Journal of Autonomous Intelligence (2023) Volume 6 Issue 3
doi: 10.32629/jai.v6i3.1123

ORIGINAL RESEARCH ARTICLE

Biomedical named entity recognition using TCN approaches and bio tagging

Thiyagu Thavittupalayam Meenachisundaram[1], Sangeetha Ramachandran[1], Sudhakaran Gajendran[2], Om Kumar Chandra Umakantham[3], Sathish Kuppani[4,*]

[1] *Department of Computer Science and Engineering, Karunya Institute of Technology and Science, Coimbatore 641114, India*

[2] *School of Electronics Engineering, Vellore Institute of Technology-Chennai Campus, Chennai 632014, India*

[3] *School of Computer Science and Engineering, Vellore Institute of Technology-Chennai Campus, Chennai 632014, India*

[4] *Department of Computer Science and Engineering, Tirumala Engineering College, Andhra Pradesh 522601, India*

*** Corresponding author:** Sathish Kuppani, skuppani@gmail.com

ABSTRACT

Biomedical named entity recognition (BNER) is to identify instances in biomedical field such as chemical compounds, drugs, genes, RNA, DNA and proteins used in extracting information. It extracts relation between various drugs and their usage, profiles of similar and related drugs with help of machine learning approach. The efficiency in biomedical field is still in research for further improvement even many supervised methods are applied. The proposed method combines two algorithms and improve performance based on features used. It uses conditional random field (CRF) for entity identification and classification of temporal conventional network (TCN) to detect and recognize subtypes in BNER. Datasets such as GENIA and CHEMDNER corpus are used for evaluation with different entity types. Results shows that proposed methods performed better compared to other machine learning approach. The detailed study of TCN has been discussed. The classification of BNER is mapped with various classification methods to enhance result of high recognition.

Keywords: named entity recognition; conditional random field; temporal conventional network (TCN)

ARTICLE INFO

Received: 8 August 2023
Accepted: 9 September 2023
Available online: 27 September 2023

1. Introduction

Named entity recognition (NER) is the sub-activity of feature or information extraction from unstructured documents into pre-defined class such as medical codes, organization, quantities etc. There are two methods for named entity recognition such as ontology-based NER and deep learning based NER. Ontology based NER is based on knowledge recognition process which collects datasets with words, terms and interrelation between them. Here, rule-based and dictionary-based methods are used where it is robust but not portable. Only for precise extraction it is widely used and does not extract information from unknown entities. Based on level of ontology, the results are broad or complete to particular domain. In medical field, detailed ontology is needed due to complexity and different terminologies in it. On the other hand, deep learning NER is more accurate than ontology due to its feature, i.e., word embedding as it understands the semantic relation between words. It automatically learns and analyze topic-based and high-level words

in domain. It performs multiple tasks at once in less time than ontology NER.

Supervised machine learning methods are used in successful NER for which corpus in annotated manually with named entity with desired type. Then machine learning models are trained to automatically recognize entity in new text based on annotated corpus. But it is more expensive as it is needed to add manual annotation every single time for new entity. So alternate to supervised machine learning is unsupervised methods developed in natural language processing (NLP) which cannot be applied in biomedical domain. In general domain NER, they do not have variation in linguistic features. But in biomedical, there are various ways to determine the same entity as it is difficult to match entity. Then multiple tokens are nested with other named entity which is highly challenge to figure out the boundary of biomedical named entity. Semi-supervised machine learning approaches are useful in entity recognition if more unannotated text available. Most common methods for feature extraction are not used and available because it varies for every entity type. In order to identity multiple entity type, different machine learning algorithms with efficiency are needed.

In this work, biomedical NER is examined, as biomedical named entity recognition (BMNER) process consists of feature extraction and discovering domain knowledge. Knowledge from unstructured text or domain is provided by NLP and machine learning algorithms. The various entities in biomedical are drugs, genes, proteins, diseases etc. The challenges in BMNER compared to other NER are variation in size, no standard name for chemical structural compounds, entity boundary definition, abbreviations in clinical notes and reports, maintenance of clinical domain vocabulary etc. NLP used for feature extraction and helps in extraction of entity with specific domain. NLP with domain specific play major role in entity identification as it adapts functionalities as text varies in biomedical domain. Dataset in NER is important for machine learning approaches for efficient and robust recognition. More number of samples increases the recognition rate.

NER is different from other classification and take place in two processes, i.e., detecting entity boundary and assigning entity to pre-defined categories. Various machine learning approaches are used in biomedical NER such as Bayesian classification, hidden Markov model (HMM), maximum entropy (ME), support vector machine (SVM), conditional random field (CRF) etc. HMM is statistical method used in NER to recognize names and classify entities. HMM on labelled data to recognize entity in biomedical text and word similarity for unlabeled data are integrated to improve performance. The major disadvantage in HMM are it is unable to identity high order correlation among protein molecule and does not show dependencies between hidden states.

Maximum entropy in NER estimates probability based on assumptions other than constraints. Constraints are obtained during training with relationship between features and outcomes. It is similar to Bayesian classification. SVM in NER is used for classification of curatable and non-curatable biomedical components. The features are extracted using NLP and finally SVM classifies entity of specific domain. CRF is mostly used due to its ability in modelling multivariate output by utilizing more features for label prediction. CRF is used for structured prediction and statistical method as model use more contextual information with more features for prediction of various named entity type.

The goal of this proposed work is to improve the BMNER process in terms of chemical compound entities. The challenges in NER are overcome with NLP and machine learning algorithm. NLP is domain specific used to extract features from unstructured text and extracted features are learned by state-of-the-art algorithms. The novel method to recognize and extract biomedical entity is proposed and examined with two datasets GENIA and CHEMDNER corpus. The proposed work uses CRF for tagger scheme and TCN for classification which is more efficient and effective in recognition system. The proposed model examines with different features available in biomedical domain such as linguistic, orthographic, morphological

features etc. Linguistic features such as similarity in document, overlapping word, entity level overlapping and topics in document. Orthographic features such as indentation of text, numerals, capitalization, caps, symbols, punctuations etc., are systematic in NER. These features are extracted and learned by machine learning model to recognize entity from new text. Finally, features and methods are integrated for classification.

2. Related work

MeSH-based mapping is used to identify hierarchically related entities[1] in biomedical domain. Mapping-based approach such as ontology medical subject headings (MeSH) is used to map bio-entities and then hierarchically related entity are recognized. Unified Medical Language System (UMLS) and Metathesaurus are two methods used for recognition of named biomedical entity. Co-Occurrence Interaction Nexus with Named-Entity Recognition (CoINNER) is web-based tool for recognizing curatable domain, i.e., genes, chemicals, diseases etc. CoINNER[2] is used with algorithms to recognize biocreative IV CTD track. Conditional random field is also used for prediction of chemicals and diseases reported in articles. Performance is measured and shows that it is comparatively achieved accuracy compared to other existing algorithms. Chemical names from biomedical survey without engineering features is difficult to identify from large biomedical domain. Long short-term memory (LSTM), dynamic recurrent neural network (RNN) and conditional random field (CRF) are used to extract only meaningful features by embedding character and word[3]. It captures orthographic and morphological features from unstructured text without the help of engineering techniques which are manual. Many systems do not involve in embedding layer features, here deep learning based bi-LSTM and CRF is used to recognize the features of entity and recognize accurately[4]. CNN is used to enhance BNER by character level embedding in extracting entity feature from biomedical field.

Existing named entity recognition system used tools such as tmTool and ezTag[5] to identify entities which are learned traditionally. It does not identify new entity which is major drawback. The BERN tool is able to identify both known and new entities. Probability-based decision rules are introduced to identify overlapping entity types. NER consists of various errors such as grammatical error, error in spelling, some sentences are truncated and abbreviation with no standard[6]. Such errors are solved by CIMIND system which is multilingual system for entity recognition based on phonetic similarity. Natural language processing (NLP) is advanced method for data processing from different domain. Information in medical health record[7–10] is increased day by day and structured forms are challenges in identifying entity[11] in biomedical field.

Multi-task model learns common features as they share some layers which faces performance degradation in terms of single task learning[12] in labelling sequence in natural language processing. To overcome this, multi-task learning along with transfer learning is combined. LSTM and CRF used to achieve better accuracy compared to single and multiple task learning. Entity system use generic type classification but it is complex even for experts. Supervised method called L2AWE is used to recognize entity based on word embedding[13]. It also improves semantic feature used in both end and overcome errors and uncertainty in domain filed. Ontology is based various information from different source which is obtained either by noun phrase extraction or named entity recognition. There are many faults in noun phrase extraction[14] which are overcome by bi-LSTM for sequence classification based on input data. POS tag also improves the performance of entity recognition. To achieve high performance in biomedical entity recognition it uses lexicons and data pre-processing to extract entity such as genes and proteins[15]. The novel neural bi-LSTM and conditional random field are used to eliminate the engineering feature extraction methods. It achieves high performance than other existing deep learning algorithms. Due to internal sequential feature of CFR, performance is low so parallel solution is needed to improve. By combining limited-memory Broyden-Fletcher-Shanno (L-BFGS) and Viterbi algorithms[16] are used to handle time-consuming CFR model.

MapReduce is used to estimate parameters in biomedical domain. Due to lack of annotation data, performance of gene entity recognition is lower[17] so NLM gene corpus is used for gene entity recognition which is developed by US.

Some NER are sentence level approaches which results in inconsistent recognition[18]. Document level approach use contextual information from document where there is no relation between sentences from different document. To overcome this, cross document NER and multiclassification auxiliary task with coarse-grained is used for entity information and achieves more than document and sentence level NER. In real-world applications, they vary in topic, text, entity distribution etc.[19], which causes mismatch between application and structure. The Med-Flair is an NER tagger used for multiple entity with integration of bi-LSTM and conditional random fields for sequence tagger. Biomedical field such as polysemy and special characters makes NER difficult[20]. A hybrid method of Bidirectional Encoder Representations from Transformers (BERT) is developed to extract features from text and learns through bi-LSTM incorporating Multi Head Attention (MHATT) for chapter level entity feature extraction. And also, CFR is used for sequence tagger.

Many biomedical NER recognize entity based on flattened structure and ignore nested entities[21]. It contains more hidden information of domain specific entity type. To identify nested entity, boundary assembly (BA) model is used which consists of three process such as boundary identification, grouping into named entity and finally classifying false entity. BERT based named entity recognition is proposed based on complexity and ambiguity of data[22]. BERT generate different vectors for same word or entity. CNN extract local features and parallelly bi-LSTM extract interior features. For selecting best entity structure in sequence, CRF is used. Stack ensemble approach with fuzzy matching is proposed in biomedical NER[23] to combine output of two classifier for greater recognition accuracy. CFR is used for extracting underlying features and fuzzy logic is for matching disease with help of Rabin Karp and Tuned Boyer Moore algorithms. Features are integrated vis hidden Markov model (HMM) with back-off modeling[24] are proposed for named entity recognition. In addition to this, method for biomedical abbreviation identification and two methods for cascaded entity recognition is proposed. It is difficult and expensive to create manually large number of high-quality corpora[25]. Novel method is proposed which automatically generate named entity from UMLS. Bootstrapping approach is also used for labelled medical entity type with LSTM and CRF.

Drug named entity recognition (DNER) is used to identify drug names from unstructured text[26]. Interactive biomedical framework is proposed to access publicly available database to extract drug entity with help of machine learning algorithm. Adapting new type of genre and different type of entity is difficult in re-annotation. Unsupervised approach is proposed to recognize entity in stepwise manner by identifying entity boundary without handcrafted engineering features. Noun phase chunker followed by filter[27] based on inverse document frequency is used to extract feature entity from text with semantic features. Italian de-identification dataset created from COVID-19 clinical records[28] two multilingual deep learning system are proposed to de-identify medical records written in various languages.

Some methods avoid sentence level semantic information and general features of semantic and syntactic[29]. A novel LSTM is proposed with sentence level reading control gate (LS-BLSTM-CRF). In this method, sentence level gate is integrated to learn more features from all sentence in the document. Character level embedding is also introduced to learn out of vocabulary words to recognize entity. Identification of appropriate feature template and selection of feature plays important role[30] in successful entity recognition. Word clustering and selection-based feature reduction approach is proposed using maximum entropy (ME). It automatically selects features using knowledge in specific domain. Most of machine learning methods fails to pay attention in certain areas while extracting features[31] the attention-based BiLSTM-CRF model is proposed in order to obtain contextual information. The NER involves two processes: named entity detection (NED) and named entity classification (NEC) for extracting entity type from domain[32]. Six classifiers with

four toolkits are integrated for entity recognition based on multiagent strategy. Stacking method to identify correct entity type and learn the features from first process.

3. Proposed work

The objective of this proposed work is to identify named entity from biomedical domain literature. Conditional random field is widely used in recent work for named entity recognition because it has ability to combine and integrate with other models for classification. The view of the proposed system is shown in **Figure 1**. The entity which given as input was processed in two stages as encoder and decoder in TCN layer. Encoder which consists convolution layer and max-pooling layer to get the entity characteristics. The next level is decoder which includes convolutional layer and up sample layer which provides the class belongs to the entity. The classes of entity given to the CRF layer recognized the entity and predict the name of the entity based on the training results.

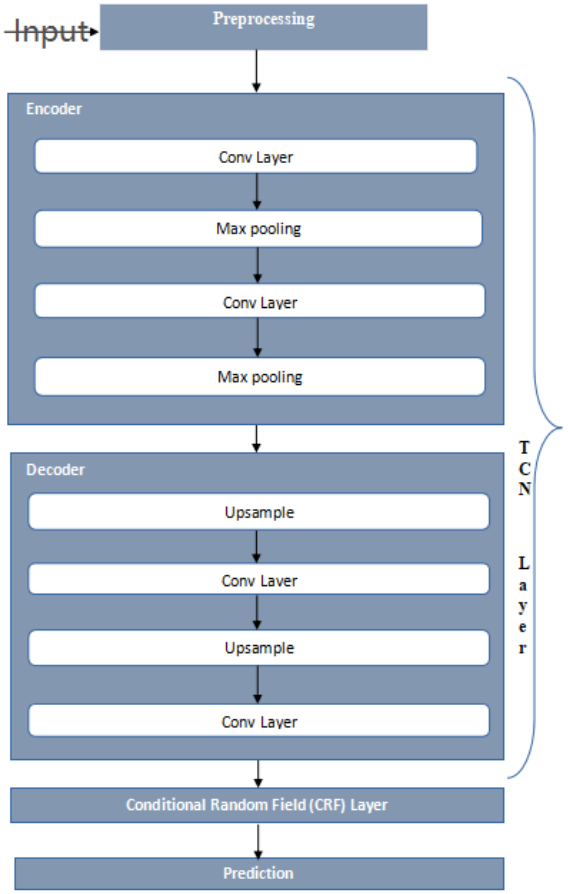

Figure 1. Proposed BNER using TCN with CRF.

Mainly conditional random field is used for sequence tagging by considering all features in input and label them with type sequentially. Temporal conventional network is state-of-art algorithm for classification with less computational time. It considers context information and identify entity boundary to detect entity and predict labels. The proposed method utilizes both methods to identify entity in biomedical field and tag in their respective entity type.

The disadvantage of CRF is it requires more computational time and space compared to other methods but TCN consumes less space and time. In this work, CRF is used to extract entity feature with sequence tagging scheme and TCN helps to identify nested and subtypes of extracted entity. TCN consists of three layers namely feature representation, TCN and softmax layer which is replaced by CRF layer. The feature representation layer has word and character vector which takes only word and character from input which is

integrated to represent semantic feature space. This integrated feature is given as input to TCN layer to extract features. Finally, the extracted feature from TCN layer is fed as input to last CRF layer.

The embedding layer is used to embed input word in a sequence to capture semantic feature, syntactic and morphological details of particular word. In the given sentence, each word is represented as vector to capture syntactic and semantic information of character and word. In this layer, feature is represented using character level and word level vector with predefined context semantic information to improve accuracy. TCN layer is convolutional layer used for solving sequence problem. In sequence tagging, integration of different features identify entity better. In the convolution layer, the kernel size is used differently to extract different features and integrate features extracted with different kernels. To handle computation by changing convolutional kernel, TCN uses dilated convolution which does not change input but only kernel size.

$$x_t = \prod_{t=1}^{T} P(y_t | y_1, y_2, \dots, y_{t-1})$$

The dilated convolution of TCN network can be written as

$$F(v) = (x \times_d f)(v) = \sum_{j=0}^{k-1} f(j) x_{v-d.i}.$$

In above equation, d represents dilated coefficient, k represents kernel size and $v - d.i$ determines which is upper layer. The dilated coefficient is used to check number of zero added between the convolutional kernel layer. The gradient vanishing is major problem that occurs in neural network and to overcome this problem TCN minimize layers in residual network by existing the dilated convolutional layer and ReLU layer within residual layer of TCN. The weight of each kernel is normalized by adding dropout. To solve sequence label problem, CRF layer in integrated to consider context information globally.

We introduce transfer score matrix $P_{i,j}$ for travelling from tag i to j which is start and end tag label in the given sentence with label type l. Let sentence type be n, then score matrix is and element of matrix determines output score under the label j. Given input sentence $X = \{x_1, x_2, \dots, x_n\}$ and tag sequence $y = \{y_1, y_2, \dots, y_n\}$, the total score $R(X, y)$ of sentence X along tag y is calculated as

$$R(X, y) = \sum_{i=0}^{n} P y_i, y_{i+1} + \sum_{i=1}^{n} M_i, y_i.$$

The probability distribution for the sequence y is given as

$$P(y|X) = \frac{e^{R(X,y)}}{\sum y_{\in y_x} e^{(R(X,y))}}.$$

Logarithmic probability of tag sequence is given by

$$\log\big(P(y^*|X)\big) = R(X|y^*) - \log\big(\sum y_{\in y_x} e^{(R(X,y))}\big).$$

Above equation used to generate correct tag sequence by using proposed model and at decoding the sequence with high score is predicted as optimal sequence as

$$y^* = argmax R(X, y).$$

Viterbi algorithm is used during prediction process for solving optimal sequence problem.

4. Result and evaluation

4.1. Dataset

In this work, for biomedical named entity recognition GENIA and CoNLL-2003 dataset corpus are used. Both datasets are used in order to show effectiveness and improvement in convolutional layer of TCN and CRF with generic field. The GENIA corpus consists of biomedical reports and text with structured and unstructured label. It has five entity types such as protein, cell line and type, gene: DNA and RNA. To avoid inconsistencies during testing, the training and testing datasets are predefined which is shown in **Table 1**.

Table 1. Entity in GENIA corpus.

	Protein	Cell line	Cell type	DNA	RNA	Total
Training	30,145	9643	851	6614	3720	50,973
Testing	5256	1076	108	1812	400	8652

CoNLL-2003 corpus is examined to showcase effectiveness and convolutional improvement in the kernel. It consists of entities such as name of person, location, organization and other medical field entities. It is divided into testing, training and validation set to improve overall performance compared to other models as shown in **Table 2**.

Table 2. Entity in CoNLL-2003 corpus.

	PER	ORG	LOC	MISC	Total
Training	5600	6521	7045	3348	22,244
Validation	1732	1452	1737	822	5743
Testing	1512	1771	1568	602	5453

To evaluate performance, known measures are used in biomedical named entity recognition namely precision, recall and F-score. Precision is measure of correctly recognized entity, recall is measure of overall correctly recognized entity and F-score is measure of both precision and recall.

4.2. CRF layer in BMER

To provide BMER system with more effectiveness, CRF layer is integrated with TCN model by using softmax layer. The integrated model on both datasets of GENIA and CoNLL-2003 corpus results are shown in **Table 3**. The effective result shows that TCN and CRF performance is marginally higher on GENIA corpus than on CoNLL-2003. The CRF layer is used mainly to consider relation between entity type and to correctly recognize entity labels which improve performance better. Usually biomedical/clinical field has more terms and alternate words where CRF extract even nested subtypes of entity type and produce better biomedical recognition.

Table 3. Performance on both dataset.

Dataset	Precision	Recall	F-score
GENIA	91.58	91.43	91.54
CoNLL-2003	86.65	84.32	85.78

Table 4. Comparison with other models on GENIA corpus.

Model	Specific feature	F-score
CNN-CRF	Word to vector, POS	71.34
SVM-CRF	-	76.98
Multilayer bi-LSTM	Word embedding	79.03
LSTM-CRF	Word, character	82.45
BERT	-	74.25
TCN-CRF	GLoVE, character	91.42
TCN-CRF (IPSO)	Word embedding, character	94.54

Table 5. Comparison with other models on CoNLL-2003 corpus.

Model	Specific feature	F-score
CNN-CRF	Word to vector, POS	83.56
SVM-CRF	-	75.34
BiLSTM-CNN	Word embedding	74.12
BiLSTM-CRF	Word, character	76.87
BERT	-	72.48
TCN-CRF	GLoVE, character	85.78
TCN-CRF (IPSO)	Word embedding, character	89.45

4.3. Comparison with other models

The proposed model is compared with existing machine learning methods such as CNN with CRF, SVM-CRF, multilayer bidirectional LSTM, LSTM-CRF and BERT on both GENIA and CoNLL-2003 corpus which is shown in **Tables 4** and **5**.

4.4. Comparison of training time

Figures 2 and **3** shows the training time of machine learning models on GENIA and CoNLL-2003 corpus. Different models take different computational time to learn features and validate the entity type. The proposed method takes less time to learn both local and nested subtype features of entity in biomedical field. Dataset vary in size, entity type and their relation between various entities based on these parameter, computational time is measured.

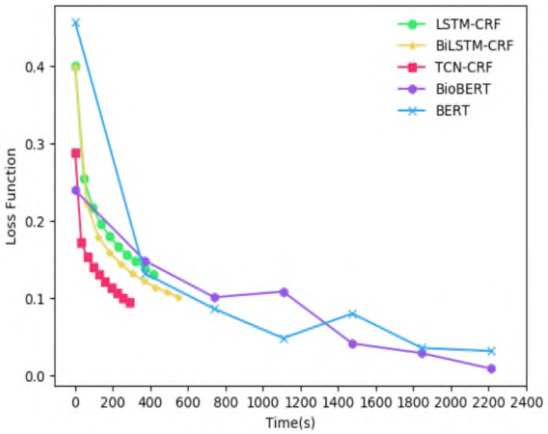

Figure 2. Comparison of training time on GENIA corpus.

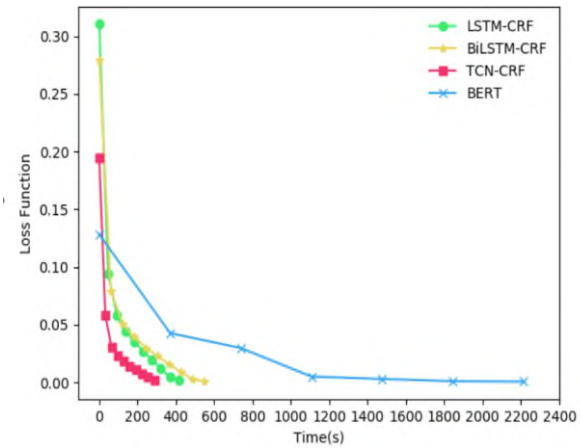

Figure 3. Comparison of training time on CoNLL-2003 corpus.

5. Conclusion

The proposed method integrates various methods to identify biomedical named entity and nested subtypes. Various features are extracted from clinical reports and learned by the supervised model to recognize entity and relation between the entity types. Efficiency of BMNER is examined with various feature sets which includes processing of data. TCN and CRF is used because of time it consumes to learn and train data is low and for sequence tagging scheme. The evaluated result shows that efficiency of proposed model extract entities from all corpuses. The performance in GENIA is higher when compared to CoNLL-2003 corpus. The training time of proposed method is compared with other machine learning models and it achieves comparative performance. Though TCN and CRF is traditional neural method, it performs better with varying in data size and type. To enhance performance, model which use unlabeled corpus to recognize and provide common solution should be adopted.

Author contributions

Conceptualization, TTM, SR, SG and OKCU; methodology, TTM and SR; software, TTM and SR; validation, TTM, SG, OKCU and SK; formal analysis, TTM; investigation, TTM; resources, SK; data curation, SK; writing—original draft preparation, TTM and SR; writing—review and editing, TTM and SR; visualization, SG and OKCU; supervision, SG and OKCU; project administration, SK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Reference

1. Yang H, Dong Y. Recognizing hierarchically related biomedical entities using MeSH-based mapping. *Tsinghua Science and Technology* 2012; 17(6): 609–618. doi: 10.1109/TST.2012.6374362
2. Hsu YY, Kao HY. Curatable named entity recognition using semantic relations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2015;12(4): 785–792. doi: 10.1109/TCBB.2014.2366770
3. Gajendran S, Manjula D, Sugumaran V. Character level and word level embedding with bidirectional LSTM-Dynamic recurrent neural network for biomedical named entity recognition from literature. *Journal of Biomedical Informatics* 2020; 112: 1036096. doi: 10.1016/j.jbi.2020.103609
4. Cho M, Ha J, Park C, Park S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics* 2020; 103: 103381. doi: 10.1016/j.jbi.2020.103381
5. Kim D, Lee J, So CH, et al. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* 2019; 7: 73729–73740. doi: 10.1109/ACCESS.2019.2920708
6. Cabot S, Darmoni S, Soualmia LF. Cimind: A phonetic-based tool for multilingual named entity recognition in biomedical texts. *Journal of Biomedical Informatics* 2019; 94: 103176. doi: 10.1016/j.jbi.2019.103176
7. Kumar CUO, Gajendran S, Bhavadharini RM, et al. EHR privacy preservation using federated learning with DQRE-Scnet for healthcare application domains. *Knowledge-Based Systems* 275; 275: 110638. doi: 10.1016/j.knosys.2023.110638
8. Kumar CUO, Gajendran S, Balaji V, et al. Securing health care data through blockchain enabled collaborative machine learning. *Soft Computing* 2023; 27(14): 9941–9954. doi: 10.1007/s00500-023-08330-6
9. Chennam KK, Maheshwari VU, Aluvalu R. Maintaining IoT healthcare records using cloud storage. In: Nath Sur S, Balas VE, Bhoi AK, et al. (editors). *IoT and IoE Driven Smart Cities*. Springer, Cham; 2021. pp. 215–233.
10. Siarry P, Jabbar MA, Aluvalu R, et al. *The Fusion of Internet of Things, Artificial Intelligence, and Cloud Computing in Health Care*, 1st ed. Springer Cham; 2021. pp. 1–23.
11. Śniegula A, Poniszewska-Maranda A, Chomatek Ł. Study of named entity recognition methods in biomedical field. *Procedia Computer Science* 2019; 160: 260–265. doi: 10.1016/j.procs.2019.09.466
12. Mehmood T, Gerevirini AE, Lavelli A, Serina I. Combining multi-task learning with transfer learning for biomedical named entity recognition. *Procedia Computer Science* 2020; 176: 848–857. doi: 10.1016/j.procs.2020.09.080
13. Nozza D, Manchanda P, Fersini E, et al. LearningToAdapt with word embeddings: Domain adaptation of named entity recognition systems. *Information Processing and Management* 2021; 58(3): 102537. doi: 10.1016/j.ipm.2021.102537

14. Santoso J, Setiawan EI, Purwanto CN, et al. Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short-term memory. *Expert Systems with Applications* 2021; 176: 114856. doi: 10.1016/j.eswa.2021.114856
15. Gridach M. Character-level neural network for biomedical named entity recognition. *Journal of Biomedical Informatics* 2017; 70: 85–91. doi: 10.1016/j.jbi.2017.05.002
16. Li K, Tang Z, Zhang F, et al. Hadoop recognition of biomedical named entity using conditional random fields. *IEEE Transactions on Parallel and Distributed Systems* 2015; 26(11): 3040–3051. doi: 10.1109/TPDS.2014.2368568
17. Islamaj R, Wei CH, Cissel D, et al. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of Biomedical Informatics* 2021; 118: 103779. doi: 10.1016/j.jbi.2021.103779
18. Wang D, Fan H, Liu J. Learning with joint cross-document information via multi-task learning for named entity recognition. *Information Sciences* 2021; 579: 454–467. doi: 10.1016/j.ins.2021.08.015
19. ElDin HG, AbdulRazek M, Abdelshafi M, Saglol AT. Med-Flair: Medical named entity recognition for diseases and medications based on flair embedding. *Procedia Computer Science* 2021; 189: 67–75. doi: 10.1016/j.procs.2021.05.078
20. Liu J, Gao L, Guo S, et al. A hybrid deep-learning approach for complex biochemical named entity recognition. *Knowledge-Based Systems* 2021; 221: 106958. doi: 10.1016/j.knosys.2021.106958
21. Chen Y, Hu Y, Li Y, et al. A boundary assembling method for nested biomedical named entity recognition. *IEEE Access* 2020; 8: 214141–214152. doi: 10.1109/ACCESS.2020.3040182
22. Zhang Q, Sun Y, Zhang L, et al. Named entity recognition method in health preserving field based on BERT. *Procedia Computer Science* 2021; 183: 212–220. doi: 10.1016/j.procs.2021.03.010
23. Bhasuran B, Murugesan G, Abdulkadhar S, Natarajan J. Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. *Journal of Biomedical Informatics* 2016; 64: 1–9. doi: 10.1016/j.jbi.2016.09.009
24. Zhang J, Shen D, Zhu G, et al. Enhancing HMM-based biomedical named entity recognition by studying special phenomena. *Journal of Biomedical Informatics* 2004; 37(6): 411–422. doi: 10.1016/j.jbi.2004.08.005
25. Kim J, Ko Y, Seo J. A bootstrapping approach with CRF and deep learning models for improving the biomedical named entity recognition in multi-domains. *IEEE Access* 2019; 7: 70308–70318. doi: 10.1109/ACCESS.2019.2914168
26. Chukwuocha C, Mathu T, Raimond K. Design of an interactive biomedical text mining framework to recognize real-time drug entities using machine learning algorithms. *Procedia Computer Science* 2018; 143: 181–188. doi: 10.1016/j.procs.2018.10.374
27. Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics* 2013; 46(6): 1088–1098. doi: 10.1016/j.jbi.2013.08.004
28. Catelli R, Gragiulo F, Casola V, et al. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. *Applied Soft Computing* 2020; 97: 106779. doi: 10.1016/j.asoc.2020.106779
29. Li L, Jiang Y. Integrating language model and reading control gate in BLSTM-CRF for Biomedical named entity recognition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2020; 17(3): 841–846. doi: 10.1109/TCBB.2018.2868346
30. Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. *Journal of Biomedical Informatics* 2009; 42(5): 905–911. doi: 10.1016/j.jbi.2008.12.012
31. Wei H, Gao M, Zhou A, et al. Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. *IEEE Access* 2019; 7: 736237–73636. doi: 10.1109/ACCESS.2019.2920734
32. Li L, Fan W, Huang D. A two-phase bio-NER system based on integrated classifiers and multiagent strategy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2013; 10(4): 897–904. doi: 10.1109/TCBB.2013.106