A COMPARATIVE STUDY OF WORD REPRESENTATION METHODS WITH CONDITIONAL RANDOM FIELDS AND MAXIMUM ENTROPY MARKOV FOR BIO-NAMED ENTITY RECOGNITION

***Maan Tareq Abd*[1] *& Masnizah Mohd*[2]**

[1,2]Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 Bangi Selangor
Malaysia

Email: maantabd@gmail.com[1], masnizah.mohd@ukm.edu.my[2]

DOI: https://doi.org/10.22452/mjcs.sp2018no1.2

ABSTRACT

Bio-Named Entity Recognition (Bio-NER) is the process of identifying and semantically classifying biomedical technical terms and named entities in Biomedicine literature. Therefore, it is a major task in biomedical knowledge acquisition. Meanwhile, Natural Language Processing (NLP) plays an important role in Bio-NER in the biomedical domain. The first and most essential biomedical literature mining task incorporates biomedical entity recognition such as protein, gene, and chemicals. The most recent Bio-NER methods rely on predefined traditional features, which attempt to capture the specific surface properties of entity types. However, these empirically predefined feature sets differ between entity types and are manually constructed and complicated, which means developing them is costly. In this paper, we systematically present a comparative evaluation study of three methods, which are: the traditional feature representation method, the continuous bag-of-words (CBOW) model, and a new prototypical representation method with two popular sequence-labeling approaches (Conditional Random Fields (CRFs) and Maximum Entropy Markov Models (MEMM)). We evaluated these models with two major Bio-NER tasks, which involve the JNLPBA and GENETAG corpora. This paper examined the prototypical word representation method and found that Word2Vec can be successfully used for Bio-NER. Our results show that the new prototypical representation method improved the performance of the two machine learning models with different datasets. Also, the new prototypical representation method performed better than the traditional feature representation method and CBOW model for both datasets. Finally, our experiment proved that the CRF classifier with the new prototypical representation method achieved the best results when 90% data was used as training data, yielding overall *F*-measure values of 0.79% and 0.85% for the JNLPBA corpus and GENETAG corpus, respectively. In comparison, the results achieved using the ME classifier yielded overall *F*-measure values of 0.76% and 0.78% for the JNLPBA corpus and GENETAG corpus, respectively.

Keywords—biomedical named entity, prototypical representation, data representation methods, Word2Vec.

1.0 INTRODUCTION

Biomedical Named Entity Recognition (Bio-NER) is the process of identifying and semantically classifying biomedical technical terms and named entities in Biomedicine literature; it is a major task in biomedical knowledge acquisition, and is challenging because of the complexity of biomedical terminology. These challenges have motivated many researchers to design efficient techniques for biomedical information extraction and text mining. Unlike general named entities (e.g. person, location, date and time), biomedical named entities have inherently complex structures. This makes it very challenging to identify and to classify these entities during biomedical information extraction. The field of Bio-NER is vast, but there is still a wide gap in performance between the general NER and existing Bio-NER [1][2][3][4][5][6][7]. Therefore, there is room for improvement in this field given that the recognition accuracy of named entities basically hovers around a 10-point F-measure. In addition, the ability of biomedical researchers to manage, integrate, and analyze biomedical data is often limited due to a lack of tools, accessibility, and training [8]. The difficulty and potential importance of this task has attracted many researchers [9].

To date, most NER tools such as (Stanford NER, Illinois NET, and OpenCalais NER WS) have been able to capture the characteristics of different entity classes using feature engineering, which finds the set of features that best help distinguish entities of a specific type from other entity classes. Currently, this procedure is manually constructed and often optimized for a specific gold standard corpus.

These traditional features do not have the powerful ability to capture words with comparable synonyms that appear in the same context [10]. The task of text recognition becomes difficult as most biomedical names contain different characteristics [11] such as:

- Descriptive nature of entity names;
- One head noun shared by two or more entity names;
- Several spelling forms for one entity name;
- Frequent use of ambiguous abbreviations; and
- Authors of biomedical texts often do not keep track of standards for naming entities and they choose abbreviated forms according to their personal preference.

All these characteristics make for trickier identification of biomedical entities compared to the identification of traditional entities. Therefore, it is necessary to explore effective word representation and learning methods to deal with the special characteristics of text in the biomedical domain. Recent works have explored and implemented new representation planners—as part of the traditional feature engineering method—with Bio-NER to address the complex structures of biomedical entities.

In this paper, we developed several Biomedical Named Entity Recognition models. The paper presents a comparative evaluation between the traditional feature representation method, continuous bag-of-words (CBOW) model, and a new prototypical representation method. The classification performance of these representation methods was scrutinized using two popular Natural Language Processing (NLP) approaches (Conditional Random Fields (CRFs) and Maximum Entropy Markov Models (MEMM)). All models were evaluated via two standard Biomedical Named Entity Recognition datasets, namely the JNLPBA corpus and GENETAG corpus. The goal of this study is to determine the representation method with the most influence on classification performance, and the classifier that most suits Bio-NER and whether or not the training data size will have an effect on overall performance.

We review the related works in this field in Section 2. Then, we discuss several techniques and approaches on this subject matter in Section 3. In Section 4, we present the implementation of the setup for the models examined in this study. We report our findings in Section 5. Finally, we discuss our conclusions in Section 6.

2.0 RELATED WORK

Rule, lexicon-based, and Machine Learning (ML) approaches are used in NER systems in the biomedical domain. These approaches depend on the linguistic criteria of the identified entities and satisfy different requirements. Moreover, these approaches have been used to identify a wide range of Bio-entities, including genes and proteins [12][13], chemicals [14][15][16], and anatomic entities [17]. Another important aspect of Biomedical Named Entity Recognition based on machine learning techniques is the feature sets used to design the classification models.

Zhang et al. [18] proposed a new neural language model to train word embeddings using ranking loss criteria with negative sampling. The experimental results showed that the ranking-based word embedding derived from the entire English Wikipedia corpus greatly helped the NER task in the general English domain. Zhou et al. [19] conducted a study to build a NER framework using a Maximum Entropy Markov model with shallow linguistic information as features. They evaluated the model using a GENIA corpus. The system obtained satisfactory results with overall F-scores of 70.0% without using any dictionary.

Levy [20] showed that the skip-gram model by Mikolov et al. (2013) may be used to implicitly factorize a word-context matrix, the values of which are the pointwise mutual information (PMI) of the several words and context pairs shifted by a global constant. Other researchers exploited similar frameworks by combining recurrent neural networks (RNNs) with CRFs and obtained the most effective results with many benchmark NER datasets [21][22]. In relation classification, 2 progressive strategies involving victimization deep neural networks, particularly RNNs and convolutional neural networks (CNNs), were used. In these studies, RNNs or CNNs determined the relation representations of the words between 2 target entities or on the words with the shortest dependency path (SDP) of 2 target entities.

Tang et al. [3] conducted a study to evaluate the different types of unsupervised word representations as part of a Bio-NER task. They systematically investigated 3 totally different types of word representation (WR) options for BNER together with clustering-based representation, distributional representation, and word embeddings by using the popular Word2Vec package to come up with the word embeddings. The study directly used the real values from the embedding matrix as features in a CRF model without any discretization and a relatively small corpus size (20,000 sentences from the BioCreAtIvE GM corpus and 22,402 sentences from the JNLPBA corpus).

Chang et al. [23] generated word embedding features from an unlabeled corpus, in which extra word features were introduced into the CRF system for Bio-NER. Li et al. [24] proposed a neural joint model to extract biomedical entities and their relations. First, their model made use of CNNs to encode character records of phrases into man- or woman-level representations. Then, the 2-D, person-stage representations, word embeddings, and part-of-speech (POS) embeddings were fed right into a bidirectional (Bi) long short-term memory (LSTM) primarily based on a Recurrent Neural Network (RNN) to examine the representations of entities and their contexts in a sentence. These representations were used to recognize biomedical entities. Segura-bedmar and Mart [25] used phrase embeddings as input for the CRF and found the most effective marginal consequences for chemical NER.

CRF and SVM were used as machine learning classifiers in Tang et al. [26], who proposed a machine learning-based system for NER in biomedical literature. Investigation was conducted on the effects of 2 types of word embeddings, which are random indexing and skip-gram, on 2 machine learning-based systems. They proved that with the same word embedding, SVM-based systems outperformed the CRF-based systems.

Bhasuran et al. [27] combined fuzzy matching and a stacked ensemble approach for disease names in Bio-NER. The simple idea behind stacked generalization is the consolidation of the yields of base-degree classifiers utilizing a second-stage meta-classifier in an ensemble. They utilized a Conditional Random Field (CRF) as the simple order method, which utilized a differing set of highlights, for the maximum component dependent on specific area, and are orthographic and morphologically related. Moreover, they applied fuzzy string matching to label rare disease names from a disease dictionary.

Wang et al. [28] presented a classifier ensemble approach for Bio-NER. Generalized Winnow, CRF, SVM, and ME were combined via three different strategies. The stacking method was proposed in the classifier ensemble method and lead to outstanding improvement in Bio-NER performance. Zhu and Shen [29] used both support vector machines (SVMs) and CRF for better performance. SVM, a binary classifier, was used to split the biological terms from non-biological phrases, and CRF was used to decide the sorting of the biological terms. Therefore, the outcomes of SVM in addition to CRF were fused and a meaningful algorithm was developed after applying 2 rules.

Murugesan et al. [30] presented a hybrid biomedical named entity tagging approach, where various kinds of features were extracted. Orthographic, morphological, prefix, and suffix features were some of the examples of the local context of each token in the study. The findings showed that the BioCreative text corpus of BCC-NER performed better compared to other open source taggers currently available at the time. Liu et al. [31] determined the best word embeddings that marginally improved the performance of the CRF-based technique using complete dictionaries as features. Habibi et al. [32] showed that a completely generic method based on deep learning and statistical word embeddings outperformed Bio-NER tools, and often by a large margin.

Ekbal et al. [33] hypothesized that the reliability of predictions of each classifier varied based on the number of diverse output training. As a consequence, they used Conditional Random Fields (CRF) and SVM frameworks to construct a number of models depending on the diverse representations of the set of features and/or feature templates.

The previous works above show the difficulty of conducting Bio-NER using the morphological, syntactic, and semantic information of words. Therefore, it is necessary to explore effective word representation and learning methods to deal with the special characteristics of text in the biomedical domain. In this paper, we conduct a comparative study of three word representation methods to execute Bio-NER tasks with two widely used machine learning methods and two different datasets.

3.0 MATERIALS AND METHODS

In the past years, several models and methodologies for Bio-NER and other biologically relevant named entities have been proposed. As shown in Fig. 1, the methodology used for the Bio-NER in this study consists of four stages.

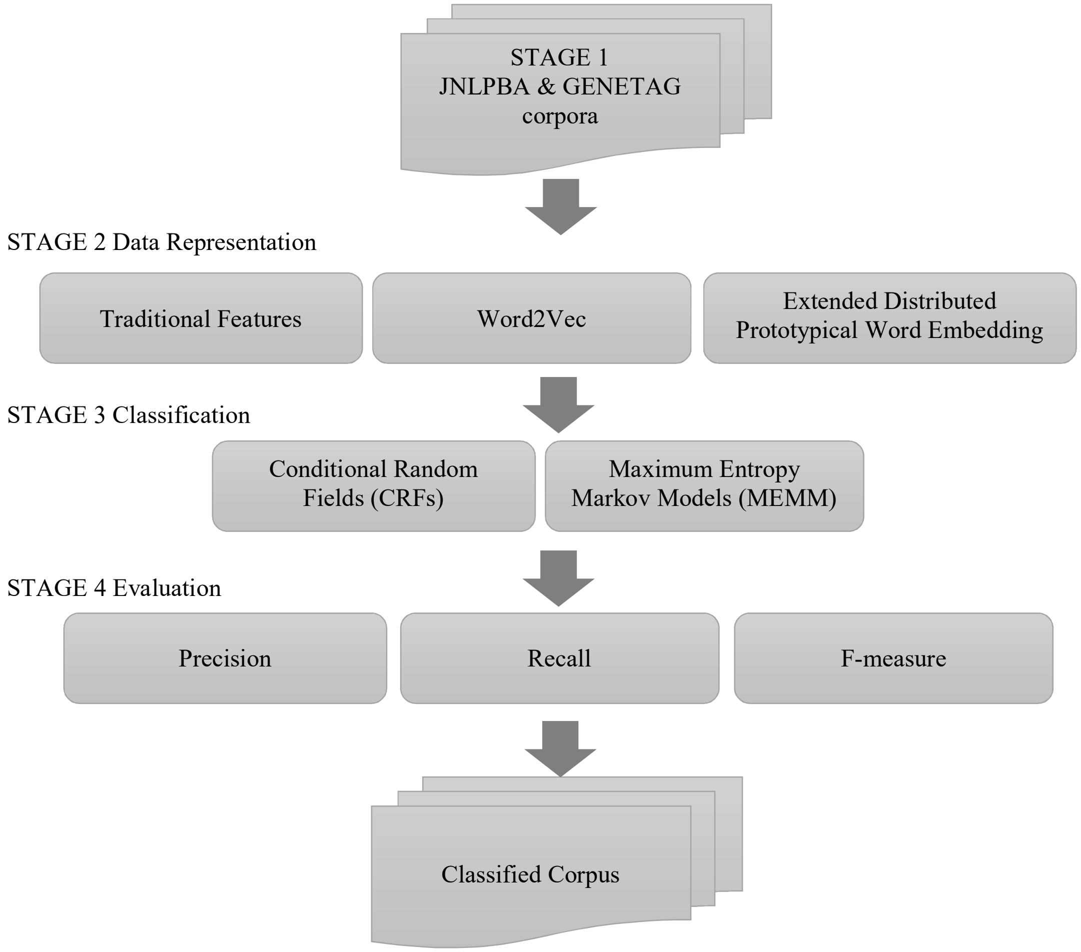
STAGE 1
JNLPBA & GENETAG corpora
STAGE 2 Data Representation
Traditional Features
Word2Vec
Extended Distributed Prototypical Word Embedding
STAGE 3 Classification
Conditional Random Fields (CRFs)
Maximum Entropy Markov Models (MEMM)
STAGE 4 Evaluation
Precision
Recall
F-measure
Classified Corpus

Fig. 1. Methodology

- Stage 1: The datasets used in this experiment, which contain a benchmark dataset of biomedical named entities, are discussed.
- Stage 2: The data representation, which consists of two main tasks, are described, as below:
  a) Feature engineering, where a set of traditional features are identified and the dataset is prepared according to these features.
  b) Word2Vec was applied; in particular, the continuous bag-of-words (CBOW) model was adopted for representation.
  c) Extended distributed prototypical data representation. This stage is crucial, and acts as input to be fed into the CRFs and ME-based Bio-NER.
- Stage 3: The goal of this stage is the recognition of the biomedical named entities using two machine learning methods.
- Stage 4: The evaluation metrics used to measure the performance of the proposed method are outlined.

3.1 Data and Corpus

The entire supervised machine-based model relies on a corpus that has been used to train it. We used the JNLPBA corpus (the most widely accepted dataset and commonly used benchmark dataset for Bio-NER) and the GENETAG corpus (used by experts in biochemistry, genetics, and molecular biology).

- The JNLPBA corpus [34] is a sub-set of the GENIA corpus. It has formatted all the abstracts of the GENIA corpus for IOB 21 notation and makes it available as a training set. The GENIA corpus is the largest annotated corpus in the molecular biology domain that is publicly available. The corpus is meant for improvement and evaluation of an information extraction and textual content mining systems in the domain of molecular biology. The JNLPBA corpus includes 22,402 sentences from Medline abstracts, selected using a PubMed question involving the 3 MeSH phrases “human,” “blood cells,” and “transcription elements.” The corpus was annotated with diverse degrees of linguistic and semantic statistics. The JNLPBA corpus contains only five classes (protein, DNA, RNA, cell line, and cell type) from the 36 classes in the GENIA corpus, but only the protein, DNA, and RNA classes were investigated in this study.

- The GENETAG corpus was used in the BioCreative II challenge [35]. This corpus is derived from the ‘MedTag’ dataset. It consists of 20,000 sentences of manually annotated gene/protein names. The primary 15K sentences were used for the BioCreative 1 (Project 1A) competition in 2004, and the other, 5K sentences was used as a test dataset for the BioCreative II (Gene Mention Task) competition in 2005. The authentic 20K sentences were run through a gene/protein name tagger, and the findings modified manually to reflect a wide definition of gene/protein name situations to a specificity constraint, a rule that requires the tagged entities to consult particular entities. Every sentence in GENETAG is turned into annotated sentences with suitable alternatives for the gene/protein names it contains, allowing for partial matching with semantic constraints.

3.2 Feature Engineering and Word Representation

Data representation is the most remarkable factor that determines supervised machine learning success in any domain and ensures that the best performance is achieved. In this section, we describe several data representation planners used in our methodology. However, up until now, no reports have used unsupervised word representation methods as features with supervised machine learning methods such as CRFs and ME for Bio-NER. Therefore, this research set out to clarify how the supervised machine learning based on Bio-NER systems could be used based on the distributed prototypical representation.

Traditional features consist of the morphological, orthographic, syntactic, and semantic facts of phrases. These features heavily account for the hassle of fact sparsity and fluctuations among entity types. Therefore, developing these features will be costly. Furthermore, these capabilities are complex and hand designed and often optimized for a selected gold standard corpus, which makes extrapolation of first-rate measures difficult.

Word2Vec is a linguistic model based on a neural network that learns the embedding of each word in a corpus. Word2Vec is a recently developed technique for building a neural network that maps words to real-number vectors, assisting words with similar meanings to map to similar vectors.

Extended prototypical representation is fully unsupervised and can automatically capture semantic and syntactic information. The obstacle inherent in traditional features could be overcome by the amalgamation of unsupervised word representation, which can derive the information automatically.

Word embeddings is a mathematical description of the word in vector form. Each position of a vector corresponds to a feature with some semantic or grammatical inference, which leads to the term word feature. For better consideration of contextual information, word embedding is used to calculate the semantic similarity between words. Word vector is a mathematical representation of one word. This field has gained much attention recently. Word embeddings contain the latent syntactic/semantic information of a word. The main objective of using word embeddings is to obtain meaningful information for the trained model. Since it is necessary to explore more evidential word representation to recognize biomedical entities against traditional features, there is a need to employ the distributed prototype word representation techniques with a powerful machine learning technique.

3.2.1 Feature Extraction

The main goal of feature extraction is to represent each word in the dataset as a vector of values for morphological, orthographical, contextual features, N-gram features, POS features, and Shallow parsing features, so that the recognition performance of Bio-named entities can improve while reducing the processing overhead. As a matter of fact, if selected features are independent and correlated with classes, the recognition performance will be improved. Otherwise, the recognition performance will drop. Therefore, in Bio-NER, it is crucial to recognize the method for selecting features and integrating features. Traditional features involve the morphological, syntactic, and semantic information of words. These features differ between entity types, which makes them costly to develop. Furthermore, these features are manually constructed and complicated and often optimized for a specific gold standard corpus.

- Morphological features (word characters) reflect common structures and/or sub-sequences of characters among several entity names, thus identifying similarities between distinct tokens. To fulfill this goal, the first and last word characters are the commonly considered morphological features. These can be used to differentiate entity names. For instance, the last three word characters such as “ase”, “ome” and “gen” frequently occur in gene and protein names.

- Context features reflect higher-level relations between words and extracted features that can be established through windows of features, reflecting the local context of each token. The application of windows consists of adding features of preceding and succeeding tokens as features of each token. In this work, a set of context features, one word before and one dynamic context feature (output tag of previous word), are considered. The context features that are used such as the surrounding words (word before, word after current word), are very effective in identifying biomedical named tags, as they reflect higher-level relations with the tag of the current word. In addition, selecting large windows of surrounding words will increase computational complexity. This work also uses one dynamic context feature (output tag of previous word), which is the most commonly considered dynamic context feature, especially because biomedical names are compounded words.

- Orthographic features, such as initial capital, all capital, includes caps, has slash, has punctuation, and has digit are used to capture knowledge about a word formation. Orthographic features are related to the orthography of the text, such as spelling rules, capitalization, digitization, and punctuation. Such features are very effective in boundary detection [30].

- N-gram features examine the existence of expressions that characterize a specific section. Bi-gram and tri-gram can be extracted from training data.

- POS features may provide useful substantiation about the boundaries of biomedical NEs, as most biomedical NEs are descriptive and very long. Verbs and prepositions usually indicate NE boundaries, while nouns not found in the dictionary are usually good nominees for NEs. We adopted POS information of the current and/or the surrounding token(s) as the features. We applied GENIA tagger V2.0.2 to provide POS information from the biomedical domain.

- Shallow parsing features may provide meaningful substantiation about the boundaries of biomedical NEs. We used a GENIA tagger to get chunks of information.

Table 1. Features used for training classifiers

Feature set	Actual features of the feature set
Context words	One word before and one word after the current word.
Dynamic feature	Dynamic feature denotes the output tag of previous words.
Orthographic features	Orthographic features: several binary features are defined: initial capital, all capital, includes caps, has slash, has punctuation, or has digit.
Word affixes	Word prefix and postfix character successions of length up to 3.

Feature set	Actual features of the feature set
N-gram	Bi-gram and tri-gram.
POS features	The correct morpho-syntactic class of a word (noun, verb, etc.).
Shallow parsing features	Chunks (noun phrases, verb phrases, prepositional phrases, complete clauses, etc.).

3.2.2 Word2Vec

This paper adopts a linguistic model called Word2Vec. This tool belongs to the class of methods called "neural language models", and is used to learn the embedding of each input word in the corpus. Word2Vec was developed by Mikolov et al. [10]. It can convert words into a distributed vector. This tool adopts two essential model architectures based on the neural network—continuous bag-of-words (CBOW) architecture and continuous skip-gram architecture—to learn the vector representations of words. This paper adopts the continuous bag-of-words (CBOW) model to convert the words into a distributed vector.

The continuous bag-of-words (CBOW) model is a Word2Vec neural network language model used to learn the embedding of each input word in a corpus. It can convert words into a distributed vector. CBOW consists of three layers: an input layer, a projection layer (also known as a hidden layer), and an output layer. The popularity of this model has risen in recent years. The CBOW model learns to foretell the target word from the words in the context window surrounding the target word. The vector portrayals of the words in the setting window are arrived at the midpoint of the process to predict the objective word. Subsequently, the CBOW treats each word in the setting window uniformly as far as its commitment to predicting the objective word.

Mikolov et al. [10] used both the n words before and after the target word w_t to foretell it, as described in Figure 2. They termed this model the continuous bag-of-words (CBOW), as it utilizes continuous representations, which have no order of importance. In this paper, we formalize the CBOW architecture and introduce the notations utilized throughout this paper.

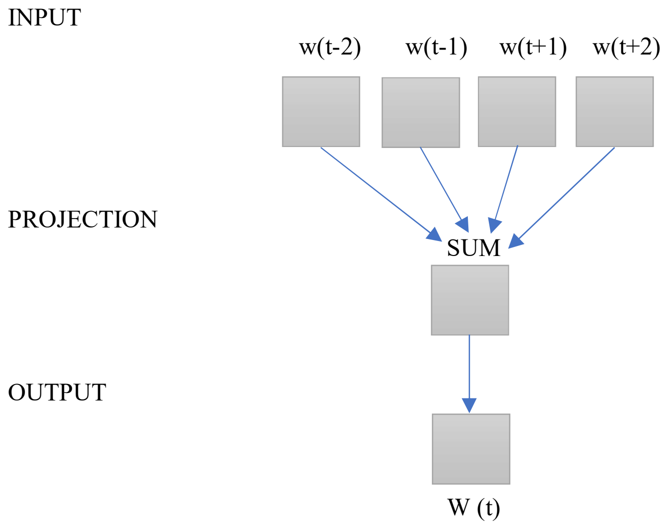

Fig. 2. Continuous bag-of-words (CBOW)

The objective function of CBOW in turn is only a little different than the language model, as shown by Equation (1):

$$J_{\emptyset} = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t | w_{t-n}, \dots, w_{t-1}, w_{t+1}, \dots\dots, w_{t+n}) \qquad (1)$$

Instead of including n previous words into the model, the model extradites a window of n words around the target word w_t at each time stride t. Then, we used the cosine similarities between the sentence representations for the sentence while other sentences were calculated in the penultimate layer.

3.2.3 Extended Distributed Prototypical Method

The performance of Bio-NER systems is always finite depending on the construction of complicated manually constructed features, which are derived from various linguistic analyses [31]. The distributed representative features was also proposed by Guo et al. [36]. As per Guo et al. [36], this work also used an association measure to extract prototypical words for each class. Unlike Guo et al. [36], however, this work introduced the extended pointwise mutual information (PMI) to overcome the non-symmetrical co-occurrence problem of PMI. In this study, an extended distributed prototypical word embedding was proposed, which can be described as follows:

Prototypical word representation method in this study selects prototypical words for each class. Similar to Guo et al. [36], the prototypical feature words were selected using the normalized pointwise mutual information (PMI) between the word and its classes using Equations (2) and (3).

$$nPMI(class, word) = \frac{PMI(class,word)}{lin\, p\ (class,word)} \qquad (2)$$

$$PMI(class, word) = lin \frac{p(class,word)}{p\ (class)\ .p(word)} \qquad (3)$$

After sets of prototypical words were constructed for each class, the co-occurrence vector was generated for each word w including representative words. Each word w_i from the dataset including prototypical words was represented using a KNN prototypical word representation. To attack the unsymmetrical co-occurrence problem of PMI, EPMI was proposed and defined to extract prototypical words based on extended mutual information (EMI) and PMI^2 [37]. 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 dataset was determined using EPMI, which is derived from extended mutual information EMI and PMI^2, as per Equations (4) and (5):

$$\mathrm{EMI}(\mathrm{w_i}, \mathrm{w_j}) = \log_2 \frac{\mathrm{P(w_i,w_j)}}{\mathrm{(P(w_i){-}P(w_i,w_j))(P(w_j){-}P(w_i,w_j))}} \qquad (4)$$

$$PMI^2 = \log_2 \frac{\mathrm{P(w_i,w_j)^2}}{\mathrm{(P(w_i))(P(w_j))}} \qquad (5)$$

Based on Zhang et al. [38] and Equations (4) and (5), the new EPMI can be written as Equation (6):

$$\begin{aligned}\mathrm{EPMI}\left(\mathrm{w_i},\mathrm{w_j}\right) &= \log_2 \frac{\mathrm{P}\left(\mathrm{w_i},\mathrm{w_j}\right)^2}{\left(\mathrm{P}(\mathrm{w_i}) - \mathrm{P}\left(\mathrm{w_i},\mathrm{w_j}\right)\right)\left(\mathrm{P}\left(\mathrm{w_j}\right) - \mathrm{P}\left(\mathrm{w_i},\mathrm{w_j}\right)\right)} \qquad (6)\\ &= \log_2 \frac{\mathrm{P}\left(\mathrm{w_i},\mathrm{w_j}\right)^2}{\left(\mathrm{P}(\mathrm{w_i})\right)\left(\mathrm{P}\left(\mathrm{w_j}\right)\right)} + \log_2 \frac{1}{\left(1 - \mathrm{P}(\mathrm{w_i})\right)\left(1 - \mathrm{P}\left(\mathrm{w_j}\right)\right)}\\ &= 2\log_2 \frac{\mathrm{P}\left(\mathrm{w_i},\mathrm{w_j}\right)}{\left(\mathrm{P}(\mathrm{w_i})\right)\left(\mathrm{P}\left(\mathrm{w_j}\right)\right)}\\ &+ \log_2 \frac{1}{\left(1 - \mathrm{P}(\mathrm{w_i})\right)\left(1 - \mathrm{P}\left(\mathrm{w_j}\right)\right)}\\ &= 2PMI + \log_2 \frac{1}{\left(1 - \mathrm{P}(\mathrm{w_i})\right)\left(1 - \mathrm{P}\left(\mathrm{w_j}\right)\right)} > EMI\left(\mathrm{w_i},\mathrm{w_j}\right) > PMI\left(\mathrm{w_i},\mathrm{w_j}\right)\end{aligned}$$

Equation (6) indicates that EPMI will amplify the association of likely co-occurring words. This kind of amplification in association is beneficial for co-occurring word calculation because it will differentiate the likely co-occurring words from those candidates, which are not real co-occurring words.

Second, a prototypical word representation was constructed for each word. Each word is represented by a vector of *n* dimensions, where *n* represents the size of selected prototypical words. For each word *w* in the training/test, the cosine similarity between *w* and all the selected prototypical words was determined using the associated embedding co-occurrence vectors. If the cosine similarity of the co-occurrence vector of *w* and the co-occurrence vector of a prototypical word were above the predefined threshold (0.50), the prototypical word will be assigned as a feature. Given the co-occurrence vector of *w (cv(w))* and the co-occurrence vector of a prototypical word *pw (cv(pw)),* the cosine similarity is defined by Equation (7):

$$sim_{cosine}(\boldsymbol{w},\boldsymbol{pw})^2 = \frac{\sum_{i=1}^{|cv|}(cv_i(w)*cv_i(pw))}{\sum_{i=1}^{|cv|}(cv_i(w))^2+\sum_{i=1}^{|cv|}(cv_i(pw))^2} \qquad (7)$$

3.3 Classification Methods

In this study, two classifier methods—CRFs and ME—were applied due to their simplicity, leverage, robustness, and reliability. A brief description of these methods is provided below:

3.1.1.1 Conditional Random Fields (CRFs) is a type of discriminative probabilistic model framework used for labeling and segmenting sequential data such as natural language text, which were firstly introduced by Lafferty and Mccallum [39]. This approach achieved experimental success in many NLP problems. CRFs are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. Conditionally-trained CRFs can easily include a large number of arbitrary non-independent features. The meaningful power of the model increases by adding new features that are conjunctions to the original features. When applying CRFs to the named entity recognition problem, a noting sequence, which is the token sequence of a sentence or document of text and a state sequence, which is its corresponding label sequence, are generated.

The conditional probability of a state sequence y given an observation sequence x is given by Equation (8):

$$\mathrm{P(y|x)} = \frac{1}{\mathrm{Z(x)}}\exp\left(\textstyle\sum_\mathrm{j} \lambda_\mathrm{j} \mathrm{F_j(y,x)}\right), \qquad (8)$$

Where, λ_j is the parameter of a corresponding feature F_j, Z(x) is a normalizing factor, and F_j can be written as:

$$F_j(y,x) = \sum_{i=0}^{n} f_i\,(y_{i-1}, y_i, x, i),$$

Where, i means the relative position in the sequence, and y_{i-1} and y_i denote the label at position i-1 and i, respectively.

3.1.1.2 Maximum Entropy (ME) is a statistical modeling technique utilized to assess the conditional probability of a target label based on given information. The ME framework is a powerful learning model, which has been successfully employed in many natural language processing tasks. The ME standard looks for the conveyance that amplifies the entropy of the dissemination subject to its known limitations. The upside of ME is that it is tough and measurably productive, while still taking into consideration the simple representation and joining of various highlights.

This technique computes the probability p(y|x), where y denotes all possible outcomes of the space, and x denotes all possible features of the space. The computation of p(y|x) depends on a set of features in x; the features are helpful for making predictions about the outcomes, y. Given a set of features and a training set, the ME estimation process produces a model, in which every feature f_i has a weight λ_i. The ME model can be represented by Equation (9):

$$\mathrm{p(y|x)} = \frac{1}{\mathrm{Z(x)}}\exp(\textstyle\sum_i \lambda_i \mathrm{f_i(x,y)}), \qquad (9)$$

$$Z(x) = \sum_{y} exp\left(\sum_{i} \lambda_i\ f_i\,(x,y)\right)$$

The probability is derived by multiplying the weights of the active features (i.e., those f_i (y,x) = 1).

4.0 EVALUATION

In this section, we evaluate the performance of the proposed Bio-NER, where we used the same evaluation metrics applied in the CoNLL-02, CoNLL-03, and JNLPBA-04 challenge tasks, which are precision, recall, and the weighted mean F-score measure.

In this stage, we determine the representation method with the most influence on classification performance, the classifier that is most convenient for Bio-NER, and could reduce the training data size to increase performance. The standard evaluation metrics are described by Equation (10):

1. Recall (R) is the ratio of number of NE words retrieved to the total number of NE words actually present in the file.

$$\mathrm{Recall} = \frac{\text{\#of correctly classified entities}}{\text{\#of entities in the corpus}} = \frac{\mathrm{tp}}{\mathrm{tp+fn}} \qquad (10)$$

2. Precision (P) is the ratio of number of correctly retrieved NE words to the total number of NE words retrieved by the system, as per Equation (11).

$$\mathrm{Precision} = \frac{\text{\#of correctly classified entities}}{\text{\#of entities found by algorithm}} = \frac{\mathrm{tp}}{\mathrm{tp+fp}} \qquad (11)$$

Where, *tp* and *fp* refer to the number of true positives and false positives reported and *tn* and *fn* denote the number of true negatives and false negatives reported, as shown in Table 2.

Table 2. Classification of document

TP	A document being categorized effectively as referring to a class.

FP	Determined as a document that is said to be related to the category incorrectly.
FN	Determined as a document that is not marked as related to a category but should be.
TN	Documents that should not be marked as being in a particular category and are not.

3. F-score measure (F) was introduced by Rijsbergen [40]. In order to obtain a better picture of the performance of the classifier, we applied the F-measure, which is computed via the weighted harmonic mean of precision and recall. This was done by combining them such that Equation (12) is derived:

$$F_{\beta=1} = \frac{(\beta^2+1) * \text{precision} * \text{recall}}{\beta^2(\text{precision} + \text{recall})} \qquad (12)$$

Where, β is a positive parameter, which represents the relative weight of recall to precision. If precision is considered to be more important that recall, then the value of β converges to zero. On the other hand, if recall is more important than precision, then β converges to infinity. Usually β is set to 1, because in this way, equal importance is given to each precision and recall.

5.0 RESULTS AND DISCUSSION

The objective of this paper is to inspect and compare the performance of traditional feature methods, the CBOW model, and a new word representation method used with two machine learning models and two datasets of Bio-NER. The findings indicate the representation method with the best recognition performance. To this end, several experiments were conducted to evaluate the different methods on different training data sizes. This study applied the same experimental settings for all word representation methods to study the effect of the size of training data and testing data on the performance of each. This means that the word representation method that was most convenient was determined when there was only small training data available. This was done to represent the poor resource problem, which is common in many Natural Language Processing and biomedical data mining tasks.

First, this paper evaluated the performance of the traditional feature method, the CBOW model, and the new word representation method with a ME classifier for different training data sizes based on the JNLPBA corpus. According to Figure 3, the extended prototypical method yielded the best macro-averaging F-measure score, and thus the best performance, compared to the traditional feature and the CBOW model. The ME classifier performed best and achieved an F-measure score of 0.76 when 90% of the JNLPBA corpus was used as training data with the extended prototypical method.

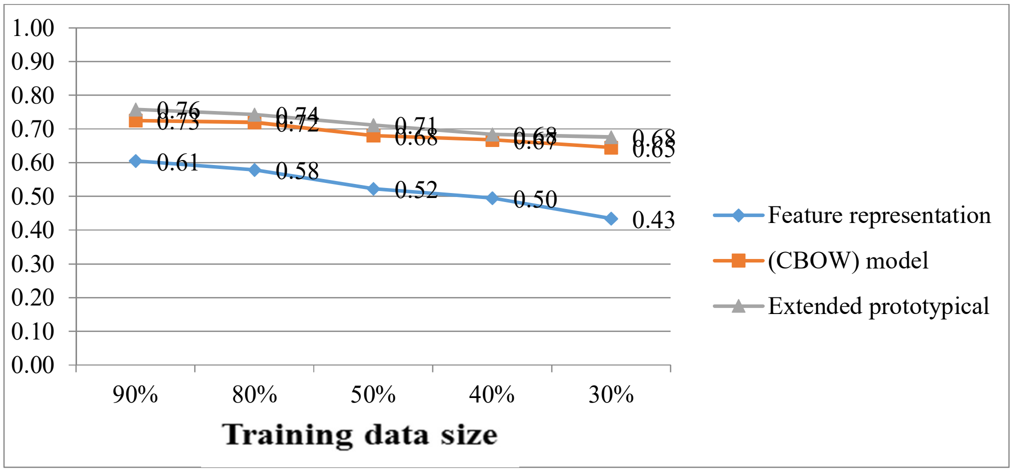

Fig. 3. Macro-averaging F-measure values for the ME classifier with the three representation methods using the JNLPBA corpus and different training data sizes.

Second, this paper evaluated the performance of the traditional feature method, the CBOW model, and the new word representation method with a ME classifier for different training data sizes using a GENETAG corpus. According to Figure 4, the extended prototypical method produced the best macro-averaging F-measure, and thus the best performance compared to the traditional feature, and the CBOW model. The best performance of the ME classifier was achieved (F-measure score = 0.78) when 90% of the GENETAG corpus was used as training data with the extended prototypical method.

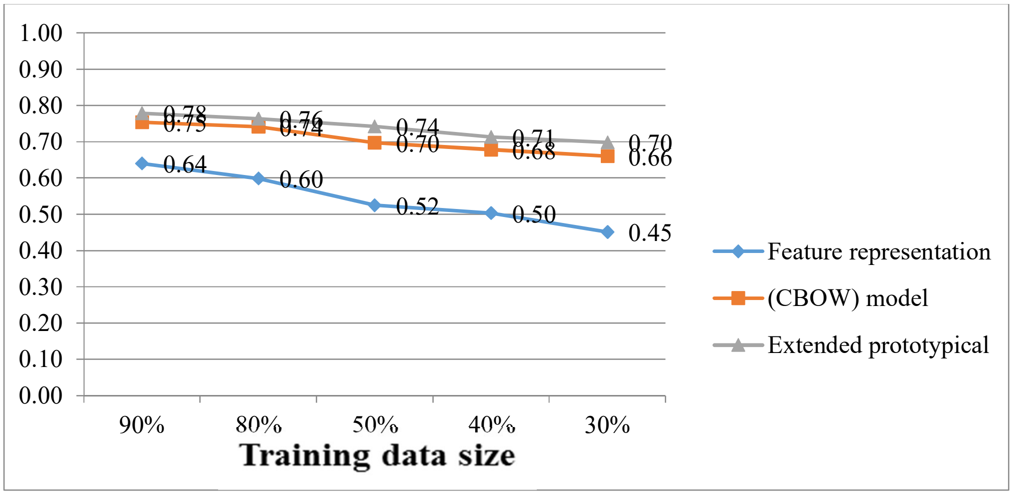

Fig. 4. Macro-averaging F-measure values for the ME classifier with the three representation methods using the GENETAG corpus and different training data sizes.

Third, this paper evaluated the performance of the traditional feature method, the CBOW model, and the new word representation methods with the CRF classifier for different training data sizes and using a JNLPBA corpus. According to Figure 5, the extended prototypical method produced the best macro-averaging F-measure score, and hence the best performance, compared to the traditional feature and the (CBOW) model. The best performance of the CRF classifier was achieved (F-measure = 0.79) when 90% of the JNLPBA corpus was used as the training data with the extended prototypical method.

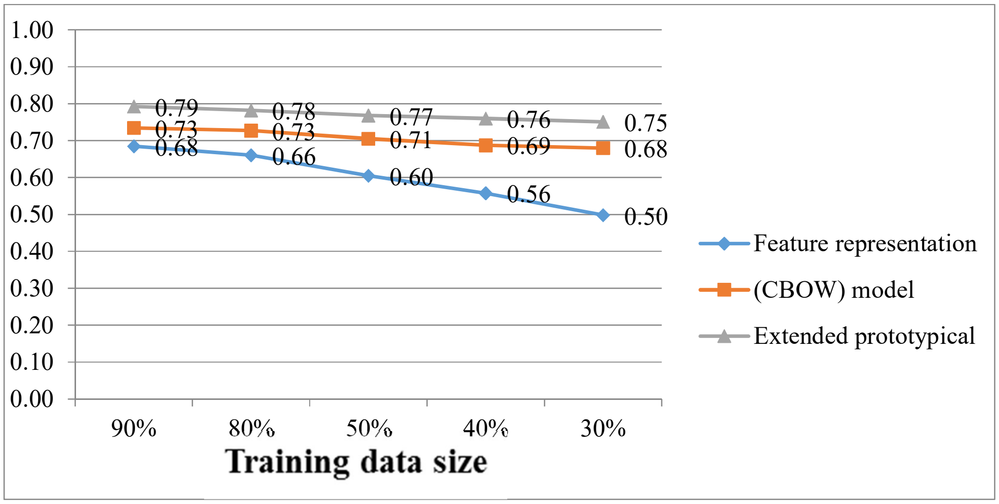

Fig. 5. Macro-averaging F-measure values for the CRF classifier with the three representation methods using the JNLPBA corpus and different training data sizes.

Fourth, this paper evaluated the performance of the traditional feature method, the CBOW model, and the new word representation method with the CRF classifier for different training data sizes and using a GENETAG corpus. According to Figure 6, the extended prototypical method produced the best macro-averaging F-measure score, and thus had the best performance compared to to the traditional feature and the CBOW model. The best performance of the CRF classifier was achieved (F-measure score = 0.85) when 90% of the GENETAG corpus was used as training data with the extended prototypical method.

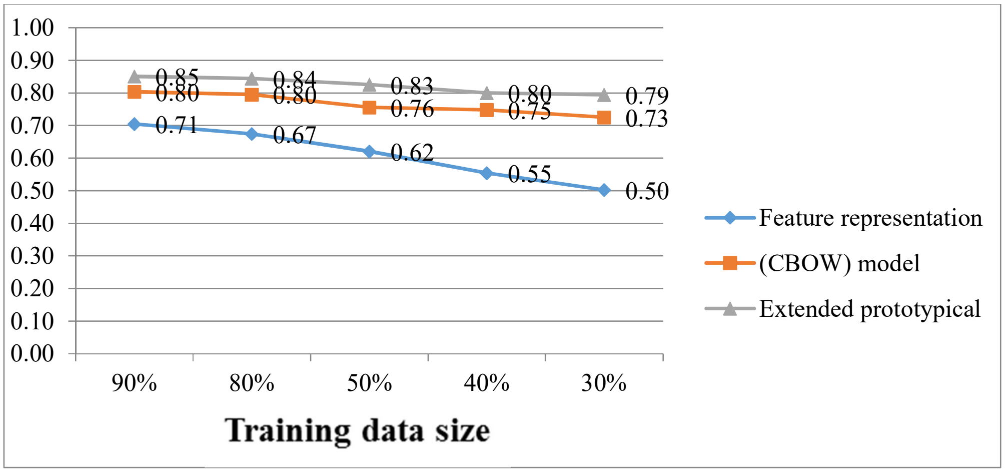

Fig. 6. Macro-averaging F-measure values for the CRF classifier with the three representation methods using the GENETAG corpus and different training data sizes.

From the comparison of the classifier performances (Figures 3, 4, 5, and 6), the CRF algorithm was found to outperform the ME algorithm for both the JNLPBA corpus and the GENETAG corpus. Furthermore, the best accuracies were achieved when the data representation was made via the extended prototypical method. According to Figures 3, 4, 5, and 6, the best result was achieved when the training size was large for both the JNLPBA corpus and GENETAG corpus and using the CRF classifier and the extended distributed representative word embedding method, yielding (0.79) & (0.85) F-measure scores, respectively. Meanwhile, the worst results were produced using the traditional feature representation with (0.68) & (0.71) F-measure scores, respectively. Comparing the behaviors of the extended distributed representative word embedding method with different training sizes for both the JNLPBA corpus and GENETAG corpus and the CRF and ME classifiers, the results show that the recognition performances increased when the size of training data was increased and the recognition performances dropped slightly with small training data sizes. The results show that the proposed extended prototypical method produced superior results over other representation methods for all training sizes and for both datasets. The main reason is that the extended prototypical method presented a distributed representation over word classes, so it was able to capture semantic and syntactic information. The syntactic information was captured through bringing together words that occur in the same syntactic structure. Each value in the vector represented the word's semantic relation with a prototypical word. Furthermore, the experimental results showed that the proposed extended prototypical method was convenient for Bio-NER with both classifiers and different datasets, even with only a very small portion of the training data size. Also, comparing the results between the JNLPBA corpus and GENETAG corpus for both classifiers in terms of macro-averaging F-measure, the F-score using the GENETAG corpus outperformed the F-score using the JNLPBA corpus. We believe that the variance in performance between the two is due to our evaluation system's lack of substitutional-tagging rules. The extended prototypical method has two main contributions. First, the extended prototypical method proved to be convenient when only small training data is available. However, this work showed that new word embedding techniques could work well with CRFs, and they are not committed to particular machine learning methods. This paper also presents a comparative study between the traditional, CBOW model, and new word representation method for Bio-NER tasks using different datasets.

6.0 CONCLUSION

In this paper, we presented a survey of the literature on Bio-NER tasks. We also presented an extensive comparative study of three data representation methods, namely the feature representation method, the continuous bag-of-words (CBOW) model, and the extended distributed prototypical method with two machine learning methods i.e. ME and CRF classifiers to execute biomedical NER tasks. The main contribution of this study is the investigation and comparison of the performance of different representation methods using different datasets and their performance with machine learning methods in terms of macro-F measure scores. Our evaluation of the JNLPBA and GENETAG corpora showed that the CRF is the best classifier of all the feature selection algorithms. Also, the extended distributed prototypical method is a superior method and showed better performance than the traditional feature representation method and CBOW model for all dataset sizes and different datasets. Our future efforts will be targeted at evaluating the extended distributed prototypical technique with other classification tasks including general NER. We plan to conduct a comparative study between many probabilistic models as well as experiment using a combination of different models. In addition, future work may extend the proposed methods and evaluate them with other advanced machine learning models such as deep learning.

REFERENCES

[1] Campos, D., S. Matos, I. Lewin, J. L. Oliveira & D. Rebholz-Schuhmann. Harmonization of gene/protein annotations: towards a gold standard MEDLINE. *Bioinformatics* vol. 28, no. 9, 2012, pp.1253-1261.

[2] Campos, D., S. Matos & J. Oliveira. Current methodologies for biomedical named entity recognition. *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data,* vol. 2013, pp.839-868.

[3] B. Tang, H. Cao, X. Wang, Q. Chen, and H. Xu, "Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks," vol. 2014, 2014.

[4] Yang, L. & Y. Zhou. "Exploring feature sets for two-phase biomedical named entity recognition using semi-CRFs". *Knowledge and information systems,* vol. 40, no. 2, 2014, pp.439-453.

[5] Li, K., W. Ai, Z. Tang, F. Zhang, L. Jiang, K. Li & K. Hwang. "Hadoop recognition of biomedical named entity using conditional random fields". *IEEE Transactions on Parallel and Distributed Systems* vol. 26, no. 11, 2015, pp. 3040-3051.

[6] Li, L., L. Jin, Z. Jiang, D. Song & D. Huang. "Biomedical named entity recognition based on extended recurrent neural networks". *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. 2015, pp. 649-652.

[7] Wang, X., C. Yang & R. Guan. "A comparative study for biomedical named entity recognition". International Journal of Machine Learning and Cybernetics, 2015, pp.1-10.

[8] L.Yao, H. Liu, Y. Liu, X. Li, and M. W. Anwar, "Biomedical Named Entity Recognition based on Deep Neutral Network," *Int. J. Hybrid Inf. Technol.*, vol. 8, no. 8, 2017, pp. 279–288.

[9] Aaron M. Cohen, William R. Hersh; "A survey of current work in biomedical text mining", *Briefings in Bioinformatics*, vol. 6, no 1, 1 March 2005, pp. 57–71.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "5021-Distributed-Representations-of-Words-and-Phrases-and-Their-Compositionality," 2013, pp. 1–9.

[11] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts : a machine learning approach," vol. 20, no. 7, 2004, pp. 1178–1190.

[12] C. H. Wei, H. Y. Kao, and Z. Lu, "GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains," *Biomed Res. Int.*, vol. 2015, 2015.

[13] S. Lee *et al.*, "BEST: Next-generation biomedical entity search tool for knowledge discovery from biomedical literature," *PLoS One*, vol. 11, no. 10, 2016, pp. 1–16.

[14] R. Alves, F. Solsona, M. Va, A. Valencia, and A. Usie, "CheNER : chemical named entity recognizer," vol. 30, no. 7, 2014, pp. 1039–1040.

[15] R. Leaman *et al.*, “tmChem: a high performance approach for chemical named entity recognition and normalization,” *J. Cheminform.*, vol. 7, no. Suppl 1, 2015, p. S3.

[16] Y. Zhang, J. Xu, H. Chen, J. Wang, and Y. Wu, “Original article Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning,” no. May 2018, 2016, pp. 1–10.

[17] S. Pyysalo and S. Ananiadou, “Anatomical entity mention recognition at literature scale,” vol. 30, no. 6, 2014, pp. 868–875.

[18] J. Weston and M. Karlen, “Natural Language Processing (Almost) from Scratch,” vol. 12, 2011, pp. 2493–2537.

[19] T. Tsai, C. Wu, and W. Hsu, “Using Maximum Entropy to Extract Biomedical Named Entities without Dictionaries,” 2005, pp. 268–273.

[20] O. Levy, “Neural Word Embedding as Implicit Matrix Factorization,” 2014, pp. 1–9.

[21] Ma, X. & Hovy, H., “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF,” 2016, *arXiv preprint arXiv: 1603.01354*.

[22] S. Subramanian, K. Kawakami, and C. Dyer, “Neural Architectures for Named Entity Recognition.”2016, pp. 1-11

[23] F. X. Chang, J. Guo, W. R. Xu, and S. R. Chung, “Application of Word Embeddings in Biomedical Named Entity Recognition Tasks,” vol. 13, no. 5, 2015, pp. 321–327.

[24] F. Li, M. Zhang, G. Fu, and D. Ji, “Open Access A neural joint model for entity and relation extraction from biomedical text,” , 2017, pp. 1–11.

[25] I. Segura-bedmar and P. Mart, “Exploring Word Embedding for Drug Name Recognition,” no. September, 2015, pp. 64–72.

[26] B. Tang *et al.*, “A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature,” vol. 7, no. Suppl 1, 2015, pp. 4–9.

[27] B. Bhasuran, G. Murugesan, S. Abdulkadhar, and J. Natarajan, “Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases,” *J. Biomed. Inform.*, vol. 64, 2016, pp. 1–9.

[28] H. Wang, T. Zhao, H. Tan & S. Zhang, “International Journal of Computer Science and Applications, Vol. 5, No. 2, pp 1- 11 © Technomathematics Research Foundation,” vol. 5, no. 2, 2008, pp. 1–11.

[29] F. Zhu and B. Shen, “Combined SVM-CRFs for Biological Named Entity Recognition with Maximal Bidirectional Squeezing,” vol. 7, no. 6, 2012, pp. 1–9.

[30] G. Murugesan, S. Abdulkadhar, B. Bhasuran, and J. Natarajan, “BCC-NER : bidirectional , contextual clues named entity tagger for gene / protein mention recognition,” *Eurasip J. Bioinforma. Syst. Biol.*, no. 1, 2017, pp.7-22.

[31] S. Liu, B. Tang, Q. Chen, and X. Wang, “Effects of Semantic Features on Machine Learning-Based Drug Name Recognition Systems: Word Embeddings vs. Manually Constructed Dictionaries,” *Information*, vol. 6, no. 4, Dec. 2015, pp. 848–865.

[32] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, “Deep learning with word embeddings improves biomedical named entity recognition,” *Bioinformatics*, vol. 33, no. 14, 2017, pp. i37–i48.

[33] A. Ekbal, S. Saha, & U. K. Sikdar, “Biomedical Named Entity extraction: Some issues of corpus compatibilities,”. *SpringerPlus*, *2*(1), 2013, pp.1–12.

[34] J. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, “Introduction to the Bio-Entity Recognition Task at JNLPBA,” 2004, pp. 70–75.

[35] L. Smith *et al.*, “Open Access Overview of BioCreative II gene mention recognition,” vol. 9, no. Suppl 2, 2008, pp. 1–19.

[36] J. Guo, W. Che, H. Wang, and T. Liu, “Revisiting Embedding Features for Simple Semi-supervised Learning,” *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, no. 2005, 2014, pp. 110–120.

[37] M. T. Abd and M. Mohd, “Extended Distributed Prototypical for Biomedical Named Entity Recognition,” vol. 6, no. 2, pp. 1–11, 2017.

[38] W. Zhang, T. Yoshida, X. Tang & T.-B. Ho. "Improving effectiveness of mutual information for substantival multiword expression extraction". *Expert Systems with Applications* vol. 36, no.8: 2012, pp. 10919-10930.

[39] J. Lafferty and A. Mccallum, "Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data Conditional Random Fields : Probabilistic Models for Segmenting and," vol. 2001, no. June, 2001, pp. 282–289.

[40] J. van Rijsbergen, "Information Retrieval, 2nd edition". Dept. of Computer Science, University of Glasgow, 1979.