticide names by Payai *et al*[11]. Rule Based NER provides results that have high precision whereas the Statistical method provides high recall. Hence, it is efficient and easier to tune statistical system to improve the precision which is better than the effort that is put to tune rule based system to increase the recall.tmChem is a very high performance approach that used ensemble machine learning method for NER, designed by Leaman *et al* [7]. DBpedia data is analyzed for named entity recognition and linking of tweets by Derczynski *et al*[2]. Biomedical Named Entity Recognition (BNER) which used clustering-based representation, distributional representation, and word embeddings for representing word features for NER designed by Tang *et al* (2014). Yang *et al*[17] used semi-markov's conditional random fields' machine learning algorithms to explore the features of two phased bio-medical NER.

Now big data is a buzz word everywhere. Li *et al*[8] designed an NER using Map Reduce paradigm for biomedical domain. Lao named entity recognition proposed by Yang *et al*[17] uses simple heuristic information along with conditional random fields' algorithm. Konkol *et al*[6] proposed a language independent NER using latent semantics which used unsupervised methods for extracting new features. Neural architecture based NER is designed by Lample *et al*(2016), which gave great performance without using any language specific knowledge base.

2. Methods and Materials

2.1 Proposed Framework for NER

The system design flow is given in figure 1. The research is focused on developing a NER for specific biomedical domain. The input is the domain specific text and the output is the keywords which are tagged with domain specific defined NER tags. The biomedical literatures are collected from the PubMed, Medline are taken as input corpus. These unstructured texts collected from web should be preprocessed and cleaned in order for further processing. The preprocessed input is tagged by Parts of Speech Tagger (POS) that help us to identify the context of the keywords such as nouns, noun phrases, and adverb adjective. The Stanford POS tagger is used which is a generic domain POS tagger. After learning the context of the words it is given as an input to the tokenizer. The tokenizer tokens the given input and stores it in a .csv file which enables us to label the tokens with the appropriate tags. This tokenized labeled input training data set is given as an input to the CRF Classification algorithm which based on the labeled input trains and models the biomedical NER system.

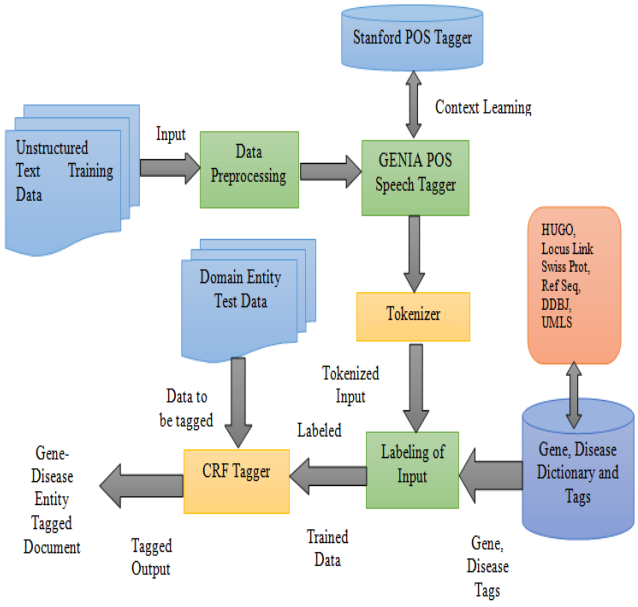

Figure 1: Domain Specific NER Workflow

When a Biomedical domain specific document is given as a test data to the CRF algorithm it labels the domain keywords with the trained tags. The detailed description of the proposed design is explained in the section below.

2.2 Proposed Work

2.2.1 Data Collection

The development corpus is a subset of PubMed and MEDLINE abstracts dealing with Huntington Disease, gene and Neurodegenerative disorder. It was annotated with disease and gene relations, based on "etiology" and

"clinical marker". The purpose of building an annotated corpus is to construct the training data for machine learning that will filter out false positives from the dictionary-based results. These data are used for training and testing purposes. The input corpus consists of text related to Huntington Disease, gene and all words related to Neurodegenerative disorder. The input corpus which is manually curated has 6996 sentences and 140481 words.

2.2.2 Data Pre-processing

The given input data is preprocessed to remove the irrelevant data. The irrelevant data is the punctuation and stop words. Since many cancer related biomedical literatures contain mentions of genes/proteins, disease names and chemical compounds one type of false positive errors was caused by mistakenly recognizing a chemical compound mention as a gene/protein. Annotated mentions in the training set were used as features to reduce such errors. The data is cleaned and given as an input to the POS Tagger.

2.2.3 Generic Domain POS Tagging

The standard Stanford Generic Domain POS Tagger is used to learn the cleaned input text. It is used to tag each word as noun, adverb, verb, adjective, etc. It can be used to learn the context of the words which is based on relationship with adjacent and related words. The named entities are filtered by considering orthographic features and syntactic features. To obtain syntactic features, we used the ENJU full parser [1] and the GENIA Part of Speech Tagger [2].

2.2.4 Tokenizer

The preprocessed cleaned data after learning the context is given as input to the Tokenizer. Given that NER is a sequential labeling problem, the optimal decision made by the NER system is based on the labels of tokens in the whole sentence. Thus, the global context surrounding a candidate term is an important factor to be considered. For this, the semantic type information present in the context of a candidate term was generated as a feature by matching the concept terms in UMLS. The job of the tokenizer is to split each of the input text tokens and save it in a excel file as .csv file. Tokenizing and saving it as a .csv file is to facilitate preparation of the training data.

2.2.5 Biomedical Domain Vocabulary

Our system first collects sentences that contain at least one pair of disease and gene names, from machine learning based methods to filter out false positives from the dictionary matching results. The named entity recognizer filtered out a lot of false positive disease and gene names. We constructed disease and gene dictionaries and used them to augment the corpus of sentences collected in step one ("disease gene pairs sentences"). Thus, all disease and gene names in the results of the proposed method have ID tags that are used in the publicly available biological databases (HUGO, LocusLink, SwissProt, RefSeq, DDBJ and UMLS). When a disease gene pair's sentence contained more than one disease name and more than one gene name, the system made sufficient copies of the sentence to accommodate all possible disease gene name pairs. Based on these keywords the tokenized input can be labeled with their respective tags.

2.2.6 NER Tags for Biomedical Domain

In this work, we considered the biomedical vocabulary and various standard ontologies for biomedical domain and the various entities / tags that are useful for information extraction are defined as Named Entities. We used them to train the Conditional Random Fields (CRF) based disease and gene name recognizer.

2.2.7 Labeling the input

The domain specific keywords are identified using the dictionary and vocabulary. Semantic and domain specific knowledge is applied on the training corpus and used for tagging the given input data. Now each and every tokenized input are labeled with their respective tags and saved. Only the domain keywords are tagged. This serves as the input data for CRF classification algorithm. Since CRF is a semi supervised algorithm the NER for Biomedical data has to be modeled with our training data.

2.2.8 CRF Tagger

The preprocessed data need to be trained, for the machine to learn the tags. Each and every preprocessed data is tokenized and input is labeled. Entities related to biomedical domain are labeled with their new tags. The non-domain entities are labeled as 'O'. BIO2 Annotation standard is used. The Beginning and intermediate positions of the entities are marked with B- and I-tags. The training set consists of words that are tokenized and labelled with appropriate tags and where the test set consists of words and sentences.

The statistical modeling method used to learn the labeled test data is Conditional Random Fields (CRFs). It is a machine learning algorithm that is widely used for pattern recognition and structured prediction. CRF is a discriminative undirected probabilistic graphical model. CRF algorithm is the most efficient than any other machine learning algorithm because it predicts sequence of labels for sequence of input text taking the context of the input into account. The given input data and its labeled NER is learnt using CRF algorithm which was used by Jenny *et al* (2005). The CRF algorithm based on the learned input tags the domain related keywords and entities of the test data.

2.2.9 Test Data

The test data related to Huntington Disease, gene and neurodegenerative disorder is collected from the PubMed abstracts and fed as input to the CRF classifier.

The classifier based on the trained input data labels the keywords of the document with the appropriate tags.

2.2.10 Output

Output is the tagged domain specific keywords which could be used for further processing for Biomedical Data Analytics.

3. Experimental Results

According to Wikipedia, the evaluation of the NER Task is done by Precision (P) and Recall (R) and F-Measure. Precision is defined as the percentage of entity labeled is correct with respect to the gold standard evaluation data. In the area of Information Retrieval the fraction of documents that are retrieved which are relevant to the query is called as Precision.

Precision in other words is the number of true positives which is the number of words correctly tagged as biomedical domain specific keywords divided by the sum of the true positives and false positives which is number of the words tagged as domain keywords by Biomedical NER.

Recall is the measures of the number of names in the gold standard that are present at exactly the same location in the predictions are correctly labeled. Precisely in the field of Information Retrieval the fraction of documents which are relevant to that has been successfully retrieved is called as Recall.

$$\text{Precision} = \frac{|(\#\text{domain specific keywords in the document}) \cup (\#\text{domain specific keywords tagged})|}{|(\#\text{domain specific keywords tagged})|}$$

$$\text{Recall} = \frac{|(\#\text{domain specific keywords in the document}) \cup (\#\text{domain specific keywords tagged})|}{|\#\text{domain specific keywords in the document}|}$$

Recall in other words is the number of true positives which is the number of words correctly tagged as biomedical domain specific keywords divided by the sum of the true positives and false negatives which is the actual number of relevant words related to biomedical domain.Precision and recall result are combined together to form F-measure of NER performance. It is calculated by the uniformly weighted harmonic mean of precision and recall.

$$F = 2 . \frac{\text{precision} . \text{recall}}{\text{precision} + \text{recall}}$$

Accuracy of our Named Entity Recognizer can be determined by the sum of the keywords correctly tagged as Travel domain and non-domain keywords by the total number of words in the document.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where 'TP' is the true positive value – number of words correctly tagged by our NER, 'TN' is the true negative value – number of words correctly classified as non-biomedical domain keywords, 'FP' is the false positive value – number of words wrongly tagged as Biomedical domain specific keywords by our NER and 'FN' is the false negative value – number of Biomedical domain specific keywords which is not tagged by our NER.

We have made an evaluation for Biomedical NER using domain related input corpus developed. The system shows more accuracy for cancer related data compared to the Stanford NER Tool

Table 1: Evaluation of Domain Specific Input

System	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)
Biomedical Domain specific NER	85	82	83	82

Table 1 Enlightens with the performance of our Domain Specific NER for domain specific output. We compared our output with the available dictionaries to find out the accuracy of our NER. Out of tagged 71645 words, 59465 words are correctly tagged as biomedical domain specific keywords. Thus our NER system has a precision of 85%. 12180 words are wrongly classified as our domain specific Keywords. 13053 keywords were not tagged as domain specific Keywords by our NER System. So our system had a Recall of 82%. The F measure which is the harmonic mean of precision and recall is 83%. Accuracy of our NER system is 82%.

4. Discussion

The input document consists of 6996 sentences and 140481 words which are specific to Biomedical Domain. In order to evaluate our NER the same domain related document is given as an input to both Stanford NER and our Domain specific NER, in order to compare the performance. The 7 class Stanford's named entity tagger tags only 22 % of the words into various entities. By using our domain specific NER, the tags increased by 29%. So there is an increase in the number of tags due to entities related to gene and disease. Our NER system was not able to classify the names of the gene mentions in most of the cases. Similarly our NER system was not able to define the names of the gene and disease due to which there was decrease in our Precision and Recall.

Table 2: Confusion Matrix

		Predicted Condition			
	Total Records 140481 words	Predicted Condition positive	Predicted Condition negative	Prevalence 52%	
True Condition	Condition Positive 72518 words	True Positive 59465 words	False Negative 13053 words	True Positive Rate (TPR), Sensitivity, Recall 82%	False Negative Rate (FNR), Miss Rate 18%
	Condition Negative 67963 words	False Positive 12180 words	True Negative 55783 words	False Positive Rate (FPR), Fallout 18%	True Negative Rate (TNR), Specificity 82%
	Accuracy 82%	Positive Predicted Value (PPV), Precision 85%	False Omission Rate 19%	Positive Likelihood Ratio 4.56	Diagnostic Odds Ratio 20.72
		False Discovery Rate 17%	Negative Predictive Rate 81%	Negative Likelihood Ratio 0.22	

Our NER has tagged 59465 out of 72518 words of the biomedical domain keywords as true positive which gives correctness of 82%. 55783 words non domain keywords are correctly classified as true negative. Also, 12180 words are tagged as false positive which are mainly due to discrepancy between names of genes/proteins, diseases, miRNA, chemical compounds. Also the false positive rate increased because the numbers were incorrectly tagged as genes/proteins and chemical compounds. 13053 words of domain keywords are wrongly classified as non-domain keywords which fall under false negatives. The name of gene/protein mentions were wrongly classified as chemical compounds. Reducing the false negative value in turn will improve the performance of our NER.

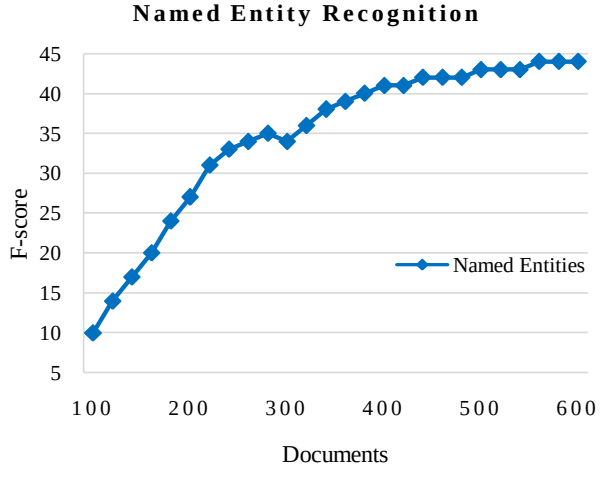

Figure 2: A 500 Fold CRF Learning curve for NER

The experiments are developed on a windows-based laptop machine with 4GB DDR3 memory and IntelI5 1.6MHz processor using Mat lab 6.0 release 12. Figure 2 depicts a 500 fold cross validation of the proposed CRF learning algorithm to identify and tag the named entities. Training samples range from 25 to 675 documents with a step of 25; testing samples are always set to 75.

A confusion matrix is used to enable us to know the performance of the classification algorithm to facilitate in defining tags for our biomedical domain. Table 2. Helps us in analyzing the performance of our system using various measures using confusion matrix. We need to improve our NER system by providing various training set to include all possible classification related to Biomedicine so as to enable the machine to learn better. But our system proved to be more effective for documents related to biomedical domain than our Generic Domain Stanford's NER tagger. The Standard Stanford NER tagger was able to tag only 30906 keywords from the document which are all Generic Domain entities. Thus our NER system proved best to identify the domain related keywords.

5. Conclusion

We have developed a domain specific Named Entity Tagger for Biomedical domain. In this work we took domain specific inputs as input corpus as an input. After various preprocessing steps, the tokenized data is labeled with the new tags defined. The input labeled data is given as an input to CRF classification algorithm which learned and builds a suitable NER model for Biomedical Domain. When a domain related input is given as test data it tags the keywords of the document with the appropriate tags. It can be used as a data preprocessing step to collect entities of a biomedical domain. It will be helpful for researchers to design a structured dataset based on the unstructured input which can be used for further analytics for the development of the biomedical analytics. This research work will facilitate the research work in the Biomedicine for mining and recommendations.

References

[1] Curran, J. R. Clark, S. Language independent NER using a maximum entropy tagger. *In Proceedings of the 7thCoNLL*. 2003: 164–167. https://aclweb.org/anthology/W/W03/W03-0424.pdf

[2] Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Bontcheva, K. Analysis of named entity recognition and linking for tweets. *Information Processing and Management.*2015;51: 32–49.

[3] Fog Computing World. 2014. Internet of Things to Generate Zettabytes of Data by 2018.

[4] John, D. L. Andrew, M. Fernando, C. N. P. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the Eighteenth International Conference on Machine Learning*, 2001; ICML '01, pages 282-289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. http://dl.acm.org/citation.cfm?id=645530.655813

[5] Kazama, J. Makino, T. Ohta, Y. Tsujii, J. Tuning Support Vector Machines for Biomedical Named Entity Recognition. *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, 2002; Philadelphia,pp. 1-8. Association for Computational Linguistics.

[6] Konkol, M. Brychcín, T. Konopík, M. Latent semantics in Named Entity Recognition. *J. Expert Systems with Applications*. 2015; 42: 3470–3479.

[7] Leama, R. Wei CH, Lu, Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *J Cheminform.*2014; 7(Suppl 1):S3.

[8] Lee, K.J. Hwang, Y. S. and Rim, H.C. Two-Phase Biomedical NE Recognition based on SVMs. *Proceedings of the ACL 2003 Workshop on NLP in Biomedicine*, 2003; 33-40.http://www.aclweb.org/anthology/W03-1305.pdf

[9] Li K, Ai W, Tang, Z. Hadoop recognition of biomedical named entity using conditional random fields. In: *IEEE transaction parallel distribution system*. 2015; pp 1–1. DOI=10.1109/TPDS.2014.2368568.

[10] Paul, M. James, M. Entity extraction without language-specific resources. *In proceedings of the 6th Conference on Natural Language Learning*2002; 20, COLING-02: 1-4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. DOI=10.3115/1118853.1118873.

[11] Payai, B. Aditi, S. Ashish, K. AGNER: Entity tagger in Agriculture domain, *2nd International Conference on Computing for Sustainable Global Development*2015.

[12] Robert, M. Christopher, D. M. Accurate unsupervised joint named-entity extraction unaligned parallel text. *In Proceedings of the 4th Named Entity Workshop*, NEWS '12, 2012: 21-29, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2392111.2392781.

[13] Saha, S.K., Sarkar, S., Mitra, P., Feature Selection Techniques for Maximum entropy based biomedical named entity recognition. *Journal of Biomedical Informatics*2009;42(5): 905-911.

[14] Srihari, R. Niu, C. Li, W. A Hybrid Approach for Named Entity and Sub-Type Tagging. *Proceedings of the sixth conference on Applied Natural Language Processing*, ACM, 2000: 247-254. http://dl.acm.org/citation.cfm?id=974181

[15] Wenhui, L. Sriharsha, V. A Simple Semi-supervised Algorithm for Named Entity Recognition, *Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing*, 2009: 58–65, Boulder, Colorado,Association for Computational Linguistics.http://dl.acm.org/ft_gateway.cfm?id=1621837&type=pdf

[16] Xavier, C. Lluis, M. Lluis, P. Named entity extraction using Adaboost. *In proceedings of the 6th Conference on Natural Language Learning* – 2002; 20, COLING-02: 1-4. Stroudsburg, PA, USA. Association for Computational Linguistics. Doi=10.3115/1118853.1118857.

[17] Yang, M. Zhou, L. Yu, Z. Gao, S. and Guo, J. Lao Named Entity Recognition based on Conditional Random Fields with Simple Heuristic Information. *12th IEEE International conference on Fuzzy Systems and Knowledge Discovery*2015.