Biomedical Named Entity Recognition Using Structured Support Vector Machine

Lobna A. Mady*, Yasmine M. Afify*, and Nagwa L. Badr*
Ain Shams University, Cairo, EGYPT

Abstract

Named entity recognition is an information extraction subtask that aims to discover named items referenced in unstructured text and classify them into predefined class labels. Identifying biomedical entities such as proteins, cell lines, cell types, DNAs and RNAs has been recognized as a challenging task in named entity recognition. In this paper, the applicability of using structured support vector machine to classify biomedical named entity recognition is thoroughly investigated. This is achieved by utilizing a combination of various types of features such as morphological, part of speech, orthographical, context and word representation to explore the classification performance. Comprehensive experiments were conducted on two popular datasets based on multiple evaluation metrics. Experimental results revealed that the performance of the structured support vector machine surpasses that of different benchmark approaches in the literature.

Key Words: Biomedical named entity recognition, machine learning, classification, natural language processing, structured support vector machine.

1 Introduction

Natural Language Processing (NLP) is a branch of artificial intelligence that aids computers to recognize and manipulate human language. Named Entity Recognition (NER) is one of the NLP linguistic communication processes. NER aims to identify the entities from text and classify them into predefined class labels such as person names, place names and numbers. Biomedical Named Entity Recognition (BNER) is one of the NER techniques which identifies biomedical instances such as proteins, viruses, cell lines, cell types, DNAs and RNAs.

According to [3], BNER has two main tasks: feature extraction and model formulating/training. Various types of features are used in BNER to classify the biomedical terms. Linguistic features are used to recognize the word language such as stemming, lemmatization and parts-of-speech tagging. Orthographic features capture the understanding of the word itself. Morphological features reflect common structures in addition to subsequences of characters among several tokens, thus identifying similarities between distinct tokens. Context features names of the target entity name), and trigger names (match names that may indicate the presence of biomedical names in the surrounding tokens).

X. Wang, et al. [35] reported that BNER has four approaches: dictionary-based models, rule-based models, machine learning models as well as hybrid approach. First, dictionary-based approach stores labeled Named Entity (NE) in a list called a gazetteer. This approach is very simple but it is also time consuming as biomedical data increases tremendously [15]. Second, rule-based approach depends on context to solve the problem of multiple NEs [10]. Every rule should be written before it is used but it is time consuming as it requires experts to build and construct rules. Also, rules created for one corpus cannot be generalized to other corpora. Third, machine learning approach tags the specified NE to words. Hidden Markov Models (HMMs) [30], Maximum Entropy Markov Models (MEMMs) [9], Conditional Random Field (CRF) [17] and Support Vector Machines (SVMs) [26] are examples of common algorithms used in this approach. Finally, hybrid-based approach which combine two or more techniques. the hybrid approaches consist of the advantages of all the other approaches [11].

According to existing literature, many researchers have faced challenges to develop various types of BNER such as: (1) specifying the boundaries of the entity names, (2) sharing the same prefix noun for many biomedical names, (3) lacking strict naming convention in the biomedical literature, (4) casual use of capitalization and hyphens, (5) having ambiguity in the classification against existing dictionaries due to the massive number of abbreviations.

To the best of the author's knowledge, no study was found in the literature where Structured Support Vector Machine (SSVM) was used in extracting biomedical entities. The main objective of this paper is to enhance the performance of BNER by using SSVM. The paper starts by reviewing the state-of-the-art research conducted in the area of BNER. Then, in the third section, the methodology including the utilized feature sets and classifiers is presented. Afterwards, in the fourth section, details of the selected datasets and evaluation metrics are provided. In the fifth section, the obtained experimental results are tabulated and discussed. Finally, the main conclusions of the performed investigation are outlined in the last section.

• Faculty of Computer and Information Sciences. Email: lobna.mady@cis.asu.edu.eg.

2 Related Work

Many research studies have been conducted on Biomedical Named Entity Recognition (BNER) to classify biomedical terms such as genes, proteins, and DNAs. Some researchers used a hybrid approach based on dictionary and machine learning to enhance the results. Y. Tsuruoka, et al. [34] proposed a two-phase hybrid approach system for recognizing protein names where a protein name dictionary was used to search for potential protein names in the first phase. Then, the candidates were filtered in the second phase using a machine learning algorithm. R. Ramachandran, et al. [27] developed another hybrid approach to identify NE from various medical literature research articles by extracting medicine names, disorders, symptoms, species, dosages, and medical instructions. The authors constructed a new dictionary to annotate the entities in the medical articles which were used as a dataset. The Blank Spacy Machine Learning model, which is a type of deep learning, was used to train the annotated entities.

Other researchers combined between CRF and neural network algorithm to generate the Word Embedding (WE) feature vectors. X. Ma, et al. [22] used character level representation as input to Convolutional Neural Network (CNN) to generate WE vector. Concatenation of the embedded vector with character level vector was implemented to feed Bidirectional Long Short-Term Memory (BLSTM) network. The output vectors from the BLSTM network were given to the CRF final layer. W. Xie, et al. [36] extracted disease entities from biomedical text by presenting CRF-based strategy using WE and cluster-based word representation features. The Brown clustering technique was used as a cluster-based approach to cluster the feature vectors. Both the WE and Brown clustering were used as features for the CRF classifier. The same technique used in [22] was proposed by [31] but byte embedding was used instead of WE to feed CNN.

A combination between CRF and other machine learning techniques was implemented by various researchers to obtain high performance for the BNER. P. T. Lai, et al. [20] used the Statistical-Principle-Based Approach (SPBA) machine-learning technique to identify the names for Genes and Proteins in biomedical corpus. The SPBA predictions were utilized as features for a CRF-based recognizer. A. Agrawal, et al. [2] proposed a sampling strategy model that combined active learning and machine learning. A modified least confidence-based query sampling strategy was used for the active learning approach. The proposed model computed the sentence confidence score and employed CRF as a classifier.

Single machine learning algorithm was utilized as a base classifier in different research studies. D. Campos, et al. [4] used a combination of different features like orthographic, morphological, context and dictionary-based features to generate feature vector and used CRF as base classifier. A feature selection methodology based on Particle Swarm Optimization (PSO) was proposed by [37]. In this methodology, the important features were selected from a set of handcrafted and WE based features. The authors adopted CRF as a learning algorithm to train the proposed features. CRFVoter was developed by [14] to transform the BNER task into a sequence tagging problem for extracting genes and proteins entities. CRFVoter used a two-stage classifier of CRF, using the output of each NER as input to 2nd level CRF which was used to label the sequence. The optimized sequence labels were integrated into one ensemble classifier. R. Ramachandran, et al. [28] constructed a model based on an optimized version of SVM where the PSO technique adjusted the weight and bias parameters in the tree based SVM model. H. Yu, et al. [39] developed an innovative model based on multistage three-way choices strategy and CRF to extract various biomedical entities.

Instead of machine learning technique, some research work was conducted using deep neural network. G. Murugesan, et al. [25] proposed Bidirectional Contextual Clues Named Entity Recognition (BCC-NER), which is a three-module technique, to deploy a bidirectional named entity tagger for gene/protein mention recognition. Text processing was covered in the first module, which contained fundamental NLP pre-processing, feature extraction, and feature selection. The second module was for bidirectional CRF and model development while post-processing was the third module. [C. Che, et al. [5] proposed a Temporal Convolutional Network (TCN) with a CRF layer to classify proteins, RNAs, DNAs, cell types and cell lines by extracting the features with TCN. These features were then decoded using CRF to predict the class labels. Efficiency of the BNER was enhanced via updating the original TCN model by merging the information generated with convolution kernels of various sizes. S. Sharma, et al. [29] proposed an embedding model framework for state-of-the-art NLP (FLAIR) to enhance the performance of BNER.

J. Lee, et al. [21] introduced BioBERT which used Bidirectional Encoder Representations from Transformers (BERT). H. Zhou, et al. [43] proposed a knowledge-enhanced system that combined entity recognition module and deep contextualized Word Representation (WR) which is a type of language model. The entity recognition module was divided into three parts: the embedding layer, the Bidirectional Long Short-Term Memory (BiLSTM) layer and the CRF layer. Z. Yuan, et al. [40] proposed Knowledge enhanced Biomedical pretrained Language Model (KeBioLM) which incorporated explicit knowledge from Unified Medical Language System (UMLS). V. Kocaman, et al. [19] proposed a BLSTM-CNN-Char framework where CNN was used to generate character level WE vector. The generated vector was used as input to BiLSTM where the output of each network was decoded by log-softmax layer and log-probabilities.

3 Proposed Approach

In the following subsections, details about the feature set and the machine learning technique, adopted in the current approach, are presented.

3.1 Feature Set

In our approach, Morphological, Orthographical, Context, Part of Speech (POS) and WR features are used. Details about

these features are presented as follows:

Morphological features study the word components and the relationships between them. These features also study the shared structure between words.

Orthographical features are commonly used in BNER to capture information about how the words are formed such as capitalization, hyphenation and punctuation. These features intend to group words with similar forms. J. Zhang, et al. [42].

Context features study the syntactic information occurrences of adjacent tokens.

POS features capture the noun phrase region. These features are useful for BNER based on the concept that NE is more likely a noun phrase**.**

WR features Based on the reviewed research work in Section 2, BNER yielded better performance when characters-level representation (words) was used to generate WR feature vector, which is used in the current approach. WR includes WE feature which is a powerful technique for representing words because WE captures both semantic and syntactic meanings of tokens. WE was developed to represent a single word by a low dimensional vector. Each vector location corresponds to a characteristic with semantic or grammatical inference. WE has the capability to detect the occurrence of similar words appearing in similar context. W. Yoon, et al. [38] and [12] demonstrated that WE trained on biomedical texts significantly improved BNER model performance. In the current approach, Skip Gram Vector [24] is adopted due to being suitable for training of rare words which commonly appear in biomedical text [41]. Table 1 shows examples of WE vector of JNLPBA tokens. More details regarding the description of features used in the proposed approach are provided in Table 2.

Table 1: Examples of WE vector of JNLPBA tokens

Token	Word embedding vector
interleukin	-0.421, 0.149, -0.725, 0.078, -0.274, 0.195, 0.007, 0.303, -0.540, -0.436, -0.189, -0.227, -0.549, 0.214, 0.845, 0.933, -0.158, -0.500, -0.509, -0.377
electrophoretic	0.274, 0.775, -0.825, -0.741, -0.080, -0.595, 1.270, 0.585, -1.150, 0.542, 0.036, 1.325, 0.690 , 0.211 , -1.193 , 0.896 , 0.475 , -0.221 , 0.493 , 1.357
t-lymphocyte	0.371, 0.025, -0.017, 0.296, -0.406, -0.550, 0.725, -0.202, -0.332, -0.320, -0.685, -0.386, -0.390, 1.003, 0.909, -0.277, -0.433, 0.392, -0.430, 0.394
il2	0.607, 0.668, 0.204, 0.656, -0.123, 0.715, 0.376, 0.747, -0.010, -0.178, -0.070, 0.515, -0.699, -0.349, 0.481, 0.428, -0.188, -0.044, 0.024, 0.379

Table 2: Description of features used in the proposed approach

Feature Type	Feature Name	Description	Notes
Morphological	**Prefix**	Refer to fixed length character sequences taken from the leftmost locations of the words	The length of the prefix from 3 to 7 depending on the size of token.
	Suffix	Refer to fixed length character sequences taken from rightmost locations of the words	The length of the suffix ranges from 3 to 7 depending on the size of token.
	Word Shape	Is defined as the assign of each token into an equivalent class	i.e., IL-2_gene is assigned to AA-0_aaaa while CD4 class is AA0
Orthographical	**ALLCAPS**	Checks if all letters of the word are upper case	i.e., LIM
	INTCAP	Checks if first letter of the word is upper case	i.e., Zta
	HASCAP	Checks if any letter of the word is upper case	i.e., polyA
	SINGLECAP	Check if word has only one upper case letter	i.e., A

	CAP&DIGIT	Check if word is a mix between numbers and upper-case letters	i.e., LSP1
	Digit&Alpha	Check if word is a mix between numbers and letters	i.e., ME1a1
	CAP&ALPHA	Check if word is a mix between lower case and upper-case letters	i.e., mRNA
	ALLDIGIT	Check if word is a number only	i.e., 280
	Alpha&Digit	Check if the word first letter is a letter and rest is numbers	i.e., DND39
	DigitSpecial	Check if the word start with digit then special character	i.e., 22-3
	AlphaDigitAlpha	Check if the word first letter is a character, then digit then character	i.e., s6k,
	DigitCommaDigit	Check if the word first letter is a number, then comma then number	i.e., 1,25
	DigitDotDigit	Check if the word first letter is a number, then dot then number	i.e., 0.6
	HasRoman	Check if word has a Roman letter	i.e., II, IV
	HasGreek	Check if word has a Greek letter	i.e., Beta, Alpha
Context	**Context feature**	Refer to tokens that appear inside a 5-word window size as proposed in [39]	i.e., 2 to the right and 2 to the left of the present token
Part of speech (POS)	**POS**	Genia Tagger [33] is used to extract POS tags of the desired entities.	i.e., NNP, VBZ, NN
	Context words POS	Genia Tagger [33] is used to extract POS tags of the context entities.	
Word Representation (WR)	**WE**	Word2vec is used for Skip Gram WE feature [Apache License] [8]	

3.2 Machine Learning Technique

The SVM classifier can be trained for regression, binary classification, and multiclass classification. Structured Support Vector Machine (SSVM) is an enhanced version of SVM model, which is used as a machine learning technique in the developed approach which can be trained for general structured output labels. SSVM combines the benefits of both CRFs and SVMs in a single algorithm and requires less training time. SVMhmm, developed by [16], is used as an implementation of SSVM.

4 Experiment and Evaluation

In the following subsections, details about the employed datasets and the evaluation metrics are presented.

4.1 Datasets

Based on the conducted literature survey, two datasets (JNLPBA and GeneTag) [18, 32], are most commonly used for extracting the biomedical entities based on the work proposed by [1, 4, 8, 31]. GeneTag has only two class labels.

GeneTag mentions the protein, DNA and RNA as belonging to the same entity type and tags them with the 'NEWGENE' tag in both the training and test datasets. If two genes or proteins are overlapped, the dataset identifies the second gene as 'NEWGENE1' tag. The training dataset includes 7500 sentences which comprise 8881 gene mentions.

On the other hand, JNLPBA dataset has five NE class labels: protein, DNA, RNA, cell line, and cell type. The IOB24 format was used to define the boundaries of the tokens. Two different class labels are applied to each entity: B-Class and I-Class donating the beginning token and intermediate token, respectively. Therefore, ten classes were generated for NEs and one additional class for non NEs. Frequencies of entities annotated in JNLPBA are shown in Table 3 [18].

4.2 Evaluation Metrics

Five widely used evaluation measures are employed in the experiments to assess the proposed SSVM approach: recall, precision, F1-measure, G-mean and MCC. The first three metrics are calculated using Equations 1, 2 and 3, respectively [23]. True Positive (TP) is the number of tokens that are correctly detected by the system. False Negative (FN) is the number of named entities that are not identified, while False

Table 3: Frequencies for NEs in JNLPBA dataset [18]

	Protein	DNA	RNA	Cell Type	Cell Line	Overall
Training	30269	9533	951	6718	3830	**51301**
Testing	5067	1056	118	1921	500	**8662**

Positive (FP) is the number of tokens that the system misidentifies.

Matthews Correlation Coefficient (MCC) is a metric used in machine learning to estimate the validity of classifications and calculated using equation 4 [6]. The Geometric Mean (G-Mean) is a metric that evaluates the equilibrium between classification performance between dominant and least classes and calculated using equation 6 [13]. Even if the negative cases are correctly labeled, a low G-Mean indicates poor performance in the categorization of positive cases. This check is necessary to avoid overfitting the negative class and underfitting the positive class. Ranges for both MCC and G-Mean are -1 to 1, with a value of -1 indicating classifier misclassifying the tokens completely and a value of 1 indicating a classifier correctly classifying the classes.

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \quad (3)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (5)$$

$$\text{G} - \text{Mean} = \sqrt{Specificity * Sensitivity} \quad (6)$$

5 Results

In the following subsections, the experimental results and discussion are presented.

5.1 GeneTag Dataset Results

Table 4 shows the obtained results for the GeneTag dataset which are obtained from the developed and other benchmark approaches. It can be noted that SSVM outperforms all other approaches by achieving a recall, precision, and F1-Measure of 97.18%, 97.19 % and 97.17%, respectively. Compared to the NERSuite model, the developed SSVM model resulted in an improvement of 13.72%. Moreover, the SSVM results in an improvement of 13.59% compared to Byte NN, 11.47% compared to Gimli and an improvement of 8.6% compared to Word-level NN.

5.2 JNLPBA Dataset Results

Table 5 presents the F1 Measure obtained from different existing approaches. One can observe that SSVM approach surpasses all other approaches. The percentage of its improvement is 30.87%, 27.54%, 25.49%, 22.65%, 17.67%, 16.82%, 11.5%, 10.54% and 9.83% compared to Byte NN, NERSuite, Gimli, Word-level NN, BioFLAIR, BioBERT, [19], KeBioLM as well as [39], respectively.

Table 6 shows the overall recall, precision, F1-Measure, G-Mean and MCC for each class label in JNLPBA & GeneTag using SSVM.

To summarize, NERSuite and Byte NN have the worst performance for both datasets. However, the CRF has a better performance when combined with other neural network techniques. SSVM outperforms all other benchmark approaches for both datasets. As shown in Table 6, SSVM achieves near optimal results for both G-mean and MCC.

6 Conclusion

This work addressed an unattended research area, which is

Table 4: Performance comparison for GeneTag dataset

Model	Recall	Precision	F1-Measure
SSVM (Our Approach)	97.18	97.19	97.17
NERSuite tool [4, 7]**,**	82.34	88.81	85.45
Gimli [4]	84.82	90.22	87.17
Word-level NN [22]	Not available	Not available	89.45
Byte NN [31]	Not available	Not available	85.54

Table 5: Performance comparison for JNLPBA dataset

Model	Recall	Precision	F1-Measure
SSVM (Proposed Approach)	89.632	91.837	90.64
NERSuite tool [4, 7]	Not available	Not available	71.07
Gimli [4]	71.62	72.85	72.23
Biomedical Named Entity Recognition Based on Multistage Three-Way Decisions[39]	80.58	84.58	82.53
Word-level NN [22, 31]	Not available	Not available	73.53
Byte NN [31]	Not available	Not available	69.26
BioFLAIR [29]	Not available	Not available	77.03
BioBERT [21.]	83.56	72.24	77.59
KeBioLM [41]	Not available	Not available	82
Biomedical Named Entity Recognition at Scale [19]	Not available	Not available	81.29

Table 6: Evaluation metrics for different entity types in JNLPBA and GeneTag datasets using SSVM

Dataset	Entity Type	Recall	Precision	F1-measure	G - Mean	MCC
JNLPBA	**DNA**	91.717	97.997	94.621	0.9563	0.9443
	Protein	96.075	93.943	94.957	0.9688	0.9317
	RNA	79.182	85.077	81.986	0.8193	0.8193
	Cell Type	93.498	95.777	94.623	0.9643	0.9392
	Cell Line	87.688	86.392	87.015	0.8656	0.8656
	Overall	**89.632**	**91.837**	**90.64**	**0.9424**	**0.90002**
GeneTag	**Genes/Proteins**	96.677	97.66	97.166	0.961	0.94367
	Not Genes/Proteins	97.68	96.71	97.195	0.9717	0.944
	Overall	**97.18**	**97.19**	**97.17**	**0.96635**	**0.9438**

applying SSVM to biomedical entities to extract genes,proteins, cell lines, cell types, DNAs and RNAs. A combination of morphological, part of speech, orthographical, context and word representation features was used to investigate the classification performance. Comprehensive evaluation was conducted using two popular datasets in respect of multiple evaluation metrics: recall, precision, F1-Measure, G-Mean and MCC. Experimental results show that SSVM achieves very promising results in BNER when it is used as a machine learning technique. It achieves near optimal results for both G-mean and MCC. It surpasses all benchmark approaches for both datasets with improvements in respect to F1-measure that ranges from 8.6% to 13.72% for GeneTag dataset and 9.83% to 30.87% for JNLBPA dataset. These promising results motivate us to explore other combinations of features with SSVM.

References

[1] M. T. Abd and M. Mohd, "A Comparative Study of Word Representation Methods with Conditional Random Fields and Maximum Entropy Markov for Bio-Named Entity Recognition", *Malaysian Journal of Computer Science*, 31:15-30, 2018.

[2] A. Agrawal, S. Tripathi, and M. Vardhan, "Active Learning Approach Using a Modified Least Confidence

Sampling Strategy for Named Entity Recognition", *Progress in Artificial Intelligence*, 10:113-128, 2021.

[3] D. Campos, S. Matos, and J. Luis, "Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools", Theory and Applications for Advanced Text Mining, 2012.

[4] D. Campos, S. Matos, and J. L. Oliveira, "Gimli: Open Source and High-Performance Biomedical Name Recognition", *BMC Bioinformatics*, 14:1-14, 2013.

[5] C. Che, C. Zhou, H. Zhao, B. Jin, and Z. Gao, "Fast and Effective Biomedical Named Entity Recognition Using Temporal Convolutional Network with Conditional Random Field", *Mathematical Biosciences and Engineering*, 17:3553-3566, 2020.

[6] D. Chicco, "Ten Quick Tips for Machine Learning in Computational Biology," BioData Mining, 10:1-18, 2017.

[7] H. Cho, "NERsuite: A Named Entity Recognition Toolkit", Tsujii Laboratory, Department of Information Science, University of Tokyo, Tokyo, Japan, [Online]. Available: http://nersuite.nlplab.org.

[8] A. Ekbal and S. Saha, "Stacked Ensemble Coupled with Feature Selection for Biomedical Entity Extraction", *Knowledge-Based Systems*, 46:22-32, 2013, Apache License, "Word2Vec." https://code.google.com/archive/p/word2vec/ (accessed Oct. 20, 2020).

[9] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting Context for Biomedical Entity Recognition", p. 88, 2004.

[10] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward Information Extraction: Identifying Protein Names from Biological Papers", Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, pp. 707-718, 1998.

[11] L. J. Gong, Y. Yuan, Y. B. Wei, and X. Sun, "A Hybrid Approach for Biomedical Entity Name Recognition" *Proceedings of the 2009 2nd International Conference on Biomedical Engineering and Informatics*, BMEI 2009, pp. 1-5, 2009.

[12] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep Learning with Word Embeddings Improves Biomedical Named Entity Recognition", *Bioinformatics*, 33:i37-i48, 2017.

[13] M. Hao, Y. Wang, and S. H. Bryant, "An Efficient Algorithm Coupled with Synthetic Minority Over-Sampling Technique to Classify Imbalanced PubChem BioAssay Data", *Analytica Chimica Acta*, 806:117-127, 2014.

[14] W. Hemati and A. Mehler, "CRFvoter: Gene and Protein Related Object Recognition Using a Conglomerate of CRF-Based Tools", *Journal of Cheminformatics*, 11:1-11, 2019.

[15] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, and B. J. A. Schijvenaars, "A Dictionary to Identify Small Molecules and Drugs in Free Text", *Bioinformatics*, 25:2983-2991, 2009.

[16] Thorsten Joachims, "SVMhmm", http://www.cs.cornell. edu/people/tj/svm_light/svm_hmm.html (accessed Aug. 16, 2020).

[17] U. Kanimozhi and D. Manjula, "A CRF Based Machine Learning Approach for Biomedical Named Entity Recognition", *Proceedings - 2017 2nd International Conference on Recent Trends and Challenges in Computational Models*, ICRTCCM 2017, pp. 335-342, 2017.

[18] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the Bio-Entity Recognition Task at JNLPBA", p. 70, 2004.

[19] V. Kocaman and D. Talby, "Biomedical Named Entity Recognition at Scale", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNCS, 12661:635-646, 2021.

[20] P. T. Lai, M. S. Huang, T. H. Yang, W. L. Hsu, and R. T. H. Tsai, "Statistical Principle-Based Approach for Gene and Protein Related Object Recognition", *Journal of Cheminformatics*, 10:1-9, 2018.

[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. Ho So, and J. Kang, "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics*, 36:1234-1240, 2019.

[22] X. Ma and E. Hovy, "End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF", *54th Annual Meeting of the Association for Computational Linguistics*, ACL 2016 - Long Papers, 2:1064-1074, 2016.

[23] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*, Cambridge University Press, 2009.

[24] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality", Advances in Neural Information Processing Systems, pp. 1-9, 2013.

[25] G. Murugesan, S. Abdulkadhar, B. Bhasuran, and J. Natarajan, "BCC-NER: Bidirectional, Contextual Clues Named Entity Tagger for Gene/Protein Mention Recognition", *Eurasip Journal on Bioinformatics and Systems Biology*, 201:1-8, 2017.

[26] Y. Peng, C. H. Wei, and Z. Lu, "Improving Chemical Disease Relation Extraction with Rich Features and Weakly Labeled Data", *Journal of Cheminformatics*, 8:1-12, 2016.

[27] R. Ramachandran and K. Arutchelvan, "Named Entity Recognition on Bio-Medical Literature Documents Using Hybrid Based Approach", *Journal of Ambient Intelligence and Humanized Computing*, 11:1-10, 2021.

[28] R. Ramachandran and K. Arutchelvan, "Optimized Version of Tree-Based Support Vector Machine for Named Entity Recognition in Medical Literature", *Proceedings of the 3rd International Conference on Intelligent Sustainable System*s, ICISS 2020, pp. 357-361, 2020.

[29] S. Sharma and R. Daniel, "BioFLAIR: Pretrained Pooled Contextualized Embeddings for Biomedical Sequence Labeling Tasks", 03:1-6, 2019, [Online]. Available: http://arxiv.org/abs/1908.05760.
[30] D. Shen, J. Zhang, G. Zhou, J. Su, and C.-L. Tan, "Effective Adaptation of a Hidden Markov Model-Based Named Entity Recognizer for Biomedical Domain", pp. 49-56, 2003.
[31] E. Sheng and P. Natarajan, "A Byte-Sized Approach to Named Entity Recognition", 2018, [Online]. Available: http://arxiv.org/abs/1809.08386.
[32] L. Tanabe, N. Xie, L. H. Thom, W. Matten, and W. J. Wilbur, "GENETAG: A Tagged Corpus for Gene/Protein Named Entity Recognition", *BMC Bioinformatics*, 6:1-7, 2005.
[33] Y. Tsuruoka, Y. Tateishi, J. D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, "Developing a Robust Part-of-Speech Tagger for Biomedical Text", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNCS, 3746:382-392, 2005.
[34] Y. Tsuruoka and J. Tsujii, "Improving the Performance of Dictionary-Based Approaches in Protein Name Recognition," *Journal of Biomedical Informatics*, 37:461-470, 2004.
[35] X. Wang, C. Yang, and R. Guan, "A Comparative Study for Biomedical Named Entity Recognition", *International Journal of Machine Learning and Cybernetics*, 9:373-382, 2018.
[36] W. Xie, S. Fu, S. Jiang, and T. Hao, "A CRFs-Based Approach Empowered with Word Representation Features to Learning Biomedical Named Entities from Medical Text", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNCS, 10676:518-527, 2017.
[37] S. Yadav, A. Ekbal, S. Saha, and P. Bhattacharyya, "Entity Extraction in Biomedical Corpora: An Approach to Evaluate word Embedding Features with PSO Based Feature Selection", Proceedings of for the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, 1:1159-1170, 2017.
[38] W. Yoon, C. H. So, J. Lee, and J. Kang, "CollaboNet: Collaboration of Deep Neural Networks for Biomedical Named Entity Recognition", *BMC Bioinformatics*, 20:55-65, 2019.
[39] H. Yu, Z. Wei, L. Sun, and Z. Zhang, "Biomedical Named Entity Recognition Based on Multistage Three-Way Decisions", *Communications in Computer and Information Science*, 663:513-524, 2016.
[40] Z. Yuan, Y. Liu, C. Tan, S. Huang, and F. Huang, "Improving Biomedical Pretrained Language Models with Knowledge", *Proceedings of the 20th Workshop on Biomedical Language Processing*, 0011603:180-190, 2021.
[41] Y. Zhang, Z. Liu, and W. Zhou, "Biomedical Named Entity Recognition based on Deep Neutral Network", *Chinese Journal of Electronics*, 29:455-462, 2020.
[42] J. Zhang, D. Shen, G. Zhou, J. Su, and C. L. Tan, "Enhancing HMM-Based Biomedical Named Entity Recognition by Studying Special Phenomena", *Journal of Biomedical Informatics*, 37:411-422, 2004.
[43] H. Zhou, S. Ning, Z. Liu, C. Lang, Z. Liu, and B. Lei, "Knowledge-Enhanced Biomedical Named Entity Recognition and Normalization: Application to Proteins and Genes", *BMC Bioinformatics*, 21:1-15, 2020.

Lobna A. Mady received her B.Sc. degrees in Computer Science in 2013 from Faculty of Computer and Information Sciences, Ain Shams University. Currently, she is working as teaching assistant at the Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University. Her research areas are Bioinformatics, Natural Language Processing and Machine Learning.

Yasmine M. Afify received her Ph.D. from the Faculty of Computer and Information Sciences, Ain Shams University, Egypt. She is now a lecturer at the Information Systems department, Faculty of Computer and Information Sciences, Ain Shams University. She is a referee for several international journals and conferences. Her main research interests include Information Retrieval, Bioinformatics, Social Networks, and Knowledge Management. She has published more than 15 research papers in international journals and conferences.

Nagwa Badr is a professor and dean at Faculty of Computer and Information Sciences, Ain Shams University. She received the B.Sc. degrees in Computer Science in 1996 and PhD from Liverpool John Moorse University, U.K. in 2003 in Software Engineering and Distributed Systems. She had done postdoctoral studies in Glasgow University, U.K. She is a head of the committee that contributed in research projects funded by national and international grants of Information Systems, Bioinformatics, Business Analytic and Health Informatics (i.e., http://www.heal-plus.eu/). Her current research areas are in Software Engineering, Cloud Computing, Big Data analytics, social networking, Arabic search engines and Bioinformatics.