Knowledge and Information Systems (2019) 60:1453–1478
https://doi.org/10.1007/s10115-018-1265-z

REGULAR PAPER

Information theoretic-PSO-based feature selection: an application in biomedical entity extraction

Shweta Yadav[1] · Asif Ekbal[1] · Sriparna Saha[1]

Received: 20 September 2016 / Revised: 13 March 2018 / Accepted: 10 May 2018 /
Published online: 21 September 2018

Abstract
Named entity recognition is a vital task for various applications related to biomedical natural language processing. It aims at extracting different biomedical entities from the text and classifying them into some predefined categories. The types could vary depending upon the genre and domain, such as gene versus non-gene in a coarse-grained scenario, or protein, DNA, RNA, cell line, and cell-type in a fine-grained scenario. In this paper, we present a novel filter-based feature selection technique utilizing the search capability of particle swarm optimization (PSO) for determining the most optimal feature combination. The technique yields in the most optimized feature set, that when used for classifiers learning, enhance the system performance. The proposed approach is assessed over four popular biomedical corpora, namely GENIA, GENETAG, AIMed, and Biocreative-II Gene Mention Recognition (BC-II). Our proposed model obtains the F score values of 74.49%, 91.11%, 90.47%, 88.64% on GENIA, GENETAG, AIMed, and BC-II dataset, respectively. The efficiency of feature pruning through PSO is evident with significant performance gains, even with a much reduced set of features.

Keywords Named entity recognition · Feature selection · Binary PSO · Correlation · Mutual information · Normalized mutual information · Particle swarm optimization

1 Introduction

Biomedical named entity recognition (NER) is the key component in biomedical text mining, which automatically recognizes and extracts biomedical entities (e.g., genes, proteins,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10115-018-1265-z) contains supplementary material, which is available to authorized users.

✉ Shweta Yadav
shweta.pcs14@iitp.ac.in

Asif Ekbal
asif@iitp.ac.in

Sriparna Saha
sriparna@iitp.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna 801103, India

chemicals, and diseases) from the text. It also serves as a primary step for some higher level task such as relation extraction, and knowledge base completion. The task of NER in the biomedical text is more complex and inherently more challenging compared to other domains because of the following reasons:

- *Long wordforms* Biomedical NEs are, in general, very long and complex. For example: *"myeloid lineage-associated cytokine receptors"*
- *Nested wordforms* Biomedical names have nested structures, such as: <RNA><DNA>CIITA< /DNA>mRNA< /RNA> Here, "CIITA" represents DNA and the entire string "CIITA mRNA" represents RNA.
- *Presence of symbols and punctuations inside the named entities (NEs)* complicates the process of their identification. For example, "5-lipoxygenase", "CD28", "INTERLEUKIN-2(IL-2)".
- *Inconsistent naming convention* Due to unavailability of standard biomedical nomenclature system, a biological entity can have different name variations. For example: "IL-2" can have several wordforms such as "IL2", "Interleukin 2", or "interleukin-2".
- *Ambiguity* Same entity with similar orthographic properties can belong to the different classes. For example, "IL-2" can be protein as well as DNA depending upon the contexts.

With the availability of the annotated data, significant advancement is observed in the data-driven techniques (supervised learning) for biomedical named entity recognition (BNER). Generally, the success of supervised systems is dependent primarily on the volume of annotated data and the feature set utilized. In the past, different classification models are developed on a large set of features after studying the properties of the data. Also without the prior knowledge, it is difficult to determine the usefulness of each feature. The large number of relevant, irrelevant, and redundant features in the dataset increases the search space that eventually leads to the problem of "the curse of dimensionality" [18].

Feature selection is a useful preprocessing step that helps in improving the performance of a classification system by selecting an optimal subset of features [42] by eliminating/reducing irrelevant and redundant features. In general, there are two standard approaches for feature selection, *viz.* **filter** and **wrapper** [42]. In the filter-based method, the assessment of feature subset is carried out by utilizing some statistical measures [38,53] without incorporating any learning/classification algorithm. In contrast, wrapper-based methods require a learning algorithm apriori to explore the feature space [38].

The existence of complex interaction among various features makes the feature selection task more challenging. It might be the case that some features relevant to one class can be irrelevant to the other features. Therefore, an optimal feature subset should be a group of complementary features that span over the diverse properties of the classes. To address this problem, a variety of feature selection techniques have been developed, such as sequential forward selection [29], sequential backward elimination [29], recursive feature elimination (RFE) [22]. However, these approaches suffer from the problem of getting stuck at local optima. Some of the works in the literature [58,59] shows that evolutionary computational technique can achieve notable performance for their global search ability.

Motivated by the strength of evolutionary algorithm, in this paper, we propose a generic feature selection technique utilizing the particle swarm optimization (PSO) [10] technique to assist the BNER models in the supervised learning framework. Particle swarm optimization (PSO) is a popular bionic algorithm where a set of randomly generated solutions move in the search space to obtain the optimal solutions. The working principle of PSO is simple (converge quickly) and requires very less number of parameters when compared with other optimization techniques like evolutionary algorithms [9,12].

The proposed approach operates in three steps: first, we define the goodness measures utilizing the concepts of information theory, such as normalized mutual information [53], entropy [7], mutual information [38], to quantify the quality of a feature subset. The formulation of objective functions is based on the idea to explore statistical properties of the feature subset. In the second step, the search capability of PSO is explored to select the best feature subset exploiting the defined objective functions. Finally, popular sequence learning technique, namely conditional random field (CRF) [32] is used for building a model using the set of optimized features obtained through the PSO-based approach. The efficacy of the proposed approach is reported on four biomedical corpora, namely GENIA, GENETAG, AIMed, and BC-II. Experimental results show that we obtain reasonable accuracy with a pruned feature set. The models developed using this reduced feature set have less complexities, and hence can be used for developing real-time applications. Our evaluation shows that compared to the conventional baseline, the use of PSO-based feature selection yields the improvements of 5.56, 2.73, and 1.98 and 2.93 F score points on GENIA, GENETAG, and AIMed and BC-II dataset, respectively. The analysis reveals that the objective function utilizing the concept of normalized mutual information is the most effective in identifying the optimal feature set.

We summarize below the main contributions of our work:

- Development of an efficient feature selection technique based on the PSO framework to assist the BNER task.
- Formulation of seven distinct information theoretic-based objective measures to exploit statistical properties of feature subset.
- Evaluation of proposed PSO-based feature selection technique on four benchmark corpora such as GENIA, GENETAG, AIMed, and BC-II for BNER task and its comparison with state-of-the-art techniques.

The remainder of the paper is structured as follows: Sect. 2 puts light on the recent works on NER task on biomedical text. In Sect. 3, we describe the proposed approach in details. Section 4 studies the computational complexity of the system. Section 5 reports experimental results and analysis. In Sect. 6, we present the comparative analysis with state-of-the-art techniques. Error analysis is shown in Sect. 7. Finally, we conclude the paper in Sect. 8.

2 Related works

The potential applicability and necessity of the NER problems have drawn the interest of many researchers in recognizing different biomedical entities like gene- and protein-related names from the available biomedical text including several shared task challenges [26,48]. In general, we can categorize the existing approaches into three classes:

2.1 Lexicon or dictionary-based NER approach

A dictionary-based approach is viewed as the simple and naive technique of extracting entity mentions in text. Existing studies reveal the technique to have a high degree of precision with the very low recall. The system developed by Friedrich et al. [17] utilizes the dictionary-based features which allocates token classes, based on the presence of token in the dictionary. The low recall value is attributed to the spelling mistake, word-level, and character level variations. The major problem with the utilization of dictionaries is that it is infeasible to maintain the huge list of the entities, and furthermore there is a continuous growth in biomedical resource. The low recall and the other attributed problems in dictionary-based method have led to

introducing several enhancements to these approaches. Some methods have enhanced the traditional dictionary-based technique by exploiting inexact matching approach, where the extended dictionary is used further for exact matching against text [24]. Despite the efforts, utilizing a dictionary reduces the scope of the system on any new entity.

2.2 Rule-based NER approach

Here, rules are defined to capture entities following some patterns and context of named entities. Rule-based method is shown to perform better compared to the dictionary-based approach. However, it is both tedious and difficult to frame each possible rule as those are handcrafted. AbGene [49] is one of the most successful rule-based NER systems to identify gene and protein names from biomedical literature with a precision of (87.00)% and recall of 67.00% on 56,000 Medline abstracts. They used Brill POS tagger learned on 7000 manually tagged biomedical sentences. Further, postprocessing was performed in the form of hand-generated rules based on lexical-statistical properties. However, the proposed system suffers from some limitations, for instance, it could miss single word gene name that occurs without contextual gene term. Moreover, they identified that the heuristics to detect an invalid combination of a compound term is not perfect. EDGAR [41] system extracts cancer relevant drug and gene information from the biomedical literature. It takes input from the parser that uses the semantic and syntactic information from the Unified Medical Language System Metathesaurus (UMLS) [5] to extract factual assertions from the text. Further, they used stochastic POS tagger to enhance syntactic parser. This approach has limited portability and as such it fails to give a comparable performance on the other domains. Another popular tool, MetaMap [2] developed by National Library of Medicine (NLM) discovers UMLS Metathesaurus reflected in the text with the precision of 85.00% and recall of 78.00%. Despite this high precision, system drastically fails for the ambiguous cases. The ProMiner [24] is also a rule-based system that extracts multiword names. It explores preprocessed synonym dictionary to extract potential name occurrences from the biomedical text which associates protein and gene database identifiers with the extracted matches. The system was evaluated on the BioCreAtIVE challenge dataset. The system obtained 80.00% F score value on the organisms mouse and fly datasets. On organism yeast, the system achieved 90.00% F score value. Bedmar et al. [45] perform drug name recognition and classification in biomedical texts. They utilized UMLS-MetaMap Transfer (MMTx) program information and defined nomenclature rules approved by the World Health Organization (WHO) to extract and classify pharmaceutical substances. The system also suffers from the ambiguity issue. The analysis revealed that the system was unable to capture genes from different organisms which are present in one abstract. Moreover, the system was unable to disambiguate because of the missing synonyms.

2.3 Supervised machine learning-based approach

With the availability of annotated data, major advancement is seen in the data-driven method for identifying the biomedical named entities (NEs). The release of biomedical benchmark corpora such as GENIA (derived from the MEDLINE) have led to the rapid advancement in biomedical text mining. Toward this end, several supervised machine learning models such as hidden Markov models (HMM) [3], conditional random fields (CRF) [32] and support vector machines (SVM) [6] have been exploited. Several shared task challenges such as BioNLP/NLPBA 2004 also used GENIA dataset in the challenge to extract the entities where a total of 9 systems was submitted.

Some of the popular state-of-art systems on this dataset include [19,36,46], where SVM and CRF were used as the popular base classifiers, respectively. In general, CRF is a popular classifier for any sequence labeling task such as Named Entity Recognition [33,61]. Zhou et al. [66] trained an HMM on a feature set such as lexical, syntactic features (i.e., prefix, suffix) and word normalization feature on GENIA dataset. Further, they included SVM to solve the data sparsity problem. Besides the lexical and syntactic features, they explored the alias and cascaded entities by utilizing the closed dictionary from the training corpus and the open dictionary from SwissProt and the alias list LocusLink. Their system reported the F score value of 72.55%. The model highly relies on the several postprocessing operation, domain-specific dictionary features which are not generic enough if evaluated on other biomedical corpus. System proposed by Settles [46] was also evaluated on the BioNLP/NLPBA 2004 shared task (GENIA) dataset. The system was trained on CRF using orthographic feature set and the semantic domain knowledge in the form of lexicon and achieved the F score value of 69.50%. This work concluded that the orthographic feature set is highly effective in capturing the entities. However, this system fails in correctly predicting RNA and cell line-type entities. It was observed that mostly misclassification was due to the low frequency terms in the corpus. System proposed by [39] used HMM-based models considering only Part-of-Speech (PoS) as feature. They achieved the F score value of 65.70%. The main drawback of this system is that it is solely dependent on PoS feature which is unable to capture all the lexical and semantic variations of biomedical text. Furthermore, the HMM model suffers from the label bias problem. Kim et al. [27] proposed two-phase system consisting of term boundary detection and semantic labeling. They used CRF and maximum entropy classifier, respectively, for boundary detection and semantic labeling. Moreover, they used finite state method to define the postprocessing rules to refine their proposed framework. The system reported 71.19% F score without domain-specific knowledge such as Gazetteers or abbreviation handling process. Finkel et al. [15] proposed NER system on GENIA dataset. They used ME as a base classifier and achieved F score of 70.06%. However, the system made use of a number of external resources, including gazetteers, web-querying, use of the surrounding abstract, abbreviation handling, and frequency counts from BNC corpus. Finkel et al. [16] further explored GENETAG dataset for BNER. They reported the F score value of 82.2%. McDonald et al. [35] employed orthographic feature set with other gene and biological term lexicons. They achieved a precision of 86.40% and recall of 78.70%. The system was identified to suffer from the boundary detection problem. Kinoshita et al. [28] proposed a system that achieved a F score of 80.90% with dictionary-based preprocessing and HMM-based PoS tagger. The SVM-based system [36]) utilized gene/protein name dictionary as the domain knowledge. It reported F score of 78.09%. These systems highly rely on external knowledge sources which fails to show reasonable performance when tested on other biomedical domain. An ensemble method developed in [55] made use of HMM, SVM and reported the F score of 82.58%. This system also utilized several external resources in the form of gazetteers to provide evidential clues. Some of the feature selection-based BNER models are available at [11,44]. Recently, in [43] authors proposed genetic algorithm (GA)-based classifier ensemble technique for identifying the biomedical entities from the texts. They employed CRF and SVM as a classifier by varying the feature combinations. Finally, different models are combined/ensemble using the genetic algorithm-based classifier ensemble technique in an efficient way. They evaluated their proposed model on JNLPBA 2004 and GENETAG datasets and obtained the F score values of 75.97% and 95.90%, respectively. However, the system was highly complex both in time & space when compared to PSO-based feature selection.

Danger et al. [8] studied the protein interaction corpus to extract 12 protein relevant entities. The system is divided into two sub-modules: a dictionary lookup which searches

for some entities in the text that can be associated with a relatively stable set of interaction terms; while the second module utilizes CRF classifier to search for the entity that cannot be described through a dictionary. They too evaluated their system on JNLPBA 2004 corpus reporting the F score value as 76.13%. The system highly relies on the dictionary-based domain-specific features and thus it is highly domain dependent. Moreover, the system was found to be confused in identifying the RNA-proteins and cell line-cell-type entities pair because of the boundary detection problem. The system also fails to capture the entity when it has strong overlaps with other entities.

In the study conducted by [64], the stepwise unsupervised solution was proposed to perform entity extraction. They utilized UMLS meta-thesaurus to collect the seed terms for each target entity class by their semantic types, semantic groups, or specific concepts. Further, they detect the boundaries by leveraging the concept of noun phrases. In their final step, all identified candidate entities are provided as the input to the classifier to predict their semantic category. They evaluated their system on i2b2 and GENIA dataset with micro F score of 53.1% and 39.5%, respectively. The major limitation of this system is the requirement of the large collection of test dataset in order to generate signature or semantics group.

PSO-based feature selection has also been popular in other domains like sentiment analysis [21], face recognition [40], spam detection [65], etc.

Neural network-based approach In recent years, neural network models have gained their popularity for solving problems in several domains [14,20,62,63]. Recently, some studied have been conducted to explore deep neural network methods for entity extraction from the biomedical corpus and clinical text (Electronic Medical Records) [30,57]. These approaches surpass the role of the manual feature engineering. The study performed by Tang et al. [50,51], Yadav et al. [60] shows that the addition of word representation features in the traditional feature set can help in improving the system performance. The system was observed to be highly efficient than machine learning-based BNER techniques.

3 Proposed approach

In this section, we present our proposed technique for feature selection utilizing the concept of information theoretic measures and PSO. We begin by first formulating the feature selection problem. Followed by that, we discuss the feature engineering phase and the proposed information theoretic-based objective functions. Finally, we present the details of our PSO-based feature selection algorithm.

3.1 Problem formulation

Given, the set of features size n: $\mathscr{F} = \{f_1, f_2 \ldots f_n\}$, a classifier C and classification performance metrics M. The feature selection problem states:

"Extract the optimal feature subset $\mathscr{F}^{'} \subseteq \mathscr{F}$ such that classification performance metrics M can be maximized on classifier C."

3.2 Feature engineering for BNER

We design diverse sets of features by analyzing biomedical texts. Descriptions of these features are presented below:

- *Local contextual feature:* Contextual information provides an important clue to identify the vicinity of the current word. For example, the words 'receptor', 'factor', and 'protein' if occurring in the local context will provide evidence in determining the protein class, and 'gene', 'promoter', and 'motif' are clues for classifying DNA.
  Context can be represented as $w_{i-1}^{i+1} = w^{i-1} \ldots . w^{i+1}$ where w_i denotes the focus word. Here, we capture the contextual information from w_{i-5}^{i+5} (i.e., preceding 5 and succeeding 5 words).
- *Word affixes:* Functional affixes are used to indicate the syntactic function which provides information in capturing the very important clues for terminology identification. We use four length word affixes as the features.
  For example, for the term 'receptor'; the prefix is 'rece' and suffix is 'ptor'.
- *Part-of-Speech (PoS) information:* Mostly biomedical NEs are noun phases, so capturing Part-of-Speech (PoS) information (extracted from[1]) provides important evidence in the identification of the NEs. Here, we use the PoS information of the current token as the feature, e.g., 'IL-2', 'NF-Kappa', and 'v-Abl'.
- *Chunk-type information:* This feature helps in properly identifying the boundary of the entity. We use the chunk information of the current and/or surrounding token(s). We use GENIA tagger v2.0.2 to extract the chunk information.
- *Dynamic NE tag(s):* This feature helps in capturing the intermediate token of NE phase by providing the better evidence of current token to be intermediate NE. It represents the output tags $t_{i-3}t_{i-2}t_{i-1}$ of the word $w_{i-3}w_{i-2}w_{i-1}$ preceding w_i in the sequence w_1^n. Here, we use the bi-gram template which takes the union of both present and preceding output class.
- *Word normalization:* We form two different variants for this feature. The first feature captures the words with the plural form, alphanumeric character, digit, hyphen, and verb which enable in transforming the word to its root form. The other form of the feature specifies the orthographic construction of the current token. It is defined as the word shape feature. For each word, the normalized form is implemented by converting the uppercase by 'A', lowercase by 'a' and digit by '0'. For example, if the token is 'NF-kappa-10', the normalized form will be 'AA-aaaaa-00'. This again is further squeezed to produce the word form of 'A-a-0'.
- *Word length:* It is observed that length of NE is generally longer compared to the other words in the text. This is based on the assumption that short words contain less information. For example, 'and', 'of', 'for' do not add any meaning to training. This binary feature is set as 1, when the length of word > 4; otherwise, the feature value is 0.
- *Infrequent word:* The words which occur more frequently have less probability of occurring as the NEs. We design the binary feature, that set 1, when the frequency of the occurrence of the current word is greater than 10; otherwise, the feature value is 0.
- *Head noun:* It represents the noun phrase which describes the functional property of the NE. Head noun phrase provides useful information in classifying them as NEs, when compared to remaining NEs words. We created a lexicon of 912 head noun, which occurs most frequently from the training dataset. This binary feature can have feature value 1 or 0, depending upon whether the word is present (1) or absent (0) in the lexicon. We have listed some examples of head nouns in Table 1.
- *Verb trigger:* There are certain action verbs (e.g., inhibit, binds, interact) that help in recognizing the presence of NEs. We created a lexicon of a verb from the training dataset.

[1] http://www.nactem.ac.uk/GENIA/tagger/.

Table 1 Examples of the head nouns

Head noun	
Clones	Assays
Extracts	Cytokines
Macrophages	Motifs
Glucocorticoids	Responses

We define a binary-valued feature that checks whether the current word appears in the list or not.

– *Informative NE information:* Usually, biomedical NEs have longer wordforms that often contain many common words which in actual do not belong to the NEs. For example, the nominals and the function words which frequently appear in the training dataset but are usually not effective in capturing the NEs. Generally, biomedical NEs contain common symbols, punctuations, common words, functional word which often are longer in length. Considering this, we developed a list of multiword NEs. We have eliminated digits and symbol from the list as these provide very less clue in the identification of the NEs.
A weight is defined to identify the important word in NEs as: $NE_{wt}(t_i)$ to be calculated as follows:

$$NE_{wt}(t_i) = \frac{\text{Frequency of } t_i \text{ as part of NE}}{\text{Frequency of } t_i \text{ in the training set}} \quad (1)$$

Finally, the most effective words were selected on the basis of two parameters: NEweight and frequency of words. We choose the threshold value of these two parameters by performing experiments on the validation set. We adopt the similar strategy as proposed by Yadav et al. [58]), to define the 5 distinct classes.

– *Orthographic features*: Based on the constructions we define various orthographic features. Generally, it is observed that in biomedical NE, special characters like (‘,’, ‘-’, ‘.’, ‘_’) are generally prominent. Therefore, these symbols provide an important clue in identifying the NEs.
We define several features such as: *InitialCapital, DigitAlpha, DigitOnly, Hyphen, CapsAndDigit, StopWord, AlphaDigit, AllCapitals, CapitalMixAlphabet, AlphaDigitAlpha, LowMixAlphabet*. We provide an example for the orthographic features in Table 2.

3.3 Proposed objective functions

This section discusses about various objective functions that we propose based on the information theoretic measures such as Entropy, Mutual Information, Correlation, Information Gain. In total, we define seven objective functions which are proved to be very useful in feature selection.

– *Objective function-1* We exploit correlation coefficient-based informative measure to derive an objective function. It follows the property: “good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other.” [23]. Correlation-based Feature Selection (CFS) evaluates a feature subset by assessing the predictive ability of each feature individually and also its degree of redundancy. Moreover, it also introduces the ‘heuristic merit’ for a feature subset instead of individual feature independently [56]. Considering this, we compute our first objective function as:

Table 2 Example of orthographic, morphologic, and prefix suffix features

Feature	Example	Feature	Example
INITCAPS	Fibrin	HAS_SLASH	P42/44
ALLCAPS	SGPT	HAS_QUOTE	gstC' mutans
UPPER-LOWER	Serum ACTH	END_PLUS	HexA+
TWOCAPS	LH	END_QUOTE	C'
THREECAPS	HMG	INITDASH	−beta
MIXEDCAPS	EcoRI	ENDDASH	CD45−
LOWERCASE	Calcitonin	2PREFIX	Fi(fibrin)
ENDDIGIT	cna1	3PREFIX	Fib(fibrin)
ALPHANUMERIC	p53	2SUFFIX	in(fibrin)
SINGLECHAR	R	3SUFFIX	rin(fibrin)
NUMBERS_LETTERS	P42	PUNCTUATION	(,).,
ROMAN	I, II, IV	HASROMAN	factor II

$$f_1(.) = \frac{\sum_{i=1}^{M} \rho_{ic}}{\sum_{i=1}^{M} \sum_{j=i+1}^{M} \rho_{ij}} \tag{2}$$

where ρ_{ic} is the correlation coefficient between feature i and the class label c and ρ_{ij} is the correlation coefficient between features, i and j.

– *Objective function-2* Motivated by the study of Hall [23], we used the objective function utilizing the concept of correlation between the features and the classes. The features having low correlation values with the class are the irrelevant features, and hence can be ignored. We compute the objective function 2 as follows:

$$f_2(.) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{3}$$

where $f_2(.)$ denotes the heuristic value, 'k' represents the feature set selected from the feature subset S. $\overline{r_{cf}}$ represents the correlation between feature and class. $\overline{r_{ff}}$ denotes the average correlation between two features. In the above equation, numerator indicates how a subset of features predicts the class. The denominator quantifies the amount of redundancy that exists between the features.

It chooses the prominent set of features that provides enough clue to distinguish a class from the other. It repeatedly draws a subset of features and, based on its neighbors, it assigns higher weight to the feature that helps to discriminate it from the neighbors of a different class. This helps in determining the (near)-optimal feature subset.

– *Objective function-3* Let, the actual feature set be represented by F. The feature F is considered to be selected if the value of the particular feature in the solution-vector is predicted as 1 by the feature selection approach. Otherwise, the feature is treated to be not selected. In this way, we categorize the feature subsets into two groups: selected feature subset represented by SF and non-selected feature subset denoted as NSF. These subsets should satisfy the following properties:

1. $F \in SF \cup NSF$
2. $SF \cap NSF \in \Phi$

On the basis of these two feature subsets SF and NSF, we define three objective functions [4] as Φ_1, Φ_2, Φ_3. Here Φ_1 represents the average dissimilarity of the selected features. Φ_1 can be stated as the average normalized mutual information between all the probable pairs of features selected. Less value of Φ_1 symbolizes that the selected features are more non-redundant with respect to each other. Φ_1 can be represented as follows:

$$\Phi_1 = \sum_{f_i, f_j \in SF, f_i \neq f_j} \frac{2NMI(f_i, f_j)}{|SF|(|SF| - 1)} \tag{4}$$

where, $NMI(f_i, f_j)$ is the normalized mutual information between two features, f_i and f_j.
Similarly, the similarity among the non-selected features is represented by Φ_2 and it is defined as the average normalized mutual information among the probable pairs of non-selected features. Φ_2 can be represented as follows:

$$\Phi_2 = \sum_{f_i \in NSF, f_j \in SF, f_j = 1NN(f_i)} \frac{NMI(f_i, f_j)}{NSF} \tag{5}$$

Here, $1NN(f)$ gives the first nearest neighbor feature where feature $f \in NSF$. Φ_3 represents the average standard deviation of the selected feature set. As our main motive is to make this objective function very robust, we try to maximize the two objective functions Φ_2 and Φ_3 while minimizing the value of Φ_1. In order to satisfy this, we combine these objective functions into a single objective function $f_3(.)$ represented as follows:

$$f_3(.) = \Phi_3(\Phi_2 - \Phi_1) \tag{6}$$

The main aim is to increase the value of $f_3(.)$. The higher value of this objective function leads to the better feature subset selection.

– *Objective function-4* This objective function utilizes all these three properties of Φ_1 (minimized), Φ_2 (maximized) and Φ_3 (maximized) and is represented as follows:

$$f_4(.) = \Phi_3\Phi_2 - \Phi_1 \tag{7}$$

The values of Φ_2 and Φ_3 should be maximized and the value of Φ_1 should be minimized in order to find out the optimal feature set.

– *Objective function-5* We experiment with one more objective function which is a slight variation of the previous objective function. Here, we did not consider Φ_3. Φ_2 should be maximized and Φ_1 should be minimized. We derive a new objective function which is represented as follows:

$$f_5(.) = \Phi_2 - \Phi_1 \tag{8}$$

– *Objective function-6* We use one more variant of the above objective function taking into the view of increasing the similarity between the non-selected features, Φ_2, and decreasing the dissimilarity value of the selected feature set, Φ_1. Greater the value of this objective function, better is the feature subset obtained. This objective function is represented as follows:

$$f_6(.) = \frac{\Phi_2}{\Phi_1} \tag{9}$$

– *Objective function-7* We also make an objective function (information gain ratio) by leveraging the concept of information gain. It can be defined as the gain over entropy. It enhances the information gain as it provides a normalized score for each feature contribution in a classification decision. The gain ratio is utilized as a disparity measure and the high gain ratio for the selected feature implies that the feature will be useful for classification. It is derived as follows:

$$f_7(.) = \frac{\text{Information Gain}}{\text{Entropy}} \tag{10}$$

3.4 Overview of particle swarm optimization

Particle swarm optimization (PSO) [10] is a stochastic population-based optimization strategy inspired by the social behavior of birds to search for the optimal path. It is a meta-heuristic model where swarm (set of particle) traverse in the search space with some velocity to obtain the best set of solutions. Every particle specifies a solution to the optimization problem.

The algorithm of feature selection that we devise here is founded on the principle of a binary version of PSO [25]. In binary PSO, each particle's position vector is represented by the binary value, i.e., 0 or 1. With the successive generations, the particle updates its position and moves toward the best solution in the search space. The overall process comprises of four steps:

1. Initialization of the population (swarm);
2. Updation of the particle's global best position and self-best position;
3. Updation of velocity vector;
4. Generation of new particles.

Algorithm 1 PSO Initialization

procedure INITIALIZATION($f_k(.), F_{train}$)
  $\overrightarrow{G} \leftarrow NULL$ ▷ Initialize the global best position
  for each particle i in $\overrightarrow{P}$ **do**
    $\overrightarrow{B}(i) \leftarrow NULL$ ▷ Initialize the personal best position
    for each feature (dimension) j in n **do**
      if $rand(0, 1) > 0.5$ **then** ▷ Initialize the particle's position
        $p(i, j) = 1$
      else
        $p(i, j) = 0$
      end if
      $v(i, j) = rand(-1, 1)$ ▷ Initialize the particle's velocity
    end for
  end for
end procedure

Below different steps of the proposed approach are described in detail.

3.4.1 Initialization of the population

Initially, each particle is encoded as the fixed length binary-valued string $\overrightarrow{P}(i) = (p(i, 1), p(i, 2), \ldots, p(i, n))$, where $p(i, j) \in (0, 1), i = 1, 2, \ldots, N$, where N is the

number of particles and $j = 1, 2, \ldots, n$, where n represents the number of features (particle dimension). The values for different bit positions $p_{(i,j)}$ of $\overrightarrow{P}(i)$ are initialized as presented in line no. 6–11 of Algorithm 1. Every particle is associated with its velocity vector $\overrightarrow{V}(i) = (v_{(i,1)}, v_{(i,2)}, \ldots, v_{(i,n)})$. Initially, we randomly set the value of the velocity vector between $(-1, 1)$ (line no. 11 of Algorithm 1). In order to obtain optimal solution, each particle also keeps track of the two variables:

1. Personal best position ($\overrightarrow{B}(i)$) which represents the best solution attained by the particle so far.
2. Global best position ($\overrightarrow{G}$) which keeps track of the best solution of entire swarm.

Initially, both $\overrightarrow{B}(i)$) and $\overrightarrow{G}$ are set to $NULL$.

3.4.2 Evaluation of goodness measure

Each PSO-selected feature subset is evaluated on the informational theoretic-based objective measures (c.f. Sect. 3.3). We have denoted the evaluation of goodness measure with the $fitness(particle\_position, obj\_func, train\_data)$. The evaluation of a particle's goodness depends on the objective function (obj_func) and the training dataset (features), which is calculated in line no. 6 and 9 of Algorithm 2.

3.4.3 Updation of the global and personal best positions

The personal best position $\overrightarrow{B}(i)$ of particle i is updated when the particle obtains a position $\overrightarrow{P}(i)$, for which the fitness value is greater than the current $\overrightarrow{B}(i)$. Similarly, the global best position $\overrightarrow{G}$ is updated based on the following:

$$fitness(\overrightarrow{B}(i), f_k(.), F_{train}) > fitness(\overrightarrow{G}, f_k(.), F_{train})$$

It is depicted in line no. (6–11) of Algorithm 2.

3.4.4 Updation of the velocity vector

In binary PSO, the particle's velocity plays a crucial role in guiding the particle to traverse in the search space to get close to the possible solution of the target problem by updating its position. Each particle update its velocity according to the line no. 16 of Algorithm 2. Inertia weight ($0 < w < 1$) is set by the user to control the velocity explosion. ϕ_1 and ϕ_2 denote the learning parameter fixed by the user. For updating the velocity, the same strategy is followed.

3.4.5 Selection of the new particles

For each particle i with dimension j, the new particle $p_{i,j}$ can be either 0 or 1 according to the following equation:

$$p_{(i,j)} = \begin{cases} 1 & if\ (rand(0,1) < S(v_{(i,j)})) \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 2 Calculatation of Global best for PSO Algorithm

1: **function** CALCULATEGLOBALBEST(ω, ϕ_1, ϕ_2, MAX_ITER, $f_k(.), F_{train}$)
2: `Initialization(`$f_k(.), F_{train}$`)` ▷ Call the PSO `Initialization()`
3: $iter = 0$ ▷ Initialize the *iteration* counter
4: **while** $iter <=$ **MAX_ITER do**
5: **for each** particle i in $\overrightarrow{P}$ **do**
6: **if** $fitness(\overrightarrow{P}(i), f_k(.), F_{train}) > fitness(\overrightarrow{B}(i), f_k(.), F_{train})$ **then**
7: $\overrightarrow{B}(i) \leftarrow \overrightarrow{P}(i)$ ▷ Update the personal best position
8: **end if**
9: **if** $fitness(\overrightarrow{B}(i), f_k(.), F_{train}) > fitness(\overrightarrow{G}, f_k(.), F_{train})$ **then**
10: $\overrightarrow{G} \leftarrow \overrightarrow{B}(i)$ ▷ Update the global best, F_{train} position
11: **end if**
12: **end for**
13: **for each** particle i in $\overrightarrow{P}$ **do**
14: **for each** feature (dimension) j in n **do**
15: ▷ Update the particle's velocity
16: $v_{(i,j)} = \omega * v_{(i,j)} + \phi_1 * (b_{(i,j)} - p_{(i,j)}) + \phi_2 * (g_{(j)} - p_{(i,j)})$
17: ▷ Update the particle's position based on the updated velocity
18: **if** $rand(0, 1) < 1/(1 + exp(-v_{(i,j)}))$ **then**
19: $p(i, j) = 1$
20: **else**
21: $p(i, j) = 0$
22: **end if**
23: **end for**
24: **end for**
25: $iter = iter + 1$ ▷ Increment the *iteration* counter
26: **end while**
27: **return** $\overrightarrow{G}$
28: **end function**

Algorithm 3 Select Optimal Features

1: **function** GETOPTIMALFEATURESSET(G, F)
2: $optimalFSet = []$ ▷ Initialize optimal feature set
3: **for each** F_i in F **do**
4: $optimalFeatures_i = []$ ▷ Initialize optimal features
5: **for each** j in $\overrightarrow{G}$ **do**
6: **if** j==1 **then**
7: ▷ Append selected features into the optimal feature list
8: $optimalFeatures_i$.append($F_{(i,j)}$)
9: **end if**
10: **end for**
11: ▷ Append optimal features into the optimal feature list
12: $optimalFSet$.append($optimalFeatures_i$)
13: **end for**
14: **return** $optimalFSet$
15: **end function**

$S(.)$ denote the Sigmoid function. The selection of the particle is depicted in line no. (18–22) of the Algorithm 2.

In Algorithm 4, we provide the overall technique of our proposed PSO-based feature selection. The optimal feature set is selected based on the final global best position $\overrightarrow{G}$. This process is described in Algorithm 3.

Algorithm 4 Information Theoretic-PSO-based Feature Selection

INPUT:
Labeled train dataset $D_{train} = \{(x^i_{train}, y^i_{train})\}_{i=1}^{N_{train}}$,
Labeled test dataset $D_{test} = \{(x^i_{test}, y^i_{test})\}_{i=1}^{N_{test}}$,
PSO parameters values: ω, ϕ_1, ϕ_2, MAX_ITER,
Objective functions: $f_k(.)$ such as $\{k \in Z : k >= 1 \text{ and } k <= 7\}$ (c.f. Section 3.3)
OUTPUT:
Performance of entity extractions on D_{test}

procedure EVALUATEENTITYEXTRACTION()
  $F_{train} = [], F_{test} = []$ ▷ Initialize the complete feature set
  $F'_{train} = [], F'_{test} = []$ ▷ Initialize the optimal feature set
  for each x^i_{train} in D_{train} **do**
    features=ExtractFeatures(x^i_{train}) ▷ As discussed in Section 3.2
    ▷ Append the features into complete feature set
    F_{train}.append(features)
  end for
  $\overrightarrow{G} \leftarrow$ `CalculateGlobalBest`(ω, ϕ_1, ϕ_2, MAX_ITER, $f_k(.)$, F_{train})
  ▷ Optimal feature set for training data based on global best position
  F'_{train}= `GetOptimalFeaturesSet`($\overrightarrow{G}$, F_{train})
  ▷ Optimal feature set for test data based on global best position
  F'_{test}= `GetOptimalFeaturesSet`($\overrightarrow{G}$, F_{test})
  `CRFModel=CRFClassifier`(F'_{train}, y_{train}) ▷ Train CRF model
  ▷ Use trained CRF model to extract and classify the entities
  $y^{predicted}_{test}$=`CRFModel`(F'_{test})
  ▷ Evaluate the system performance based on gold and predicted entities
  $precision, recall, f$ score = `Evaluate`($y^{predicted}_{test}$, y^{gold}_{test})
end procedure

4 Computational complexity analysis

The total computation cost required to run the PSO-based framework using the individual objective function can be computed in three steps. The very first computation cost $\mathcal{O}(n^2)$ is to evaluate the objective function. The next computation cost is to run the PSO algorithm and finally the complexity to build the model using CRF-based classifier has to be considered. Consider, for $N =$ no. of particles, $n =$ no. of features, $I =$ no. of iterations, the average no. of bits selected for a given particle is F_{avg} . For the S size of training samples and L label set

The total cost of the proposed framework will be $\mathcal{O}((F_{avg} * S^2 * L^2) + (I * N * n^2))$ in our case, $S >> n$, i.e., the size of training set, S, is much larger than the total feature set n. Therefore, the dominant computational complexity comes from the CRF algorithm. But this is required in case of any classification model built using CRF. Moreover, the cost of building the model using the full features set is $(\mathcal{O}(n * S^2 * L^2))$ which is much higher than the cost of building the model using the reduced number of features obtained by the proposed feature selection technique $(\mathcal{O}((F_{avg} * S^2 * L^2)))$. Using the proposed filter-based feature selection model complexity has reduced significantly as compared to the wrapper-based feature selection model. Thus, the gain in terms of time complexity is significant.

5 Datasets, experiments, and analysis

This section provides an overview of datasets, evaluation scheme, experimental results and thorough analysis of the results. The task is formulated as a sequence labeling problem, and we use CRF [32] for the experiments. For our implementation, we use a $C++$ based $CRF++$ package.[2] To properly denote the boundaries of NE, we follow the BIO encoding scheme, where B, I and O denote the beginning, intermediate and outside of NE tokens, respectively.

5.1 Datasets

We perform our experiments on four biomedical benchmark datasets, namely GENIA version 3.02 corpus, GENETAG, AIMed, and BC-II. GENIA corpus was derived from MEDLINE using the MeSH terms such as *'human'*, *'blood cells'* and *'transcription factors'*. Datasets used while training & testing have been extracted from the GENIA version 3.02 corpus.

The NE types to be identified from this dataset are 'DNA', 'RNA', 'Cell line', 'Cell-type', 'Protein'.

The AIMed corpus contains protein–protein interaction information. The aim here is to identify and classify the entity as NE of type 'Protein'. This dataset is developed from Database of Interacting Protein (DIP), containing a total of 197 abstracts.

The other dataset we explored was GENETAG dataset derived from 'MedTag'. It consists of correct and incorrect genes and protein names. For the evaluation of our approach on GENETAG datasets, we consider the dataset available at.[3] Genes described in the datasets were annotated with 'NEWGENE' tag and the overlapping genes are annotated by another term, namely 'NEWGENE1'.

The dataset consists of total 20,000 sentences with the gene/protein mention, where 7500, 2500, and 5000 sentences are used for training, validation, and test set.

We also evaluate our proposed approach on widely adopted BioCreative II GENE Mention dataset. The dataset consists of MEDLINE abstracts which were manually annotated for gene mention. The dataset consists of total 15,000 sentences, where 7500 sentences are used as a training set, 2500 sentences as validation set, and 5000 sentences were utilized for the testing.

5.2 Evaluation scheme

We evaluate our systems on precision, recall, and F score which are defined as follows:

$$precision = \frac{|\{\text{Ground truth NE chunks}\} \cap \{\text{System predicted NE chunks}\}|}{|\{\text{System predicted NE chunks}\}|} \tag{11}$$

$$recall = \frac{|\{\text{Ground truth NE chunks}\} \cap \{\text{System predicted NE chunks}\}|}{|\{\text{Ground truth NE chunks}\}|} \tag{12}$$

F score metric value is calculated on the precision and recall values as follows:

$$F_\beta = \frac{(1+\beta^2)(recall * precision)}{\beta^2 * precision + recall} \tag{13}$$

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

[3] ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/.

Table 3 Results of baseline systems on all the biomedical datasets

Objective dataset	GENIA	GENETAG	AIMed	BC-II
Baseline-1				
No. of features	57	57	57	57
Recall	70.19	95.70	88.32	80.57
Precision	67.31	80.83	87.64	86.20
F score	68.72	87.64	87.98	83.29
Baseline-2				
No. of features	57	57	57	57
Recall	70.41	88.08	88.94	81.29
Precision	67.52	88.68	88.04	90.63
F score	68.93	88.38	88.49	85.71

Here $\beta = 1$, we consider the script available here,[4] for the evaluation on all three datasets. This script is the updated form of CONLL-2003 shared task [54] evaluation script. The script generates three different types of F measures according to the matching of exact, right and left boundary. For evaluating BC-II dataset, we used the same script as released by Biocreative shared task.[5]

5.3 Experimental results

We define the following two baseline systems to compare our proposed feature selection approach.

- *Baseline 1* This baseline model is developed by training CRF with the full feature set as discussed in subsection 3.7.
- *Baseline 2* This baseline model is built by manually varying the contextual feature combination.[6]

Results of these two baselines are reported in Table 3. For our rest of the experiments, we have performed threefold cross-validation on the training data to set the values of the PSO parameters. We present in Table 4 the values of different parameters used in the experiments.

Thereafter, we evaluate our proposed approach on these four datasets.

5.3.1 Analysis of results

We carried out deep analysis of the results both in terms of the performance (F score) (c.f. Table 5) and the number of features selected for all the datasets (c.f. Table 6). The obtained results (Table 5) show that our proposed approach significantly outperforms the baseline systems. The objective functions $f_3(.)$, $f_4(.)$, $f_5(.)$, and $f_6(.)$ which are based on the concept of normalized mutual information outperform in terms of F score value on the GENIA, AIMed, GENETAG, and BC-II datasets, respectively, compared to the objective functions based on the correlation ($f_1(.)$, $f_2(.)$) and information gain ($f_7(.)$). Moreover, in terms of the number of features selected, objective function $f_2(.)$ outperforms other objective functions

[4] http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html.

[5] https://sourceforge.net/projects/biocreative/files/biocreative2entitytagging/1.1/.

[6] The details are provided in Supplementary.

Table 4 Results of the proposed approach with different PSO parameter settings on validation dataset

PSO-RUN	Parameter settings			GENIA	GENETAG	AIMed	BC-II
	Inertia weight	ϕ_1	ϕ_2	F score	F score	F score	F score
PSO-1	0.7298	1.49618	1.49618	73.57	90.81	91.73	87.44
PSO-2	0.3925	2.5586	1.3358	73.18	90.09	90.51	88.03
PSO-3	−0.4349	−0.6504	2.2073	72.94	90.13	89.92	88.61
PSO-4	0.4091	2.1304	1.0575	73.15	89.54	88.93	85.09
PSO-5	−0.3593	−0.7238	2.0289	71.08	88.51	88.26	87.24

Table 5 Results of the proposed approach on the defined objective functions for each dataset

Objective function	# of selected feature				F score			
	GENIA	GENETAG	AIMed	BC-II	GENIA	GENETAG	AIMed	BC-II
$f_1(.)$	22	9	16	19	73.65	90.60	88.65	86.52
$f_2(.)$	19	5	13	15	73.65	88.42	88.94	87.11
$f_3(.)$	22	22	21	18	74.49	89.22	88.95	87.59
$f_4(.)$	29	19	27	24	73.93	88.24	90.47	87.70
$f_5(.)$	22	25	27	20	74.21	91.11	88.82	88.64
$f_6(.)$	26	21	26	25	74.08	90.13	89.42	88.61
$f_7(.)$	27	24	21	22	74.41	87.98	88.94	87.19

by selecting 19, 5, 13, 15 features from GENIA, AIMED, GENETAG, and BC-II dataset, respectively. From the evaluation results, we observe the followings:

- The feature subset having a high NMI with respect to the target output is likely to reduce the uncertainty on the values taken by the output.
- NMI-based objective functions were able to detect the nonlinear relationships between the variables while the correlation derived objective functions are restricted to only linear dependencies.
- NMI-based objective function is found to be more relevant for the jointly redundant or relevant features which make univariate criteria useless.
- The objective functions formulated on correlation and information gain allow overestimation of the relevance of some features.
- As NMI derived objective measures is the KL distance between the joint density and the product of the individual densities. Therefore, NMI can measure non-monotonic relationships and other more complicated relationships, when compared to correlation-based measures.

From the obtained results and the above-mentioned claims, NMI derived objective function $f_6(.)$ can be selected among 7 proposed objective functions. Though the use of $f_6(.)$ does not lead to the highest performance for all the datasets but the obtained results by $f_6(.)$ as the objective function are consistent and are very near to the optimal solution. Note that $f_5(.)$ and $f_6(.)$ are two standard ways of combining two functions, ϕ_1 and ϕ_2. In order to select a good subset of feature values, ϕ_1 should be minimized and ϕ_2 should be maximized. Our ultimate goal is to select the optimal feature set by maximizing the given objective functions. The objective function $f_5(.)$ can lead to zero value ($\phi_2 \approx \phi_1$) and can also have negligible

Table 6 Optimal feature selected through the proposed PSO-based framework on all the datasets

Dataset	Features Context features (1–6)	Content features (7–16)	Orthographic features (17–34)	Prefix (35–38)	Suffix (39–42)	Word normalization (45–46)	Informative word (47–50)	Dynamic NE tag (51)	Word length (52)	Infrequent word (53)	Head noun (54)	Verb trigger (55)	PoS information (56)	Chunk information (57)
GENIA	1, 2	8, 15	18, 19	37	42	45, 46	48	–	–	–	✓	–	✓	✓
	6	16	21, 25											
			28, 31											
			33, 34											
GENETAG	1, 3	7, 9	18, 20	36, 38	41, 42	46	50	✓	–	–	✓	✓	✓	✓
	6	10, 12	22, 31											
		13, 16	34											
AIMed	1, 3	11, 12	17, 18	–	38, 41	46	–	–	✓	-	✓	✓	✓	✓
	6	13, 14	20, 23											
		15, 16	24, 25											
			26, 29											
			32, 33											
BC-II	1, 2	8, 13	18, 20	38	39	–	✓	-	✓	–	–	✓	✓	✓
	5	14, 16	26, 28											
			31, 33											

The feature indices are presented within the brackets. The features selected are listed by their index numbers. For example, feature index 51, '✓' denotes the Dynamic NE tag feature has been selected. Here '–' represent the feature that is not selected by the proposed feature selection technique

Table 7 p Value for different objective functions on each datasets

Objective functions	GENIA	GENETAG	AIMed	BC-II
$f_1(.)$	0.02405	0.03123	0.01747	0.04520
$f_2(.)$	0.01141	0.02289	0.01977	0.01304
$f_3(.)$	0.01747	0.01001	0.00178	0.01772
$f_4(.)$	0.00863	0.01952	0.02789	0.02119
$f_5(.)$	0.02778	0.03005	0.00144	0.02586
$f_6(.)$	0.03811	0.04178	0.01477	0.01952
$f_7(.)$	0.04430	0.03910	0.02967	0.02248

effect ($\phi_2 \gg \phi_1$),*but this is not the problem with objective function*, $f_6(.)$. *Because of these reasons, we obtain consistent performance by* $f_6(.)$ *over various datasets.*

5.4 Statistical significance test

We carried out the statistical significance test on the obtained results (through each objective functions) in the PSO-based framework. For different datasets and for each objective function, experiments were executed for 20 independent runs and we utilized t statistic to examine the obtained results. Table 7 shows the t test results, where p value is the probability of the improvement to occur by chance. For all the datasets, the obtained p value is less than 0.05. It verifies that the model is statistically significant and is unlikely that the improvement in classification accuracy happened by chance.

6 Comparative analysis with state-of-the-art techniques

In this section, we present the comparisons of our proposed NER system with the existing state-of-the-art techniques. We compare our system with both domain-dependent and -independent techniques. For our experiments, we have considered only PoS and chunk information as the domain-dependent features. As evident from Table 8, our system almost outperforms existing state-of-the-art techniques. However, our system was unable to beat Danger et al. [8]. They have adapted supervised machine learning techniques using handcrafted features, which cover both domain-dependent as well as domain-independent features. In their first phase, system was learned on CRF with the prominent domain-independent feature set such as word shape, brief word shape; affixes; PoS; and chunk information. The obtained F score with this phase was reported as 71.29%. In the second phase, they incorporated some domain-dependent features (cell line lexicon) with other handcrafted features such as the DNA sequence, head noun, distance to the head noun, roman, Greek, leading to F score improvement by 5.16 points. In contrast, our proposed system was generated using only few PSO-selected features which are domain independent in nature and the system also does not rely on any gazetteer. The success of Danger et al. system is attributed to the use of the cell-lexicon whose removal leads to obtaining 71.29% F score.

We further compare our system with the existing techniques on the GENETAG, AIMed, and BC-II datasets as reported in Tables 9, 10, and 11, respectively. The proposed NER system utilizing PSO-based feature selection outperforms the other state-of-art systems on all the datasets except AIMed. On AIMed dataset, our system lacks by 3.13 F score points to [13]) system. This is because of different parameter setting (population size). [13]) used genetic

Table 8 Comparisons with state-of-the-art techniques: GENIA dataset

System	Approach	Domain knowledge	F score
Proposed approach	PSO + information theoretic measure + CRF	PoS	74.49
Danger et al. [8]	CRF	Cell line lexicon, PoS	76.45
GuoDong and Jian [19] Final	HMM & SVM	Name duplication, cascaded NEs dictionary, PoS	72.55
GuoDong and Jian [19]	HMM & SVM	POS, phrase	64.1
Kim et al. [27]	Two-phase model with ME and CRF	PoS, rule-based component	71.19
Park et al. [37]	ME	PoS, domain-salient words using WSJ,phrase, morphological patterns, collocations from Medline	66.91
Finkel et al. [15]	ME	Gazetteers, PoS	70.06
Settles [46]	CRF	PoS, semantic lexicons	70.00

Table 9 Comparisons with state-of-the-art techniques: GENETAG dataset

System	Approach	Domain knowledge	F score
Proposed approach	PSO + CRF	PoS	91.11
Kinoshita et al. [28]	Trigrams tags	postprocessing + PoS	80.9
Mitsumori et al. [36]	SVM	dictionary (names of protein and gene)	78.09
Finkel et al. [16]	ME + postprocessing	–	82.2
McDonald and Pereira [35]	CRF	–	82.4
GuoDong et al. [55]	HMM, SVM, Ensemble technique	Postprocessing	82.58

Table 10 Comparisons with state-of-the-art techniques: AIMed dataset

System	Approach	Domain knowledge	F score
Proposed approach	PSO + CRF	PoS	90.47
Ekbal et al. [13]	Feature selection (SVM and GA)	PoS	93.60
Sikdar et al. [47]	Multiobjective DE-based ensemble	PoS	90.56

algorithm-based framework on the population size of 200. In contrast, a very less population size (20) has been used in our proposed framework. We also performed the experiment with the same population size as 20 for GA, and achieved the comparable performance (90.35 F score).

7 Error analysis

Here, we analyze the results to get an idea of the possible errors. We made the following observations:

Table 11 Comparisons with state-of-the-art techniques: BC-II

System	Approach	Domain knowledge	*F* score
Proposed approach	PSO + CRF	PoS	88.64
Li et al. [34]	Extended recurrent Neural network	–	81.87
Tang et al. [52]	Word embedding + CRF	–	80.96
RK Ando [1]	Semi-supervised learning alternating structure optimization (ASO)	Word strings and character types of the current and neighboring words, domain lexicon lookup	87.21
Kuo et al. [31]	CRF	PoS abbreviations of biological chemical compounds	86.83

– *Boundary detection problem* We observed that our system on GENIA and GENETAG datasets suffer from the problem of the boundary detection. The classifier is largely confused among the classes: 'I-PROTEIN' and 'B-PROTEIN', where 164 instances were wrongly classified. The misclassifications of NEs to "Other-than-NE" class amounts to 1305 instances. For the GENETAG dataset, we analyze that majority of the classes were wrongly predicted as 'I-NEWGENE'. In total, 487 instances were misclassified as 'I-NEWGENE'.
  The main cause of boundary detection problem is due to the descriptive naming convention, especially in case of the entity type Protein". One of the misclassified examples is "T cell activation-specific enhancer, where the boundary was not detected properly. It is even hard for the biologist to identify that the descriptive words like "normal, activation" would be a part of entity.
  NE disambiguation, often, is a problem for improper identification of the NEs. For example, the NE "T3 binding sites" is a protein term and was ambiguated by the NE term "DNA" which is not acceptable. The system had difficulties in identifying NEs containing parentheses.
– *Short words* This error was mostly predominant in AIMed. These were mainly misclassified as a part of NEs. The probable reason behind this might be that in training data, many short words appear in the training as part of the NEs, and our model fails in identifying the context. It was also observed that our system lacks in identifying the instances with lowercase or symbols which are therefore tagged as "Other-than-NE".
– *Acronyms* These words either refer to the non-gene entity acronyms or some value. For example, 'HAT' is the abbreviation for the entity name "hepatic artery thrombosis" but actually it was referring to "histone acetyl transfer", a non-entity name. Errors were encountered due to the false negative cases, where gene names in the test set were not known. The classifier lacks in knowledge and sufficient contextual clues.

8 Conclusion

In this paper, we have proposed a novel filter-based method for feature selection using information theoretical concepts for solving NER task in multiple biomedical corpora. In particular, we have used PSO as an optimization technique and determine the best fitting feature set for the problems. The main focus of this paper was to compare the several information theory-based objective functions in the PSO-based feature selection framework. To

this end, we have defined seven goodness measures that were highly effective in identifying features relevant for solving the NER problems in biomedical domain. We have exploited the concept of normalized mutual information, gain ratio, and correlation to design our objective functions.

We have evaluated the proposed technique on multiple biomedical corpora. As a base classifier, we have used CRF. Experimental results show that our models achieve good performance levels for all the datasets without using heavy domain-specific resources and/or tools. The obtained results by the proposed method are as good as the state of the arts, and the most appealing characteristic of the proposed method is that we are able to reduce the complexity of the model significantly by minimizing the use of features. Comparisons among several objective functions reveal that information gain-based metric is very helpful in determining the best subset of features.

Correlation as the objective function was observed to be significant in improving the accuracy. In future, we would like to build a feature selection technique by optimizing all the objective functions simultaneously using the concept well known as multiobjective optimization.

References

1. Ando RK (2007) Biocreative II gene mention tagging system at IBM watson. In: Proceedings of the second biocreative challenge evaluation workshop, Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, vol 23, pp 101–103
2. Aronson AR (2001) Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of the AMIA symposium, American Medical Informatics Association, p 17
3. Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. Ann Math Stat 41(1):164–171
4. Bhadra T, Bandyopadhyay S (2015) Unsupervised feature selection using an improved version of differential evolution. Expert Syst Appl 42(8):4042–4053
5. Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 32(suppl 1):D267–D270
6. Cortes C, Vapnik V (1995) Support vector machine. Mach Learn 20(3):273–297
7. Cover TM, Thomas JA (1991) Elements of information theory. Wiley-Interscience, New York
8. Danger R, Pla F, Molina A, Rosso P (2014) Towards a protein-protein interaction information extraction system: recognizing named entities. Knowl Based Syst 57:104–118
9. Deb K, Agrawal S, Pratap A, Meyarivan T (2000) A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: International conference on parallel problem solving from nature. Springer, pp 849–858
10. Eberhart RC, Kennedy J (1995) A new optimizer using particle swarm theory. Proceedings of the sixth international symposium on micro machine and human science, New York, NY 1:39–43
11. Ekbal A, Saha S (2013) Stacked ensemble coupled with feature selection for biomedical entity extraction. Knowl Based Syst 46:22–32
12. Ekbal A, Saha S, Garbe CS (2010) Feature selection using multiobjective optimization for named entity recognition. In: 20th international conference on pattern recognition (ICPR), 2010. IEEE, pp 1937–1940
13. Ekbal A, Saha S, Sikdar UK (2013) Biomedical named entity extraction: some issues of corpus compatibilities. SpringerPlus 2(1):1
14. Ekbal A, Saha S, Bhattacharyya P et al (2016) A deep learning architecture for protein-protein interaction article identification. In: 23rd international conference on pattern recognition (ICPR), 2016. IEEE, pp 3128–3133
15. Finkel J, Dingare S, Nguyen H, Nissim M, Manning C, Sinclair G (2004) 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. Association for Computational Linguistics, pp 88–91
16. Finkel J, Dingare S, Manning CD, Nissim M, Alex B, Grover C (2005) Exploring the boundaries: gene and protein identification in biomedical text. BMC Bioinf 6(Suppl 1):S5

17. Friedrich CM, Revillion T, Hofmann M, Fluck J (2006) Biomedical and chemical named entity recognition with conditional random fields: the advantage of dictionary features. In: Proceedings of the second international symposium on semantic mining in biomedicine (SMBM 2006), vol 7. BioMed Central Ltd, London, UK, pp 85–89
18. Gheyas IA, Smith LS (2010) Feature subset selection in large dimensionality domains. Pattern Recognit 43(1):5–13
19. GuoDong Z, Jian S (2004) Exploring deep knowledge resources in biomedical name recognition. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics, pp 96–99
20. Gupta D, Tripathi S, Ekbal A, Bhattacharyya P (2016) A hybrid approach for entity extraction in code-mixed social media data. MONEY 25:66
21. Gupta DK, Reddy KS, Ekbal A et al (2015) Pso-asent: Feature selection using particle swarm optimization for aspect based sentiment analysis. In: International conference on applications of natural language to information systems. Springer, pp 220–233
22. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422
23. Hall MA (1999) Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato
24. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J (2005) Prominer: organism-specific protein name detection using approximate string matching. BMC Bioinf 6(Suppl 1):S14
25. Kennedy J, Eberhart R (1997) A discrete binary version of the particle swarm algorithm. In: 1997 IEEE international conference on systems, man, and cybernetics, 1997. Computational cybernetics and simulation, vol 5, pp 4104–4108
26. Kim JD, Ohta T, Tsuruoka Y, Tateisi Y, Collier N (2004) Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics, pp 70–75
27. Kim S, Yoon J, Park KM, Rim HC (2005) Two-phase biomedical named entity recognition using a hybrid method. In: Natural language processing–IJCNLP 2005. Springer, pp 646–657
28. Kinoshita S, Cohen KB, Ogren PV, Hunter L (2005) Biocreative task1a: entity identification with a stochastic tagger. BMC bioinf 6(Suppl 1):S4
29. Kittler J (1978) Feature set search algorithms. In: Chen CH (ed) Pattern recognition and signal processing. Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, pp 41–60
30. Kumar A, Ekbal A, Saha S, Bhattacharyya P et al (2016) A recurrent neural network architecture for de-identifying clinical records. In: Proceedings of the 13th international conference on natural language processing, pp 188–197
31. Kuo CJ, Chang YM, Huang HS, Lin KT, Yang BH, Lin YS, Hsu CN, Chung IF (2007) Rich feature set, unification of bidirectional parsing and dictionary filtering for high f-score gene mention tagging. In: Proceedings of the second biocreative challenge evaluation workshop. Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, vol 23, pp 105–107
32. Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML, pp 282–289
33. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C (2016) Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360
34. Li L, Jin L, Jiang Z, Song D, Huang D (2015) Biomedical named entity recognition based on extended recurrent neural networks. In: IEEE international conference on bioinformatics and biomedicine (BIBM), 2015. IEEE, pp 649–652
35. McDonald R, Pereira F (2005) Identifying gene and protein mentions in text using conditional random fields. BMC Bioinf 6(Suppl 1):S6
36. Mitsumori T, Fation S, Murata M, Doi K, Doi H (2005) Gene/protein name recognition based on support vector machine using dictionary as features. BMC Bioinf 6(Suppl 1):S8
37. Park KM, Kim SH, Rim HC, Hwang YS (2006) Me-based biomedical named entity recognition using lexical knowledge. ACM Trans Asian Lang Inf Process (TALIP) 5(1):4–21
38. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
39. Ponomareva N, Pla F, Molina A, Rosso P (2007) Biomedical named entity recognition: a poor knowledge hmm-based approach. In: Natural language processing and information systems. Springer, pp 382–387
40. Ramadan RM, Abdel-Kader RF (2009) Face recognition using particle swarm optimization-based selected features. Int J Signal Process Image Process Pattern Recognit 2(2):51–65

41. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L (2000) Edgar: extraction of drugs, genes and relations from the biomedical literature. In: Pacific symposium on biocomputing. Pacific Symposium on Biocomputing, NIH Public Access, p 517
42. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. bioinformatics 23(19):2507–2517
43. Saha S, Ekbal A, Sikdar UK (2015) Named entity recognition and classification in biomedical text using classifier ensemble. Int J Data Min Bioinf 11(4):365–391
44. Saha SK, Sarkar S, Mitra P (2009) Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inf 42(5):905–911
45. Segura-Bedmar I, Martínez P, Segura-Bedmar M (2008) Drug name recognition and classification in biomedical texts: a case study outlining approaches underpinning automated systems. Drug Discov Today 13(17):816–823
46. Settles B (2004) Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics, pp 104–107
47. Sikdar UK, Ekbal A, Saha S (2015) Mode: multiobjective differential evolution for feature selection and classifier ensemble. Soft Comput 19(12):3529–3549
48. Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, Lin YS, Klinger R, Friedrich CM, Ganchev K (2008) Overview of biocreative ii gene mention recognition. Genome Biol 9(Suppl 2):S2
49. Tanabe L, Wilbur WJ (2002) Tagging gene and protein names in biomedical text. Bioinformatics 18(8):1124–1132
50. Tang B, Cao H, Wu Y, Jiang M, Xu H (2012) Clinical entity recognition using structural support vector machines with rich features. In: Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics. ACM, pp 13–20
51. Tang B, Cao H, Wang X, Chen Q, Xu H (2014) Evaluating word representation features in biomedical named entity recognition tasks. BioMed Res Int 2014: https://doi.org/10.1155/2014/240403
52. Tang B, Cao H, Wang X, Chen Q, Xu H (2014) Evaluating word representation features in biomedical named entity recognition tasks. BioMed Res Int
53. Thang ND, Lee YK et al (2010) An improved maximum relevance and minimum redundancy feature selection algorithm based on normalized mutual information. In: 10th IEEE/IPSJ international symposium on applications and the internet (SAINT), 2010. IEEE, pp 395–398
54. Tjong Kim Sang EF, De Meulder F (2003) Introduction to the Conll-2003 shared task: language-independent named entity recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, vol 4. Association for Computational Linguistics, pp 142–147
55. Wang H, Zhao T, Tan H, Zhang S (2008) Biomedical named entity recognition based on classifiers ensemble. IJCSA 5(2):1–11
56. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW (2005) Gene selection from microarray data for cancer classificationa machine learning approach. Comput Biol Chem 29(1):37–46
57. Yadav S, Ekbal A, Saha S, Bhattacharyya P (2016) Deep learning architecture for patient data de-identification in clinical records. In: Proceedings of the clinical natural language processing workshop (ClinicalNLP), pp 32–41
58. Yadav S, Ekbal A, Saha S (2017a) Feature selection for entity extraction from multiple biomedical corpora: a PSO-based approach. Soft Comput. https://doi.org/10.1007/s00500-017-2714-4
59. Yadav S, Ekbal A, Saha S (2017b) Feature selection for entity extraction from multiple biomedical corpora: a PSO-based approach. Soft Comput 21:1–24
60. Yadav S, Ekbal A, Saha S, Bhattacharyya P (2017c) Entity extraction in biomedical corpora: An approach to evaluate word embedding features with pso based feature selection. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: volume 1, Long Papers, vol 1, pp 1159–1170
61. Yadav S, Ekbal A, Saha S, Pathak PS, Bhattacharyya P (2017d) Patient data de-identification: a conditional random-field-based supervised approach. In: Handbook of research on applied cybernetics and systems science. IGI Global, pp 234–253
62. Yadav S, Ekbal A, Saha S, Bhattacharyya P, Sheth A (2018a) Multi-task learning framework for mining crowd intelligence towards clinical treatment. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 2 (short papers), vol 2, pp 271–277
63. Yadav S, Kumar A, Ekbal A, Saha S, Bhattacharyya P (2018b) Feature assisted bi-directional LSTM model for protein–protein interaction identification from biomedical texts. arXiv preprint arXiv:1807.02162
64. Zhang S, Elhadad N (2013) Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inf 46(6):1088–1098

65. Zhang Y, Wang S, Phillips P, Ji G (2014) Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowl Based Syst 64:22–31
66. Zhao S (2004) Named entity recognition in biomedical texts using an hmm model. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics, pp 84–87

Shweta Yadav Shweta Yadav is a Ph. D. student in the Department of Computer Science and Engineering, Indian Institute of Technology Patna, India. Her current research interests include natural language processing, information extraction, machine learning applications, and text mining in the biomedical and clinical domain. In these areas, she has authored or coauthored 14 papers in the journals like Soft Computing and Knowledge and Information System and took part in conferences like EACL, NAACL, ICPR, NLDB, and LREC.

Asif Ekbal Asif Ekbal is an Associate Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Patna, India. Prior to joining in IIT Patna, he worked as a post-doctoral research fellow in University of Trento, Italy, and Heidelberg University, Germany. His current research interests include natural language processing, information extraction, machine learning applications, opinion mining, and text mining. In these areas, he has authored or coauthored around 100 papers in the journals like ACM TALIP, Knowledge Based Systems, and Knowledge Engineering and took part in conferences like ACL, COLING, EACL, IJCNLP, and ECAI. Google Scholar citation, which is the benchmark of Computer Science, shows his citation count of 1593 with h5-index of 22. He is the recipient of the Best Innovative Project Award from the Indian National Academy of Engineering (INAE), JSPS Invitation Fellowship from the Govt. of Japan and Visesvaraya Young Faculty Research Fellowship Award from Govt of India.

Sriparna Saha Dr. Sriparna Saha received M.Tech and Ph. D. degrees in computer science from Indian Statistical Institute Kolkata, Kolkata, India, in 2005 and 2011, respectively. She is currently an Associate Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Patna, India. She is the author of a book published by Springer-Verlag. She has authored or coauthored more than 120 papers. Her current research interests include pattern recognition, multiobjective optimization, and biomedical information extraction. Her h-index is 20, and the total citation count of her papers is 2548 (according to Google Scholar). She is also a senior member of IEEE. She is the recipient of the Google India Women in Engineering Award 2008, NASI YOUNG SCIENTIST PLATINUM JUBILEE AWARD 2016, BIRD Award 2016, IEI Young Engineers' Award 2016, SERB WOMEN IN EXCELLENCE AWARD 2018, and SERB Early Career Research Award 2018. She is the recipient of Humboldt Research Fellowship, Indo-U.S. Fellowship for Women in STEMM (WISTEMM) Women Overseas Fellowship program 2018, and CNRS fellowship. She had also received India4EU fellowship of the European Union to work as a Post-doctoral Research Fellow in the University of Trento, Italy, from September 2010 to January 2011. She was also the recipient of Erasmus Mundus Mobility with Asia (EMMA) fellowship of the European Union to work as a Post-doctoral Research Fellow in the Heidelberg University, Germany, from September 2009 to June 2010.