RESEARCH ARTICLE OPEN ACCESS

Identification of Named Entities in Biomedical Articles Based On Bidirectional LSTM Model and Semantic Entity Probability

Ms. M. Parvathavarthini [1], Dr. S.K. Jayanthi [2], Mrs. R. Swathi [3]

[1](Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

[2](Head and Associate Professor, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

[3](Assistant Professor, Department of Computer Applications, Sree Abiraami Arts and Science for Women, Gudiyatham, Vellore, Tamilnadu, India

ABSTRACT

The present abundance of biomedical literature data is followed by new information being created and advanced. The massive information of knowledge about biomedical data is represented only in the text format of biomedical literature, web pages, and other sources. Biomedical Named Entity Recognition (Bio-NER) system is the task that focuses to recognize automatically the named entities in biomedical texts. In the biomedical field, drug discovery leads to ambiguity while recognizing the entities. Currently there is no complete biomedical named entity dictionary to contain various entity types. In addition, as the new named entities continue to emerge and the names of entities which has not been recognized are increasing. Hence, this paper aims to solve the semantic conflicts of new biomedical entities using semantic entity probability based on Bi-directional Long-Short Term Memory (Bi-LSTM) method and Bayes Theorem. The performance of classifying biomedical named entities in biomedical articles has been evaluated using precision, recall and F1 measures and demonstrated that the deep learning-based Bi-LSTM model in Bio-NER depicts better accuracy compared to dictionary-based Bio-NER.

Keywords:- Biomedical Named Entity Recognition, Identification, Bi-LSTM Model, Classification, Semantic Entity Probability, Evaluation Measures.

I. INTRODUCTION

Named Entity Recognition (NER) is the method of classifying entities semantically into predefined groups such as individual, place, name, organisation, etc. in a text. In the biomedical field, the purpose of Biomedical Named Entity Recognition (Bio-NER) is to find and classify text sequence in the biomedical terms such as names of genes, proteins, drugs and diseases. In future, it is a crucial step for automatic processing of biomedical literature to be mounted on a large scale, and further to perform high-level information extraction and retrieval in biomedical science. In the discovery of the human gene and drugs, the biomedical domain generated huge volumes of medical and biological information. It also includes a wealthy number of medical organizations that provide important data access information. Bio-NER comprises of two procedures, identification of terms and classification of terms.

The method of identification discovers the region of an entity named in a text. Its primary objective is to distinguish between medical terms and non-terms without looking at a medical term's semantic significance. However, word classification determines that entity's semantic notion and assigns it to a biomedical class, such as genes, proteins or DNA. Due to the development of internet, number of articles and documents published in the biomedical domain is growing rapidly. Thus, it could be very time-consuming to extract information from this vast collection of articles on a specific subject.

This paper is to identify the biomedical named entities in biomedical literature articles and also to increase the accuracy of each biomedical text by affix features. The main objective is to solve the semantic conflict problem for new biomedical words and validate the categorized and non-categorized biomedical named entities.

The rest of this paper is organized as follows: Research inference is made based on the related work shown in Section II. Methodology is described to identify and classify the biomedical named entities using data mining approaches in Section III. Results and performance evaluation is done in Section IV. Conclusion and future scope is given in Section V.

II. RELATED WORK

XinglongWang and Jun’ichiTsujii, [2009], has proposed named entity disambiguation method for classifying each individual occurrence of an entity. It classifies pair- wise relations between the entity tags and species words. Heuristics Baseline method is introduced to denote the names of model organisms. To improve the performance of multi-class entity classification, Predicate Argument Structure (PAS) method and Syntactic Parser is performed. The experimental result shows that the relation classification system cannot cover all entity tags and it only classifies the informative keywords co-occurring in the same sentence [9].

Ahmed Sultan Al-Hegami and Fuad Tarbosh Bagash [2010], has proposed machine learning classification approach for biomedical named entity recognition. They use different feature sets to synthesize a more accurate classification procedure. Support Vector Machine method tries a decision surface, in order to separate the training data nodes into two main classes, and makes decisions based on the existing support vectors. Multi-variate Bernoulli models based on Naïve Bayes model is used to classify entity names. They conclude that the machine learning techniques have been used for building and developing biomedical named recognition which requires feature selection and extraction. The results determine that the K-Nearest Neighbor trained with suitable features is more suitable to recognize named entities of biomedical texts than other models [1].

Asma Ben Abacha and Pierre Zweigenbaum [2011], has proposed and compared three methods based on domain-knowledge and machine-learning techniques in clinical text corpus. The domain specific tools such as Meta-Map recognizes and categorizes medical terms by associating them to concepts and semantic types of the UML Meta thesaurus and Semantic Network. The classification result of medical entity recognition in word level is implemented by conditional random field classifier. The experimental results show that the identification of entity boundaries is identified by noun phrases chunking segmentation and the best performance was obtained with a conditional random field classifier using lexical and semantic features [2].

Hye Jeong Song, Byeong-Cheol Jo1, et al [2018], proposed a Recurrent Neural Network (RNN) that uses a Jordan-type algorithm and an Elman-type algorithm. They also measure the performance of conditional random fields using word embedding as their features. Word embedding has increased performance in natural language processing, machine translation, voice recognition. Word embedding has been used as features in natural language processing and is mapped from a word in the higher-dimensional space into a real-numbered vector in the lower-dimensional space. They compared the two RNN algorithms with Word2Vec, Canonical Correlation Analysis (CCA), and embedding methodologies. RNN has a drawback of storing long-short term dependencies in long text sequence. Hence, Long Short-Term Memory will be used to predict or preprocessing the text sequence [6].

WonjinYoon, JaewooKang, et al [2019], introduced CollaboNet, which consists of multiple BiLSTM-CRF models, for biomedical named entity recognition. While existing models were only able to handle datasets with a single entity type, CollaboNet leverages multiple datasets and achieves the highest F1 scores. Unlike recently proposed multi-task models, CollaboNet is built upon multiple Single-Task Models (STMs) that's end information to each other for more accurate predictions using BIOES tags. The experimental result shows that the model achieved state-of-the-art performance on four Bio-NER datasets in terms of F1 score, precision and recall. Although the model requires a large amount of memory and time, which existing multi-task models require as well, the simple structure of CollaboNet allows researchers to build another expert model for different entity types without using the BIOES tags [8].

Based on the review given above, it is observed that machine learning and deep learning algorithms for classifying the biomedical tags achieves better identification than dictionary-based approaches. The reviews state that there is no complete biomedical named entity dictionary that includes different entity types. Therefore, the new named entities are continuing to arise and the names of new entities which has not been recognized are increasing. In order to find the new biomedical text terms in entity, deep learning algorithms have been used to achieve better classification of new text entities than machine learning algorithms.

III. METHODOLOGY

Due to the change of the development of clinical practice, the evidence-based medicine is a new approach to health care for promoting the collection, interpretation, and integration of valid, important and applicable research-derived evidence. The best available evidence is applied to improve the quality of clinical judgements.

The medical literature articles use various strategies to improve the yield of medical line searches, systematic reviews, and software tools for clinical decision making. Hence, the biomedical named entity recognition identifies the biomedical entity that helps to discover new information about medicine.

The proposed work has a sequence of steps to implement the objectives. The figure 1 shows the overall architecture of the current work. The phases involved in this process are as follows,

- Dataset collection
- Bidirectional Long Short-Term Memory Model

- Biomedical entity dictionary
- Syntactic word-based approach
- Semantic entity probability
- Performance evaluation

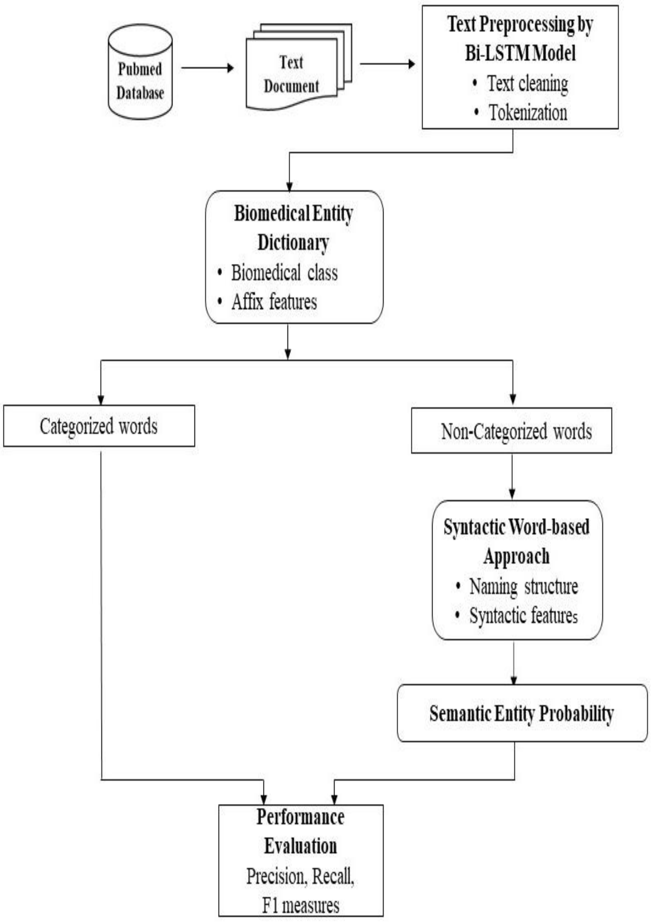


Fig. 1 System Architecture

A. Dataset Collection

PubMed is a free digital repository search engine that primarily accesses the full-text research articles that have been published within the biological and medical literature. It maintains the biomedical database which is used in information retrieval and extraction for biologists. The research articles enhance development in science and health by providing access to biomedical and genomic information. An extensive collection of articles about biomedical science has been widely utilized for the discovery of new medicine. For the current work, the biomedical literature database downloaded from the PubMed repository have been used.

B. Bidirectional Long Short-Term Memory Model

The first step of text mining is text preprocessing. The text preprocessing in biomedical text is done by using Bi-LSTM model. The preprocessing module splits the given literature into single sentences and tokens. The following steps for preprocessing the biomedical texts have been implemented.

The text cleaning and tokenization in biomedical texts is done by using Bidirectional Long Short-Term Memory (Bi-LSTM) model. It provides Bi-LSTM method to generate character-level to word-level representation of biomedical texts. The Keras library provides one hot encoding function that works with the following steps,

- Converts the text into lowercase characters,
- Filter out punctuation marks,
- Generate words by using sequence of characters,
- Splits the words by space.

The Keras library in the long short-term memory model splits the biomedical text into a list of words. The Keras Sequential model builds the Bi-LSTM network to implement character to word level representations. Bidirectional Long Short-Term Memory model are made up by two separate LSTM layers that scan the input text sequence in the opposite direction. The two hidden layers are connected to the same output layer, so that the context information can be accessed in two opposite directions. The context information actually used by the network is learned during network training and it does not need to be specified in advance. Therefore, the information in the contexts are learnt independently of each other.

C. Character-level to Word-level representation in biomedical texts

The architecture of Bi-LSTM model process is similar to implement character to word level representations. The hidden layer of each character at the time t of the forward LSTM is affected by the hidden layer at time $t - 1$ of the backward LSTM of another character and the each input character at time t. The LSTMs in two different directions do not share the cell state value. The two hidden layers of each character are connected to the same output layer, so that the context information can be accessed in two opposite directions to predict the word. The status value of the cell output by the forward LSTM is only transmitted to the forward LSTM. The input character of each time node is respectively transmitted to the forward and reverse LSTMs.

During the specific calculation, Bi-LSTM first scans the entire sequence. It calculates the hidden state of the forward LSTM character along the direction from 1 to t, and it then calculates the hidden state of the reverse LSTM from t to 1. The forward and reverse output characters are combined and connected to the Bi-LSTM output character to synthesize the final output word.

The Bi-directional Long Short-Term Memory (Bi-LSTM) model improves the LSTM by feeding the input to the LSTM network twice, once in the original direction and once in the reversed direction. The output characters from both directions are concatenated to represent the final word. This model allows for detection of dependencies from both previous and subsequent words in a sequence to generate the word.

D. Biomedical Entity Dictionary

The biomedical entity dictionary, also sometimes called a bio thesaurus, or entity gazetteer, consists of a list of all the entities and their possible synonyms with common spelling, or orthographic, variations. The natural language text is compared with dictionary entry which include a large number of pre-existing biomedical entity name, based on the matching strategy. The matched words are referred by main categorical entity. The mismatched words are referred by non-categorical entity. It can be a significant investment of work to create a new dictionary from scratch, so most dictionaries are built either from existing databases or ontologies.

Dictionary based name recognition is used for extracting information from biomedical articles as it can provide information on recognized biomedical entity terms. This method identifies Named Entities by matching terms. The dictionaries will have huge collection of names of the entities. The phrases to be matched from the given text will be compared with the existing entries in the dictionaries.

The prefix and suffix can provide good clues for classifying named entities in biomedical texts. To represent the target entity names properly, it is necessary to have a rich set of affix features. The tokens are most essential basic inner features. The suffix features such as I, II, III, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, i, ii, iii, iv, +ve, -ve are used in this work.

Step 1: The sequence of resultant output words from the Bi-LSTM model checks the biomedical words that matches with biomedical entity dictionary.

Step 2: The biomedical entity dictionary checks each word in the document by matching strategy. The results of matched words are categorized as biomedical entity class and the mismatched words are categorized in the other category.

Step 3: The suffix features have been combined with biomedical entity to identify biomedical words in a text sequence. The biomedical words are identified in entity types and the resultant non-categorized biomedical words are to be classified based on semantic entity probability.

E. Syntactic word-based approach

In Syntactic word-based approach, the named entities are recognized by predefined rules that describe typical naming structure. The 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. While identifying the words, the semantic meaning of an entity is to be clarified using the steps of rules as follows,

***Rule 1:* Previous word:** If the word preceding an entity is a biomedical word, assign the class indicated by that word to the entity.

***Rule 2:* Categorical word in the same sentence:** If a biomedical word and an entity appear in the same sentence, assign its class to the entity. When more than one biomedical word co-occurs in the sentence, priority is given to the species word to the entity's left with the smallest distance. If all biomedical words occur to the right of the entity, take the nearestone.

***Rule 3:* Majority vote:** assign the most frequently occurring word in the text based on the categorical word count occurrence to all entity classes.

The non-categorical entities are extracted to find the entity class type and the syntactic word-based approach identifies the biomedical word using the previous species word and majority vote. The main categorical entity is also predicted by using this method. Hence, the resultant biomedical words are not semantically categorized in rule-based approach. The semantic conflict of a new biomedical word is solved by using conditional probability. This method splits the word to find the category of multiple biomedical words that are categorized in other entities. It predicts the category and word count of new biomedical word in a biomedical text sequence.

F. Semantic Entity Probability

The discovery of new biomedical words in named entities leads to semantic conflict problem in biomedical texts. The one entity class type words are similar to another entity class type words and this leads to ambiguity to recognize the biomedical entity. The semantic entity probability is used to identify the correct entity based on semantic meaning of the entity class. The conditional probability in Bayes Theorem is used to solve this semantic problem of new biomedical words.

In Bayes Theorem, the conditional probability is a measure of the probability of an event occurring given that another event has occurred. Based on the dependent rule, the semantic probability is measured based on the number of word counts occurring in the entity class. The semantic entity probability is obtained by the conditional probability of dependent entity class.

The new biomedical word given in the class is determined by counting its number of occurrences in particular given entity. The entity which has high probability results categorize the new word in biomedical entity class. From this definition, the semantic entity probability of a new biomedical word is defined as P (word | entity) and it is easily obtained by the probability of occurring the number of word counts in the entity by dividing the particular biomedical entity.

$$\mathbf{P\ (word \mid entity) = \frac{P\ (both\ word\ and\ entity)}{P(entity)}}$$

Here, the above equation is implemented in biomedical text to find the probability of semantic named entities.

- P(word | entity) – it represents the conditional probability, the probability of new biomedical word occurring in the given particular biomedical entity class is depended upon the occurrences of previous entity words in the biomedical entity,
- P(word and entity) – it represents the joint probability of biomedical class and word; the probability that both biomedical words occurring in mutually exclusive events,
- P(entity) – the probability of a biomedical entity class for new biomedical word.

IV. RESULTS AND DISCUSSION

The results and analysis are indented to be of use to researchers from all fields to study algorithms and techniques experimentally. The results are discussed in this chapter through experimental analysis.

G. *Experimental Analysis*

In the experimental work, five literature articles datasets (GENEIDAB1, NCBIDISAB2, BC4DRUG, DRGIDAB4, BC2GM) was used, all of which were collected from National Centre for Biotechnology Information (NCBI) repository.

Each of the five datasets were constructed from biomedical abstracts and each dataset focuses on all four biomedical entity types: gene, protein, drug and disease. All the datasets contain training and test data sets for biomedical text sequences. The existing and proposed work identifies the biomedical named entities in text document. The process of existing and proposed work contains the following steps:

Step 1: The biomedical literature articles has been downloaded from PubMed central in National Center for Biotechnology Information (NCBI).

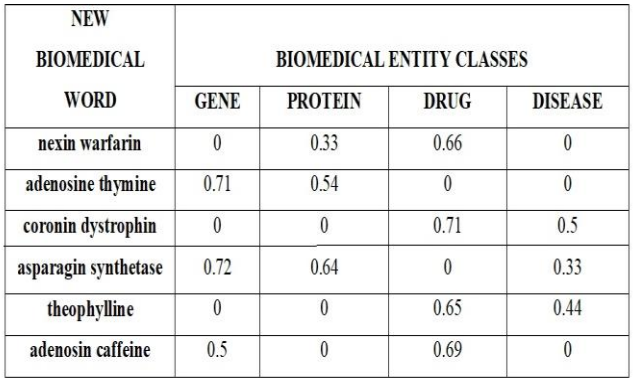
Step 2: The literature document is to be processed for text preprocessing. The text preprocessing is done by using Bidirectional Long Short-Term Memory to clean the text and for tokenization.

Step 3: The resultant tokenized words from Bi-LSTM model matches the entity string through biomedical entity dictionary and affix features.

Step 4: In addition, the syntactic word-based approach defines rules for typical naming structure to identify the main categorical entity of biomedical texts.

Step 5: The mismatched entities of false positive is determined by semantic entity class probability to classify the named entities based on semantic meaning of the biomedical text entities. The semantic entity probability based on Bi-LSTM model categorize the new biomedical word in entity class based on the semantic context of the word.

TABLE I
SEMANTIC ENTITY PROBABILITY VALUES IN NEW BIOMEDICAL WORDS

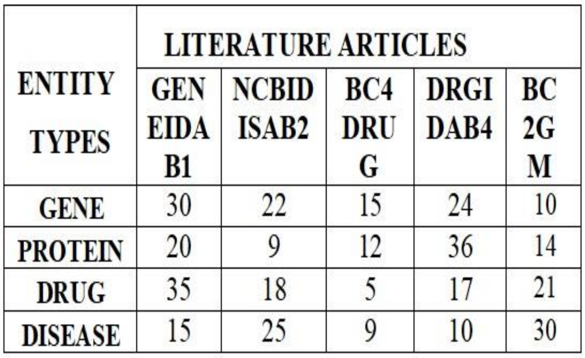


NEW BIOMEDICAL WORD	BIOMEDICAL ENTITY CLASSES			
	GENE	PROTEIN	DRUG	DISEASE
nexin warfarin	0	0.33	0.66	0
adenosine thymine	0.71	0.54	0	0
coronin dystrophin	0	0	0.71	0.5
asparagin synthetase	0.72	0.64	0	0.33
theophylline	0	0	0.65	0.44
adenosin caffeine	0.5	0	0.69	0

The table 1 has shown the semantic entity class probability of new biomedical words that results in class of higher probability values in entity classes.

In table 2, The categorical word counts of biomedical entity types in literature articles are given below.

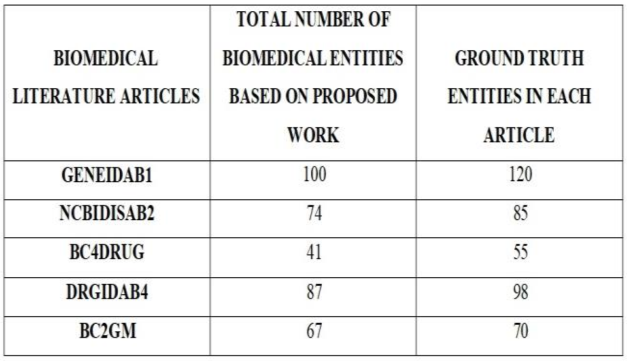
TABLE IIIII
ENTITY TYPE COUNTS IN BIOMEDICAL ARTICLES



ENTITY TYPES	LITERATURE ARTICLES				
	GEN EIDA B1	NCBID ISAB2	BC4 DRU G	DRGI DAB4	BC 2G M
GENE	30	22	15	24	10
PROTEIN	20	9	12	36	14
DRUG	35	18	5	17	21
DISEASE	15	25	9	10	30

In table 3, the total number of biomedical entities in each article are identified and classified based on Bi-LSTM model and Semantic entity probability. The ground truth entities in each article are based on BIOES (Beginning Intermediate Other End Simple) tag notation.

TABLE IVVVI
GROUND-TRUTH ENTITIES IN BIOMEDICAL ARTICLES



BIOMEDICAL LITERATURE ARTICLES	TOTAL NUMBER OF BIOMEDICAL ENTITIES BASED ON PROPOSED WORK	GROUND TRUTH ENTITIES IN EACH ARTICLE
GENEIDAB1	100	120
NCBIDISAB2	74	85
BC4DRUG	41	55
DRGIDAB4	87	98
BC2GM	67	70

H. Evaluation Metrics

For the evaluation of biomedical named entity recognition task, true positives are counted from exact matches between predicted biomedical entity and ground truth entities based on the BIOES notation. Precision, Recall and F1 score were used to evaluate the models.

Precision (P) = C/M (4.1)

Recall (R) = C/N (4.2)

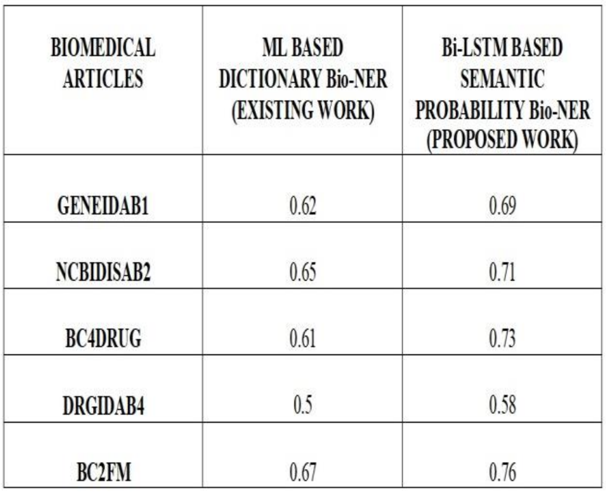
F1 Score = 2 P*R / P+R (4.3)

- M = total number of predicted entities in the biomedical text sequence,
- N = total number of ground truth entities in the biomedical text sequence,
- C = total number of correct entities.

The dataset used the above three evaluation metrics equations 4.1, 4.2, 4.3 to validate the biomedical named entities and the proposed work results are compared to the existing machine learning Bio-NER model by using this measure. The values of C, N, M in proposed work are, C = 90, M = 100, N = 120.

In table 4, the comparison results of existing and proposed work for Biomedical Named Entity Recognition in Literature Articles are given below,

TABLE VIIIV
EVALUATION RESULT ANALYSIS



BIOMEDICAL ARTICLES	ML BASED DICTIONARY Bio-NER (EXISTING WORK)	Bi-LSTM BASED SEMANTIC PROBABILITY Bio-NER (PROPOSED WORK)
GENEIDAB1	0.62	0.69
NCBIDISAB2	0.65	0.71
BC4DRUG	0.61	0.73
DRGIDAB4	0.5	0.58
BC2FM	0.67	0.76

V. CONCLUSION AND FUTURE WORK

This paper mainly concentrates to identify the biomedical named entities in medical literature articles. The main reason for the development of proposed system is to classify named entities based on semantic meaning of the biomedical text terms. The current work uses Bi- LSTM model to extract the word features. The matching strategy uses biomedical entity dictionary, syntactic word-based approaches for biomedical entity text terms. As a result, the deep learning extracts better word features for tokenization than normal pre-processing methods and biomedical entities are classified based on semantic class meaning of the entity.

When the new biomedical entity exists in false recognition or it is categorized in other entities, the conditional probability classifies the predicted entity of new biomedical word. Hence, semantic conflict problem is occurred whether the new word exists in similar entity classes. The conditional probability predicts the entity class for the new word and the entity which has high probability results in correct biomedical entity. The categorical word counts of entity classes in literature articles is determined based on Bi-LSTM model and semantic entity probability method. In addition, the biomedical entities are also identified and classified based on the semantic entity probability.

In future, text features can be extracted automatically using deep learning networks can be used. In future, the research articles dataset can be of any type such as pdf, html or other formats. This work can also be extended for the development of deep artificial neural network specialized on this problem. The linguistic resources for domain knowledge should be built for performance development. Gazetteers, word embedding methods, and other resources will also be developed to cover more target entity types and datasets.

REFERENCES

[1] Ahmed Sultan, Al-Hegami, *A Biomedical Named Entity Recognition using Machine Learning Classifiers and Rich Feature Set*, International Journal of Computer Science and Network Security, VOL.17 No.1, January 2017.

[2] Asma Ben Abacha and Pierre Zweigenbaum, *Medical Entity Recognition: A Comparison of Semantic and Statistical Methods*, Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, USA, June 2011.

[3] Dai H, et al, New *Challenges for Biological Text-mining in the Next Decade*, International Journal of Biomedical Engineering, August 2004.

[4] Gerner M, Nenadic G, Bergman CM, *A Species Name Identification System for Biomedical Literature*, BMC Bioinformatics, International Journal of Computer Science, November 2010.

[5] GuoDong Zhou and Jie Zhang, *Recognizing Names in Biomedical Texts: A Machine Learning Approach*, International Journal of Biomedical Engineering, Vol. 20 no. 7 October 2004.

[6] Hye-Jeong Song, Byeong Cheol Jo, et al *Comparison of Named Entity Recognition Methodologies in Biomedical Documents*, from International Conference on Biomedical Engineering Innovation (ICBEI), Taiwan. November 2016.

[7] Krallinger M, Rabal O, et al. *The Corpus of Chemicals and Drugs and its Annotation Principles*, *Cheminformatic*, International Journal of Bioinformatics, July 2015.

[8] WonjinYoon, ChanHoSo, et al, *CollaboNet: Collaboration of Deep Neural Networks for Biomedical Named Entity Recognition* from International Workshop on Data and Text Mining in Biomedical Informatics Turin, Italy October 2018.

[9] Xinglong Wang and Junichi Tsujii, *Classifying Relations for Biomedical Named Entity Disambiguation*, International Conference on Empirical Methods in Natural Language Processing, Singapore, August 2009.

[10] Xuan Wang, Yu Zhang, *Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning*, Department of Biomedical Engineering, Stanford University, USA, October 2018.