Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN : 2348-4470
p-ISSN : 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue on Recent Trends in Data Engineering

Volume 4, Special Issue 5, Dec.-2017

Bio Medical Named Entity Recognition Using Machine Learning Algorithms

Dr. K.S. Wagh[1], Aishwarya Kulkarni[2], Pratiksha Pawar[3], Neha Kirange[4], Shraddha Kashid[5]

Department of Computer Engineering,
All India Shri Shivaji Memorial Society's,
Institute of Information Technology,
Kennedy Road -411001

Abstract —*Named-entity recognition system (NER) [1] identifies different entities in many ways such name of person locations, and organizations from news articles, reports, blogs, tweets. Main steps of named entity recognition are boundary detection of entities and classification of entities into already defined classes. This results of recognition and classification is widely used in information retrieval and extraction.*

The main component of biomedical natural language processing is named entity recognition system which extracts information from the text and finally does the knowledge discovery. As amount of health and biomedical text being available is huge and since much of the data is recorded in non-structured text, like in clinical notes and biomedical publications the bottleneck of biomedical information processing is how to make use of the knowledge resources and build scalable models to process large amounts of text. Biomedical named-entity recognition (BM-NER), also known as biomedical concept identification or concept mapping, is a key step in biomedical language processing.

In this paper, we are proposing a Biomedical Named Entity Recognition System for extracting Name, Problem and Test from the Textual Clinical Lab Reports using two widely accepted datasets, i2b2 [2] and GENIA corpora [3]and we are attempting to correlate the appropriate Treatment associated with it using Machine Learning Algorithms [4] and Natural Language Processing [5].

Rest of the paper is organized as follows, section 2 gives in depth literature survey, in section 3 we discuss different approaches used for Bio-NER. In section 4 describes proposed system architecture.

Keywords*: BM-NER, Machine Learning, K-nearest neighbor, N-gram, NP-Chunker, IDF.*

I. INTRODUCTION

Named-entity recognition system (NER) [1] identifies different entities in many ways such name of person locations, and organizations from news articles, reports, blogs, tweets. Main steps of named entity recognition are boundary detection of entities and classification of entities into already defined classes. This results of recognition and classification is widely used in information retrieval and extraction.

The main component of biomedical natural language processing is named entity recognition system which extracts information from the text and finally does the knowledge discovery. As amount of health and biomedical text being available is huge and since much of the data is recorded in non-structured text, like in clinical notes and biomedical publications the bottleneck of biomedical information processing is how to make use of the knowledge resources and build scalable models to process large amounts of text. Biomedical named-entity recognition (BM-NER), also known as biomedical concept identification or concept mapping, is a key step in biomedical language processing.

In this paper, we are proposing a Biomedical Named Entity Recognition System for extracting Name, Problem and Test from the Textual Clinical Lab Reports using two widely accepted datasets, i2b2 [2] and GENIA corpora [3]and we are attempting to correlate the appropriate Treatment associated with it using Machine Learning Algorithms [4] and Natural Language Processing [5].

Rest of the paper is organized as follows, section 2 gives in depth literature survey, in section 3 we discuss different approaches used for Bio-NER. In section 4 describes proposed system architecture.

II. LITERATURE SURVEY

1. **A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set**

Ahmed Sultan Al-Hegami et al [6] proposed that This paper gives the comparison between different subsets of features and three classification approach (Naïve Bayes, K-Nearest Neighbour and decision tree) for biomedical named entity recognition. The proposed model uses Support Vector Machine(SVM) and Naive Bayes algorithms which are satisfactory and effective for BNER. More time required to train complex algorithm. Does not perform well on multiple class tasks.

2. A Survey on Biomedical Named Entity Extraction

Almas Tasneem et al [7] proposed a method containing Rule based, Dictionary based and Machine learning based methods. This paper includes the challenges perceived by the researchers in BIO-NER task and investigates the works done in the field of BIO-NER by using the multiple approaches available for the task. Supervised machine learning based approaches face the problem in creating large enough training sets.

3. Biomedical Named Entity Recognition - a swift review

S.Vijaya et al [8] proposed an idea that Hybrid approaches as the results showed are greater than using the approaches alone. Using classifiers with Hybrid approaches can be used to improve the precision and recall rate. Hybrid approach is used to improve efficiency of system.

4. A Comparison of Named Entity Recognition Tools Applied to Biographical Texts

Samet Atdağ et al [9] proposed a model that giving description about NER tools. Ner tools mainly focus is on selection of an appropriate named entity recognition (NER) tool for biographic texts.

5. A Comparative Study of Segment Representation for Biomedical Named Entity Recognition

H.L.Shashirekha et al [10] proposed the model with the three phases- Tokenization, Boundary detection, Type classification. Support Vector Machines (SVMs) and Conditional Random fields (CRFs) are used to train different Bio-NER models used in this paper. The performance of f-score degrades with the complexity of model. Use limited size of context rather on whole text. Affected by data distribution.

6. A Comparative Study of Biomedical Named Entity Recognition Methods Based Machine Learning Approach

Mohammed RAIS et al [11] proposed a model comparing CRF, HMM, MEMM, SVM, ME, NB and DT models. CRF surpass all the other Machine-Learning methods on both corpora. It contributes to faster and better research in the field.

7. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts

Shaodian Zhang et al [12] gives possible improvements on the approach including nested NPs as candidates, better chunker for medical text, better domain representations, and improved IDF values of phrases. The paper gives solution to tackle the challenges of entity boundary detection and entity type classification.

The authors in all the papers have given a brief about various methods for Named Entity Recognition from traditional ways to new concepts. Machine Learning is the best AI technique to extract data from Bio Medical texts through Hybrid Approaches using Supervised and Unsupervised ways. CRF, HMM, K-nearest are the best algorithms for implementing a Bio-NER because they handle large amount of datasets with efficiency and giving exact output as needed.

III. APPROACHES USED FOR BIO-NER EXISTING METHODOLOGIES

Named Entity Recognition is an important task of Information Extraction(IE), which identifies Named Entities(NEs) and classifies them into predefined classes of NEs. Biomedical Named Entity Recognition aims at automatically recognizing Biomedical Named Entities(BioNEs) such as genes, proteins, cells, drugs, diseases, etc. Named entity recognition is very important component of biomedical natural language processing, as it extracts information and ultimately discovers knowledge from text. Sometimes the same name is shared by different types of bio-entity types which increases the ambiguity. It is necessary to find out the eligible classes for which a classifier is suitable.
The figure 1. shows the taxonomy of various approaches for extraction of named entities.

1. **Dictionary based approaches:**

Dictionary based name recognition is used in [8] for extracting information from biomedical documents as it can provide ID information on recognized terms. This method identifies Named Entities by matching terms. As Dictionary based approaches have limitations such as false positive recognition and lack of a unified resource that covers newly published names. This method is found to have a high degree of precision [7] but it has a poor recall.

2. **Rule Based Approaches:**

In Rule based approach [8] the named entities are recognized by predefined rules that describe typical naming structure. Here [7] rules are defined in an attempt to recognize entities which describe the formation patterns and context of named entities. In this approach, the rules are developed manually using lexical-syntactic features or using existing information lists.

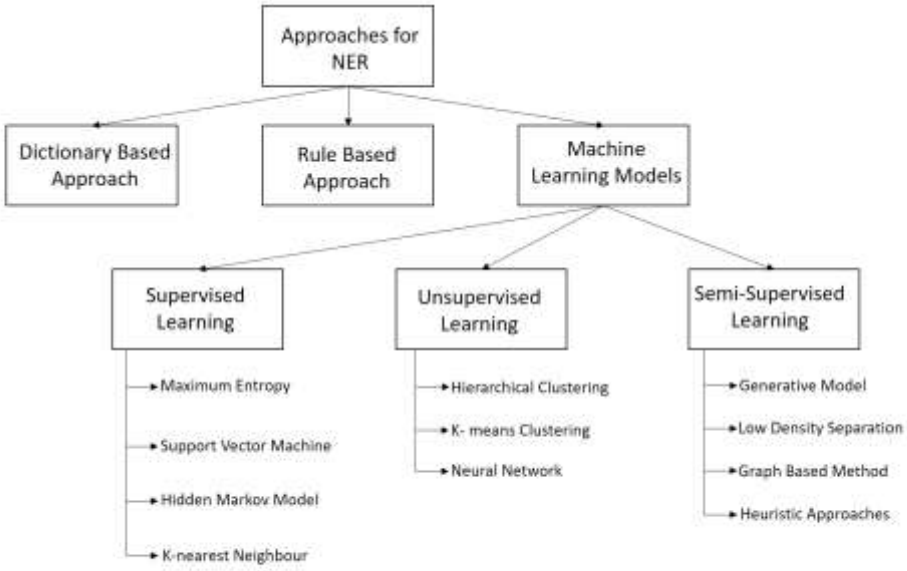

Figure 1. Taxonomy of approaches for Named Entity Recognition

3. **Machine Learning Model:**

To train data set [8] supervised Machine Learning methods are used widely. Machine learning methods give better performance and it can be easily used on another domain. Among many approaches used in machine learning approaches the Maximum Entropy, SVM and HMM methods are being used by many researchers because of their results shown outperform than other approaches. In this technique [7] a system learns automatically by using negative and positive training.

As BioNER is efficiently done by Machine Learning Model, the proposed system is based on Machine Learning Model. Machine Learning algorithms are generally classified into three types:

1) Supervised Learning
2) Semi-supervised Learning
3) Unsupervised Learning.

1. Supervised Learning:

Supervised learning methods in NER task require a large amount of training, usually data which is manually annotated demands a lot of cost and time investment. All data is labeled and the algorithms learn to predict the output from the input data.

2. Semi-supervised Learning:

Semi-supervised learning uses both [7] labeled data and unlabeled data for the learning process to reduce the dependence on training data. The system is first trained on an initial small set of examples and unlabeled data is tagged. The resulting annotations are then highlighted to increase the initial training set. The added training set is then used to re-train the system.

3. Unsupervised Learning:

In the unsupervised learning, decisions [7] are made on unlabeled data. The methods of unsupervised learning are mostly built upon clustering techniques, similarity based functions and distribution statistics. Unsupervised learning methods make decisions on a large unannotated data.

There are several supervised learning techniques:

Maximum Entropy:

Naïve Bayes method and Nearest neighbor method is used with [8] Maximum Entropy. Constraints are derived from training data, expressing some relationship between features and outcome. The ME principle seeks the distribution that maximizes the entropy of the distribution subject to the known constraints.

Support Vector Machine:

SVM approaches are most successful in [8] classifying text automatically and is considered [6] one of the classification techniques with a very high efficiency. Based on the idea of structural-risk minimization, from the computational-

learning theory, SVM tries a decision surface, to separate the training data nodes into two main classes, and makes decisions based on the existing support vectors.

Hidden Markov Model:
This model consists of [8] states and observations. Hidden Markov Model is used for representing sequential data. HMM is a generative model. The model assigns the joint probability to paired observation and label sequence. Then the parameters are trained to maximize the joint likelihood of training sets.

Name **Parameters**	**Banner**	**Abner**	**Dnorm**	**Gimli**	**TaggerOne**
Datasets	I2b2, Genia Corpus	NLPBA 7 BioCreative corpora.	NCBI disease corpus & MEDIC vocabulary (MeSH,OMIM)	GENETAG & JNLPBA corpus.	---
Algorithm/ Approach	Machine Learning, CRF.	Machine Learning, CRF.	Machine Learning.	Machine Learning, CRF.	Machine Learning.
Input	Free text (paste), Sentences.	Free text (local), Sentences.	Plain text.	Scientific text.	Biomedical text.
Output	Gene,protein names, Bio-entity tagged text, Acronyms, Abbreviations.	Gene/protein names, Bio-entity tagged text, Semantically labelled text, gene/Protein labelled text.	Diseases mentioned in biomedical text.	Biomedical names, gene/protein DNA, RNA, cell line and cell type names	Diseases and chemicals in biomedical text.
Advantages	Open source, Useful as an extensible NER implementation and standard for comparing innovative techniques,	Open source, variety of orthographic and contextual features with an intuitive graphical interface and Java application programming interface.	High-performing and mathematically principled framework which employs pairwise learning for the task of disease normalization.	Can be used as command line tool, offering full functionality, such as orthographic, morphological and domain knowledge features.	Trainable and not limited to a specific concept type, can handle multiple concept types simultaneously.

Table 1. Study of existing Bio-NER tools

IV. PROPOSED SYSTEM ARCHITECTURE

The system is evaluated using 2 datasets i.e. 2b2 and GENIA corpora. The i2b2 corpus contains clinical notes which contains Problems, Tests, and Treatments which are annotated entities. GENIA corpus is a collection of biomedical related words which contains biological entities such as DNA, RNA, and protein.

Figure 2. gives a general overview of Bio-NER system. It generally consists of 3 phases namely

1. Feature Extraction
2. Boundary Detection
3. Entity Classification

1. Feature Extraction: Transformation of input data into a set of features. Features are distinctive properties of input patterns that help in differentiating between the categories of input patterns.

Tokenization: Given a character sequence, tokenization is the process of breaking up a sequence of string into pieces called tokens. It includes all the important tasks which are based on the quality of the tokens generated. In biomedical domain abbreviations, apostrophes, hyphenation, multiple formats and various sentence boundaries such as period, colon, semi-colon, explanation mark or dash cause problem while tokenization. N-gram method and simple approach are used to overcome such problems in tokenization. Tokenization also includes removal of white space, removal of stop words,

Part of Speech tags. POS tags defines the lexical category of the word like noun, verb, adjective etc. POS tag is a very important feature in BioNER.

Context words: The context window is information about the prior and posterior words to the current word. If wi is the current word, then the context window of size 3 is wi+3 i−3 = wi−3....wi+3. This feature is used under the principle that surrounding words carry effective information for classifying BioNEs.

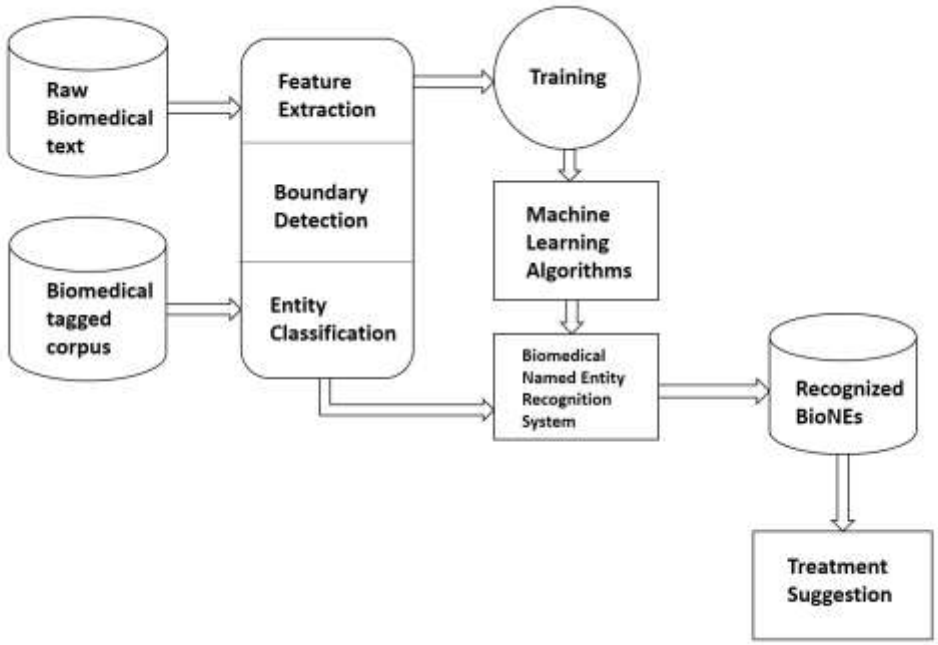

Figure 2. Proposed architecture of Bio-Ner

2. Boundary Detection: The second step is for detecting the boundaries of entities. In this step candidates are collected for entity classification. The NP chunker and Inverse document frequency (IDF) based technique is employed to remove the noun phrases which are of no use.

3. Entity classification: Entity classification is done by using signature generation method. The similarity based method is mainly used in word sense disambiguation (WSD), which assumes that meaning of a word is closely related to the distribution of different words surrounding it. Finally, we use a TF-IDF. TF-IDF gives how important a word or term is to a document than raw frequency. Words that are more important will have larger weights in signature generation method.

V. RESULT

The proposed system gives extracted bio medical terms and classification according to gene, protein, DNA, RNA, blood cells, diseases, problem, test etc. According to the recognized disease, problem and test an appropriate treatment is expected to be suggested.

VI. CONCLUSION

Named Entity Recognition is a well-developed and a lot of research is available. Lots of biomedical recognition tools are implemented such as BANNER, AB-NER, Dnorm, Gimli, TaggerOne, MedLee. They mainly include recognizing gene, protein names, disease mentions, cell line and cell type names, DNA, RNA from textual clinical reports. In this paper, stepwise solution to BM-NER is provided including a seed term extractor, an NP chunker, an IDF filter, and a entity classifier based on distributional semantics. Our proposed method does not rely on any rules or training data, which allows it to be applied in different settings and applications

VII. FUTURE WORK

The project includes extraction of Biomedical entities using Machine Learning and Natural Language Processing which include Name, Problem and Test from Lab Reports and mapping appropriate Treatment using widely accepted datasets. The previously implemented Bio-NER tools are only available for text reports, the Bio-NER can also be implemented for images, by extracting information by scanning the reports such as X-Rays, ECG.

REFERENCES

[1] Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat (2008), Named Entity Recognition Approaches.
[2] Weiyi Sun, Anna Rumshisky, Ozlem Uzuner (2013), Evaluating temporal relations in clinical text: 2012 i2b2 Challenge.
[3] J.-D. Kim, T. Ohta, Y.Tateisi, J. Tsujii (2003), GENIA corpus—a semantically annotated corpus for bio-textmining.
[4] Anish Talwar, Yogesh Kumar(2013), Machine Learning: An artificial intelligence methodology.
[5] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, Parve Kuksa (2011), Natural Language Processing (ALmost) from Scratch.
[6] Ahmed Sultan Al-Hegami,Ameen Mohammed Farea Othman,Fuad Tarbosh Bagash,(2017). A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set.
[7] Almas Tasneem,Archana B,(2016). A Survey on Biomedical Named Entity Extraction.
[8] S.Vijaya, Dr.R.Radha,(2017).Biomedical Named Entity Recognition - a swift review.
[9] Samet Atdağ and Vincent Labatut. A Comparison of Named Entity Recognition Tools Applied to Biographical Texts.
[10] H. L. Shashirekha,Hamada A. Nayel,(2016) A Comparative Study of Segment Representation for Biomedical Named Entity Recognition.
[11] Mohammed RAIS,Abdelmonaime LACHKAR,Said EL ALAOUI OUATIK(2014),A Comparative Study of Biomedical Named Entity Recognition Methods Based Machine Learning Approach.
[12] Shaodian Zhang , Noémie Elhadad(2013),Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts.

