Entity Extraction in Biomedical Corpora: An Approach to Evaluate Word Embedding Features with PSO based Feature Selection

Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya
Indian Institute of Technology Patna
Bihar, India
{shweta.pcs14,asif,sriparna,pb}@iitp.ac.in

Abstract

Text mining has drawn significant attention in recent past due to the rapid growth in biomedical and clinical records. Entity extraction is one of the fundamental components for biomedical text mining. In this paper, we propose a novel approach of feature selection for entity extraction that exploits the concept of deep learning and Particle Swarm Optimization (PSO). The system utilizes word embedding features along with several other features extracted by studying the properties of the datasets. We obtain an interesting observation that compact word embedding features as determined by PSO are more effective compared to the entire word embedding feature set for entity extraction. The proposed system is evaluated on three benchmark biomedical datasets such as GENIA, GENETAG and AiMed. The effectiveness of the proposed approach is evident with significant performance gains over the baseline models as well as the other existing systems. We observe improvements of 7.86%, 5.27% and 7.25% F-measure points over the baseline models for GENIA, GENETAG, and AiMed dataset respectively.

1 Introduction

The tremendous amount of information accumulated in the domains of molecular biology has drawn the attention of biomedical natural language processing (BioNLP) community in order to facilitate the development of various tools for various text processing applications, curation and organization of ever-growing biomedical literature etc. Entity extraction is crucial step for solving several pipelined applications such as information extraction, automatic summarization, question-answering, word sense disambiguation etc. Biomedical entities mostly refer to the biological sequences of protein & gene such as DNA, RNA, cell_type, cell_line etc. (Kim et al., 2004). The way of extracting these information from biomedical and clinical texts refers to as entity extraction. An automatic system which can extract biomedical names such as gene, protein or any disease name from text can substantially reduce the human efforts. However, extracting these entities from text poses several challenges which are presented as follows:

1. Named entities are very generative in nature, i.e. many new names are continuously being generated. Any dictionary can not capture all the various forms of a given name.

2. Similar words convey different meanings, and therefore, a word can have multiple NE types. For example, gene names often contain alphabets, digits, hyphens, and other characters, thus having many variants (e.g., "HIV-1 enhancer" versus "HIV 1 enhancer"). Moreover, many abbreviations (e.g., "IL2" for "Interleukin 2") constitute integral parts of biomedical named entities (NEs).

3. Biomedical names are usually of long length, and contains different types of symbols, and hence boundary detection becomes problematic.

4. Ambiguity: Same name could be used to represent variety of biological entities which further worsen the problem.

The challenges as of these kinds are the primary causes behind the low accuracies of the systems developed for entity extraction in biomedi-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1159–1170,
Valencia, Spain, April 3-7, 2017.

cal text. The research challenges have been addressed in the literature including in some shared-task challenges, such as JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and its Applications) in 2004 (Kim et al., 2004) and BioCreative (Critical Assessment for Information Extraction in Biology Challenge) II GM (gene mention) subtask in 2007 (Smith et al., 2008). Over the years several benchmark corpora have been created that do not conform to the uniform annotation guidelines. Therefore the system, developed by targeting a specific domain, often fails to show reasonable accuracy when it is evaluated for some other domains. In our work we attempt to build a system for entity extraction that performs well across various biomedical corpora.

Popular existing system mostly rely on rule-based system or supervised machine learning technique to automatically extract entities. They looked upon this problem as in terms of sequence labeling and used algorithm such as hidden markov models (HMM) (Zhao, 2004), support vector machines (SVM) (Kazama et al., 2002; GuoDong and Jian, 2004), maximum entropy Markov model (MEMM) (Finkel et al., 2005) and conditional random fields (CRF) (Ekbal et al., 2013; Settles, 2004; Kim et al., 2005). These supervised learning models is fully dependent on the features that we use for training. Some of the popular features used in the existing studies include linguistic features such as morphological, syntactic and semantic information of words and domain-specific features from biomedical ontologies such as BioThesaurus (Liu et al., 2006) and UMLS (Unified Medical Language System) (Bodenreider, 2004). However, these features heavenly account to the problem of data sparsity.

In the recent past, there has been huge interest in using large unlabeled corpus to generate word representation feature using deep neural network technique. We are motivated by the strength of deep learning concepts to build our model. We use the well-known word embedding model that is a robust framework to incorporate word representation features (Mikolov et al., 2013b). Word representation feature is a mathematical description of the word in vector form. Each position of vector corresponds to a feature with some semantic or grammatical inference which leads to the term word feature. Word representation features contains latent syntactic/semantic information of a word. The main objective to use word embedding is to provide more useful information to the model being trained. Vector based word representation has powerful capability that captures the phenomenon that words having the similar meanings should appear together (Mikolov et al., 2013b). In traditional machine learning, data sparsity is a problem that often causes the degradation in performance. This drawback could be overcome by the incorporation of word embedding with the presumption that similar type of word (as to semantics) appear in the similar context (Mikolov et al., 2013b).

The aim is to exploit the usefulness of neural network based word embedding (Bengio et al., 2003) as a feature for entity extraction in biomedical text. In addition we also make use of a very diverse feature set that exploits the properties of data and problem specific knowledge. We restrict ourselves from using much domain-specific information for feature extraction, keeping in view easy adaptability of the system to more than one biomedical corpora.

However, the huge dimensionality of the word representation vector often contributes to the complexity of the system. This motivated us to apply feature selection technique to reduce the dimensionality contributed by word embedding as well as to improve the system performance. Our algorithm for feature selection is based on wrapper based approach, which is formulated as an optimization problem. We use Particle Swarm Optimization(PSO) (Kennedy and Eberhart, 1997) as the underlying optimization strategy. Particle Swarm Optimization is an evolutionary technique, inspired by the social behavior of birds. Some recent studies show that PSO converges faster compared to some other widely used optimization techniques (Bansal et al., 2011). Inspired by this observation we use PSO in our current study.

To analyze the effect of pruned word embedding, we have carried out an experiment with all the handcrafted features and the reduced features as determined by PSO. We perform experiments on three standard datasets, namely GENIA, GENETAG and AiMed. Evaluation results show that we achieve significant performance gains with the use of pruned word embedding feature set. The best performance of the system was obtained when we apply PSO based feature selection technique on combination of handcrafted features set and word

embedding features. The key contribution of this paper are, **(i)** proposal of PSO based feature selection technique in bio-medical entity extraction. **(ii)** analysis of feature selection on only word representation features. **(iii)** impact of feature selection on word representation features with handcraft features.

2 Related Works

There has been quite a significant number of existing works available for biomedical named entity recognition (BNER). These approaches can be divided into three major categories: (1) dictionary based, (2) rule based and (3) machine learning based techniques. Among these existing approaches, machine learning based techniques have gained a lot more attention due to the availability of sufficiently good amount of annotated corpus. For example, majority of the systems submitted to the JNLPBA challenge made use of machine learning algorithms which have been observed to significantly outperform the dictionary based methods.

Some of the recent works in BNER includes the unsupervised model as proposed in (Zhang and Elhadad, 2013), and the system based on CRF (Li et al., 2015a). A two-phase approach based on semi-Markov CRF is proposed in (Yang and Zhou, 2014). In the first phase boundaries of entities are identified while in the second phase semantic labeling is performed to label the detected entities. A CRF based system has been proposed by (Tang et al., 2015), where in the first step boundaries of NEs are identified and in the second step appropriate labels are assigned. (Grouin, 2014) performed experiments on the i2b2/VA-2010 challenge dataset to detect bacteria and biotopes names. They developed a model based on CRFs. An unsupervised approach is proposed in (Han et al., 2016) that made use of clustering based active learning. They have used Shared Nearest Neighbor (SNN) clustering technique. The work reported in (Li et al., 2015a), authors have proposed a parallel CRF algorithm (MapReduce CRF) which provides a mechanism to minimise the time taken for CRF learning. They showed that the proposed approach outperforms other traditional models in terms of time and efficiency. While, most of the proposed system used CRF, recently (Patra and Saha, 2013) proposed a an entity extraction system based on SVM. Particularly, they have introduced a tree kernel based function that can efficiently solve the full NER task. The work proposed in (Tohidi et al., 2014) aims to improve the performance of entity extraction using statistical character-based syntax similarity (SCSS) algorithm. This algorithm computes the similarity between the identified candidate entities and a known set of well-known NEs. This set of NEs is created by extracting the most frequently occurring NEs in the GENIA V3.0 corpus. In recent times deep learning based approaches such as Recurrent Neural Network and Bi-directional LSTM have also used for entity extraction(Li et al., 2015b; Limsopatham and Collier, 2016). It is well known that relevant features play an important role for building a high accurate system. In our work, in addition to the standard features we also use the features extracted from the word embedding model.

Bengio et al.(Bengio et al., 2003) have proposed a neural network based model for vector representation of words. Distributed representation (also known as word embedding) of a word has been used to improve the performance of various NLP tasks like Part-of-Speech (POS) tagging, NER in news-wire domain (Collobert et al., 2011), parsing (Socher et al., 2013; Turian et al., 2010) etc. Word cluster has been used used by Miller et al.(Miller et al., 2004) to boost the performance of a NER system. Tang et al. (Tang et al., 2012; Tang et al., 2013) have reported that performance of biomedical entity extraction can be improved when word representation is used as a feature to CRF and SVM classifiers.

Here we propose a PSO based feature selection technique that determines the most relevant features from a full word embedding set, and use this subset as feature for classifier's training. Feature selection has been widely used for many tasks such as gene expression (Ding and Peng, 2005), face recognition (Seal et al., 2015) and signal processing (Alamedine et al., 2013). Dealing with biomedical text is, however, more difficult and challenging as the features have non-numeric values and the texts are heavily unstructured. Except the few works such as NER (Ekbal and Saha, 2016), co-reference resolution (Sikdar et al., 2015) and sentiment analysis (Gupta et al., 2015), systematic methods of feature selection using metaheuristics algorithms are very rare. Nevertheless, the importance of using pruned neural language

model based word representation features with effective feature selection have not been exploited so far in the literature.

2.1 A Brief Introduction to Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is a meta-heuristic intelligent technique inspired by social behavior of the swarm for its survival (Eberhart and Shi, 1998; Kennedy and Eberhart, 1997). This is a population based technique which is perceived in birds and fishes for the search of the best path. In general, PSO consists of the swarm of the particle where each particle has its particular position in the search space with which it moves around the search space by some velocity. The particle selects the best path on each iteration by using its memory and by learning the effective path that was followed previously by the swarm. The new position is chosen on the basis of the knowledge gained previously by its self-best position and the best position of the swarm. PSO, being a meta-heuristic model, makes few or no assumption about the problem being optimized and can search very large spaces of candidate solutions. This makes PSO highly efficient for the optimization purpose (Yan et al., 2013). The algorithm iterates by keeping track of two variables:
Global best position represents the most promising vector found so far, and Personal best position denotes the particle's own personal best solution.

2.1.1 Algorithm: PSO based Feature Selection

1. Initially, we randomly set the swarm population. Each particle of the swarm is represented by binary-valued features of length n (total no. of feature) and has its position and velocity with which it moves in search space. Mathematically, particle position and particle velocity are represented as: $\overrightarrow{P}(i)$ and $\overrightarrow{V}(i)$ respectively:

$$\overrightarrow{P}(i) = (p(i,1), p(i,2), ..., p(i,n))$$

$$\overrightarrow{V}(i) = (v(i,1), v(i,2), ..., v(i,n))$$

where $p(i,j) \in \{0,1\}, i = 1,2,...,N$ and $j = 1,2,...,n$ where N is no. of particle. Particle maintains its best position $(\overrightarrow{B}(i))$ that they have achieved so far and also the global best position $(\overrightarrow{G})$i.e., the best position of the particle having the best solution.

2. Particle's position $\overrightarrow{P}(i)$ value is set either $\{0,1\}$ on the basis of following expression:

$$p_{(i,j)} = \begin{cases} 1 & \text{if} \quad random \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

3. Each particle is evaluated on the basis of fitness function (F-measure value) $f(\overrightarrow{P}(i))$. The memory is updated by keeping track of the best position and global best position.

4. Initially, the value of best position $(\overrightarrow{B}(i))$ of every particle is set to 0. At every epoch(ep) the value of the best position is updated as follows:

$$f(\overrightarrow{B}(i))_{ep} = max(f(\overrightarrow{P}(i))_{ep}, f(\overrightarrow{B}(i)_{ep-1})$$

5. Update in the global best position value is done when the fitness function $f(\overrightarrow{B}(i))$ in the swarm is superior than the existing $f(\overrightarrow{G})$.

6. Originally, the velocity vector is generated randomly. At each iteration, velocity of a particle is updated according to the following equation:

$$v_{(i,j)} = \omega * v_{(i,j)} + \phi_1(b_{(i,j)} - p_{(i,j)}) + \phi_2(g_{(j)} - p_{(i,j)}) \quad (1)$$

where $\omega (0 < \omega < 1)$, ϕ_1 and ϕ_2 are known as inertia weights. These parameters are initialized with an uniformly generated random numbers in the range (0,1). The $b_{(i,j)}$, $p_{(i,j)}$, and $g_{(j)}$ denote the j^{th} components of $\overrightarrow{B}(i)$, $\overrightarrow{P}(i)$ and $\overrightarrow{G}$, respectively.

7. The position of a particle is updated by the following mathematical expression:

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

where 0$\leq$ random $\leq$ 1 is an uniform random number.

$$\mathrm{S}(\mathrm{v}_{(i,j)}) = \frac{1}{1+exp(-\overrightarrow{v}_{(i,j)})}$$

This represents the sigmoid function. Thus, we update the particle position value of 0 or 1 on the basis of the value of velocity.

8. Repeat steps 4-7 until convergence.

2.2 Learning Word Representations

Word embedding (also known as distributed word representations) persuade a real-valued latent semantic or syntactic vector for each word from a large unlabeled corpus by using continuous space language models (Tang et al., 2014). Better word representation can be obtained if we have a large amount of training data as the obtained real-valued vectors of words become more representative. We use the popular *word2vec*[1] tool proposed by Mikolov et al. (Mikolov et al., 2013a) to extract the vector representations of words. Owing to its simpler architecture which reduces the computational complexity, this technique can be used for large corpus. Two models have been proposed in (Mikolov et al., 2013a) to learn vector representation known as Continuous Bag-of-Words Model (CBOW) and Skip-gram model. Since skip-gram model is able to capture the semantic information of a word, we adapt this to train the model for vector representation. The Skip-gram architecture tries to maximize the classification of a word based on the other words in the same sentence. More formally, given a sequence of training words w_1,w_2,....., w_T, the objective of the Skip-gram model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\leq j\leq c\\ j=0}}\log p(w_{t+j}|w_t) \qquad (2)$$

where c is the window size. Here, we show few words that are more nearby to any biomedical entity: *'antigen'*, *'lymphocytes'* and *'inhibited'*. If we look at the most similar words for the word *'lymphocytes'*, we observe that apart from syntactically similar words like *'T-lymphocytes'*, *'B-lymphocytes'*, it is also able to capture the words which are semantically similar like *'CD3+'*, *'PBLs'* and *'T-cells'*.

3 Features for Entity Extraction

The features being extracted are described as follows:

1. Contextual feature: It is the local contextual feature which refers to the tokens which appear within the window size of 10 words, i.e 5 to the left and 5 to the right w.r.t current token.

2. Word prefixes and suffixes: These features refer to the fixed length character sequences stripped either from the left or rightmost positions of the words.

3. Word length: It is observed that short words are rarely the NEs. We define a binary-valued feature that triggers the value 1 if the length of current word is greater than the threshold value specified. The threshold value is set as 5 in this case.

4. Part-of-Speech (PoS) information: PoS provides useful syntactic evidence for detecting named entities (NEs).We use PoS information of the current and/or the surrounding token(s) as the feature. The PoS information was extracted from the GENIA tagger[2] V2.0.2.

5. Chunk information: We use GENIA tagger V2.0.2 corpus to extract the chunk information. We employ the chunk information of the present and neighboring tokens as the features.

6. Word shape: Word shape is defined as the mapping of each word to its equivalent class. In order to implement this feature we normalize the words by converting every capital character by 'A', small character to 'a' and digit to '0'. After this conversion, we squeeze the consecutive characters into a single character. For example, if we consider the token 'Ly-49', the normalized word for this token would be 'Aa-00'.

7. Word class feature : This feature is based on the concept that entities present in the same class are mostly similar. Here, all the capital letters are converted to 'A', small letters to 'a', numbers to 'O' and non-English characters to '-'. After this conversion, we squeeze the consecutive characters into a single character. For example, the word class feature for the token 'IL-2-mediated' is 'AA-O-aaaaaaaa', which is further reduced to 'A-O-a'.

8. Orthographic features: We use several orthographic features that consider capitalization and digit information. These features are:

[1] https://code.google.com/p/word2vec/

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

initial capital, all capital, capital in inner, initial capital then mix, only digits, digit with special character, initial digit then alphabet, digit in inner. It is observed that some symbols like (',', '-', '.', '_') are very common in the biomedical text. Some symbols like ',' are also very helpful for the identification of NE boundaries.

4 Methodology

We propose a PSO based feature selection technique that determines the most relevant features from a set of features, containing both handcrafted as well as word embedding based features. We use Conditional Random Field (CRF) (Lafferty et al., 2001) as a base learning algorithm. For each token, a feature vector is generated from the training and test dataset using the features as described in the previous section. Basic steps of our algorithm are as follows:

1. Initially, we design 32 features (listed in Section-3) for three datasets, namely GENIA, GENETAG and AiMed. These features are used for the classifier's training. The models built using these features are termed as the baseline models.

2. We generate the word embedding feature vector of 200 dimensions based on the model trained on a large corpus like Wikipedia and the biomedical corpora such as PubMed [3] and PubMed Central Open Access (PMC OA)[4].

3. A new feature set is generated by combining both word embedding based features and handcrafted features.

4. PSO based feature selection is performed to determine the most relevant feature set.

5. CRF classifier is trained with the features selected by PSO. The model, thus generated, is evaluated on all the three datasets.

Figure-1 depicts the various steps of our proposed approach.

5 Datasets and Experiments

[3]http://www.ncbi.nlm.nih.gov/pubmed

[4]http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist

5.1 Dataset

Our system is evaluated on three distinguished biomedical datasets, namely GENIA[5], AiMed[6] and GENETAG[7]. The GENIA corpus is derived from the MEDLINE corpus. It comprises of 500,000 and 100,000 words in training and test dataset, respectively. These datasets are manually annotated with five NE tags, namely *Protein*, *DNA*, *RNA*, *Cell_line* & *Cell_type*.

AiMed corpus was created using 20,000 sentences having gene/protein names extracted from the Database of Interacting Protein (DIP). We use 7,500 labeled sentences for training and 2,500 sentences for validation. For evaluation we use a test set consisting of 5,000 sentences.

GENETAG dataset is derived from the 'MedTag' dataset. Training and test datasets comprise of 118K and 142K words, respectively. In order to properly denote the boundaries of NE, we use the IOB2[8] encoding scheme. We evaluate our system in terms of recall, precision and F-measure values. For evaluation we use the script, which was made available with the JNLPBA 2004 shared task [9].

5.2 Baseline Models and Analysis

We start experiments with the first baseline (i.e. Baseline-1) by developing the model trained with all the features as discussed in Section-3. We evaluate the presence of word embedding features trained on various unlabeled data sets obtained from the different text sources. In order to realize the effect of each trained word representation model, we augment the word vector obtained from the respective model one by one to the baseline feature set. In order to obtain word embedding, we use four different models trained on the unlabeled data extracted from PubMed [10], PubMed Central Open Access (PMC OA)[11] and the latest English Wikipedia dump[12]. Corpus statistics of PubMed and PMC OA are provided in Table-1.

The extracted text upon which four different models are trained are as follows:

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

[6]ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/

[7]ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/GENETAG.tar.gz

[8]I, O and B represent the intermediate, outside and beginning token of a NE

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

[10]http://www.ncbi.nlm.nih.gov/pubmed

[11]http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist

[12]http://en.wikipedia.org/wiki/Main_Page

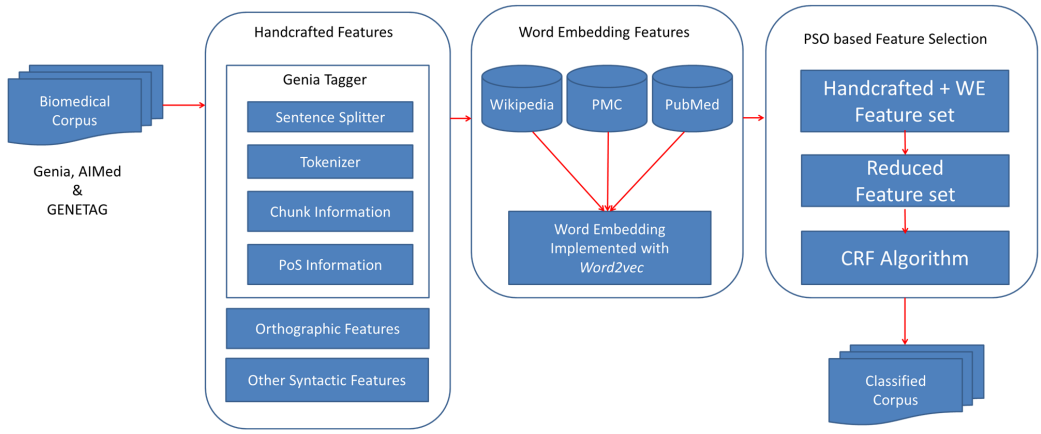

Figure 1: Proposed model architecture for biomedical entity extraction

1. Model developed using extracted data from PubMed biomedical literature: denoted as WE(1).
2. Model built on the extracted text from PMC biomedical literature: denoted as WE(2).
3. Model developed using the combination of extracted text from PubMed and PMC biomedical texts: denoted as WE(3).
4. Model trained using the extracted text from PubMed, PMC and Wikipedia: denoted as WE(4).

We develop the second baseline (i.e. Baseline-2) by executing the best word embedding model in combination with the hand-crafted feature set. We further develop the third baseline, i.e. Baseline-3 by merging word embedding feature set as determined by PSO along with the full handcrafted feature set. We observe that selection of relevant word embedding features helps in improving performance over the whole word embedding feature set.

We generate 200-dimensional word vectors using the parameters [13] as follows: *skip-gram* model with a window size of 5, hierarchical soft-max training, and a frequent word sub-sampling threshold of 0.001. In order to make our proposed system generic, i.e. not biased to any particular domain of data, we use the same parameters of PSO in all our settings. We fine-tune the parameters ω, ϕ_1 and ϕ_2 by performing 3-fold cross validation experiments. We keep the number of particles and the number of iterations as 10 and 100, respectively throughout all the experiments.

[13] We use same parameters for training of all the four models

Corpus	Documents	Sentences	Tokens
PubMed	22,120,269	124,615,674	2,896,348,481
PMC	672,589	105,194,341	2,591,137,744
PubMed+PMC	22,792,858	229,810,015	5,487,486,225

Table 1: Corpus statistics (Pyysalo et al., 2013) of PubMed and PMC OA openly available biomedical literature; PubMed abstracts for articles that are also present in PMC OA were discarded while creating the data

Effectiveness of PSO based feature selection is evident with performance improvement as shown in Table-5.

5.3 Comparison with Existing Feature Selection Techniques

Here we compare our PSO based feature selection technique with other existing feature selection techniques. We perform experiments with both filter and wrapper based models. For filter based model, we use univariate feature selection based on information theoretical concept like Information Gain. While for multivariate filter model we use correlation based feature selection. Our results indicate that PSO performs better than univariate by 3.03 % and multivariate by 2.60 % F-measure points for the GENIA dataset. We also observe quite similar behaviors for the other two datasets.

In addition, we also explore two popular wrapper based feature selection techniques, Genetic Algorithm (GA) (Holland, 1975) based feature selection (Ekbal et al., 2010) technique and Recursive Feature Elimination (RFE) (Guyon et al., 2002)

Feature Selection	GENIA			GENETAG			AiMed		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Filter (Information Gain)	69.02	71.28	70.13	88.25	94.47	91.25	87.46	90.47	88.93
Filter (Correlation)	70.19	70.95	70.56	88.89	93.68	91.22	87.18	89.85	88.49
Wrapper (GA)	72.48	71.98	72.22	89.19	95.04	92.02	89.07	91.11	90.07
Wrapper (RFE)	71.28	71.54	71.40	89.25	94.81	91.94	88.67	91.66	90.14
CRF[PSO]	72.48	73.87	73.16	89.33	96.42	92.74	89.77	92.09	90.92

Table 2: Comparison of PSO with other filter (Information Gain & Correlation) and wrapper (G.A & RFE) based feature selection technique

based approach. Genetic algorithm belongs to the class of randomized wrapper model where feature selection is always classifier dependent and is less prone to stuck at local optima. The RFE is categorized under the deterministic type wrapper model which is computationally less complex than randomized type but has the disadvantage to stuck at local optima.

Results show that PSO performs better than RFE for all the datasets and GA for two datasets (GENIA & GENETAG) in terms of F-measure and the number of features selected. Results are depicted in Table-2. On AiMed dataset, GA and RFE based feature selection techniques perform quite comparable to our PSO based method. It is to be noted that PSO based feature selection yields better performance even with a smaller set of features. The pruned and compact feature set incurs less less computational complexity.

6 Result and Discussion

Table-3 shows the extensive results of our proposed system on all three datasets, namely GENIA, AiMed and GENETAG by augmenting word embedding features. It seems that word embedding features generated from the model which is trained on the combined datasets of PubMed, PMC and Wikipedia [WE(4)] perform better than the other models. The unsupervised word representation features help in detecting unseen entities, i.e. those not appearing in the training data set.

We augment word embedding WE(4) features to the hand-crafted features, and then apply feature selection using PSO on this combined set. Feature selection through PSO not only helps in improving the performance, but at the same time it reduces the feature dimensionality. Evaluation results as reported in Table-4 reveal this fact. Table-3 clearly depicts the effectiveness of word embedding features in BNER (biomedical NER) system. We observe improvements of 7.86%, 5.27% and 7.25% F-measures over the first baseline (i.e. Baseline-1) for GENIA, AiMed and GENETAG data sets, respectively by using PSO based feature selection on PubMed-PMC-wikipedia trained word embedding and handcrafted features. Evaluation also suggests that performance does not degrade significantly, even when we use word embedding features obtained only from Pubmed & PMC OA. It seems that word embedding features obtained from the combination of Pubmed and PMC are more representatives compared to the individual one.

We also show evaluation of some of the existing approaches that attempt to make use of word representation features. A F-measure of 71.39% is reported in the work (Tang et al., 2014). Word representation feature was also used in (Chang et al., 2015) that reported to have achieved F-measure value of 71.77%.

We perform statistical significance (t-test) test on the results obtained by our proposed model. For different datasets, experiments are executed for 10 independent runs and the t-statistic is adopted to analyze the obtained experimental results. Using the known distribution of the test statistic, p-value is calculated. It is observed that p values are less than 0.04 for all the three data sets, which signify that our obtained results are statistically significant.

6.1 Error Analysis

Here, we analyze the outputs obtained for each dataset in order to identify the possible errors.We categorize the errors in three ways as follows:

1. Wrong boundary: This error occurs due to the incorrect boundary identification of entities.These types of cases are observed

System	GENIA			AiMed			GENETAG		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Baseline-1: Handcraft feature	66.78	63.89	65.30	84.11	87.25	85.65	83.88	87.17	85.49
WE1(Handcraft feature + PubMed)	70.72	72.29	71.50	88.42	89.21	88.81	81.92	95.82	88.33
WE2(Handcraft feature + PMC OA)	70.72	72.29	71.50	88.56	89.02	88.78	82.01	95.62	88.29
WE3(Handcraft feature + PubMed + PMC)	70.79	72.47	71.62	89.48	89.01	89.24	82.41	95.89	88.64
WE4(Handcraft feature + PubMed + PMC + Wikipedia)	70.88	72.64	71.75	89.07	90.11	89.59	82.78	95.70	88.77
Baseline-2: Best of WE model	70.88	72.64	71.75	89.07	90.11	89.59	82.78	95.70	88.77
Baseline-3: (PSO with only WE(4)) + Handcraft feature	71.92	72.62	72.26	89.63	90.34	89.98	83.46	95.69	89.15
Proposed: PSO with (handcrafted features + WE)	72.48	73.87	**73.16**	89.77	92.09	**90.92**	89.33	96.42	**92.74**
WE model by Tang et al.(Tang et al., 2014))	70.78	72.00	71.39	-	-	-	-	-	-
WE model by Chang et al.(Chang et al., 2015))	71.36	72.18	71.77	-	-	-	-	-	-

Table 3: Performance evaluation on GENIA, AiMed and GENETAG data sets using various word embedding (WE) features trained on different unlabeled data.

Approach	Dataset		
	GENIA	GENETAG	AiMed
Handcraft Features + W.E Features	232	230	232
PSO based feature selection	129	136	121

Table 4: Comparison of no. of features being used to train the model: Before feature selection and after feature selection

Approach	Dataset		
	GENIA	GENETAG	AiMed
Only W.E features	57.78	55.63	41.22
PSO selected W.E features	58.96	57.41	42.74

Table 5: Comparisons (in terms of F-score) between whole word embedding features using WE(4) and the PSO selected word embedding features excluding handcrafted features. Here, W.E: Word embedding

mostly with the entities having long and compounded wordforms such as 'T cell activation-specific enchance'. We also observe that our system lacks in correctly classifying the instances which includes brackets.

2. Incorrect entity type: This error is obtained when the entity is properly identified but it belongs to some other entity class. This error is more prominent in case of GENIA and GENETAG datasets. For GENIA dataset, classifier is mostly confused with 'Protein' vs. 'Cell_line' or 'Cell_type'. In total 126 Protein words are wrongly classified either as the 'Cell_line' or 'Cell_type'. While with the use of PSO, the rate of mis-classification was reduced to 97. In GENETAG, majority of classes are predicted as 'I-NEWGENE'. This may be due to the fact that majority of the instances belongs to the 'I-NEWGENE' category. While after applying PSO, we observe that mis-classification of 'I-NEWGENE' is significantly reduced from 325 to just 129.

3. Missed entity: Our system misses significant number of NE instances.It is found that number of false negatives count to 1357, 155 and 40 for GENTIA, AiMed and GENETAG, respectively. All these NEs are mis-classfied to belong to the other-than-NE category.

7 Conclusions & Future work

In this paper we have investigated the effect of word embedding features in addition to the handcrafted features for entity extraction from three benchmark biomedical data sets, namely GENIA, AiMed & GENETAG. We have evaluated the system using four different word representation schemes trained on extracted texts from PubMed, PMC OA biomedical literature and Wikipedia dump datasets. In addition to this we have performed PSO based feature selection on the whole feature set for the different data sets. We can conclude that instead of using a full word representation feature, if only prominent features are used, it could help in improving the performance of the system. In future work, we would like to perform additional experiments to fine-tune the dimensions of vectors and the parameters of CRF through cross-validation on the training set. The applicability of feature selection on word embedding features need to be explored in other domain also. In addition we want to compare the performance of representation obtained through *word2vec* to the others such as GloVe. We would also like to explore deep learning techniques replacing CRF.

References

Dima Alamedine, Catherine Marque, and Mohamad Khalil. 2013. Binary particle swarm optimization for feature selection on uterine electrohysterogram signal. In *Advances in Biomedical Engineering (ICABME), 2013 2nd International Conference on*, pages 125–128. IEEE.

J.C. Bansal, P.K. Singh, Mukesh Saraswat, Abhishek Verma, Shimpi Singh Jadon, and Ajith Abraham. 2011. Inertia weight strategies in particle swarm optimization. In *Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on*, pages 633–640. IEEE.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

F.X. Chang, J. Guo, W.R. Xu, and S. Relly Chung. 2015. Application of word embeddings in biomedical named entity recognition tasks. *Journal of Digital Information Management*, 13(5).

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205.

Russell C. Eberhart and Yuhui Shi. 1998. Comparison between genetic algorithms and particle swarm optimization. In *Evolutionary Programming VII*, pages 611–616. Springer.

Asif Ekbal and Sriparna Saha. 2016. Simultaneous feature and parameter selection using multiobjective optimization: application to named entity recognition. *Int. J. Machine Learning & Cybernetics*, 7(4):597–611.

Asif Ekbal, Sriparna Saha, Utpal Kumar Sikdar, and Md Hasanuzzaman. 2010. A genetic approach for biomedical named entity recognition. In *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, volume 2, pages 354–355. IEEE.

Asif Ekbal, Sriparna Saha, and Utpal Kumar Sikdar. 2013. Biomedical named entity extraction: some issues of corpus compatibilities. *SpringerPlus*, 2(1):1–12.

Jenny Finkel, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2005. Exploring the boundaries: gene and protein identification in biomedical text. *BMC bioinformatics*, 6(1):1.

Cyril Grouin. 2014. Biomedical entity extraction using machine-learning based approaches. *substance*, 6:1–611.

Zhou GuoDong and Su Jian. 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*, pages 96–99. Association for Computational Linguistics.

Deepak Kumar Gupta, Kandula Srikanth Reddy, Shweta, and Asif Ekbal. 2015. Pso-asent: Feature selection using particle swarm optimization for aspect based sentiment analysis. In *Natural Language Processing and Information Systems - 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015 Passau, Germany, June 17-19, 2015 Proceedings*, pages 220–233.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Xu Han, Chee Keong Kwoh, and Jung-jae Kim. 2016. Clustering based active learning for biomedical named entity recognition. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 1253–1260. IEEE.

John H. Holland. 1975. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press.

Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. 2002. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pages 1–8. Association for Computational Linguistics.

James Kennedy and Russell C. Eberhart. 1997. A discrete binary version of the particle swarm algorithm. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, volume 5, pages 4104–4108. IEEE.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 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*, pages 70–75. Association for Computational Linguistics.

Seonho Kim, Juntae Yoon, Kyung-Mi Park, and Hae-Chang Rim. 2005. Two-phase biomedical named entity recognition using a hybrid method. In *Natural Language Processing–IJCNLP 2005*, pages 646–657. Springer.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289.

Kenli Li, Wei Ai, Zhuo Tang, Fan Zhang, Lingang Jiang, Keqin Li, and Kai Hwang. 2015a. Hadoop recognition of biomedical named entity using conditional random fields. *IEEE Transactions on Parallel and Distributed Systems*, 26(11):3040–3051.

Lishuang Li, Liuke Jin, Zhenchao Jiang, Dingxin Song, and Degen Huang. 2015b. Biomedical named entity recognition based on extended recurrent neural networks. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 649–652. IEEE.

Nut Limsopatham and Nigel Collier. 2016. Learning orthographic features in bi-directional LSTM for biomedical named entity recognition. *BioTxtM 2016*, page 10.

Hongfang Liu, Zhang-Zhi Hu, Jian Zhang, and Cathy Wu. 2006. Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22(1):103–105.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Scott Miller, Jethran Guinness, and Alex Zamanian. 2004. Name tagging with word clusters and discriminative training. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Rakesh Patra and Sujan Kumar Saha. 2013. A kernel-based approach for biomedical named entity recognition. *The Scientific World Journal*, 2013.

S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44.

Ayan Seal, Suranjan Ganguly, Debotosh Bhattacharjee, Mita Nasipuri, and Consuelo Gonzalo-Martin. 2015. Feature selection using particle swarm optimization for thermal face recognition. In *Applied Computation and Security Systems*, pages 25–35. Springer.

Burr Settles. 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*, pages 104–107. Association for Computational Linguistics.

Utpal Kumar Sikdar, Asif Ekbal, Sriparna Saha, Olga Uryupina, and Massimo Poesio. 2015. Differential evolution-based feature selection technique for anaphora resolution. *Soft Comput.*, 19(8):2149–2161.

Larry Smith, Lorraine K. Tanabe, Rie J. Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(Suppl 2):S2.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria, August. Association for Computational Linguistics.

Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 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*, pages 13–20. ACM.

Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. *BMC medical informatics and decision making*, 13(Suppl 1):S1.

Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu. 2014. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014.

Zhuo Tang, Lingang Jiang, Li Yang, Kenli Li, and Keqin Li. 2015. Crfs based parallel biomedical named entity recognition algorithm employing mapreduce framework. *Cluster Computing*, 18(2):493–505.

Hossein Tohidi, Hamidah Ibrahim, and Masrah Azrifah Azmi Murad. 2014. Improving named entity recognition accuracy for gene and protein in biomedical text literature. *International journal of data mining and bioinformatics*, 10(3):239–268.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Xuesong Yan, Qinghua Wu, Hanmin Liu, and Wenzhi Huang. 2013. An improved particle swarm optimization algorithm and its application. *International Journal of Computer Science Issues (IJCSI)*, 10(1).

Li Yang and Yanhong Zhou. 2014. Exploring feature sets for two-phase biomedical named entity recognition using semi-crfs. *Knowledge and information systems*, 40(2):439–453.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098.

Shaojun Zhao. 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*, pages 84–87. Association for Computational Linguistics.