http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 6 No. 2, December 2017: 1 - 11
e-ISSN: 2289-2192

EXTENDED DISTRIBUTED PROTOTYPICAL FOR BIOMEDICAL NAMED ENTITY RECOGNITION

MAAN TAREQ ABD
MASNIZAH MOHD

ABSTRACT

Biomedical Named Entity Recognition (Bio-NER) is an essential step of biomedical information extraction and biomedical text mining. Although, a lot of researches have been made in the design of rule-based and supervised tools for general NER, Bio-NER still remains a challenge and an area of active research, as still there is huge difference in F-score of 10 points between general newswire NER and Bio-NER. The complex structures of the biomedical entities pose a huge challenge for their recognition. To handle this, this paper explores different effective word representations with Support Vector Machine (SVM) to deal with the complex structures of biomedical named entities. First, this paper identifies and evaluates a set of morphological and contextual features with SVM learning method for Bio-NER. This paper also presents an extended distributed representation word embedding technique (EDRWE) for Bio-NER. These models are evaluated on widely used standard Bio-NER dataset namely GENIA corpus. Experimental results show that EDRWE technique improves the overall performance of the Bio-NER and outperforms all other representation methods. Results analysis shows that the new EDRWE is satisfactory and effective for Bio-NER especially when only a small-sized data set is available.

Keywords: word representation, word embedding, biomedical named entity.

INTRODUCTION

The exponential growth in the size of available biomedical literature has motivated a lot of interest in designing efficient techniques for biomedical information extraction and text mining. There are about 22 million abstracts in the domains of medicine, bio-medical sciences and laboratory sciences in the MEDLINE database (Lee et al. 2016). Biomedical named entity recognition (Bio-NER) is an essential step of biomedical information extraction and biomedical text mining. Biomedical named entity recognition aims to identify and classify technical entities in the domain of molecular biology. These entities are of interest to biologists and scientists such as protein and gene names, cell types, virus name, DNA sequence, and others.

Unlike general named entities (e.g. person, location, date and time), biomedical named entities have inherently complex structures which poses a big challenge for their identification and classification in biomedical information extraction. The Bio-NER is vast, but there is still a wide gap in performance between the general NER and the existing Bio-NER. Therefore, there is a room for improvement as recognition accuracy of name entity has basically hovered around 10 points in their F-measure. In addition, the ability of biomedical researchers to manage, integrate and analyse biomedical data is often limited due to a lack of tools, accessibility, and training dataset (Yao et al. 2015). The difficulty and potential importance of this task has attracted many researchers.

The Bio-NER is more difficult than general NER because of the complex situations such as irregular expression, consist of long compounded words, hardly distinguished boundaries, same

word or phrase can refer to different named entities and daily changing group members. To handle this, this paper explores and implements several new word representation schemes with SVM to Bio-NER to deal with the complex structures of biomedical entities.

First, this paper identifies a very rich feature set that includes variety types of features based on orthography, morphological and contextual features. Then, it introduces and implements extended distributed representative word embedding technique (EDRWE) for biomedical named entity recognition. It uses SVM framework to build a number of models depending upon various representations. Experimental results demonstrate that extended distributed representative word embedding technique (EDRWE) significantly improves the overall performance over traditional features representation technique.

RELATED WORK

In the past years, several models and methodologies have been proposed for Bio-NER such as to extract gene, protein, chemical, drug and other biological relevant named entities. The release of the GENIA corpus has pushed forward related research using various supervised learning models, including Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and Conditional Random Fields (CRFs). This approach has been used to identify a wide range of Bio entities, including genes and proteins (Wei et al. 2015; Lee et al. 2016), chemicals (Usié et al. 2014; Leaman et al. 2015; Zhang et al. 2016) and anatomic entities (Pyysalo & Ananiadou 2014).

Ekbal et al. (2013) hypothesize that the reliability of predictions of each classifier differs among the various output classes. Thus, they use CRF and SVM frameworks to build a number of models depending upon various representations of the set of features and/or feature templates. Sikdar et al. (2014) propose a single objective optimisation based classifier ensemble technique using the search capability of genetic algorithm for Bio-NER texts. Here, the genetic algorithm is used to quantify the amount of voting for each class in each classifier. CRF and SVM are used to build a number of models. Bhasuran et al. (2016) developed a hybrid classifier using stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. For stacked ensemble, they used of several CRFs as their base level classifier that combines outputs as a second-level meta classifier in an ensemble.

Chang et al. (2015) generate word embedding features from an unlabelled corpus, which as extra word features are induced into the CRFs system for Bio-NER. Li et al. (2017) propose a neural joint model to extract biomedical entities and their relations. First, their model uses CNNs to encode character information of words into their character-level representations. Second, character-level representations, word embeddings and POS embeddings are fed into a bi-directional (Bi) long short-term memory (LSTM) based RNN to learn the representations of entities and their contexts in a sentence. These representations are used to recognize biomedical entities.

This paper introduces and implements extended distributed representative word embedding technique (EDRWE), and induce these features into a SVM-based Bio-NER system. This paper also conducts a comparative study of three word representation methods for Bio-NER tasks.

MATERIALS AND METHODS

In the past years, several models and methodologies have been proposed for Bio-NER such as gene, protein, chemical, drug and other biological relevant named entities. As shown in Fig. 1, the

framework of the proposed method comprises of four stages namely corpus, data representation, SVM and evaluation.

The first stage discusses the dataset used in the study which contains a benchmark data set of biomedical named entities. The second stage describes the data representation which consists of two main tasks which are, (I) feature engineering where a set of traditional features are identified and the data set are prepared according to these features (II) extended distributed representative word embedding technique for word representation. This stage is crucial as the input to be fed into a SVM-based Bio-NER. The third stage aims to apply the recognition of the biomedical named entities using SVM-based Bio-NER. The fourth stage discusses the evaluation metrics used to measure the performance of the proposed method.

DATA AND CORPUS

The dataset that has been used in this study is GENIA corpus which is the most widely accepted dataset and commonly used benchmark dataset for biomedical named entity recognition. The GENIA corpus also has been adopted by many research groups as an assessment.

The GENIA corpus is the primary collection of biomedical literature compiled and annotated within the scope of the GENIA project. The GENIA corpus is the largest annotated corpus in the molecular biology domain publicly available. The corpus was created to support the development and evaluation of information extraction and text mining systems for the domain of molecular biology. The corpus contains Medline abstracts, selected using a PubMed query for the three MeSH terms “human,” “blood cells,” and “transcription factors.” The corpus has been annotated with various levels of linguistic and semantic information. The original GENIA corpus contains 36 classes of entities. A more widely used version of GENIA corpus is the one simplified for the BioNLP/NLPBA shared task, in which entities are grouped into only five major classes: protein, DNA, RNA, cell line, cell type.

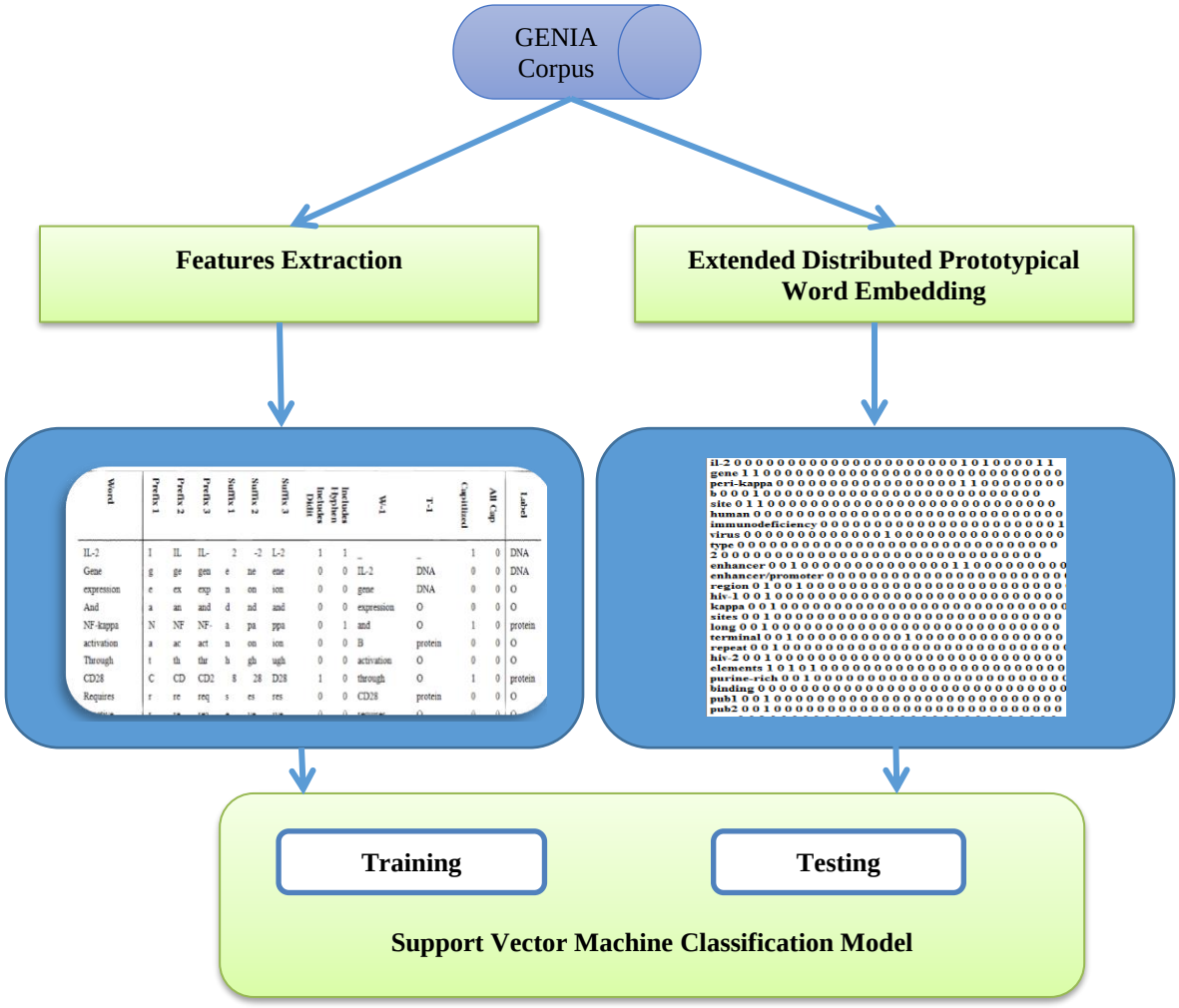

FIGURE 1. Overflow of the used Methodology

FEATURE ENGINEERING AND WORD REPRESENTATION

Data representation is the most important factors in supervised machine learning success in any domain to achieve the best performance. In this section, we described several data representation schemes used in our methodology.

FEATURE EXTRACTION

The main aim of feature extraction is to represent each word in the data set as a vector of values for morphological, orthographical and contextual features that can improve the recognition performance of Bio named entities while reducing the processing overhead. As a matter of fact, if selected features are independent and correlated with classes, the performance of the recognition will be improved. Otherwise, the performance of the recognition will be dropped. Thus, how to select features and integrate features is important in Bio-NER. Traditional features such as morphological, syntactic and semantic information of words. These features differ between entity types, which make their development costly. Furthermore, these features are complex hand

designed and often optimized for a specific gold standard corpus. As shown in Table 1, this work represents each word using a set of morphological, orthographical and contextual features.

TABLE 1. Features used for training classifiers

Feature set	Actual features in the feature set	Number of features
Context words	One word before and one word after current word	2
Dynamic feature	Dynamic feature denotes the output tag of previous words	1
Orthographic features	Orthographic features : Several binary features are defined: initial capital, all capital, • Includes caps, has slash, has punctuation, has digit	6
Word affixes	Word prefix and suffix character sequences of length up to 3.	6

EXTENDED DISTRIBUTED REPRESENTATIVE WORD EMBEDDING (EDRWE)

The performance of Bio-NER systems is always limited to the construction of complex hand-designed features which are derived from various linguistic analyses(Li et al. 2015).The distributed representative features was also proposed by Guo et al. (2014) for the general English domain. As in Guo et al. (2014), this work uses association measure to extract representative (prototypical) words for each class. Unlike, Guo et al. (2014), this work introduces a second order representation method that maps words to co-occurrence vectors and then represents each word as a vector of its vector similarity values with all representative words vectors. This works also introduces the extended pointwise mutual information to overcome the unsymmetrical co-occurrence problem of PMI. In this work, an extended distributed representative word embedding is proposed which can be described as follow:

Prototypical words extraction: The prototype feature method selects representative (prototypical) words for each class of the five GENIA classes: protein, DNA, RNA, cell line and cell type. The prototypical words will be assigned to each class according to co-occurrence measure between a word w and a biomedical named entity class. As in (Guo et al. 2014), the prototypical feature words are selected using the normalized pointwise mutual information (PMI) between the word and its classes using the following equations.

$$nPMI(class, word) = \frac{PMI(class,word)}{lin\ p\ (class,word)} \quad (1)$$

$$PMI(class, word) = lin \frac{p(class,word)}{p\ (class)\ .p(word)} \quad (2)$$

A set of representative (prototypical) words are constructed for each biomedical entity classes: protein, DNA, RNA, cell line, cell type as shown in Table 2.

TABLE 2. Sample of Representative (Prototypical) Words

Class	Representative (Prototypical) Words
DNA	RIC ; CIS-ACTING ; LTR ; CONSTRUCTS ; GIRE ; REPORTER ; TATA ; BOX ; BP
O (Others, not named entity)	ACTIVITY ; SIGNALING ; INDUCED ; INFECTION ; PHOSPHORYLATION ; STIMULATION ; RESPONSE ; ROLE ; LEVELS
PROTIEN	FACTORS ; P95VAV ; TCR ; EPO ; STAT ; CD28 ; ER ; B1 ; B2 ; B-ALPHA ;
Cell Type	LYMPHOCYTES ; THYMOCYTES ; PERIPHERAL ; BLOOD ; PROGENITOR ; TYPES ; LGL ; PBL ; NON-ERYTHROID ; MONOCYTES
Cell line	U937 ; LINE ; LINES ; NK3.3 ; K562 ; LYMPHOCYTE ; 32DC13 ; CELLS ; JURKAT ; UNPRIMED
RNA	MRNAS ; MRNA ; TNF ; CYTOPLASMIC ; IL-2R ; IFN-GAMMA ; GM-CSF ; AND ; ALPHA

Prototypical words representation: After sets of representative (prototypical) words are constructed for each class, then co-occurrence vector is generated for each word w including representative (prototypical) words. Each word w_i from the data set including representative (prototypical) words is represented using K prototypical words representation. EPMI is proposed and defined to representative (prototypical) words based on extended mutual information EMI and PMI^2 to overcome the unsymmetrical co-occurrence problem of PMI. In order to generate the co-occurrence vector v for the word w_i, the co-occurrence relation between the word w_i and every word w_j from the data set using EPMI which is derived from extended mutual information EMI and PMI^2 as follows:

$$\mathrm{EMI}(w_i, w_j) = \log_2 \frac{P(w_i,w_j}{(P(w_i)-P(w_i,w_j))(P(w_j)-P(w_i,w_j))} \quad (3)$$

$$PMI^2 = \log_2 \frac{P(w_i,w_j}{(P(w_i))(P(w_j))} \quad (4)$$

$$\mathrm{EPMI}(w_i, w_j) = \log_2 \frac{P(w_i,w_j)^2}{(P(w_i)-P(w_i,w_j))(P(w_j)-P(w_i,w_j))} \quad (5)$$

Second, a prototypical words representation is constructed for each word. Each word is represented by a vector of $\boldsymbol{n}$ dimensions, where **n** represents the size of selected prototypical words. For each word $\boldsymbol{w}$ in the training/test, the cosine similarity between $\boldsymbol{w}$ and all the selected prototypical words using the associated embedding co-occurrence vectors. If the cosine similarity of the co-occurrence vector of $\boldsymbol{w}$ and the co-occurrence vector of a prototypical word is above the predefined threshold (0.50), the prototypical word will be assigned as a feature. Given the co-occurrence vector of $\boldsymbol{w}$ ($cv(w)$) and the co-occurrence vector of a prototypical word $\boldsymbol{pw}$ ($cv(pw)$), the cosine similarity is defined as:

$$sim_{cosine}(\boldsymbol{w}, \boldsymbol{pw})^2 = \frac{\sum_{i=1}^{|cv|}(cv_i(w) * cv_i(pw))}{\sum_{i=1}^{|cv|}(cv_i(w))^2 + \sum_{i=1}^{|cv|}(cv_i(pw))^2} \quad (6)$$

Support Vector Machine (SVM)

SVM is generally a popular technique for named entity recognition, which is used in the machine learning area. SVM is considered one of the recognition techniques with a very high efficiency. It is application in biomedical NER presents a major problem. SVMs are primarily binary classifiers and are often trained using a one against rest approach. To handle this problem, two levels of classification are used here. Their first level is the identification level which consisted of training an SVM classifier to simply identify biomedical entities from non-biomedical entities and post biomedical entities to the next level. The second level is the Classification level which involves the classification of biomedical entities into one of the five GENIA classes: protein, DNA, RNA, cell line, cell type through five separated of SVM classifiers.

SVM 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, which are selected as the only components that are efficient in the training set.

Given a training data in the form of n k-dimensional real vectors xi, and integers yi , where yi is either 1 or 1. Whether yi is positive or negative point to the class for the vector i. The aim of the training phase of the SVM is to plot the vectors in a k-dimensional hyperspace and draw a hyperplane which as evenly as possible separates points from the two categories.

$$\vec{\alpha} = argmin\left\{-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\langle\overrightarrow{x_i}, \overrightarrow{x_j}\rangle\right\} \qquad (7)$$

$$\text{Subjectto:}\sum_{i=1}^{n}\alpha_i y_i = 0;\ 0 \le \alpha_i \le C$$

EVALUATION

To evaluates the performance of the proposed Bio-NER, this work uses the same evaluation metrics used in the CoNLL-02, CoNLL-03 and JNLPBA-04 challenge tasks which are precision, recall, and the weighted mean Fβ=1-score. Precision is the percentage of biomedical named entities found by Bio-NER method that are correct. Recall is the percentage of biomedical named entities present in the corpus that are found by the Bio-NER method. For each biomedical entity class, precision and recall are defined as follows:

$$\text{Recall} = \frac{\#\text{ of correctly classified entities}}{\#\text{ of entities in the corpus}} = \frac{tp}{tp + fn} \qquad (8)$$

$$\text{Precision} = \frac{\#\text{ of correctly classified entities}}{\#\text{ of entities found by algorithm}} = \frac{tp}{tp + fp} \qquad (9)$$

F-measure is the most standard measure for evaluating recognition systems as it to consider a trade-off between precision and recall, where it combines precision and recall by function. F1-measure is the harmonic mean of precision and recall and is defined as follows:

$$F_1 = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}} \qquad (10)$$

RESULTS AND DISCUSSION

Since the aim of this paper is to study the ability of several word representation methods to generate a reliable recognition model for biomedical named entity and to examine which representation lead to the best recognition performance, several experiments are conducted to perform evaluation of different methods at different training data sizes. Although, different training data sizes are used, this work uses the same experiment settings for all word representation methods to study the impact of the size of the training data and test data on their performances. This means this work studies which word representation methods is suitable when there is only small training data is available. The poor resources problem is a common in many natural language processing and biomedical data mining tasks.

First, this paper evaluates the performance of the feature representation method where each word in the data set as a vector of values for morphological, orthographical and contextual features and studies their effect on the performance of the SVM. The effects of these features on Bio-NER performance are examined given different training sizes. Three word representation methods for Bio-NER tasks are (a) feature representation method with SVM classifier, (b) baseline prototypical word embedding method with SVM classifier and (c) extended distributed representative word embedding (EDRWE) with SVM classifier.

The experimental results of the feature representation method are shown in Table 3. Comparing the behaviors of the feature representation method with different training sizes, the results show that the recognition performances are increased when the size of training data is increased.

TABLE 3. Performance of Feature Representation Method with SVM Classifier on Different Training Sizes from GENIA Dataset

Training size	Test Size	Precision	Recall	F-Measure
90	10	57.40	91.84	70.65
80	20	54.18	89.28	67.44
50	50	42.75	84.58	56.79
40	60	43.44	90.16	58.63
30	70	36.33	82.60	50.47

Second, this paper evaluates the performance of the baseline prototypical word embedding representative method as in Guo et al. (2014) and studies their effect on the performance of the SVM. The Bio-NER performances are examined given different training sizes. The experimental results are shown in Table 4. Comparing the behaviors of the baseline prototypical word embedding representative method results with different training sizes, as in feature representation method, results show that the recognition performances are increased when the size of training

data is increased and the recognition performances are dropped when small training data sizes. The size of prototypical words is determined experimentally. The best size of prototypical words is fixed to 20 prototypical for each class. However, results show that baseline prototypical word embedding representative method performs slightly better than the feature representation method. The main reason is that baseline prototypical word embedding representative method can capture more semantic and syntactic information than morphological, orthographical and contextual features.

TABLE 4. Performance of Baseline Prototypical Word Embedding Method with SVM Classifier on Different Training Sizes from GENIA Dataset

Training	Test	Precision	Recall	F-Measure
90	10	72.06	82.62	76.98
80	20	70.56	78.08	74.13
50	50	63.88	84.00	72.57
40	60	63.78	80.83	71.30
30	70	63.88	82.08	71.85

Finally, this paper evaluates the performance of the extended distributed representative word embedding (EDRWE) method and studies their effect on the performance of the SVM. The Bio-NER performances are examined given different training sizes. The experimental results are shown in Table 5. Comparing the behaviors of its results with different training sizes, as in feature representation method, results show that the recognition performances are increased when the size of training data is increased and the recognition performances are dropped slightly when small training data sizes. Table 5 show that the proposed EDRWE produce superior results to other representation methods for all training sizes. Experiments also indicate that the EDRWE method produced the best results. The main reason is that EDRWE method presents a distributed representation over word classes so it can capture semantic and syntactic information. The syntactic information can be captured by brings together words that occur in a same syntactic structure (LEBRET 2016). Each value in the vector represents word's semantic relation with a prototypical word. Furthermore, experimental results show that the proposed EDRWE is suitable for Bio-NER even when only a very small fraction of the training data size.

Although the results obtained by (Zhu 2016) is better than that obtained here, the proposed EDRWE has two main contributions. First, EDRWE proves to be suitable when only small training data is available. Unlike, this work shows that new word embedding techniques can work well with SVM and they are not committed to particular machine learning method. This paper also presents a comparative study between traditional and new word representation methods for Bio-NER tasks.

TABLE 5. Performance of Extended Distributed Representative Word Embedding (EDRWE) with SVM Classifier on Different Training Sizes from GENIA Dataset

Training	Test	Precision	Recall	F-Measure
90	10	80.62	85.12	82.81
80	20	77.06	89.42	82.78
50	50	70.28	90.37	79.07
40	60	67.42	90.17	77.15
30	70	69.64	88.26	77.85

CONCLUSION

The main contribution of this work is that it proposes a new word representation methods namely extended distributed representation word embedding technique (EDRWE) for Bio-NER. This paper presents a comparative study of three word representation methods for Bio-NER tasks. The results indicate that the proposed EDRWE produce superior results to other representation methods for all training sizes. Experimental results show that the proposed EDRWE is suitable for Bio-NER even when only a very small fraction of the training data size is available.

Our future efforts will be targeted at evaluating extended distributed representation word embedding technique (EDRWE) with other classification tasks including general NER. In addition, future work may extend the proposed methods and evaluates them with other advanced machine learning models such as deep learning.

REFERENCES

Bhasuran, B., G. Murugesan, S. Abdulkadhar & J. Natarajan 2016. Stacked Ensemble Combined with Fuzzy Matching for Biomedical Named Entity Recognition of Diseases. *Journal of Biomedical Informatics* 64: 1-9.

Chang, F., J. Guo, W. Xu & S. R. Chung 2015. Application of Word Embeddings in Biomedical Named Entity Recognition Tasks. *Journal of Digital Information Management* 13(5): 321-327.

Ekbal, A., S. Saha & U. K. Sikdar 2013. Biomedical named Entity Extraction: Some Issues of Corpus Compatibilities. *SpringerPlus* 2(1): 601.

Guo, J., W. Che, H. Wang & T. Liu 2014. Revisiting Embedding Features for Simple Semi-supervised Learning. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Leaman, R., C.-H. Wei & Z. Lu 2015. Tmchem: A High Performance Approach for Chemical Named Entity Recognition and Normalization. *Journal of Cheminformatics* 7(1): 1-10.

Lebret, R. P. 2016. Word Embeddings for Natural Language Processing.Tesis PhD Thesis Laboratoire De L'idiap,

Lee, S., D. Kim, K. Lee, J. Choi & S. Kim 2016. BEST: Next-Generation Biomedical Entity Search Tool for Knowledge Discovery from Biomedical Literature. *PloS one* 11(10):1-16.

Li, F., M. Zhang, G. Fu & D. Ji 2017. A Neural Joint Model for Entity and Relation Extraction from Biomedical Text. *BMC bioinformatics* 18(1): 198.

Li, L., L. Jin, Z. Jiang, D. Song & D. Huang 2015. Biomedical Named Entity Recognition based on Extended Recurrent Neural Networks. *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. pp. 649-652.

Pyysalo, S. & S. Ananiadou 2014. Anatomical Entity Mention Recognition at Literature Scale. *Bioinformatics* 30(6): 868-875.

Sikdar, U. K., A. Ekbal & S. Saha 2014. Differential Evolution based Multiobjective Optimization for Biomedical Entity Extraction. *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. pp. 1039-1044.

Usié, A., R. Alves, F. Solsona, M. Vázquez & A. Valencia 2014. CheNER: Chemical Named Entity Recognizer. *Bioinformatics* 30(7): 1039-1040.

Wei, C.-H., H.-Y. Kao & Z. Lu 2015. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed research international* 2015.

Yao, L., H. Liu, Y. Liu, X. Li & M. W. Anwar 2015. Biomedical Named Entity Recognition based on Deep Neutral Network. *International Journal of Hybrid Information Technology* 8(8): 279-288.

Zhang, Y., J. Xu, H. Chen, J. Wang, Y. Wu, M. Prakasam & H. Xu 2016. Chemical Named Entity Recognition in Patents by Domain Knowledge and Unsupervised Feature Learning. *Database* 2016: 1-10.
Zhu, F. S., Bairong 2016. Combined SVM-CRFs for Biological Named Entity Recognition with Maximal Bidirectional Squeezing. *PLoS ONE* 7(6).

Maan Tareq Abd
Masnizah Mohd
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor
MALAYSIA
maan.tariq97@yahoo.com, masnizah.mohd@ukm.edu.my

Received: 18 August 2017
Accepted: 30 November 2017