IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BIOMEDICAL NAMED ENTITY RECOGNITION (BNER) USING WORD REPRESENTATION FEATURES BASED ON CRF

Sudhakaran Gajendran[1], Manjula D[2]

[1]Research Scholar,[2]Professor
[1,2]Department of Computer Science and Engineering,
[1,2]Anna University, Chennai, India

Abstract: With the rapid advancement of technology and the necessity of processing large amounts of data, biomedical Named Entity Recognition (NER) has become an essential technique for information extraction in the biomedical field. In this work, we present a technically simple architecture for Biomedical Named Entity Recognition (BNER) using CRF algorithm. The proposed system uses clustering based word representation and word embedding based word representation as the only features along with the basic feature set to identify the entities. In addition, we incorporated BIO entity tag representation to effectively identify the biomedical entities. The proposed system is evaluated on two publicly available dataset BioCreAtIvE II GM corpus and JNLPBA corpus. The experimental results on both the corpus shows that the system outperforms all the existing system with the relatively high F - score of 85.92 % on BioCreAtIvE II GM corpus and 76.64 % on JNLPBA**.**

***Index Terms* - Biomedical Named Entity Recognition (BNER), CRF, Bioinformatics, Word Representation.**

I. INTRODUCTION

Biomedical literature is expanding at a faster pace with the rising advancement in the field of computer and biological technology. An potential for data mining approaches in this area has been created by the explosion of biomedical literature. BNER is a crucial phase in bioinformatics to classify terms or entities pointing to a particular individuals in biomedical literature. Much harder functions like Drug - Drug Interaction (DDI) and Drug - Target Interactions could be carried out only when the names are identified properly. While several algorithms for this task have been proposed, BNER still a daunting task and there is a huge gap comparing the BNER models and newswire applications. An F-score of over 96 percent can be obtained by the best NER systems in newswire articles (Sundheim, 1995), while the performance of benchmarking BNER models is still between 75 percent and 85 percent in F1-score (Cohen and Hersh, 2005).

Present BNER frameworks can be split into three approaches: methods on dictionaries (Yang et. al., 2008), heuristic rules based approaches (Olsson et.al., 2002) and methods of machine learning. Systems based on Machine learning framework are more robust compared to other approaches and the advantage of using machine learning approaches is that they can identify the new BNER entities apart from those that are available in dictionaries. Machine learning techniques are tried in all possible ways to identify the chemical entities. Hidden Markov Model (HMM) (Zhou and Su,2004), Support Vector Machine (SVM) (Lee et. al., 2004) Maximum Entropy Markov Model (MEMM) (Finkel et.al., 2004) and CRFs include these techniques (McDonald and Pereira, 2005, Tsai et.al., 2006, Settles, 2004). However, as with any machine learning method, most benchmarked systems follow a huge number of features. An output label is also represented by combining position information as complex entity tags, increases the features list due to the increased labels list. This in turn increases the cost of training.

The system proposed is a simple but technically strong architecture for the recognition of biomedical entities using clustering based word representation and word embedding based word representation as the only key features. An output label is also defined by combining BIO region content with a different class C and using a CRF algorithm. Experiments performed on datasets from JNLPBA2004 (Kim et. al., 2004) clearly explains that the model not only reduces the training time but also boost the accuracy on identifying the BNER tags.

The remaining part of this paper is structured as follows: related works are listed in detail in Section 2. The systems proposed are described in detail in Section 3. The experimental setup and dataset are explained in Section 4. Comparisons between our method and other models are made in section 5. Finally, in Section 6, conclusions and future work are given.

II. RELATED WORKS

In several forms of large-scale biomedical data analysis, text mining applications in the biomedical field are useful, such as network biology, gene prioritisation, drug repositioning or database development. Recognition of chemical names, such as drugs, diseases, proteins, species, chemicals or mutations, is the most important role in biomedical text mining (Al-Hegami et. al. 2017). Different approaches to biomedical named entity identification have been established in general biomedical systems. In general, BNER frameworks can be split into three approaches: rule-oriented models, dictionary-oriented models, and in the biomedical NER world, machine learning frameworks are more discussed recently as shown in Fig.1. Most benchmarking systems have used a machine learning frameworks by feature engineering to identify and make use of the features identified (Al-Hegami et. al. 2017).

The earlier stages of BNER system used rule based approaches where those systems depends on the pre derived set of rules. Those rules focuses on covering all the aspects like alphanumeric, special characters, and symbols so as to identify the entities (Eltyeb and Salim, 2014). Small repositories with both positive and negative tags are also included to enhance the performance of the systems. But the rule based models cannot improve or enhance with the pace of emerging needs in the BNER frameworks. Also, there is a need to frame new rules each time when there is a substantial change in the events and that created the huge time lag in the developments of rule based systems.

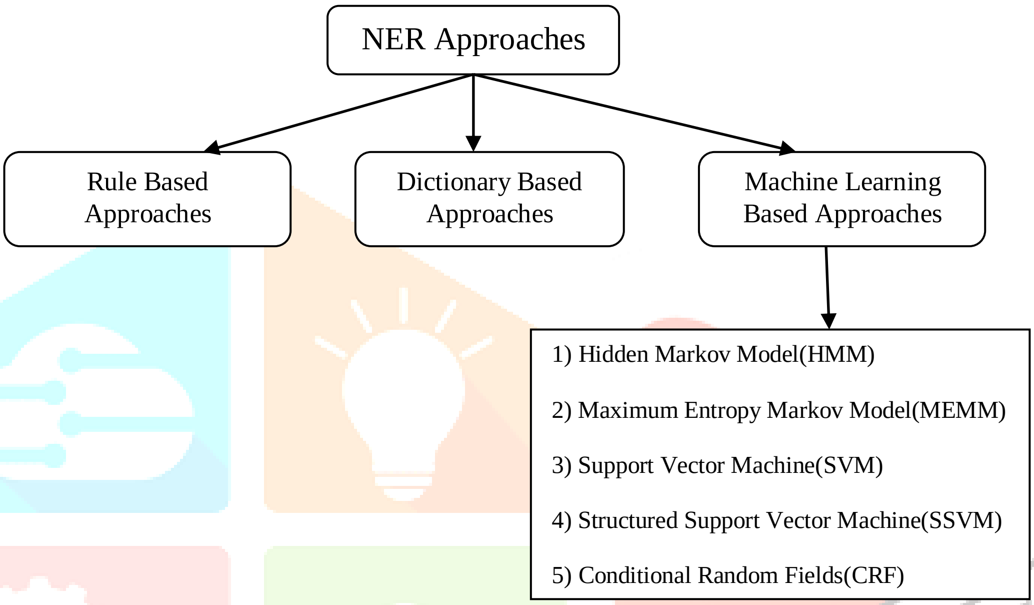
NER Approaches
Rule Based Approaches
Dictionary Based Approaches
Machine Learning Based Approaches
1) Hidden Markov Model(HMM)
2) Maximum Entropy Markov Model(MEMM)
3) Support Vector Machine(SVM)
4) Structured Support Vector Machine(SSVM)
5) Conditional Random Fields(CRF)

Fig.1. Different Approaches for Named Entity Recognition

The dictionary based models focuses on building the huge repository to hold as many entities as possible to extract all the matched entities for a biomedical literature. Dictionary approaches mostly used lemmas to match the term by identifying the most identical entity from the repository (Eltyeb and Salim, 2014). But it is very difficult to organize and maintain the huge repository and also it is infeasible to keep track of the new entities to be added into the repository (Akhondi, et. al., 2015).

Most of the powerful BNER framework available today is based on machine learning based algorithms to identify the biomedical entities from the text. The machine learning based framework converts the entity recognition as a sequence labeling task so as to identify the best tag for each word. Some of the powerful machine learning algorithm includes, Maximum Entropy Markov Models(MEMMs) (Eltyeb and Salim, 2014), Hidden Markov Models (HMMs) (Umare and Deshpande, 2015), Support Vector Machines (SVMs) (Lee et. al., 2004), Conditional Random Fields (CRFs) (Xu et. al., 2004) and Structural Support Vector Machines (SSVMs) (Tang et. al., 2015, Lyu et. al., 2017). The systems based on these frameworks focuses to extract the semantic and syntactic information from the text and use those information's to identify the best tag for each entity from the research text. The machine learning based frameworks mainly depends on the features included in the model so as to perform as expected. These features list includes context features, syntactic and semantic features, and also other features such as monographic and orthographic features, bag of words, POS tags etc. In common, the performance of these systems truly based on the selected features list and the training the model based on these features.

However, as with any machine learning method, most benchmarked models follow a huge number of features. An output label is also represented by combining position information with a different class C as complex entity tags, increases the features list due to the increased labels list. The price of training would also be considerably increased. This work presents a simple but technically strong framework for the recognition of biomedical entities using clustering based word representation and word embedding based word representation as the only key features. An output label is also defined by combining BIO region content with a semantic class C and using a CRF algorithm.

To overcome these issues, the proposed system focuses on a simple but technically strong architecture for the recognition of biomedical entities using clustering based word representation and word embedding based word representation as the only key features. An output label is also defined by combining BIO region content with a different class C and using a CRF algorithm. The CRF algorithm are a class of statistical modeling framework that often predicts a tag for each entity by taking into the account of ”neighboring” entities information. The CRF algorithm produced the best results in BNER among the other machine learning frameworks.

III. PROPOSED WORK

The proposed system architecture is explained in detail in Fig.2. The system accepts the input as the sentences from the NER corpus and preprocesses it using sentence detection and tokenization. In addition, the model uses the BIO tag representation to represent the tokens as either beginning, inside or outside tags. Then, clustering based Word Representation (WR) and word embedding based WR are used as the only features along with the basic features set. The tokens with the tags and the features selected is then passed to CRF algorithm to train the model. It then identifies the entities as the biomedical entity based on the training and the features.

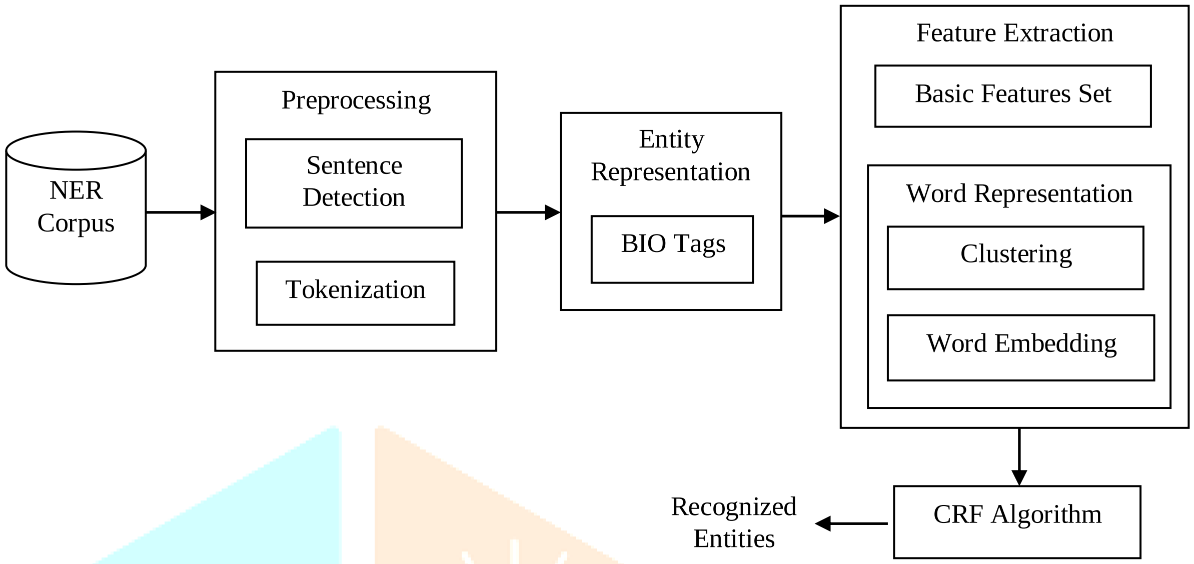

Fig. 2 Architecture of BNER System

3.1 Preprocessing: Sentence Detection & Tokenization

The proposed system uses a sentence detector to detect whether the sentence ends with a punctuation mark. The sentence detector in OpenNLP is used here. Sentence boundary recognition, however is difficult since finding out the punctuation mark is uncertain mostly (Read et. al., 2012). In order to further enhance the efficiency of the detection of sentences, we collected several abbreviations from the training and production sets, such as var., sp., cv., syn., etc. We then created several laws, such as if the current sentence ends with these abbreviations or with a comma, or whether the next sentence begins with a lower case letter. Then, the next two preceding sentences are combined into one sentence. Later, the sentences are split into tokens using the tokenizer producing words, numbers or special characters. On a finer stage, we initiated tokenization. As a matter of fact, we have combined certain special characters with the same meaning before any pre-processing, such as " ≥ " vs. " ≥ ", "*" vs. "*", etc.

3.2 Tags for entities

Once the tokenization is performed, the tokens are then annotated with the suitable tags to handle the BNER as a classification task. The BIO format (Sang and Veenstra 1999) is a widely used entity tag representation in which each term is assigned to a label as follows: B = entity beginning, I = entity within and O = entity outside. As shown in Table.1, the suitable BIO tags are annotated for each token. In Table.1, Example 1 and 2 corresponds to JNLPBA and BioCreAtIvE II GM datasets respectively.

3.3 Feature Selection

3.3.1 Basic Features

The proposed system used four features as the baseline features which includes 1) Monographic and orthographic features and bag-of-word; 2) POS tags; 3) UMLS features; and 4) discourse information such as parts in clinical identifiers. In (Tang et. al., 2012), we also tried combinations of features to enhance the accuracy of identifying the entities. Consequently, our baseline approach in this analysis included all four forms of features in (Jiang et. al., 2015) and the combined features in (Tang et. al., 2012). This research then centred on comparing the contribution of clustering based WR and word embedding WR along with the baseline features.

Table.1 Examples of BIO tags

Example 1	Token	IL-2	gene	expression	and	NF-kappa	B	activation	. . .
	Label	B-DNA	I-DNA	O	O	B-protein	I-protein	O	. . .
Example 2	Token	Comparison	with	alkaline	phosphatases	and	5	—	nucleotidase
	Label	O	O	B-GM	I-GM	O	B-GM	I-GM	I-GM

3.3.2 Clustering-Based WR.

In an unlabeled corpus, the clustering-based WR produces clusters over words and represents a word per cluster(s) to which it belongs. The principle is that semantically/syntactically related words appear to be in the same or near clusters. The Brown clustering algorithm (Brown et. al., 1992) (https:/github.com/percyliang/brown-cluster/), a hierarchical clustering algorithm, was adopted similarly to (Tang et. al., 2013). We implemented the Brown clustering framework which represents the tokens in the form of binary tree with all the tokens as the leaf nodes producing the hierarchical clusters. Fig.3 represents a hierarchical cluster containing 7 tags from the JNLPBA dataset. The numbers in the squares (e.g., 00) represent the subpaths encoded with a binary sequence starting from the root of the cluster and terms that share more similar subpaths are semantically closer. All subpaths from the root to a term i.e., a leaf node) were used as features in our experiments. "For example, for the word "for" (001010), the following characteristics were extracted: {"0," "00," "001," "0010," "00101," and "001010"}. {"0," "00," "001," "001," "001010, "001010," and "00101010"}. From the range of {50, 100, 200, 500, 1000, and 2000}, the number of clusters for running Brown clustering algorithms was chosen. The optimised cluster numbers for BioCreAtIvE II GM and JNLPBA datasets were 500 and 200, respectively.

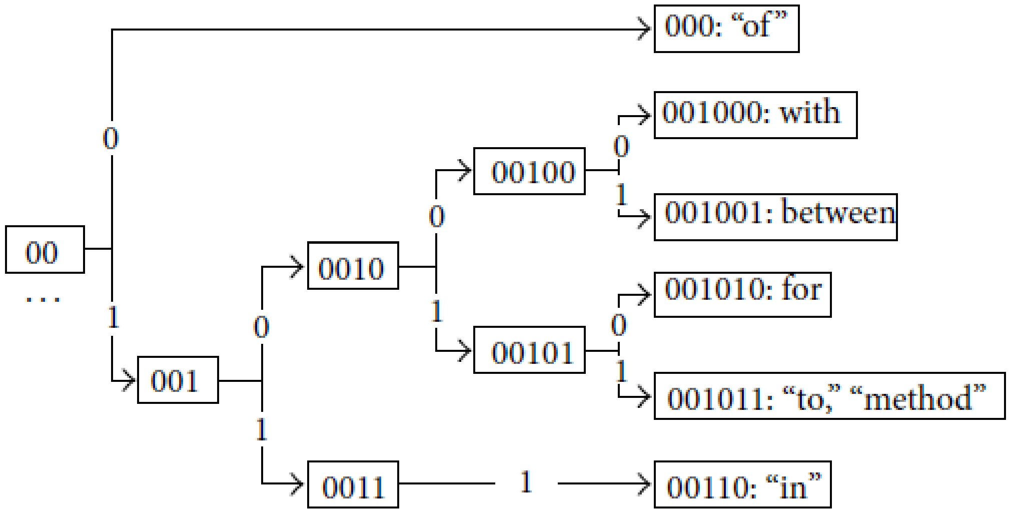

Fig. 3. Example of Brown Clustering for 7 words

3.3.4 Word Embeddings Based WR.

Word embeddings converts each entity from the dataset into real valued vector by using a continuous space language models. A term can be represented by its vector directly, and related terms are likely to have comparable vectors. We followed the approach in our experiments in (Mikolov et. al., 2013) (https:/code.google.com/p/word2vec/), a neural network language model for word embedding generation. From the {50, 100, 200, and 300} set, the dimension of each word vector was chosen. For the BioCreAtIvE II GM and JNLPBA datasets, the optimised measurements of each word vector were 50 and 100, respectively.

3.4 CRF Algorithm

For building probabilistic models for segment and labelled sequence data, CRFs are a recently introduced formalism (Lafferty et. al., 2001). In NER tasks, CRFs have achieved good results as they are free from the so-called label bias problem by the use of global normalisation. CRFs are undirected graphical models that use a given set of characteristics to encode a conditional distribution of probability. CRFs represent the $P_r(y|x)$ conditional model that uses the Markov random field, with nodes corresponding to elements of the y-structured entity, and possible functions that are X-conditional. Linear-chain CRFs are also used within the NER framework. A segment of words may be a given observed segment X, and a state segment y may be a segment in $\{I, O\}^{|x|}$, where y_i = I represents "word x_i is within a entity" and y_i = O indicates the other. The conditional probability of the state segment as $Y = \{y_i \,|1 \le i \le n\}$ for a given input segment X is given in Eq.1.

$$P(Y|X,\lambda) = \frac{1}{Z(X)} \exp \sum_{i=1}^{n} \sum_k \lambda_k f_k(y_{i-1}, y_i, X, i) \qquad (1)$$

where fk (yi−1, yi , X, i) is a feature function, λk is a weight of a feature and Z(X) is the normalization factor over all the state sequences for the sequence X, yi−1 and yi denote the previous state and the current state, respectively.

The model of CRFs is used to find the maximum probability of a set of pairs (x, y) given as training information. The labelling can be achieved using a modified Viterbi algorithm for a new unlabeled sequence.

IV. Experiments and Evaluation

4.1 Data Sets

The proposed system used the BioCreAtIvE II GM and JNLPBA datasets for conducting the experiments and evaluating the performance of the model. Table.2 represents the statistics of the datasets used for evlaution. In the JNLPBA, sentences are tokenized in the earlier stage whereas the tokenization is not performed in BioCreAtIvE II GM dataset. In our research, to perform tokenization for the BioCreAtIvE II GM corpus, we used GENIAtagger.

Table 2. Statistics of Datasets

Dataset	Training	Testing	Entity Type
BioCreAtIvE II GM	15,000 sentences	5,000 sentences	Gene/Protein
JNLPBA	18,546 sentences	3,856 sentences	protein, DNA, RNA, cell line, and cell type

4.2 Experimental Setup

In this analysis, we began with a framework that uses the basic features listed in the previous section, such as bag-of-word and POS. Then by incorporating each one separately to the baseline scheme, we assessed the impact of clustering-based, and word embeddings. In addition, we tested various combinations of two kinds of WR characteristics. Both WR characteristics were extracted from BioCreAtIvE II GM and JNLPBA's entire datasets. As an implementation of CRF, we used CRFsuite (http:/www.chokkan.org/software/ crfsuite/) and optimised its parameters by 10-fold crossvalidation on the training batch of each dataset. During 10-fold cross-validation, the optimum number for each type of word representation characteristic is also calculated. Using the test set of each corpus, the output of different baselines was assessed and recorded as precision, recall and F1-measure, measured using the formal evaluation method as given in Eq. 2 - 4 (Kim et. al., 2004, Smith et. al., 2008).

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{4}$$

Where TP is the list of entities correctly identified, FP is the list of entities wrongly identified as entities and FN is the list of missed out tags.

V. RESULTS AND DISCUSSION

5.1 Comparison with Variants of Proposed Systems

The performance of CRF-based BNER approaches on the BioCreAtIvE II GM corpus is displayed in Table 2, where two word representation have been incorporated separately or as the mixture of two representation. As displayed in the table, the performance of BNER systems has been enhanced by each individual type of WR functionality. The F-measures were enhanced by 1.41 percent and 0.85 percent respectively on the BioCreAtIvE II GM corpus when the two word representations were separately applied to the baseline method. It seemed that two representations enhance the performance and also act as the added advantage to another. When comparing the performance of the different models of the proposed system, BNER system with two representations outperformed the other models. For instance, the F1- measures on the BioCreAtIvE II GM corpus were improved by 176%. when the mixture of two representations are used (versus improvements of 1.41 percent and 0.85 percent when either clustering-based or word embedding were added to the basic feature framework). The proposed system achieved the highest F1-measures of 85.92 percent when the mixture of two representations are used. The BNER model with the highest accuracy is used to evaluate the performance on JNLPBA corpus.

Table.2 Results on BioCreAtIvE II GM corpus

	BioCreAtIvE II GM		
	Precision	Recall	F1 Score
Base Model	85.92	79.76	84.16
Base + Clustering WR1	86.82	80.95	85.57
Base + Word Embedding WR2	86.25	80.51	85.01
Base + WR1 + WR2	87.16	81.23	85.92

5.2 Comparison with the existing systems

In this section we compare the variant model with the highest accuracy obtained from the previous section is evaluated using JNLPBA corpus. Table.3 shows the comparison between our framework with the other existing systems. In Specific, the proposed framework is compared with the most recent literature such as Saha et al. (2009), Ju et al., (2011), Yang et al. (2014) and Tang et al. (2014).

Yang et. al., (2014) obtained the reasonable score next to our system using Semi CRF approach in JNLPBA corpus. Semi CRF model obtained the second highest result with a two phase model namely term boundary detection and semantic labeling. However, the proposed framework obtained the highest accuracy when comparing with the existing systems with the F1 measure of 76.64 in the JNLPBA dataset, achieving the benchmarking performance by 2.1 % in JNLPBA dataset defending the massively used feature selection models and multiple layers of CRF models.

Table. 3. Comparison with State-of-art-art Systems on JNLPBA Corpus

Systems	JNLPBA		
	Precision	Recall	F1 Score
Saha et. al.	67.86	66.94	67.41
Ju et. al.	72.01	76.76	74.31
Yang et. al.	72.83	76.54	74.64
Tang et. al.	70.78	72.00	71.39
Our System	74.63	77.69	76.74

The precision and recall measure obtained in the proposed system clearly states that the clustering based WR and word embedding based WR. Adding BIO entity tagging allowed the system to easily identify the entities from the literature text without any massive set of features. The proposed framework with a simple but powerful architecture model with word embedding and clustering based WR model obtained marginally good performance than massive feature set systems. \\

VI. CONCLUSION

The proposed system presents a technically simple architecture for Biomedical Named Entity Recognition (BNER) using CRF algorithm. The proposed system uses clustering based word representation and word embedding based word representation as the only features along with the basic feature set to identify the entities. In addition, we incorporated BIO entity tag representation to effectively identify the biomedical entities. The proposed system is evaluated on two publicly available dataset BioCreAtIvE II GM corpus and JNLPBA corpus. The experimental results on both the corpus shows that the system outperforms all the existing system with the relatively high F - score of 85.92 % on BioCreAtIvE II GM corpus and 76.64 % on JNLPBA corpus. Thus the clustering based WR and word embedding based WR along with the CRF algorithm proven to be an effective method to identify biomedical entities.

REFERENCES

[1] Akhondi, S. A., Hettne, K. M., Horst, E. V. D., Mulligen, E. M. V. and Kors, J. A. 2015. Recognition of chemical entities: combining dictionary-based and grammar-based approaches. Journal of Cheminformatics, 7: 1-11.

[2] Al-Hegami, A. S., Othman, A. M. F. and Bagash, F.T. 2017. A biomedical named entity recognition using machine learning classifiers and rich feature set. International Journal of Computer Science and Network Security, 17(1): 170.

[3] Brown, P. F., DeSouza, P. V., Mercer, P. V., Pietra, V. J. D. and Lai, J. C. 1992. Class-based n-gram models of natural language. Computational Linguistics, 18: 467–479.

[4] Cohen, A. M. and Hersh, W. R., 2005. A survey of current work in biomedical text mining. Brief Bioinformatics. 6: 57–71.

[5] Eltyeb, S. and Salim, N. 2014. Chemical named entities recognition: a review on approaches and applications. Journal of Cheminformatics.

[6] Finkel, J., Dingare, S., Nguyen, H., Nissim, M. and Manning, C. 2004. Exploiting context for biomedical entity recognition: from syntax to the web. In: JNLPBA, pp. 88–91.

[7] Jiang, M., Chen, Y., Liu, M., Rosenbloom, S. T., Mani, S., Denny, J. C. and Xu, H. 2011. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc,18:601-606.

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

[9] Kim, J D., Ohta, T., Tsuruoka,Y., Tateisi, Y. and Collier, N. 2004. Introduction to the bio-entity recognition task at JNLPBA. in Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp. 70–75.

[10] Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on, machine learning (ICML ’01), pp 282–289

[11] Lee, C., Hou, W. J. and Chen, H.H. 2004. Annotating multiple types of biomedical entities: a single word classification approach. In: JNLPBA, pp. 80–83.

[12] Lee, K., Hwang, Y., Kim, S. and Rim, H. 2004. Biomedical named entity recognition using two-phase model based on SVMs. Journal of Biomedical Informatics, 37(6): 436-447.

[13] Lyu, C., Chen, B., Ren, Y. and Ji, D. 2017. Long short-term memory RNN for biomedical named entity recognition. BMC bioinformatics, 18(462).

[14] McDonald, R. and Pereira, F. 2005. Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics, 6, S6.

[15] Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient estimation of word representations in vector space. CoRR, vol. abs/1301.3781.

[15] Olsson, F., Eriksson, G., Franzén, K., Asker, L. and Lidén, P. 2002. Notions of correctness when evaluating protein name taggers. In: COLING, pp. 765–771.

[16] Read, J., Dridan, R., Oepen, S. and Solberg, L. J. 2012. Sentence boundary detection: A long solved problem? In Proceedings of the 24nd International Conference on Computational Linguistics, 985-994.

[17] Saha , S k., Sarkar, S. and Mitra, P. 2009. Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of Biomedical Informatics. 42, 905–911.

[18] Sang, E. F. T. K. and Veenstra, J. 1999. Representing text chunks. Ninth Conference of the European Chapter of the Association for Computational Linguistics, pp. 173-179.

[19] Settles, B. 2004. Biomedical named entity recognition using conditional random fields and novel feature sets. In: JNLPBA, pp. 104–107.

[20] Sundheim, B.M. 1995. Overview of results of the MUC-6 evaluation. In: MUC-6, pp. 13–31.

[21] Tang, B., Cao, H., Wang, X., Chen, Q. and Xu, H. 2014. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. BioMed Research International. 240403.

[22] Tang, B., Cao, H., Wu, Y., Jiang, M. and Xu, H. 2012. Clinical entity recognition using structural support vector machines with rich features. Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics, 13-20.

[23] Tang, B., Cao, H., Wu, Y., Jiang, M. and Xu, H. 2013. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC Medical Informatics and Decision Making, vol. 13, no. supplement 1, p. S1.

[24] Tang, B., Feng, Y., Wang, Wu, Y., Zhang, Y., Jiang, M., Wang, J. and Xu, H. 2015. A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature. Journal of Cheminformatics, 7(1): 232–240.

[25] The Apache OpenNLP Library. [http://opennlp.apache.org/index.html].

[26] Tsai, R.T., Sung, C. L., Dai, H. J., Hung, H. C., Sung, T. Y. and Hsu, W. L. 2006. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC Bioinformatics 7, S11.

[27] Umare, S. P. and Deshpande, N. A. 2015. A survey on machine learning techniques to extract chemical names from text documents. (IJCSIT) International Journal of Computer Science and Information Technologies , 6(2):1263-1266.

[28] Wei, C-H., Harris, B. R., Kao, H-Y. and Lu, Z. tmVar: A text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 2013, 129(11):1433-1439.

[29] Xu, K., Zhou, Z., Hao, T. and Liu, W. 2017. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Sep 9-11.

[30] Yang, L. and Zhou, Y. 2014. Exploring feature sets for two-phase biomedical named entity recognition using semi-CRFs. Knowledge Information System. 40:439–453.

[31] Yang, Z. H., Lin, H. F. and Li, Y.P. 2008. Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature. Computational Biology and Chemistry., 32: 287–291.

[32] Zhou, G. D. and Su, J. 2004. Exploring deep knowledge resources in biomedical name recognition. In: JNLPBA, pp. 96–99.