Nilanjana Das
Rakesh Dutta
Uttam Kumar Mondal
Mukta Majumder
Jyotsna Kumar Mandal

ENHANCED CLUSTER MERGING AND DEEP LEARNING TECHNIQUES FOR ENTITY NAME IDENTIFICATION FROM BIOMEDICAL CORPUS

Abstract *For mining biomedical information identifying names is the prime task. Complex and uncertain naming styles of biomedical entities are the major setbacks here. Thus, state-of-the-art accuracy of biomedical name identification is reasonably inferior compared to general domain. This study includes Machine Learning and Deep Learning techniques to recognize names from biomedical corpus. In supervised classification, a classifier is built by finding required statistics from training corpus. Accordingly, performance of the system is primarily dependent on quantity and quality of training corpus. But manually preparing a large training dataset with enriched feature samples is laborious and time-taking. Therefore, various techniques were adopted in the literature to make effective use of raw corpora. We have incorporated a novel Cluster Merging technique and Attention Mechanism with BERT embedding for boosting Machine Learning and Deep Learning classifiers respectively. The suggested results outpour that profound techniques are competent and delineate signifying improvement over surviving methods.*

Keywords biomedical named entity recognition, conditional random field, support vector machine, cluster merging, BERT, bidirectional GRU, attention mechanism

Citation Computer Science 26(1) 2025: 49–75

1. Introduction

The task of information extraction (IE) is a critical aspect of natural language processing (NLP) that involves automatically retrieving organized data from unstructured text. Two fundamental tasks involved in information extraction (IE) are named entity recognition (NER) and relation extraction (RE). Further relation extraction between two or more entities necessitate identification of the entity names as prerequisite. A named entity is a succession of words (n-grams) that represents a name of a specific type. Named entity recognition (NER) involves recognizing and categorizing entity names within unstructured text into predetermined classes, which may include person, organization, location, protein, DNA, cell-type, and others. During the past two decades, the evolution of NER systems in multiple languages and fields has been a widely explored research topic. A Significant amount of investigations have been conducted for the development of NER systems in biomedical domain too. However, complex and uncertain naming styles of biomedical entities (such as T-cell, IL-8, mRNA, human immunodeficiency virus (HIV-1) etc.) are the major challenges for identifying biomedical names. These entity names are often long and include hyphen, numeric value, uneven capitalization and uncommon words; these make the classification and boundary identification of biomedical named entities (NEs) rather intricate. And the general NER system outperforms the biomedical NER system in terms of accuracy [69]. Therefore, researchers are still devoting their effort on improving the accuracy of biomedical NER system.

NER approaches can be divided into two primary classes: rule-based approaches, and supervised approaches that rely on Machine Learning or Deep Learning techniques. Rule-based approach uses handcrafted rules for extracting names. This approach yields good results if the rule set has high coverage and can detect complex names that might be difficult with supervised approach. However, rule-based approach is language and domain dependent and not portable. On the other hand, in supervised approach, a classifier is trained. The Machine Learning (ML) algorithm employs annotated samples to train a statistical classifier [22,67]. Thus, the system's effectiveness is majorly dependent on the quality and quantity of the training corpus. Generating a vast quantities of labelled training data with rich feature samples is a challenging and time-taking task. However, there is a huge collection of external raw corpora available that contain valuable information and can be used to augment the annotated training data [37]. To make use of this raw data, effective processing is necessary so that only the relevant parts of the data have a significant influence on the NER results. At this point the significance of word clustering becomes relevant. It is a method of grouping similar words into cluster. This is an efficient method to improve upon the accuracy of labelling and sequencing task like identifying named entity and parts-of-speech tagging etc. Deep learning has appeared as a successful technique, achieving state-of-the-art performance in numerous applications [30], including speech recognition [18], image segmentation [31], image classification [17], and NER [28,35]. Deep learning techniques commonly require a substantial quantity

of labelled data for supervised learning, as well as longer training times and greater computing resources compared to traditional machine learning methods. However, despite these requirements, they have demonstrated highly effective results [6].

This paper presents a biomedical Named Entity Recognition (NER) technique that uses several machine learning classifiers and an efficient cluster merging technique to achieve higher accuracy. Support Vector Machine (SVM) and Conditional Random Field (CRF), two widely used ML classifiers are incorporated for GENIA corpus version 3.02 (size 492K words) collected from JNLPBA-2004. The accuracy of a supervised ML algorithm radically relies on the affluence of feature values. With an enhanced combination of candidate features our CRF based baseline NER system gets the highest F-measure of 64.40 with Precision 65.27% and Recall 63.56% and baseline SVM classifier achieves the maximum F-measure of 64.30 along with Precision 62.72% and Recall 65.97%.

Word Clustering has been well anticipated in recent NLP tasks and achieved a greater success; it involves searching for a specific structure or pattern within a dataset that has not been processed or analyzed [41]. So, it is decided to use clustering technique in this Bio-medical NER task. It has been found in the literature that different clustering techniques are available. Farley and Raftery classified clustering into two major categories: partitioning and hierarchical [12]. Han and Kamber suggested additional three types of clustering: model-based, grid-based, and density-based approaches [16]. The explanation of having various clustering techniques is that the perception of "cluster" is not exactly defined [10]. To choose a particular clustering algorithm for any work depends on several parameters like: type of data, size of data and purpose. In this work, both partitioning and hierarchical clustering methods have been used. In partitioning method, a large object is sub divided and grouped into several clusters with every cluster containing at least one element. It is a repetitive process where the objects may be repositioned into other groups based on their relevance. There are two major techniques exist in partitioning clustering method: K-medoids and K-means. The K-means clustering is a straightforward centroid-based technique where overall dataset is divided into 'K' number of clusters, each of which consist similar type of words. On the other side, K-medoids clustering, nearly similar to K-means algorithm; here overall dataset is partitioned into 'K' number of mutually exclusive clusters but clusters are well fitted onto a data point (cluster representative). These representatives are chosen to the closest one (data points) from the cluster's centre rather than the centre point of the clusters (which is used in K-means technique). But as the data points increase the K-means requires maximum time and the K-medoids works convincingly superior to the K-means and also better at scalability for large dataset [36, 54]. So, it is decided to use K-medoids technique from partitioning clustering method. The hierarchical clustering methods organize the overall dataset in a tree like structure. These methods construct clusters by iteratively dividing the data either in a bottom-up or top-down manner. Here brown clustering technique is adopted as a hierarchical clustering. The Brown clustering [3] is a widely used hierarchical clustering method that has found application in

various Natural Language Processing (NLP) tasks [32, 39, 45, 55, 58]. It shows better performance for the data which require special care (corpus which includes special symbols and noisy text etc.). Clusters are constructed here based on the statistical analysis of bigram words. However, in the task of named entity recognition (NER), larger contextual words also contain crucial information. To successfully acquire distant information effectively, partitioning clustering technique, specifically K-medoids has been employed. Thus two type of clusters are acquired from two different clustering techniques. Subsequently these are merged to get improve accuracy from the developed biomedical NER system. For using the clustering technique, extra 238K words have been collected from MEDLINE, bio-medical corpora and used in different word representations. After using word clustering, the modified CRF based approach archives the modest F-measure of 77.09 with Precision 75.43% and Recall 78.81% and the modified SVM based system gets the highest F-measure of 75.17 along with Precision 74.41% and Recall 75.94%.

In parallel to machine learning approaches deep learning methods also have been explored for the NER task. A number of deep learning efforts were made by researchers concentrating on NER researches. These methods leveraged neural networks to identify entities by extracting word features from a vector sequence. Deep learning models offer superior performance compared to classical machine learning and rule-based approaches, but still have challenges. Deep learning techniques are classified into two classes: Convolutional Neural Network (CNN) [26] and Recurrent Neural Network (RNN) [8]. CNN is capable of obtaining local features but not context information. Although RNN is able to capture the context information from text because of its sequential nature, recognizing entities becomes challenging due to the prevalence of non-entity words within sentences and the presence of redundant information in a significant portion of the text. To deal with the aforementioned issues, this article introduces a new model named A-BiGRU that tries to identify the names from biomedical corpus. First, the BERT model is applied to generate vectors from words, which can improve the expressiveness of context information. The next phase involves training of the BiGRU network. An attention mechanism is incorporated to overcome the information redundancy. Finally, Softmax function is used to predict the entity label associated with each word, and achieves highest F-measure of 81.20 with Precision 82.89% and Recall 79.57%. In the present study, the suggested model is evaluated on the GENIA corpus version 3.02 and assess its outcomes compared to modern state-of-the-art techniques. The salient contributions of the paper that makes a number of value additions to the literatures are highlighted below.

1. The main aim of this article is to enhance the accuracy of biomedical Named Entity Recognition (NER) task compared to the current state-of-the-art methods. The suggested systems demonstrate superior performance compared to existing approaches.
2. The article presents an innovative approach for merging two distinct classes of clusters, namely hierarchical cluster (Brown cluster) and partitioning cluster (K-medoids cluster), in an efficient manner.

3. The introduced cluster merging technique improves the effectiveness of machine learning classifiers.
4. The BERT model is employed to generate word-level vectors, thereby enhancing the capability of neural networks to extract crucial information.
5. The approach uses an Attention mechanism-based Bidirectional Gated Recurrent Unit (A-BiGRU) to address the issue of information redundancy and enable detailed interpretation of text vectors. This model is directly able to access contextual information from both preceding and succeeding parts of the text.

Outside introduction, this article has been subdivided into 6 other sections. Section 2 discusses relevant previous research. Section 3 explains the baseline NER models. Section 4 presents cluster merging based enhancement of baseline models. Section 5 introduces proposed Deep Learning based NER Model. Analysis and discussion are summarized in Section 6, while Section 7 concludes the paper.

2. Related work

The NER task can be regarded as a problem of labelling sequences. The popular methods used for developing NER systems are rule-based [7], machine learning-based [5], and deep learning-based [21]. The rule based approach relies on handcrafted rules that are created manually and requires a deep understanding of the specific domain; therefore, not easy to pursue for complex NEs [13, 14]. In this section, first, we have mentioned some of the works which used ML classifiers then some deep learning based approaches for identifying the biomedical NEs.

Machine learning classifiers have been extensively used in numerous NER systems to detect named entities in the biomedical field; some of these were Hidden Markov Models [43, 53, 68], Maximum Entropy Classifier [11, 34, 50], Support Vector Machine [22, 40, 47], Conditional Random Field [29, 52, 57] etc.

A machine learning approach leverages labelled training samples to construct a statistical classifier. In the biomedical field there are several publicly available corpora for NER task. Some of the popularly used corpora are GENIA [23], JNLPBA [24], BioCreative [65], and BioInfer [44] etc. In this biomedical NER system development we have used GENIA corpus version 3.02 collected from JNLPBA-2004[1].

In the recent literature, it has been found that there is a growing of interest among the researchers to use raw external corpus to enhance ML based baseline biomedical NER system [41, 55]. Jain et al. proposed a Maximum Entropy based NER system on the Hindi Health domain corpus [20]. Their system incorporated Context Pattern-based extension and other features like POS, synonyms, gazetteers etc. to enhance the performance. Word embedding representation or word clustering is a widely used technique to incorporate external context or global information from unlabeled corpora without direct human intervention.

[1] http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004

We have adapted K-medoids technique from partitioning clustering and Brown clustering technique as a hierarchical clustering for our NER task. Several relevant studies on these two clustering techniques were made in the previous researches. Toh et al. proposed a system which used noisy user generated text for NER from Twitter corpus by using CRF [56]. Performance of their baseline system was further enhanced by brown clustering and K-means. We came across to a few studies which presented comparative studies between K-means and K-medoids algorithms [1, 54]. It has been also observed that several scholars used brown clustering strategies in their researches [2, 3, 38, 42, 58–60]. Most of these techniques were intended for processing and extracting information from large unlabelled corpus. Brown clustering was also popularly used for identifying names [32, 45, 49, 58]. It was mainly used in a number of NER systems for post processing tasks or to improve their performances [4, 41, 46, 55].

Deep learning has been widely regarded for sequence labelling problems like NER, POS tagging, sentiment analysis etc. Hammerton et al. [15] suggested a Long Short Term Memory (LSTM) for the NER system on the Reuters Corpus and European Corpus. They designed the LSTM network using SARDNET, a self-organizing map for sequences. Their model achieved the F-measure of 72.88. Huang et al. [19] applied various deep learning methods (LSTM, BiLSTM) along with combination of deep learning and CRF (LSTM-CRF and BiLSTM-CRF) to a number of NLP applications, such as POS tagging, Chunking and NER. They also showed that the efficiency of BiLSTM-CRF was better compared to other models. Katiyar et al. [21] proposed multi-layer BiLSTM to address the problem of nested NER. They executed the model on the ACE2004 and ACE2005 datasets. They obtained F-measure of 69.7 and 70.2 respectively. Yu et al. [66] introduced a method to identify NEs by using multi-layer BiLSTM with a biaffine dependency parsing model. Their system obtained the best F-measure of 80.50 on GENIA dataset.

3. Baseline NER system

The study has explored two popularly applied supervised classifiers, namely Conditional Random Field and Support Vector Machine, in order to build baseline NER model.

Conditional Random Field (CRF)

CRF is a probabilistic classifier that is capable of classifying and organizing sequential data, such as natural language texts [27]. For the NER system development, we have used the CRF++ toolkit[2], an open-source executable framework that is straightforward to use and can be customized. It is capable of handling various NLP tasks, including part-of-speech tagging, named entity recognition (NER), and relation extraction.

[2]https://taku910.github.io/crfpp/

Support Vector Machine (SVM)

SVM, proposed by Vapnik [61], is a widely employed machine learning classifier. Since it is a boolean classifier, a pairwise or one-vs-rest method is applied for multi-class classification. In this NER system development, SVM is utilized as one of the machine learning classifiers. It performs classification by establishing an N-dimensional hyperplane that effectively separates the data into two categories. The NER task consists of two main phases: training and classification, which are executed using YamCha[3]. YamCha is a widely-used, open-source, and modifiable SVM implementation that is broadly applied in numerous NLP applications, including Information Retrieval, NER, Text Mining, etc. Additionally, SVM toolkit, YamCha supports kernel functions. In this experiment the 2nd degree of polynomial kernel has been used.

3.1. Feature set

The literature has demonstrated that use of different features can boost the performance of a supervised NER system. So, it has been experimented with various types of probable features (Word Window, Affix, Parts-of-Speech (POS), Digit, Capitalization, Symbolic, Dictionary, Word Shape etc.) and chosen the best possible feature combinations depending on result.

3.2. Data set

The dataset has been taken from 'Bio-Entity Recognition Task BioNLP/JNLPBA 2004' named GENIA Corpora 3.02 version. The GENIA corpora package consist readymade training and testing data. Dimensions of the training corpus is approximately 492K words and these consist of a total number of named entities 150K. The corpora consist of only 5 NE classes' Protein, Cell-Type, Cell-Line, RNA, and DNA. We have also collected an extra 238K words from MEDLINE bio-medical corpora for using in clustering technique at post processing phase.

3.3. Performance metrics of NER system

We have evaluated the system using F-measure (F) which represents the harmonic mean of precision and recall.

$$F = \frac{\left(1+\beta^2\right)(p \cdot r)}{\beta^2 \cdot (p+r)} \tag{1}$$

Recall is percentage of total NEs which are retrieved accurately whereas precision is percentage of precise observation. β determines the relative weight between recall and precision and typically set to a value of 1.

[3]http://chasen.org/~taku/software/yamcha/

3.4. Result of baseline models

Table 1 presents the accuracies obtained in CRF and SVM based baseline NER models. The maximum F-Measure of 64.40 with 65.27% precision and 63.56% recall are obtained from CRF based model and an F-Measure of 64.30 with 62.72% precision and 65.97% recall from the SVM based model.

From the results of the two baseline models, shown in Table 1, it is apparent that, many of the names are not identified. This may be due to quantity and quality of labelled training corpus. But, manually preparing quantitatively large training data with annotated feature samples is laborious and burdensome. To enhance the performance of these models by intelligently processing the raw external corpus without any human effort, clustering techniques have been incorporated which have been discussed in the next section.

Table 1

The outcomes of the NER models based on CRF and SVM on the GENIA dataset

Combination of different features set	Word Window 7			Word Window 5			Word Window 3		
	P	R	F	P	R	F	P	R	F
CRF: AFFIX-POS-NUM-SYM-CAP-DICT-SHAP	67.75	60.57	63.96	65.27	63.56	64.40	64.27	63.89	64.08
SVM: AFFIX-POS-SHAP-NUM-SYM-CAP-DICT	63.54	64.76	64.02	62.72	65.97	64.30	63.06	64.96	64.00

Precision => P, Recall => R, F-Measure => F

4. Proposed word clustering based enhancement of the NER system

As, the baseline system does not produce satisfactory result, clustering methods have been used to improve the system. Clustering is the most common and popular technique of unsupervised learning. Here, the primary idea is to cluster data points depending on their higher levels of similarity among each other compared to the remaining data points.

It requires no human intervention to group the data samples which means that the grouping is performed on the basis of inherent similarity and difference among the data set. While working with text data set, clustering algorithm is segmenting the text documents into partitions (clusters) where each cluster consists of similar type of words like synonyms, syntactically similar words, different tense form of a particular word, etc. There are several clustering methods available as discussed earlier, here two types of word clustering techniques have been integrated to boost up the performance of the system, one is partitioning method and another is hierarchical method. There are also, several techniques exist depending on partitioning and hierarchical methods, from where K-medoids and Brown clustering have been chosen.

4.1. K-medoids clustering

To perform K-medoid clustering (see Fig. 1), the whole document have been converted into vector points, by assigning weights to each word based on the term frequency-inverse document frequency (tf-idf) calculation and then form matrix by the idea of Vector Space Model (VSM) [51].

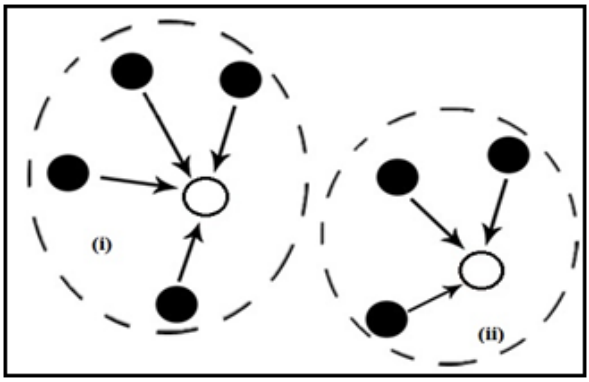

Figure 1. K-medoids cluster

Then k number of random vector points have been chosen as the initial representatives and calculate the distance from all other remaining vector points to find the closest-to-representative for initial mapping (to form a cluster). After that, we have again selected a random non-representative vector point and calculated the total cost. This random selection procedure of non-representative vector points will continue until we get the better value of the total cost from the previous selection and finally, we have swapped the best non-representative point with the initial representative point. As for example, according to Figure 1 there are two clusters where the filled points are non-representative vector points and bare points are representative vector point. The K-medoid approach followed in this study is described below.

Procedure of K-Medoid clustering technique:

1. Randomly pick 'k' samples as preliminary representative points of k clusters, according to the distance of residual samples and the representative point objects, remaining samples are allocated to nearby cluster.
2. Randomly choose a non-representative point.
3. Replace the representative point by non-representative point and check if quality of the cluster is improved. If so, then preserve the substitution; else discard it.

Repeat step 2 to 3 till no change.
The quality of clustering can be assessed using a Silhouette cost function, which calculates the average variation between objects and their representative points.

4.2. Brown clustering

The Brown clustering has been adopted as hierarchical clustering technique. This clustering technique is based on bigram statistics of sequential data such as input text and the output of this technique is a binary tree. The Figure 2 depicts an example of a binary tree as a result of the hierarchical clustering. This binary tree contains 7 words at the leaf nodes from GENIA corpus. Every word is depicted by a binary string within the rectangle, which is generated from the sub-paths string originating from the root of the tree. The size of the word representation is equal to the height of the tree. Now to calculate the number of clusters from this tree, it is necessary to perform tree pruning at a specific level. For example, if pruning is performed at level 2, then we get 4 clusters, which are [CD28, NF-Kappa], [a, the], [HIV-1] and [of, for]. Now the noticeable fact is that, among these four clusters, each cluster's words have the same prefix of bit string with the length of 2 (at the cutting level). These prefixes represent each unique cluster, so we have used these prefixes of the bit strings as training feature. For the experimental purpose, we have used prefix length of 4, 8 and 12.

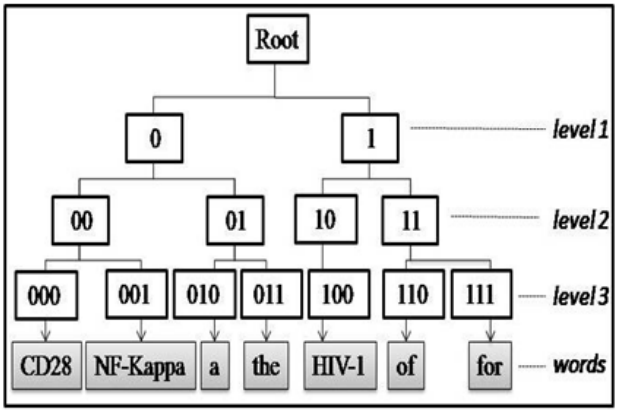

Figure 2. Brown cluster

4.3. Cluster merging

Now, two different sets of clusters have been obtained by pruning Brown cluster and K-Medoids Clustering. Next, these two techniques are combined to attain higher efficiency from the NER system. The proposed cluster merging technique is based on the perspective of similarity of tokens in the candidate clusters. The cluster merging algorithm decides whether to merge those two clusters into one unique cluster; if the test clusters are of higher similarity then merge them into one single cluster; as a result, we get a new one. The proposed cluster merging technique is described in Algorithm 1.

Algorithm 1 Cluster Merging Procedure

Require:
1. Set of K-Medoids Clusters: $KM = \{km_1, km_2, \cdots\cdots\}$
2. Set of Brown Clusters: $B = \{b_1, b_2, \cdots\cdots\}$

Ensure:
1. Merged Clusters: $MC = \{mc_1, mc_2, \cdots\cdots\}$

Begin:

for each b_i from B **do**
  T=0; ▷ Set Initial Threshold
  for each km_i from KM **do**
    $C = (b_i \cap km_j)$; ▷ Common Words in Current 2 Clusters
    A=(No. of Words in km_j + No. of Words in b_i)/2;
    /*Average No. of Words in Current 2 Clusters */
    $P = (C/A * 100)\%$; ▷ Calculated New Threshold
    if $(P > T)$ **then**
      T=P;
      Temp=j; ▷ Store Index or ID of K-Medoids cluster
    end if
  end for
  $mc_i = MERGE(b_i, km_{Temp})$; ▷ Merge these two clusters
end for

For each brown cluster b_i, it checks each and every K-medoids cluster km_j to find the highest similarity. This highest similarity is judged by the similarity threshold value (T) which is initially taken as zero. First, we have extracted the common or similar words (C) from the two current clusters (b_i, km_j). Since all clusters are not of same size, so we have taken the average (A) of the two adjacent clusters and calculate the percentage of similarity (P). For each iteration, the threshold value T will be updated, if the value of P is higher; subsequently mark the location of km_j (j) into 'Temp' for further merging operation. Finally, the two adjacent clusters (b_i, km_{Temp}) are merged into a new cluster mc_i. This process is repeated for each and every brown cluster b_i to obtained the corresponding merged clusters. This merged cluster set is employed as a feature in the experiment.

By integrating word clustering technique CRF based approach attains an F-Measure of 77.09 with 75.43% precision and 78.81% recall and SVM based approach accomplishes an F-Measure of 75.17 with 74.41% precision and 75.94% recall. Two comparative studies of the CRF and SVM based baselines have been shown with their clustering based enhancements in Figure 3 and Figure 4. These two figures justify the importance of incorporating cluster margin technique as a post processing technique.

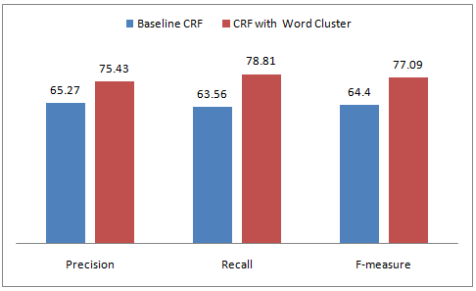

Figure 3. Comparative result of CRF based baseline model with its clustering based enhancement

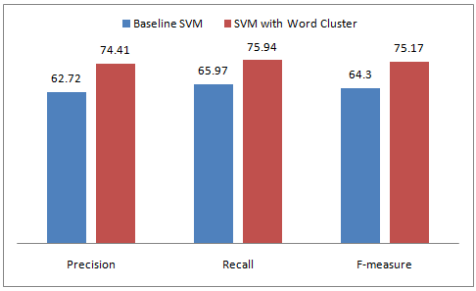

Figure 4. Comparative result of SVM based baseline model with its clustering based enhancement

5. Proposed deep learning based NER model

This section outlines the architecture of the proposed deep learning model (A-BiGRU), as depicted in Figure 5.

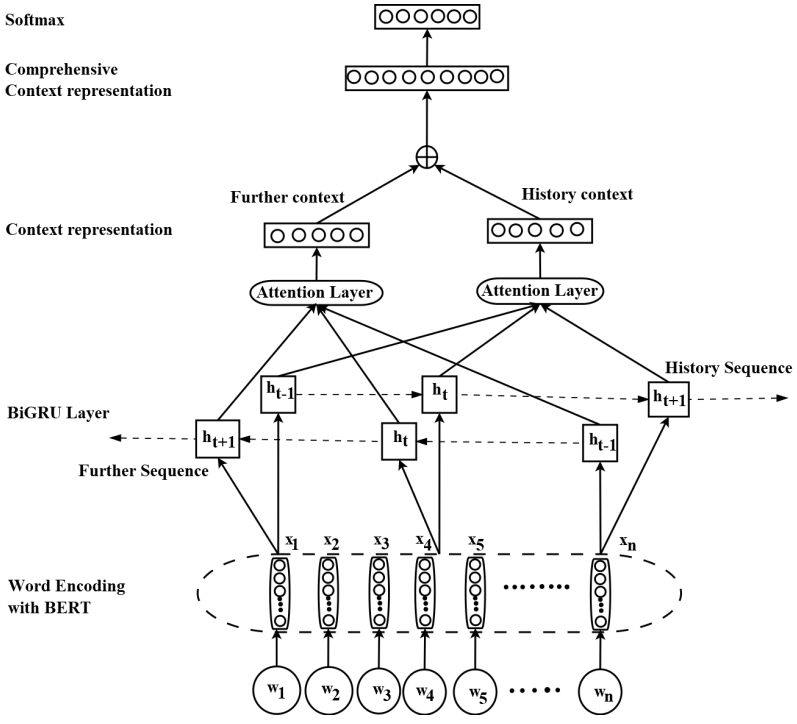

Figure 5. The architecture of the A-BiGRU

The A-BiGRU model includes three essence layers: the input feature layer, BiGRU layer, and attention layer, which are used for entity recognition. The BERT model is used in the input feature layer for representing the word into vector form. The BiGRU network mainly collects texts' context information. The attention mechanism allows the model to gain a greater comprehension of the text's detailed features, while the Softmax classifier is employed to predict the entity label associated with every word.

5.1. Embedding layer

The BERT approach employs bidirectional transformers as encoders which combine the contextual knowledge from both sides of the present word. During the process of learning word vector representations, the encoder is no longer follows a sequential approach to identify words within sentences. Instead, it masks or replaces some words at random in specific proportion of the context and generates predictions for the original words. Moreover, in the BERT model, there are training tasks that operate at the sentence level, focusing on comprehending the contextual relationship that exists within sentences. The specific technique is employed to randomly substitute certain sentences, and the encoder leverages the preceding sentence to determine whether the subsequent sentence is the original one. These two tasks cooperatively obtain vector expressions at word level and sentence level.

The BERT includes a mechanism for fine-tuning parameters. The input sentence sequence is denoted as $w = ([CLS], w_1, w_2, ..., w_n, [SEP])$, in this given context, $[CLS]$ represents the beginning of a sample sentence sequence, while $[SEP]$ is used to indicate the space that separates the sentences. They are used for training at the sentence level. Each word vector is represented by three components: a word embedding vector, a sentence embedding vector, and a position embedding vector. The word embedding is demonstrated as

$$e^w = \left(e^w_{[CLS]}, e^w_{w_1}, e^w_{w_2}, ..., e^w_{w_n}, e^w_{[SEP]}\right)$$

the sentence embedding (segment embedding) is demonstrated as

$$e^s = (e^s_A, e^s_A, e^s_A, ..., e^s_A, e^s_A)$$

and the word position embedding is demonstrated as

$$e^p = \left(e^p_0, e^p_1, e^p_2, ..., e^p_n, e^p_{n+1}\right)$$

To obtain the word feature input for the BERT model, we merge these three embedding vectors together, as shown in Figure 6.

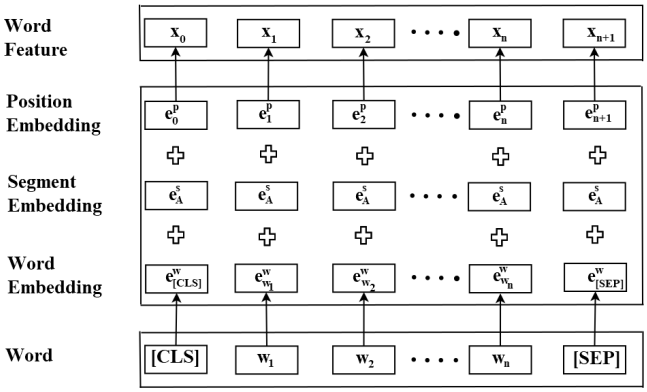

Figure 6. The architectural design of the A-BiGRU

The final representation of the word vector obtained after training is performed by the Formula (2) and serves input for the BiGRU and self-attention layers.

$$x = [x_1, x_2, x_3, ..., x_n] \tag{2}$$

5.2. BiGRU and Attention Mechanism (AM)

Basically, entity recognition involves the task of sequence labeling. In this case, BiGRU is utilized for sequence modeling, which enables the extraction of contextual knowledge from a series of word vectors. BiGRU consists of two GRUs, namely forward and backward. Each GRU includes an update and a reset gate. In order to address the issue of gradient disappearance or explosion in RNNs, the gate structure is employed to selectively retain context information. The GRU architectural design is demonstrated in Figure 7.

The GRU computations are shown as follows:

$$r_t = \sigma\left(w_r \cdot [h_{t-1}, x_t] + b_r\right) \tag{3}$$

$$z_t = \sigma\left(w_z \cdot [h_{t-1}, x_t] + b_z\right) \tag{4}$$

$$\widetilde{h_t} = tanh\left(w_h \cdot [r_t \cdot h_{t-1}, x_t] + b_h\right) \tag{5}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \widetilde{h_t} \tag{6}$$

Where the sigmoid function is denoted by σ and $\cdot$ symbolizes dot product. The symbol w represents weighted matrices and b symbolizes bias. The input vector is x_t, the hidden state is h_t at time t. The update gate is z_t which regulates the impact of the preceding output on the present state. The reset gate is r_t which is applied to regulate the significance of h_{t-1} to $\widetilde{h_t}$. $\widetilde{h_t}$ indicates the information in the current unit that has to be updated. The length of a sequence is captured by both gates. In contrast to LSTM, GRU offers a simpler structure and faster training speed.

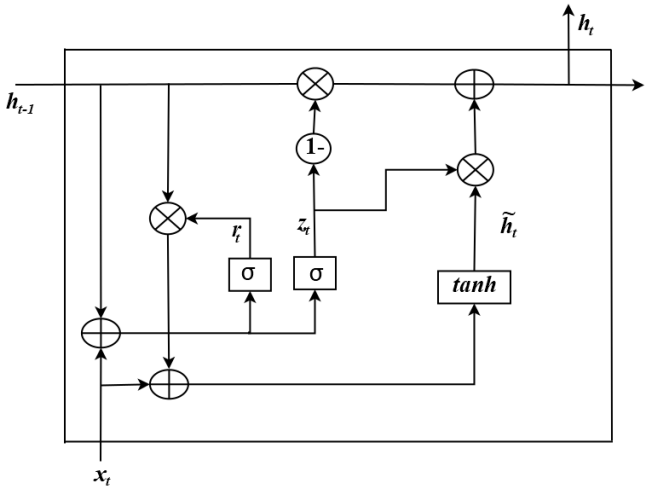

Figure 7. The architectural design of the A-BiGRU

The BiGRU employed in this article consists of both forward and backward GRU. The forward layer of GRU is denoted by $\overrightarrow{h_t}$ and the backward layer of GRU is denoted by $\overleftarrow{h_t}$. BiGRU is used to fetch the future contextual features $\overrightarrow{h_t}$ and the historical contextual features $\overleftarrow{h_t}$ from the context, using formula (7) and formula (8). The output of BiGRU layer is obtained by merging the outputs of forward and backward layer at time t, as illustrate in formula (9).

$$\overrightarrow{h_t} = GRU\left(x_t, \overrightarrow{h_{t-1}}\right) \tag{7}$$

$$\overleftarrow{h_t} = GRU\left(x_t, \overleftarrow{h_{t-1}}\right) \tag{8}$$

$$h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] \tag{9}$$

The objective of BiGRU is to preserve the contextual properties of the sentence sequence. However, the significance of semantic relationship of every word has a different importance in the sentence sequence for the NER task. Normally the text (dataset) contains a huge amount of unnecessary words (not name words), resulting in redundant information. It is challenging for the BiGRU model to fetch crucial information from sentence sequences. Consequently, attention mechanism is incorporated in the BiGRU model to capture further significant information. It can give more weightage to important information and understand the meaning of sequences better. In the A-BiGRU model (BiGRU with attention layer), two attention layers are employed to handle the preceding and subsequent contextual information separately. These pieces of information are then concatenated in the AM (Attention Mechanism) module and fed into the Softmax classifier. The architecture of the A-BiGRU model is shown in Figure 5.

A-BiGRU employs the dropout layer and the Softmax classification layer to obtain the probability distribution required for classification. Dropout layer is employed to mitigate overfitting. The cross-entropy loss function is applied here, which is preferable over the mean square error approach. Adam optimizer is widely recognized as an efficient and successful method for fine-tuning model parameters through backpropagation [25]. By employing the cross-entropy loss function in the stochastic gradient descent technique, the likelihood of gradient disappearance is minimized. The loss function is represented in formula (10) as follows.

$$L_{total} = -\frac{1}{num} \sum_{Sp} [ylno + (1-y)\, ln\, (1-o)] \tag{10}$$

Where, 'num' represents the number of words in the training data, 'Sp' denotes batch sample size used for training, 'y' corresponds to predicted sample label, and 'o' represents actual sample label.

The complete technique for learning (A-BiGRU) is presented in Algorithm 2.

Algorithm 2 A-BiGRU

Require:

1. $w = w_1, w_2, \cdots\cdots, w_n$; ▷ w represents sequence of words.
2. $o = o_1, o_2, \cdots\cdots, o_n$; ▷ o is the actual sample label.

Ensure:

1. $y = y_1, y_2, \cdots\cdots, y_n$; ▷ y represents predicted named label.

Begin:

Step1: The BERT model is used to generate word vector as $x = [x_1, x_2, x_3, \cdots\cdots, x_n]$;

Step2: By using BiGRU, the current and preceding contextual features, denoted as $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, are extracted from the feature vectors, using formula 7 and formula 8;

Step3: The current and past context representations are combined to obtain comprehensive context representations denoted as $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$, using formula 9;

Step4: The h_t is inputted into the Softmax function to obtain the appropriate class designations, denoted as (y).

Step5: The cross entropy function is employed to optimize the parameters of model and sequence of entity labels, using formula 10;

Step6: return $y = y_1, y_2, \cdots\cdots, y_n$;

5.3. Parameter setting

The model parameter settings include an embedding size of 768 and a hidden state dimension of 100 for the GRU. The rate of learning is set to 0.001, while the rate of dropout is 0.5. The batch size used during training is 32. The proposed model has a total of 121 million trainable parameters. Back Propagation and Adam's Optimizer are used to train the network. Table 2 displays parameter configuration of the suggested model.

Table 2
Parameter selection

Parameter	Values
Embedding Size	768
GRU unit dimension	100
Learning	0.001
Optimizer	Adam's Stochastic Optimizer
Dropout	0.5
Batch size	32
Activation function	Softmax
Number of trainable Parameters	∼121M

6. Discussion and analysis

Noticeable improvements have been observed in both CRF and SVM based models, after incorporating clustering technique, but the deep learning based BIGRU with Attention Mechanism outperforms all of them by achieving highest F-Measure of 81.20 with 82.89% Precision and 79.57% Recall. Figure 8 demonstrates a comparative analysis of the baseline outcomes and the improvements achieved by incorporating word clustering in the different stages of developing the NER systems along with A-BiGRU model.

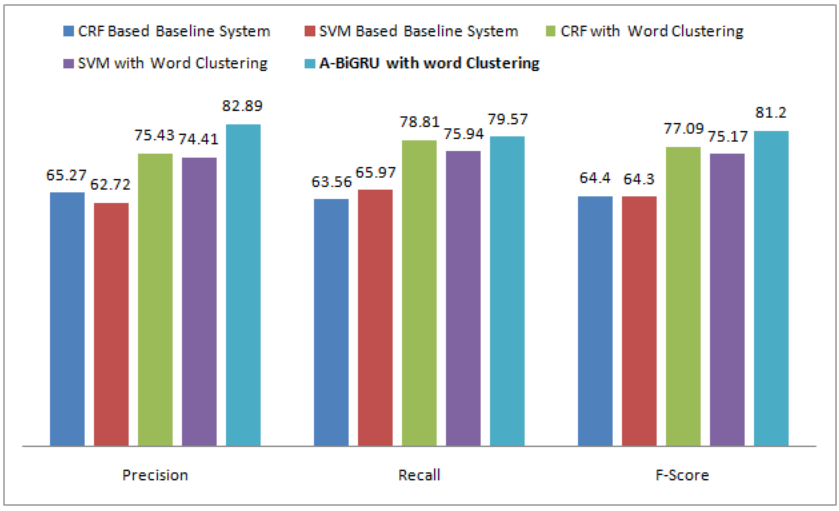

Figure 8. Comparative study of baseline models
with their clustering based enhancements and A-BiGRU

A number of researchers worked on biomedical corpus to develop NER systems for identifying NEs like: Protein, Cell-Type, Cell-Line, RNA, DNA etc. Although none

of these systems achieved ample accuracy like general domain NER systems. We have listed several advanced systems that utilized the GENIA dataset, demonstrating their state-of-the-art performances, and presented a comparison of the proposed systems with them in Table 3.

Zhou and Su introduced the top-performing model in the JNLPBA 2004 shared task [68], which attained an impressive F-measure of 72.55. Their system utilized a combination of deep knowledge about domain and various features to achieve such results. Tang et al. developed machine learning-based BNER systems with 3 distinct word representation (WR) features; including word embeddings, clustering and distributional representation [55].

Table 3
Comparison with existing approaches

NER System	Precision	Recall	F1-Score
[55]	70.78	72.00	71.39
[68]	75.99	69.42	72.55
Proposed SVM with Word Clustering	**74.41**	**75.94**	**75.17**
[9]	74.17	77.87	75.97
[48]	74.17	77.87	75.97
Proposed CRF with Word Clustering	**75.43**	**78.81**	**77.09**
[21]	76.7	71.1	73.8
[33]	76.2	73.6	74.9
[63]	79.0	77.3	78.2
[64]	78.6	79.3	78.9
[62]	77.9	80.7	79.3
[66]	81.80	79.30	80.50
Proposed A-BiGRU	**82.89**	**79.57**	**81.20**

This system reached an F-Measure of 71.39. Ekbal, Saha and Sikdar used genetic algorithm and proposed a SOO based ensemble system on GENIA dataset which attained an F-measure of 75.97; it currently produces the superior result [9, 48] for any machine learning based NER. The proposed system (CRF with cluster merging technique) outperforms the existing machine learning techniques by accomplishing the highest F-measure of 77.09.

Katiyar et al. introduced an innovative method based on recurrent neural networks to address both nested named entity recognition and nested entity mention detection. Their approach achieved the highest F-measure (F1-Score) of 73.80 on the GENIA dataset [21]. Lin et al. proposed an Anchor-Region Networks (ARNs), a sequence-to-nuggets architecture for nested entity mention detection. Their technique produced the greatest F-measure of 74.90 on the GENIA dataset [33]. Xia et al. suggested a Multi-Grained Named Entity Recognition (MGNER) technique to detect and recognize entities on multiple granularities. Their approach achieved an F-measure of 78.20 [63]. Yan et al. proposed pre-trained Seq2Seq model BART

to performed Named Entity Recognition (NER) task. They achieved F-measure of 78.90 on the GENIA dataset [64]. Wan et al. suggested span-based graph method for identifying the nested named entity. They obtained F-measure of 79.30 on this dataset [62]. Yu et al. introduced a deep learning method by incorporating multi-layer BiLSTM with Biaffine dependency parsing model. Their system obtained the best F-measure of 80.50 on GENIA dataset [66]. The proposed deep learning model A-BiGRU surpasses it by accomplishing the highest F-measure of 81.20.

Figure 9, and Figure 10 present comparative analysis of A-BiGRU with Multilayer BiLSTM [66] on the validation dataset regarding accuracy and loss function respectively.

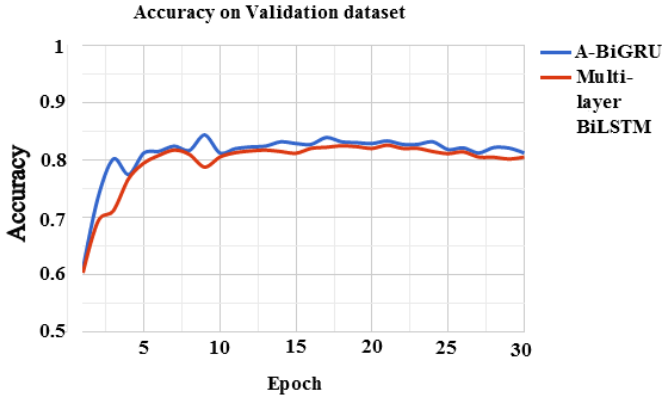

Figure 9. A comparison between A-BiLSTM and Multi-layer BiLSTM [66] models based on the accuracy observed on the validation dataset

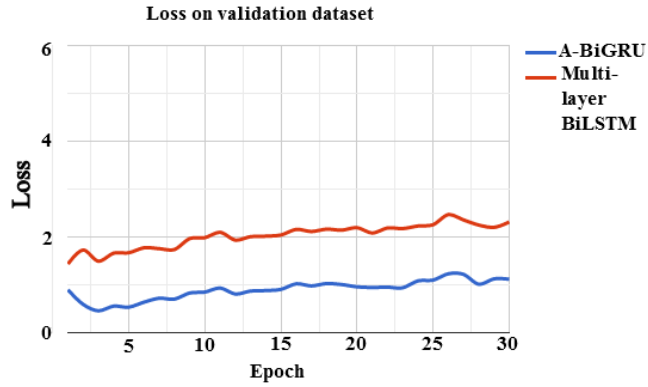

Figure 10. A comparison between A-BiLSTM and Multi-layer BiLSTM [66] models based on the loss observed on the validation dataset

7. Conclusion

This study has demonstrated the construction of a biomedical NER system using a comprehensive set of features incorporating two widely used machine learning classifiers, CRF and SVM. Quality and quantity of tagged training data with enriched set of feature samples are prerequisite criteria for a better performing supervised system. But while sufficiently large such natural language corpus is not available; intelligent processing of raw external corpus can play an imperative role to obtain required statistics to improve the classifier. The clustering technique is one of such widely adopted promising approach. In this biomedical NER system development two type of word clustering techniques, partitioning and hierarchical have been incorporated, where K-medoids and Brown clustering have been chosen respectively. A combination of these two clustering methods has achieved much improve performance over the baseline SVM and CRF based models. In parallel to these we have also prepared a deep learning based NER system (A-BiGRU) which combines BERT language model, BiGRU with Attention Mechanism. The proposed model intelligently captures the context information by incorporating the BERT model to improve the expressiveness of context information by generating word vectors and the Attention Layer is incorporated to overcome the information redundancy. While comparing the proposed system with the state-of-the-art methods, it is found that the result of the proposed system sidelines the surviving NER systems in this domain.

One of the tricky issues in the task is dealing with ambiguities, in which the same term might represent multiple entities (such as a word "Cell" can refer to a biological cell or can refer to an experimental or laboratory setting. Again, another word "Protein" can refer to specific type of body substances or dietary components.) in different situations. As a result, the accuracies of these systems are diminished and certain entities may be overlooked. In future we would like to resolve these issues and extracting the relation among these entities.

Author contributions

The manuscript has been written, edited, and evaluated by all authors.

Conflict of interest

The authors affirm that they have no financial or commercial affiliations that could potentially create a conflict of interest in relation to the research.

References

[1] Balabantaray R.C., Sarma C., Jha M.: Document Clustering using K-Means and K-Medoids, *arXiv preprint arXiv:150207938*, 2015. doi: 10.48550/arXiv.1502.07938.

[2] Biemann C.: Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: *Proceedings of TextGraphs: the first workshop on graph based methods for natural language processing*, pp. 73–80, 2006. doi: 10.3115/1654758.1654774.

[3] Brown P.F., Della Pietra V.J., de Souza P.V., Lai J.C., Mercer R.L.: Class-based n-gram models of natural language, *Computational Linguistics*, vol. 18(4), pp. 467–480, 1992.

[4] Cherry C., Guo H.: The unreasonable effectiveness of word representations for twitter named entity recognition. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 735–745, 2015. doi: 10.3115/v1/n15-1075.

[5] Chieu H.L., Ng H.T.: Named entity recognition: a maximum entropy approach using global information. In: *COLING 2002: Proceedings of the 19th International Conference on Computational Linguistics*, 2002. doi: 10.3115/1072228.1072253.

[6] Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P.: Natural language processing (almost) from scratch, *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011. http://jmlr.org/papers/v12/collobert11a.html.

[7] Deng J., Cheng L., Wang Z.: Self-attention-based BiGRU and capsule network for named entity recognition, *arXiv preprint arXiv:200200735*, 2020.

[8] Dutta R., Majumder M.: Attention-based bidirectional LSTM with embedding technique for classification of COVID-19 articles, *Intelligent Decision Technologies*, vol. 16(1), pp. 205–215, 2022. doi: 10.3233/idt-210058.

[9] Ekbal A., Saha S., Sikdar U.K.: Biomedical named entity extraction: some issues of corpus compatibilities, *SpringerPlus*, vol. 2, 601, 2013. doi: 10.1186/2193-1801-2-601.

[10] Estivill-Castro V., Yang J.: Fast and robust general purpose clustering algorithms. In: R. Mizoguchi, J. Slaney (eds.), *PRICAI 2000. Topics in artificial intelligence. 6th Pacific Rim International Conference on Artificial Intelligence, Melbourne, Australia, August/September 2000. Proceedings*, Lecture Notes in Artificial Intelligence. Subseries of Lecture Notes in Computer Science, vol. 1886, pp. 208–218, 2000. doi: 10.1007/3-540-44533-1_24.

[11] Finkel J.R., Dingare S., Nguyen H., Nissim M., Manning C.D., Sinclair G.: Exploiting context for biomedical entity recognition: From syntax to the web. In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pp. 88–91, 2004. doi: 10.3115/1567594.1567614.

[12] Fraley C., Raftery A.E.: How many clusters? Which clustering method? Answers via model-based cluster analysis, *The Computer Journal*, vol. 41(8), pp. 578–588, 1998. doi: 10.1093/comjnl/41.8.578.

[13] Fukuda K., Tsunoda T., Tamura A., Takagi T.: Toward information extraction: identifying protein names from biological papers. In: *Pacific Symposium on Biocomputing*, pp. 707–718, 1998. https://psb.stanford.edu/psb-online/proceedings/psb98/abstracts/p707.html.

[14] Grishman R.: The NYU system for MUC-6 or where's the syntax? In: *Proceedings of the 6th Conference on Message Understanding*, pp. 167–175, 1995. doi: 10.3115/1072399.1072415.

[15] Hammerton J.: Named entity recognition with long short-term memory. In: *CONLL '03: Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003*, vol. 4, pp. 172–175, 2003. doi: 10.3115/1119176.1119202.

[16] Han J., Pei J., Tong H.: *Data mining: concepts and techniques*, Morgan Kaufmann, 4th ed., 2022.

[17] He K., Zhang X., Ren S., Sun J.: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/cvpr.2016.90.

[18] Hinton G., Deng L., Yu D., Dahl G.E., Mohamed A.r., Jaitly N., Senior A., *et al.*: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine*, vol. 29(6), pp. 82–97, 2012. doi: 10.1109/msp.2012.2205597.

[19] Huang Z., Xu W., Yu K.: Bidirectional LSTM-CRF models for sequence tagging, *arXiv preprint arXiv:150801991*, 2015. doi: 10.48550/arXiv.1508.01991.

[20] Jain A., Yadav D., Arora A., Tayal D.K.: Named-entity recognition for Hindi language using context pattern-based maximum entropy, *Computer Science*, vol. 23(1), 2022. doi: 10.7494/csci.2022.23.1.3977.

[21] Katiyar A., Cardie C.: Nested named entity recognition revisited. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 861–871, 2018. doi: 10.18653/v1/n18-1079.

[22] Kazama J., Makino T., Ohta Y., Tsujii J.: Tuning support vector machines for biomedical named entity recognition. In: *BioMed '02: Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, vol. 3, pp. 1–8, 2002. doi: 10.3115/1118149.1118150.

[23] Kim J.-D., Ohta T., Tateisi Y., Tsujii J.: GENIA corpus – a semantically annotated corpus for bio-textmining, *Bioinformatics*, vol. 19(suppl_1), pp. i180–i182, 2003. doi: 10.1093/bioinformatics/btg1023.

[24] Kim J.-D., Ohta T., Tsuruoka Y., Tateisi Y., Collier N.: 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 (NLPBA/BioNLP)*, pp. 70–75, 2004. doi: 10.3115/1567594.1567610.

[25] Kingma D.P., Ba J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:14126980*, 2014. doi: 10.48550/arXiv.1412.6980.

[26] Kong J., Zhang L., Jiang M., Liu T.: Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition, *Journal of Biomedical Informatics*, vol. 116, p. 103737, 2021. doi: 10.1016/j.jbi.2021.103737.

[27] Lafferty J., McCallum A., Pereira F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, Morgan Kaufmann, 2001. https://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf.

[28] Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C.: Neural architectures for named entity recognition. In: K. Knight, A. Nenkova, O. Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, 2016. doi: 10.18653/v1/n16-1030.

[29] Leaman R., Gonzalez G.: BANNER: an executable survey of advances in biomedical named entity recognition. In: R.B. Altman, A.K. Dunker, L. Hunter, T. Murray, T.E. Klein (eds.), *Pacific Symposium on Biocomputing 2008. Kohala Coast, Hawaii, USA, 4–8 January 2008*, pp. 652–663, World Scientific, 2008.

[30] LeCun Y., Bengio Y., Hinton G.: Deep learning, *Nature*, vol. 521(7553), pp. 436–444, 2015. doi: 10.1038/nature14539.

[31] Li Y., Qi H., Dai J., Ji X., Wei Y.: Fully convolutional instance-aware semantic segmentation. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4438–4446, 2017. doi: 10.1109/cvpr.2017.472.

[32] Liang P.: *Semi-supervised learning for natural language*, Ph.D. thesis, Massachusetts Institute of Technology, 2005.

[33] Lin H., Lu Y., Han X., Sun L.: Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. In: A. Korhonen, D. Traum, L. Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5182–5192, 2019. doi: 10.18653/v1/p19-1511.

[34] Lin Y.F., Tsai T.H., Chou W.C., Wu K.P., Sung T.Y., Hsu W.L.: A maximum entropy approach to biomedical named entity recognition. In: *BIOKDD'04: Proceedings of the 4th International Conference on Data Mining in Bioinformatics*, pp. 56–61, Citeseer, 2004.

[35] Ma X., Hovy E.: End-to-end sequence labeling via Bi-directional LSTM-CNNs-CRF. In: K. Erk, N.A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1064–1074, 2016. doi: 10.18653/v1/p16-1101.

[36] Madhulatha T.S.: Comparison between K-Means and K-Medoids Clustering Algorithms. In: D.C. Wyld, M. Wozniak, N. Chaki, N. Meghanathan, D. Nagamalai (eds.), *Advances in Computing and Information Technology: First International Conference, ACITY 2011, Chennai, India, July 15–17, 2011. Proceedings*, pp. 472–481, Springer, 2011. doi: 10.1007/978-3-642-22555-0_48.

[37] Maity S., Das N., Majumder M., Dasadhikary D.R.: Word Embedding and String-Matching Techniques for Automobile Entity Name Identification from Web Reviews, *EAI Endorsed Transactions on Scalable Information Systems*, vol. 8(33), pp. 1–11, 2021. doi: 10.4108/eai.14-5-2021.169918.

[38] Matsuo Y., Sakaki T., Uchiyama K., Ishizuka M.: Graph-based word clustering using a web search engine. In: *EMNLP '06: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 542–550, 2006. doi: 10.3115/1610075.1610150.

[39] Miller S., Guinness J., Zamanian A.: Name tagging with word clusters and discriminative training. In: *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 337–342, 2004.

[40] Patra R., Saha S.K.: A kernel-based approach for biomedical named entity recognition, *The Scientific World Journal*, vol. 2013, 2013. doi: 10.1155/2013/950796.

[41] Patra R., Saha S.K.: A novel word clustering and cluster merging technique for named entity recognition, *Journal of Intelligent Systems*, vol. 28(1), pp. 15–30, 2019. doi: 10.1515/jisys-2016-0074.

[42] Pereira F., Tishby N., Lee L.: Distributional clustering of English words. In: *31st Annual Meeting of the Association for Computational Linguistics*, pp. 183–190, Columbus, Ohio, USA, 1994. doi: 10.3115/981574.981598.

[43] Ponomareva N., Pla F., Molina A., Rosso P.: Biomedical named entity recognition: A poor knowledge HMM-based approach. In: Z. Kedad, N. Lammari, E. Métais, F. Meziane, Y. Rezgui (eds.), *Natural Language Processing and Information Systems: 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France, June 27–29, 2007. Proceedings*, Lecture Notes in Computer Science, vol. 4582, pp. 382–387, Springer, 2007. doi: 10.1007/978-3-540-73351-5_34.

[44] Pyysalo S., Ginter F., Heimonen J., Björne J., Boberg J., Järvinen J., Salakoski T.: BioInfer: a corpus for information extraction in the biomedical domain, *BMC Bioinformatics*, vol. 8, pp. 1–24, 2007. doi: 10.1186/1471-2105-8-50.

[45] Ratinov L., Roth D.: Design challenges and misconceptions in named entity recognition. In: *CoNLL '09: Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147–155, 2009. doi: 10.3115/1596374.1596399.

[46] Ritter A., Clark S., Mausam, Etzioni O.: Named entity recognition in tweets: an experimental study. In: *EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1524–1534, 2011. https://aclanthology.org/D11-1141.pdf.

[47] Rössler M.: Adapting an NER-system for German to the biomedical domain. In: *JNLPBA '04: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 92–95, 2004. doi: 10.3115/1567594.1567615.

[48] Saha S., Ekbal A., Sikdar U.K.: Named entity recognition and classification in biomedical text using classifier ensemble, *International Journal of Data Mining and Bioinformatics*, vol. 11(4), pp. 365–391, 2015. doi: 10.1504/ijdmb.2015.067954.

[49] Saha S.K., Mitra P., Sarkar S.: A comparative study on feature reduction approaches in Hindi and Bengali named entity recognition, *Knowledge-Based Systems*, vol. 27, pp. 322–332, 2012. doi: 10.1016/j.knosys.2011.09.015.

[50] Saha S.K., Sarkar S., Mitra P.: Feature selection techniques for maximum entropy based biomedical named entity recognition, *Journal of Biomedical Informatics*, vol. 42(5), pp. 905–911, 2009. doi: 10.1016/j.jbi.2008.12.012.

[51] Salton G., Wong A., Yang C.S.: A vector space model for automatic indexing, *Communications of the ACM*, vol. 18(11), pp. 613–620, 1975. doi: 10.1145/361219.361220.

[52] Settles B.: Biomedical named entity recognition using conditional random fields and rich feature sets. In: *JNLPBA '04: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 107–110, 2004. doi: 10.3115/1567594.1567618.

[53] Shen D., Zhang J., Zhou G., Su J., Tan C.L.: Effective adaptation of hidden Markov model-based named entity recognizer for biomedical domain. In: *BioMed '03: Proceedings of the ACL 2003 workshop on natural language processing in biomedicine*, vol. 13, pp. 49–56, 2003. doi: 10.3115/1118958.1118965.

[54] Soni K.G., Patel A.: Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data, *International Journal of Computational Intelligence Research*, vol. 13(5), pp. 899–906, 2017. https://www.ripublication.com/ijcir17/ijcirv13n5_21.pdf.

[55] Tang B., Cao H., Wang X., Chen Q., Xu H.: Evaluating word representation features in biomedical named entity recognition tasks, *BioMed Research International*, vol. 2014, 2014. doi: 10.1155/2014/240403.

[56] Toh Z., Chen B., Su J.: Improving twitter named entity recognition using word representations. In: *Proceedings of the Workshop on Noisy User-generated Text*, pp. 141–145, 2015. doi: 10.18653/v1/w15-4321.

[57] Tsai T.h., Chou W.C., Wu S.H., Sung T.Y., Hsiang J., Hsu W.L.: Integrating linguistic knowledge into a conditional random fieldframework to identify biomedical named entities, *Expert Systems with Applications*, vol. 30(1), pp. 117–128, 2006. doi: 10.1016/j.eswa.2005.09.072.

[58] Turian J., Ratinov L.A., Bengio Y.: Word representations: A simple and general method for semi-supervised learning. In: *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 384–394, 2010.

[59] Ushioda A.: Hierarchical clustering of words and application to NLP tasks. In: E. Ejerhed, I. Dagan (eds.), *Fourth Workshop on Very Large Corpora*, University of Copenhagen, Copenhagen, 1996. https://aclanthology.org/W96-0103.pdf.

[60] Uszkoreit J., Brants T.: Distributed word clustering for large scale class-based language modeling in machine translation. In: *ACL 2008. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15–20, 2008, Columbus, Ohio, USA*, pp. 755–762, 2008.

[61] Vapnik V.N.: *The nature of statistical learning theory*, Springer, New York, NY, 1999. doi: 10.1007/978-1-4757-3264-1.

[62] Wan J., Ru D., Zhang W., Yu Y.: Nested named entity recognition with span-level graphs. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 892–903, 2022. doi: 10.18653/v1/2022.acl-long.63.

[63] Xia C., Zhang C., Yang T., Li Y., Du N., Wu X., Fan W., *et al.*: Multi-grained named entity recognition. In: A. Korhonen, D. Traum, L. Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1430–1440, 2019. doi: 10.18653/v1/p19-1138.

[64] Yan H., Gui T., Dai J., Guo Q., Zhang Z., Qiu X.: A unified generative framework for various NER subtasks. In: C. Zong, F. Xia, W. Li, R. Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5808–5822, 2021. doi: 10.18653/v1/2021.acl-long.451.

[65] Yeh A., Morgan A., Colosimo M., Hirschman L.: BioCreAtIvE task 1A: gene mention finding evaluation, *BMC Bioinformatics*, vol. 6, S2 (2005), 2005. doi: 10.1186/1471-2105-6-s1-s2.

[66] Yu J., Bohnet B., Poesio M.: Named entity recognition as dependency parsing. In: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6470–6476, 2020. doi: 10.18653/v1/2020.acl-main.577.

[67] Zhou G., Su J.: Named entity recognition using an HMM-based chunk tagger. In: P. Isabelle, E. Charniak, D. Lin (eds.), *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 473–480, 2002. doi: 10.3115/1073083.1073163.

[68] Zhou G., Su J.: Exploring deep knowledge resources in biomedical name recognition. In: *JNLPBA '04: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pp. 96–99, 2004. doi: 10.3115/1567594.1567616.

[69] Zhou H., Ning S., Liu Z., Lang C., Liu Z., Lei B.: Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes, *BMC Bioinformatics*, vol. 21(1), 35, 2020. doi: 10.1186/s12859-020-3375-3.

Affiliations

Nilanjana Das
Department of Computer Science, Vidyasagar University, Midnapore, West Bengal, India

Rakesh Dutta
Department of Computer Science and Application, Hijli College, Kharagpur, India

Uttam Kumar Mondal
Department of Computer Science, Vidyasagar University, Midnapore, West Bengal, India

Mukta Majumder
Department of Computer Science & Technology, University of North Bengal, Siliguri, India

Jyotsna Kumar Mandal
Department of Computer Science & Engineering, University of Kalyani, Kalyani, West Bengal, India

Received: 8.07.2023
Revised: 7.06.2024
Accepted: 9.06.2024