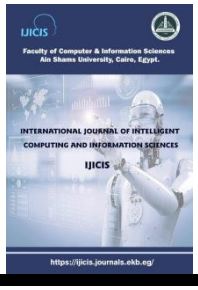
IJICIS, Vol.22, No.1, 98-107
DOI: 10.21608/ijicis.2022.104170.1134

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



Nested Biomedical Named Entity Recognition

Lobna A. Mady*
Information Systems, Faculty of Computer and Information Sciences, Ain shams University, Cairo, Egypt
lobna.mady@cis.asu.edu.eg

Yasmine M. Afify
Information Systems, Faculty of Computer and Information Sciences, Ain shams University, Cairo, Egypt
yasmine.afify@cis.asu.edu.eg

Nagwa L. Badr
Information Systems, Faculty of Computer and Information Sciences, Ain shams University, Cairo, Egypt
nagwabadr @cis.asu.edu.eg

Received 2021- 11-03; Accepted 2022-01-18

Abstract: Named entity recognition has been regarded as an important task in natural language processing. Extracting biomedical entities such as RNAs, DNAs, cell lines, proteins, and cell types has been recognized as a challenging task. Most of the existing research focuses on the extraction of flat named entities only and ignores the nested entities. Nested entities, on the other hand, are commonly used in real world biomedical applications due to their ability to represent semantic meaning of the named entity. This paper proposes an approach to improve the performance of nested biomedical named entity recognition by using a combination of diverse types of features namely morphological, orthographical, context, part of speech and word representation features while using Structured Support Vector Machine as a machine learning technique. The results obtained from the proposed approach were compared with those from popular benchmark approaches. The popular dataset "Genia" is utilized to evaluate the proposed approach which achieved Recall, Precision and F1-Measure of 84.033%, 85.946 %, and 84.113% respectively.

Keywords: Machine Learning, Nested Entities, Classification, Biomedical Named Entity Recognition.

1. Introduction

Information Extraction is the process where unstructured text, such as newspaper articles and research articles, is used to extract structured information. Recognizing information components for example names (organizations, locations, and people), numeric terms (date and time), and percent expressions (money) is a key sub-task for the IE process. Named Entity Recognition (NER) is the task of distinguishing references to these entities in text. Biomedical Named Entity Recognition (BNER) is one

* Corresponding author: Lobna A. Mady
Information Systems, Faculty of Computer and Information Sciences, Ain shams University,Cairo, Egypt
E-mail address: lobna.mady@cis.asu.edu.eg

of the NER techniques which detects biomedical entities such as Proteins, RNAs, Cell lines, DNAs, Cell types, and Viruses. There are two types of biomedical Named Entities (NEs): flat entities and nested entities. Nested entities are entities that are embedded in other entities such as shown in Figure 1. Recognizing nested biomedical NEs aids in the extraction of more accurate semantic information from text.

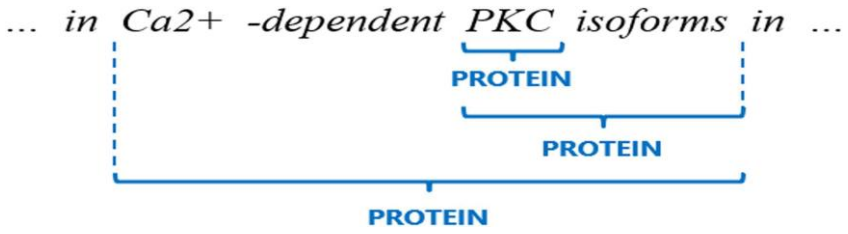


Figure 1: Example of biomedical nested entities

According to [1], BNER has four primary methodologies: Rule-based approaches where several rules are developed depending on entity class label [2]. However, it is impossible to apply rules developed for one corpus to other corpora. Second, dictionary-based approaches, which extract NEs in predefined dictionaries that consist of large collections of entities for each class label [3]. Building those dictionaries is an effort and time-consuming process as the numbers of biomedical entities increase in a rapid speed. Third, Machine Learning (ML) approaches, and finally Hybrid based approaches, which combine one or more from previously mentioned methodologies [4].

ML approaches have different types of algorithms such as Conditional Random Fields (CRF) [5], Hidden Markov Model [6] and Support Vector Machines (SVMs) [7]. ML approaches are used in many fields such as predicting behavior of users [8], Emotion detection [9], and Disease Detection [10]. Feature selection is a main task when applying ML techniques for NE extraction from biomedical text. Selecting the appropriate feature set to characterize the tokens and entities is usually a difficult task. The aim of this selection is to find the best collection of attributes for developing models of the phenomena under investigation. Different types of features were utilized in NER such as linguistic, orthographic, morphological and context features.

Nevertheless, the majority of the aforementioned methodologies are unable to handle nested NEs. BNER has traditionally been treated as a series of labelling problem where each entity in the sentence receives a single class label. These methodologies are usually developed based on the incorrect assumption that the entities do not overlap (flat entities). However, entities in real world languages can be highly nested or overlapped. This task has received a lot of attention from research community. However, there is still potential development in terms of enhancing the performance of nested BNER.

Taking into consideration the importance of nested BNER, the primary goal of this paper is to enhance the performance of nested BNER by using Structured Support Vector Machine (SSVM) and a combination of various types of features. In order to assess proposed approach, we used the popular dataset Genia [11–13] and three commonly used evaluation metrices namely Recall, Precision, and F1-measure [14]. The paper begins with examining the most recent research in the field of Nested BNER. Then, in the third section the proposed approach is presented including selected features and the adopted ML technique. Afterwards, in the fourth section the selected dataset and evaluation metrics are provided. Then, in the fifth section the obtained experimental results are then tabulated and reviewed. Finally, the main conclusion of the conducted study is outlined.

2. Related Work

Different research studies have been conducted on nested BNER in the literature [11-17]. A novel technique to solve the problem of nested BNER was proposed by [11] where each sentence was converted into a tree. This approach precisely depicts the nested structure and allows influencing the entities by labels of the adjacent words and the entities that are contained in them. The authors of [11] visualized each sentence as a parse tree, with words and sentences simulated as leaves and entities, respectively. Each node is then marked with both its parent and grandparent labels. Word Embedding (WE), word shape, context, and Part of Speech (POS) features on each label are then employed to enhance the performance.

Nested BNER problem was represented by [12] as a directed hypergraph. The authors formulated the structured prediction issue as the construction of a hypergraph encoding all entities in the input sentence's token-level. Edge probabilities were calculated by assigning probabilities to all possible edges from a tail node, which aided in the hypergraph's greedy construction. Long Short-Term Memory Networks (LSTM) based sequence labeling model was used to learn the nested biomedical NE hypergraph for an input sentence. To extract entity mentions, softmax was used to assign probabilities to the various types of edges in the hypergraph. Then, the edge with the highest probability and hyperarcs with probability over a predetermined threshold for each token was chosen.

Authors of [13] used Bidirectional LSTM (BiLSTM) to generate WE which finds contextualized representation of each token. Then, a multitask learning technique was employed to manage the nested NE. WE was retrieved by generating position embedding dimension using a shared trainable embedding matrix and corresponding syntax representation for each sentence. This syntax representation contains the POS and constituency relations of each token. Using those syntax and position embedding, the attention module assessed the context representation value which were used to feed a CRF layer to predict the possible class labels.

The task of nested NER was divided into three models [15]: Identifying boundaries, assembling candidates, and distinguishing actual NEs. In identifying boundaries model, the tokens were used as input to character-level and word level embedding layers. The output from those embedding layers was used as to feed the Bi-LSTM layer. The output was then used to feed CRF layer for detecting the boundaries of NEs. Following the detection of NE boundaries, they were reassembled into NE candidates using greedy matching method for further investigation. In the distinguishing actual NEs model, a Multiple Convolutional Neural Networks (MultiCNN) model was designed to predict whether the candidates were a correct NEs.

Layered model for nested NER was proposed by [16], which was called Pyramid and made up of a group of interconnecting layers. Each layer determined if a text region was a complete entity. The hidden state sequence was used as input into a Convolutional Neural Networks (CNN). The higher layer between each two consecutive layers aggregated two nearby hidden states from the lower layer, resulting in the pyramid shape. The authors represented each word by concatenating the character embeddings, generated by a LSTM, and the word embeddings to get the morphological and orthographic features.

Authors of [17] treated entity recognition as a span classification task where entities were represented using word embedding, character level embedding, contextualized word embedding, and POS embedding to generate seed spans. These spans were sampled from a sequence of words where the suggested spans have higher overlap with entities while contextual spans have lesser overlap. To maintain the proposal spans and remove the contextual spans, the authors employed a filter and

calculated the chance that the span belongs to the span suggestions. Meanwhile, a regressor identified each span's border to locate the entity left and right the boundaries. Based on the regressor's output, the limits of the span suggestions were altered and used as input into the entity classifier module based (Soft Non-Maximum Suppression, Soft-NMS) method.

3. Proposed Approach

The proposed approach is explained in full in this section, together with the feature set and the ML algorithm.

3.1. Feature Set

In our approach, Morphological, Orthographical, Context, POS, and WR features are used. These features were chosen because they are frequently used in BNER [18,19]. Details about these features are presented as follows:

Morphological features examine the word components and their interactions and look at how words have similar structures.
Orthographical features group words that have similar forms together and commonly used to capture information about how the words are created.
Context features detect words that come before or after the token to figure out what its class label.
Part Of Speech features recognize NEs POS information. Nouns are usually strong candidates for NEs, whereas verbs and prepositions usually reveal NEs bounds.
Word Representation (WR) features take a string representing a word as input and produce a collection of values that represent a word in a vector space. It is one of the basic building components in Natural Language Processing.
WE feature is a type of WR which is a powerful technique for representing words because WE represents words with comparable meanings in the same way. Semantically similar tokens are assigned similar vectors. [20,21] proved that WE trained on biomedical tokens significantly improved BNER model performance. In the current approach, Skip Gram Vector [22] is chosen because it is well-suited to the training of uncommon terms that frequently appear in biomedical contents [23]. Table 1 shows examples of WE vector of nested Genia tokens.

Table 1 Examples of WE vector of Nested Genia tokens

Token	Word embedding vector
NF - kappa B	0.901, -0.389 ,0.453 ,0.307, 0.070, -0.013, 0.059, -0.176, -0.494, 0.058, 0.001, -0.122, 0.142, 0.022, 0.107, -0.033, -0.714, 0.396, 0.517, 0.298
resting normal human PBL	0.026, 0.025, 0.023, 0.005, -0.006, -0.155, 0.068, 0.161, 0.310, -0.015, 0.058, 0.043, -0.121, -0.012, -0.026, -0.045, -0.094, -0.262, -0.070, 0.061
caspase - 3	-0.106, 0.047, -0.213, 0.431, -0.040, 0.241, 0.147, -0.034, 0.013, 0.257, -0.055, 0.087, -0.121, 0.323, 0.212, 0.085, 0.569, 0.259, 0.382, -0.032
normal human monocytes	-0.102, 0.147, -0.760, -0.313, 0.256, -0.014, -0.482, -0.222, 0.154, 0.203, -0.525, 0.184, -0.004, -0.082, -0.012, 0.231, 0.167, 0.605, 0.687, 0.234

Clustering Based feature is a type of WR which creates clusters over tokens and represents each token by its the cluster. The concept is that semantically or syntactically comparable words tend to cluster

together in the same or close clusters. More details regarding the set of features utilized in the proposed approach are provided in Table 2.

3.2. Machine learning technique

The generated feature vector from the previously mentioned features is used as input to Structured Support Vector Machine (SSVM) that is an enhanced version of SVM algorithm [31] which is used as a ML technique in the proposed approach. The SSVM can be trained for general structured output labels, while the SVM classifier can be trained for regression, binary classification, and multiclass classification. SSVM combines the benefits of both CRFs and SVMs in a single algorithm and requires less training time. SSVM translates the sequence labeling problem by Viterbi algorithm and Markov chain. SSVMs model sequence labelling issues using the big margin method, which has strong generalization capacity. SVMhmm, developed by [31], is used as an implementation of SSVM.

4. Experiment and Evaluation

In the following subsections, the employed dataset, and the evaluation metrices are presented.

4.1. Dataset

A commonly used GENIA dataset Version 3.0.2 [32], [11–13] is empolyed for nested BNER in the current paper. Seventeen percent of biomedical NEs in the GENIA corpus are embedded within a different NE with a maximum nested level of three. Genia dataset contains 1999 abstract records which were extracted from the MEDLINE database. The Genia dataset includes the following five class labels: RNA, protein, cell type, DNA, and cell line. As proposed by [11] and [33], the first ninety percent of the sentences in the dataset were utilized for training, while the left ten percent were utilized for testing.

4.2. Evaluation Metrics

Three widely used evaluation measures are employed in the experiments to assess the proposed SSVM approach: Recall, Precision, and F1-measure. The three metrics are calculated using Eqs. 1, 2, and 3, respectively [14]. The number of NEs accurately recognized by the system is known as True Positive (TP). The number of NEs that are not identified is known as False Negative (FN), while the number of NEs that the system misidentifies is known as False Positive (FP).

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

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

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

Table 2 Explanation of features employed in the proposed approach

Feature Type	Feature Name	Description	Example
Morphological	**Prefix**	Refers set of characters that are taken from the leftmost location of the words. The prefix length rage from 3 to 7 depend on the number of characters of the tokens	"CD28 surface receptor" prefix is (CD28)
	Suffix	Refers set of characters that are taken from the rightmost location of the words. The suffix length rage from 3 to 7 depend on the number of characters of the tokens	"CD28 surface receptor" suffix is (receptor)
	Word Shape	Provides static methods for mapping any String to another String based on its "word shape" for example, whether it's capitalized, numeric, or otherwise.	"tal–1" is assigned to aaa-1
Orthographical	**ALLCAPS**	Checks if the word's letters are all capitalized	G - CSF
	INTCAP	Checks if the word's first letter is capitalized	mature B lymphocytes
	HASCAP	Checks if any of the word's letters are capitalized	mRNA
	SINGLECAP	Checks if the token only has only one upper case letter	glutathione S - transferase
	CAP&DIGIT	Check to see if the word has a combination of digits and uppercase letters.	E2
	Digit&Alpha	Check if the token has a mix of digits and letters.	HIV - 1 tat
	CAP&ALPHA	Check if the token has a combination of lowercase and uppercase letters	mRNA
	ALLDIGIT	Checks that the token only contains digits	42
	Alpha&Digit	Check if the token starts with an alphabet and the remainder of the characters are integers	p53
	DigitSpecial	Check if the word starts with number, then special character	5' - flanking region
	AlphaDigitAlpha	Check if the token begins with a character, then a digit, and finally a character.	557 and - 417 zeta
	DigitCommaDigit	Check if the first letter of the word is a digit, then a comma, and finally a digit.	32 , 36 to 42 and 110 kD proteins
	DigitDotDigit	Check if the first letter of the word is a digit, dot then digit.	0.6
	HasRoman	Check if word has a Roman letter	II, IV
	HasGreek	Check if word has a Greek letter	Beta, Alpha
Context	**Context feature**	Refer to tokens that appear within a 5 word window size as proposed in [24]	i.e., Two tokens to the right and two tokens to the left of the token
Part of speech (POS)	**POS**	Genia Tagger [25] is used to extract POS tags of each token	i.e., NNP, VBZ, NN
	Context words POS	Genia Tagger [25] is used to extract POS tags of each token	i.e., NNP, VBZ, NN
Word Representation (WR)	**WE**	Word2vec is used for Skip Gram WE feature for each token [26]	
	Clustering Based feature	Brown clustering [27] is used as implementation for cluster based WR to enhance performance of BNER, as proposed by [28–30].	Example of Brown Cluster classes: "noncycling b cells" cluster is "0111100" while "g - protein - coupled serpentine receptors" cluster is "01111010"

5. Results

Figure 2 displays the recall, precision, and F1-measure of each class label achieved from the proposed approach. Table 3 demonstrates the obtained results from proposed approach against other benchmark approaches. It can be noted that SSVM surpasses all other benchmark approaches by achieving a Recall, Precision, and F1-Measure of 84.033%, 85.946 %, and 84.113%, respectively.
Compared to the benchmark approach [11], the proposed approach showed an improvement of 19.6% in F1-Measure as well as an improvement of 11.46 % and 30.42% in the Recall and Precision, respectively. Moreover, the SSVM results in improvement of 5.3%, 26%, and 13.97 % in Recall, Precision and F1-Measure, respectively compared to [12]. Furthermore, an improvement of 3.8%, 16%, and 9.5% for Recall, Precision and F1-Measure, respectively compared to [13]. And compared to [16] an improvement of 6.9%, 11.59%, and 8.1% in Recall, Precision and F1-Measure, respectively. While an improvement of 4.4% in F1-Measure was employed by our approach compared to [17] as well as an improvement in the Recall and Precision by 4.8% and 6.2%, respectively. Finally, an improvement in Recall, Precision and F1-Measure by 1.3%, 7.7%, and 3.4%, respectively compared to benchmark approach [15] .

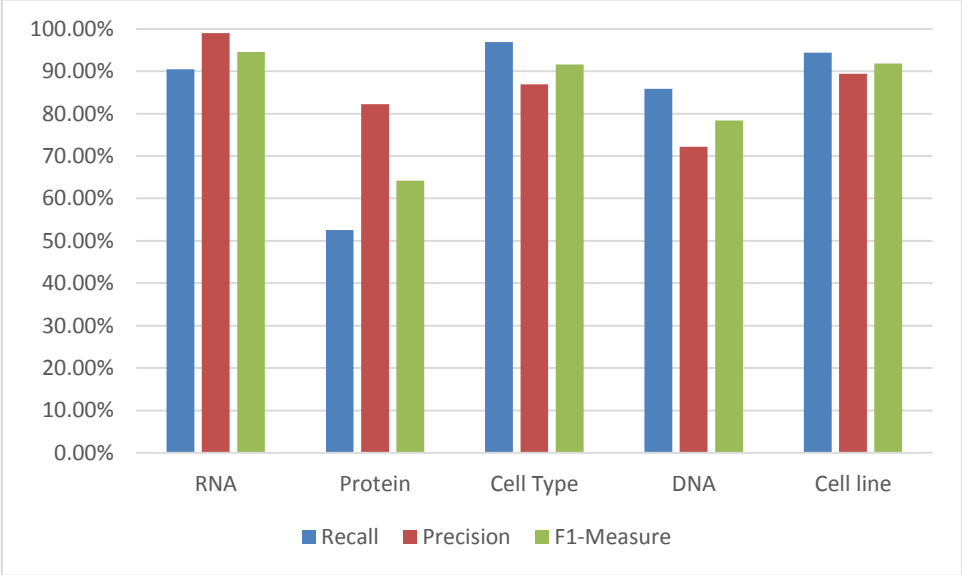


Figure 2: The overall Recall, Precision and F1-Measure for each class label in Genia using SSVM

Table 3 Performance Comparison of the proposed approach against benchmark approaches

Approach	Recall	Precision	F1-Measure
SSVM (Our Approach)	**84.03**	**85.95**	**84.11**
Nested Named Entity Recognition [11]	75.39	65.90	70.33
Nested Named Entity Recognition Revisited [12]	79.8	68.2	73.8
Recognizing Nested Named Entity in Biomedical Texts: A Neural Network Model with Multi-Task Learning [13]	80.9	73.8	76.8
A Boundary Assembling Method for Nested Biomedical Named Entity Recognition [15]	82.96	79.77	81.34
A Layered Model for Nested Named Entity Recognition [16]	78.60	77.02	77.78
Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition [17]	80.19	80.89	80.54

6. Conclusion

In biomedical data, nested entities are very prevalent and has significant influence in enhancing the performance of BNER. In this paper, SSVM was utilized to enhance the performance of nested BNER such as genes, protein, cell line and cell types. SSVM was utilized in a combination of different types of features such as morphological, orthographical, context, part of speech, and word representation features to improve the extraction performance. Comprehensive evaluation was conducted using three popular evaluation metrics namely Recall, Precision, and F1-measure. Using Genia dataset, the proposed approach surpassed six benchmark approaches by an improvement percentage of 3.4% ~ 19.6%.

References

[1] S. Suman, A. Dash, S.S. Rautaray, A Literature Survey on Biomedical Named Entity Recognition, Advances in Power Systems and Energy Management. 690 (2021) 109–119. https://doi.org/10.1007/978-981-15-7504-4_12.

[2] F. Olsson, G. Eriksson, K. Franzén, L. Asker, P. Lidén, Notions of correctness when evaluating protein name taggers, in: Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, 2002: pp. 1–7. https://doi.org/10.3115/1072228.1072338.

[3] K.M. Hettne, R.H. Stierum, M.J. Schuemie, P.J.M. Hendriksen, B.J.A. Schijvenaars, E.M. Van Mulligen, J. Kleinjans, J.A. Kors, A dictionary to identify small molecules and drugs in free text, Bioinformatics. 25 (2009) 2983–2991. https://doi.org/10.1093/bioinformatics/btp535.

[4] R.T.-H. Tsai, A hybrid approach to biomedical named entity recognition and semantic role labeling, in: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006: pp. 243–246. https://doi.org/10.3115/1225797.1225806.

[5] B. Settles, Biomedical named entity recognition using conditional random fields and rich feature sets, in: International Conference on Fuzzy Systems and Knowledge Discovery, 2004: p. 104. https://doi.org/10.3115/1567594.1567618.

[6] N. Collier, C. Nobata, J. Tsujii, Extracting the names of genes and gene products with a hidden Markov model, in: COLING 2000 Volume 1: The 18th International Conference on

Computational Linguistics, 2000: pp. 201–207. https://doi.org/10.3115/990820.990850.
[7] Z. Ju, J. Wang, F. Zhu, Named entity recognition from biomedical text using SVM, in: 5th International Conference on Bioinformatics and Biomedical Engineering, ICBBE 2011, 2011. https://doi.org/10.1109/icbbe.2011.5779984.
[8] K. Ibrahim, M. Aborizka, F. Maghraby, Prediction of users charging time in cloud environment using machine learning, International Journal of Intelligent Computing and Information Sciences. 18 (2018) 39–57. https://doi.org/10.21608/ijicis.2018.30121.
[9] S. Ibrahiem, K. Bahnasy, M. Morsey, M. Aref, Feature Extraction Enchancement in Users' Attitude Detection, International Journal of Intelligent Computing and Information Sciences. 18 (2018) 1–13. https://doi.org/10.21608/ijicis.2018.30115.
[10] B. Elshoky, O. Ibrahim, A. Ali, Machine Learning Techniques Based on Feature Selection for Improving Autism Disease Classification, International Journal of Intelligent Computing and Information Sciences. 21 (2021) 65–81. https://doi.org/10.21608/ijicis.2021.61582.1058.
[11] J.R. Finkel, C.D. Manning, Nested named entity recognition, in: In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: pp. 141–150. https://doi.org/10.3115/1699510.1699529.
[12] A. Katiyar, C. Cardie, Nested named entity recognition revisited, in: In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: pp. 861–871. https://doi.org/10.18653/v1/n18-1079.
[13] H. Fei, Y. Ren, D. Ji, Recognizing Nested Named Entity in Biomedical Texts: A Neural Network Model with Multi-Task Learning, in: Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, 2019: pp. 376–381. https://doi.org/10.1109/BIBM47256.2019.8982966.
[14] C.D. Manning, P. Raghavan, H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.
[15] Y. Chen, Y. Hu, Y. Li, R. Huang, Y. Qin, Y. Wu, Q. Zheng, P. Chen, A Boundary Assembling Method for Nested Biomedical Named Entity Recognition, IEEE Access. 8 (2020) 214141–214152. https://doi.org/10.1109/ACCESS.2020.3040182.
[16] J. WANG, L. Shou, K. Chen, G. Chen, Pyramid: A Layered Model for Nested Named Entity Recognition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: pp. 5918–5928. https://doi.org/10.18653/v1/2020.acl-main.525.
[17] Y. Shen, X. Ma, Z. Tan, S. Zhang, W. Wang, W. Lu, Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition, ACL/IJCNLP. (2021) 2782–2794. https://doi.org/10.18653/v1/2021.acl-long.216.
[18] J. Zhang, D. Shen, G. Zhou, J. Su, C.L. Tan, Enhancing HMM-based biomedical named entity recognition by studying special phenomena, Journal of Biomedical Informatics. 37 (2004) 411–422. https://doi.org/10.1016/j.jbi.2004.08.005.
[19] D. Campos, S. Matos, J. Luis, Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools, 2012. https://doi.org/10.5772/51066.
[20] W. Yoon, C.H. So, J. Lee, J. Kang, CollaboNet: Collaboration of deep neural networks for biomedical named entity recognition, BMC Bioinformatics. 20 (2019) 55–65. https://doi.org/10.1186/s12859-019-2813-6.
[21] M. Habibi, L. Weber, M. Neves, D.L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics. 33 (2017) i37–i48. https://doi.org/10.1093/bioinformatics/btx228.
[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on

Neural Information Processing Systems, 2013: pp. 3111–3119.
[23] Y. Zhang, Z. Liu, W. Zhou, Biomedical Named Entity Recognition based on Deep Neutral Network, Chinese Journal of Electronics. 29 (2020) 455–462. https://doi.org/10.1049/cje.2020.03.001.
[24] H. Yu, Z. Wei, L. Sun, Z. Zhang, Biomedical named entity recognition based on multistage three-way decisions, Communications in Computer and Information Science. 663 (2016) 513–524. https://doi.org/10.1007/978-981-10-3005-5_42.
[25] Y. Tsuruoka, Y. Tateishi, J.D. Kim, T. Ohta, J. McNaught, S. Ananiadou, 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). 3746 LNCS (2005) 382–392. https://doi.org/10.1007/11573036_36.
[26] Apache License, Word2Vec, (2013). https://code.google.com/archive/p/word2vec/ (accessed October 20, 2020).
[27] P.~Brown, V.~Della Pietra, P. de Souza, J.~Lai, R.~Mercer, Class-based n-gram models of natural language, Computational Linguistics. 18 (1992) 467–479.
[28] N. Perera, M. Dehmer, F. Emmert-Streib, Named Entity Recognition and Relation Detection for Biomedical Information Extraction, Frontiers in Cell and Developmental Biology. 8 (2020). https://doi.org/10.3389/fcell.2020.00673.
[29] D. Song, L. Li, L. Jin, D. Huang, Biomedical named entity recognition based on recurrent neural networks with different extended methods, International Journal of Data Mining and Bioinformatics. 16 (2016) 17–31. https://doi.org/10.1504/IJDMB.2016.079799.
[30] B. Tang, H. Cao, X. Wang, Q. Chen, H. Xu, Evaluating word representation features in biomedical named entity recognition tasks, BioMed Research International. 2014 (2014). https://doi.org/10.1155/2014/240403.
[31] Thorsten Joachims, SVMhmm, (2008). http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html (accessed August 16, 2020).
[32] J.D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA corpus - A semantically annotated corpus for bio-textmining, Bioinformatics. 19 (2003) 180–183. https://doi.org/10.1093/bioinformatics/btg1023.
[33] W. Lu, D. Roth, Joint mention extraction and classification with mention hypergraphs, in: Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015: pp. 857–867. https://doi.org/10.18653/v1/d15-1102.