BIOMEDICAL NAMED ENTITY RECOGNITION WITH BILSTM-EDA: A DEEP LEARNING APRROACH

ShanthaKumari R[1], Roopa Devi E M[2], Vinothkumar S[3], Asifaa Sulthana N[4], Fahima Begum B[5] and Kaushik G[6]

[1] Kongu Engineering College, Erode 638060, TamilNadu, India

[2] Kongu Engineering College, Erode 638060, TamilNadu, India

[3] Kongu Engineering College, Erode 638060, TamilNadu, India

[4] Kongu Engineering College, Erode 638060, TamilNadu, India

[5] Kongu Engineering College, Erode 638060, TamilNadu, India

[6] Kongu Engineering College, Erode 638060, TamilNadu, India

Abstract. In the medical field, dealing with a large volume of transcription is a difficult task, and it may take some time to read every line. Named Entity Recognition (NER) technology demonstrates the ability to scan entire documents and identify people, groups, and places. Identification of medical named entities from an unstructured natural language text is one of the most crucial topics in the field of Natural Language Processing (NLP), such as a medicine, disease, or treatment, using the medical NER model. This research presents a named entity recognition for a recurrent neural network structure based on the RNN version known as LSTM (Bi-LSTM), which makes use of Exploratory Data Analysis (EDA) to better comprehend the dataset. The experiments show the result of harmonic mean of 93.6%F1 score, Precision 95%, Recall 91%, and Accuracy of 92.89%.

Keywords: Named Entity Recognition (NER), Natural Language Processing (NLP), Long Short-Term Memory(LSTM) and Recurrent Neural Network.

1 Introduction

Due to the challenges in extracting meaningful information, an abundance of data that is readily available online is always less dependable. To get important and practical stuff from data, information extraction is required. It is a necessary and important task in Natural Language Processing (NLP) that involves categorizing and identifying specific entities or objects in text [1]. Basically, it is used to get related information from large amounts of data, such as social media posts, scientific paper and news articles [3] .Entity Recognition is used in a diverse of practicalities, such as machine interpretation, question answering, including informational analysis [5].

Since entity recognition has been successfully used in extracting real-world entities, different research approaches holding an extensive variety of topics have been proposed. A text's named entities will be recognized by NER, which will group them into predefined categories such as organizations (ORG), individuals (PER), work of art (WOA) [6]. Example of NER: “J.K. Rowling [PER], the author of the Harry Potter

series [WOA], was born on July 31, 1965 [DATE], in Yate, Gloucestershire, England [LOC]" [6]. The individual characteristics and groupings attributed to the identified entities can shift within different domains due to the wide-ranging investigations conducted by various researchers. Traditional NER approaches rely heavily on manual procedures to generate entity rules and rule-based NLP algorithms [13][14]. It is difficult to search them up manually and come up with them because there are so many textual transcriptions in the medical profession. This has increased the demand for automated information extraction. Text mining can be utilized to completely automate a time-consuming operation [2]. New insights in the medical sciences are predicted, and numerous research initiatives have already been proposed related to this field.

As technology advances, there are more and more opportunities to glean insightful information from vast amounts of unstructured data. Medical researchers publish a great amount of knowledge in their writings. The use of medical textual data is challenging due to the abundance of patient information [8]. This makes it difficult, arduous, and time-consuming for health workers to accomplish activities like information retrieval. In reality, hospitals generate vast amounts of data that can be used to advance understanding. The exponential growth of available data has rendered it nearly unfeasible for medical professionals to independently grasp and derive novel insights. Employing efficient computational techniques to generate knowledge representations can aid in information retrieval within the medical domain. In order to ease workload stress and reinvent workflows, healthcare organizations and practitioners now have access to AI-enabled technologies [7].The proposed model demonstrates named entity recognition for an RNN-based Bidirectional LSTM-based recurrent neural network structure (Bi-LSTM), which use Exploratory Data Analysis (EDA) for better visualization of document length distribution, sentence count distribution, distribution of entity types, get the top 20 most and their counts and heatmap of overlapping entities

2 Literature Review

The majority of contemporary neural BioNER systems use external knowledge for pre- or post-editing rather than including it during training because the model cannot learn it [9]. They provide a unified multi-task Machine Reading Comprehension (MRC) architecture for BioNER to incorporate prior information into the model. At query sequences, they introduce three different types of prior knowledge: Wikipedia, annotated schemes, and entity dictionaries. Then, to jointly train the primary task BioNER and the auxiliary task MRC, their model utilizes a multi-task learning technique.

The goal of [10] is to exclude or encode Personally Identifiable Information (PII) from IME reports written by doctors. Using the shared default parameters for each model, some of the NER toolkits of OpenNLP and spaCy, two free NLP platforms, are compared for their performance at recognizing five types of PII across trials of randomly selected IME reports.

Emphasize the significance of understanding microbial communities through interactions among microorganisms, particularly bacteria's role in human diseases. It highlights the importance of mining and organizing small-scale data from medical literature on bacterial interactions to support microbiome research. The paper [15]

introduces a language model-based method for bacterial Named Entity Recognition (NER), achieving an impressive F1 score of 96.14%, surpassing previous results in bacteria NER. This approach advances the field of microbiome research by automating the identification of bacterial entities in text, facilitating a deeper understanding of microbial interactions.
The goal of [14] suggests a novel hybrid-based method for identifying named entities in medical literature texts. To annotate the entities in the medical documents, a new dictionary for symptoms, dosage forms, and routes of administration has been created. The blank Spacy machine learning model trains on the annotated items. Comparing the trained model to the current model reveals a respectable level of accuracy.

T-RoBERTa-BiLSTM-CRF, a transfer learning-based electronic medical record entity recognition model, to address the issue of data scarcity[11]. The model aggregates the traits of medical data from various sources and uses a small amount of electronic medical record data as target data for additional training. Some of the issue of use of deep neural networks and pre-trained language models in the field of biomedical NER models is constrained by the absence of annotated electronic medical record datasets.

To enhance the performance of NER on scientific papers, a novel NER model that makes use of deep learning techniques is proposed. It uses domain-specific embedding and Bi-LSTM [12]. The model greatly outperforms earlier approaches used on the same data-set, receiving a 98% F1-score on a curated data-set of Covid-related scientific publications published in different Web of Science and PubMed indexed journals, and the results demonstrate the effectiveness of the methodology in accurately identifying and categorizing named entities (disease and drug) in scientific literature, paving the path for upcoming advancements in biomedical text mining.

Most of these related works concentrate on medical NER models in various ways, such as pre- and post-sentence mapping, comparison of predefined models like spacy opensource NLP with various datasets, and building models with T-RoBERTa-BiLSTM-CRF. However, these works have encountered some limitations, such as a shortage of medical terms etc. To tackle these natural language processing tasks, the proposed model employs a deep learning approach with LSTM layers to capture sequential dependencies with annotated text file by analysing the texts with IOB tagging and map the corresponding entities.

3 Proposed Methodology

In recent Bio-NER systems, incorporating external knowledge for editing is common due to the model's limited learning capacity. A poly-tasking architecture has been proposed, integrating prior knowledge effectively into Bio_NER, facilitating better performance. Another objective is to handle Personally Identifiable Information (PII) in IME reports authored by doctors, assessing NER toolkits for this purpose. For improved NER performance in scientific papers, an innovative model using deep learning methods and specialized embeddings, along with Bi-directional Long Short-Term Memory (LSTM) Networks, has been introduced. Addressing data scarcity challenges in the biomedical domain, the T-RoBERTa-BiLSTM-CRF model, built using transfer learning, focuses on entity recognition in electronic medical records.

However, challenges persist, such as the scarcity of medical terms in existing work. The proposed model adopts a deep learning approach with LSTM layers, utilizing IOB tagging and entity mapping to address natural language processing tasks effectively.

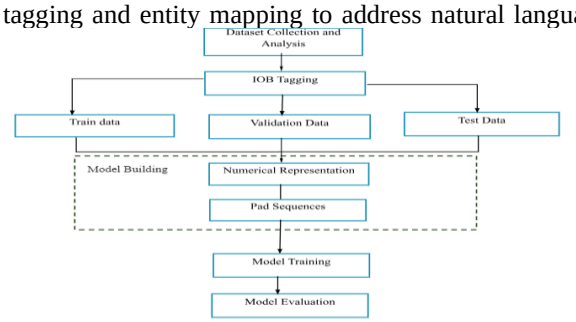

Figure 1: Proposed Methodology

3.1 Dataset Description

"MACCROBAT2020" dataset can be downloaded as a ZIP-compressed file for free from the figshare.com. 200 source documents which contain transcriptions in the form of plain text and 200 annotation documents are both included in this Zip file. Documents are titled using PubMed document IDs. The content was modified to solely include the specifics of a clinical case report and is taken from full-text papers on PubMed Central. All the annotated files are created manually. For example, "15939911.txt" contains material from the document (Figure 2) and the annotated data looks as below "15939911.ann" file (Table 1).

"CASE: A 28-year-old previously healthy man presented with a 6-week history of palpitation. The symptoms occurred during rest, 2–3 times per week, lasted up to 30 minutes at a time and were associated with dyspnoea."

Figure 2 : Picture of Text File

Table 1 : Structure of Annotated text file

Entity	Start index	End Index
Age	0.76	0.77
History	0.84	0.82
sex	0.83	0.82
….	….	….

3.2 Dataset Analysis

Exploratory data analysis (EDA) examines data sets, summarizes their key features, and frequently makes use of data visualization techniques. Its main goal is to encourage data analysis before making any assumptions which helps assist in finding better understanding data patterns, glaring errors, discovering intriguing relationships

between the variables, and spotting outliers or unusual occurrences. The proposal of the dataset with EDA is tabulated in Table 2.

Table 2: EDA Analysis of Dataset

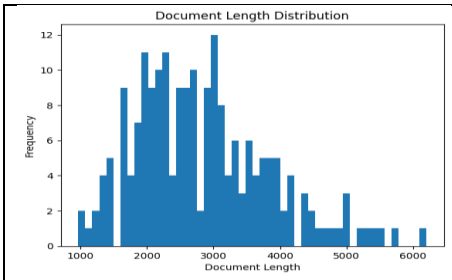

Figure 3: Document Length Distribution

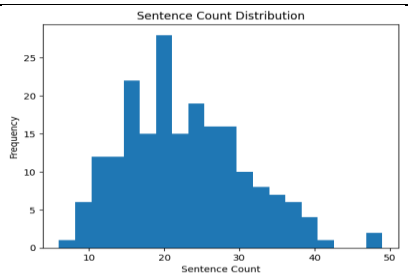

Figure 4: Sentence Count Distribution

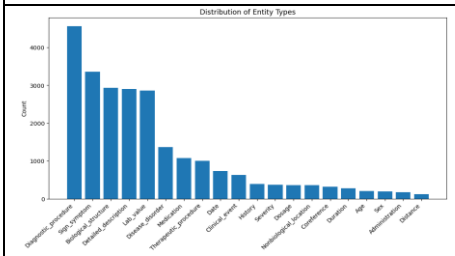

Figure 5: Distribution of Entity Types

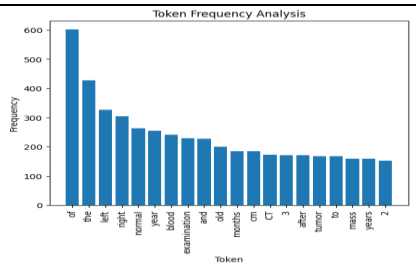

Figure 6: Token Frequency Analysis

4 IOB Tagging

This process illustrates step-by-step how each word in a sentence can be marked with labels that indicate whether they are a part of a named entity or not with three tags namely B, I, O in which B-Tag indicates beginning of an entity, I-Tag indicated inside entity, and O-Tag indicates stop words in which those words are no need for further processing to build the model.

4.1 Assembling json file

To provide IOB tagging, initially load the annotated file(.ann) and split the annotations corresponding to the text and separate the entities, store it in a separate list. Later the list is passed on-to predefined set function to identify the unique entities available throughout all the transcriptions. Then create a list of dictionaries, where each dictionary represents a piece of text data along with its annotations. Afterwards, create a JSON object containing this list and save it to a separate file.

4.2 Create bio_file

The JSON data is loaded and prepared for conversion into BIO format for Named Entity Recognition. The annotated JSON data is processed, and tokens are cleaned of trailing punctuations. Sentences are split, and tags are assigned to each token based on whether

it's the beginning (`B-tag`) of a new entity, part (`I-tag`) of an existing entity, or a common stop word (`O-tag`). The resulting tagged sequences and entity labels are stored in an output directory. To convert .ann files to BIO format, the text and annotations are extracted, and tokens are tagged accordingly, resulting in 200 (.bio) files for Named Entity Recognition which is figured in Figure 7.

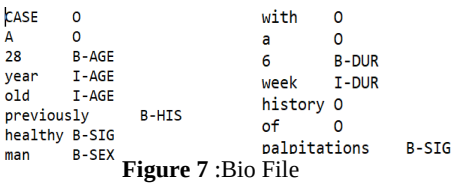

Figure 7 :Bio File

4.3 Process the Bio-files

The .bio files which was created in previous step is taken to find a stop word is being wrongly tagged ('B' tag), later the text files are processed finding non-alphanumeric characters, extra whitespaces, converting to lowercase, and lemmatizing it using spaCy to check if the resulting lemma is a stop word being removed else returns the lemma or an empty string accordingly. The mentioned procedure is processed for every .bio file available.

5 Data Splitting Strategy

The BIO formatted dataset consists of a total of 4341 sentences. These sentences are split in the ratio of 70:10:20 (train: validation: test).
The number of sentences of train split is set to $\mathbf{3038} \longrightarrow \boldsymbol{int}(\mathbf{4341} * \mathbf{0.7})$.
The number of sentences of validation split is set to $\mathbf{434} \longrightarrow . int(4341 * 0.1)$
The number of sentences of test split is set to $\mathbf{868} \longrightarrow \boldsymbol{int}(\mathbf{4341} * \mathbf{0.2})$.
The labels for the sentences are also split into three sets: train_labels, valid_labels, and test_labels. This is done in the same way as for the sentences themselves.

6 Numerical Representation

Two dictionaries, label_to_index and index_to_label, are generated for efficient label mapping in numerical form. These dictionaries enable the conversion of labels to numerical indices and vice versa. Labels are assigned unique indices in a sorted manner, starting from 1, with an added '<PAD>' token for empty positions in sequences. These dictionaries facilitate streamlined sequence classification tasks.

6.1 Allocate Vocabulary

Now convert text files into numerical format for model training. This is done by tokenizer object which creates a vocabulary from training data and transforms sentences into sequences of numerical indices. To ensure uniform sequence lengths, padding or truncation is applied typically to a length of 100 tokens. This ensures consistent input for effective model training and evaluation. For each sentence default

value 101 added initially and 102 added at the end of each sentence to identify the start of the sentence.

For Example: A Case Study about year old ……
[[101 231 232 431 518 621 716 102] ……

6.2 Label Encoding

After the above process do ensures uniformity across all datasets, including valid_labels and test_labels. Finally, labels(entities) are one-hot encoded for effective label prediction. The processed data, including vocabulary and labels, along with label indexing dictionaries, are saved in a compressed .npz file.

For Example: Entity: [[Age, Loc, previously healthy, ...…]]
One-hot Encoding: [[1,0, 0, ...] [0,1,0, ...] [0,0, 1...] .....]

7 Model Building

i. **Data Loading and Preparation:** Load pre-processed data for training, testing, and validation sets, including text, labels, and entity labels. Reuse the tokenizer to tokenize new text efficiently.

ii. **Model Design:** The Figure 8 denotes the structural representation of LSTM model which is composed of layers like Embedding layer, Dense layer, LSTM layer. These layers are fully connected and applies activation function thus forms a complete model and learns each word of a sentence.

iii. **Embedding Layers:** Responsible for converting integer-encoded words into fixed-size dense vectors using word embedding. Parameters include INPUT_DIM (unique categories count), EMBEDDING_DIM (embedding vector dimension), and INPUT_LENGTH (input sequence length).

iv. **Dense Layers:** Essential for feature extraction, dimensionality reduction, and non-linear transformations. Configured with specific units and activation functions. The last dense layer uses SoftMax activation for text classification.

v. **Bi-LSTM Layers:** Incorporate two LSTM layers in the sequential model, allowing processing of sequences in both forward and backward directions.

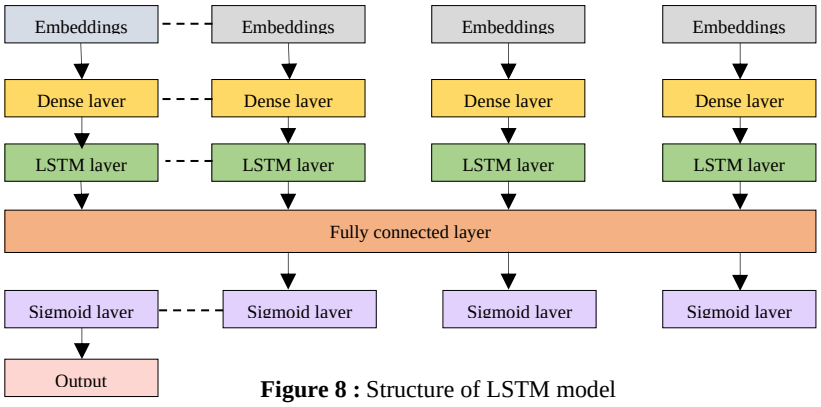

Figure 8 : Structure of LSTM model

These layers are stacked sequentially in the Keras Sequential model, with each layer serving a specific purpose in the neural network architecture for text classification.

8 Model Fitting

The Bi-LSTM model, consisting of 5 layers, including an Embedding layer, is trained on the provided data for a set of 100 epochs. During training, the model fine-tunes its internal parameters using backpropagation and gradient descent to minimize the categorical cross-entropy loss. After each epoch, the model's performance on the validation dataset is assessed, allowing us to monitor generalization and detect overfitting.

9 Proposal of The Output

Once the model is trained, highlighting the important entities is an essential task. For this spacy model is used, in which spacy model provides the ability to integrate custom models and pipelines, allowing for additional functionalities like text classification or sequence labelling. With the help of spacy's displacy modules, assign each entity with certain desired colour. When the inferred text is given, based on the entity and the entity colour those text will be coloured or highlighted which could be a necessary one and or a medical related term which is figured in Figure ***9***.

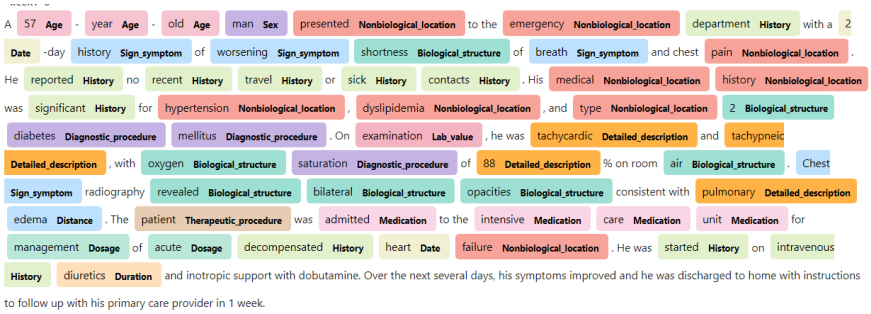

Figure 9 : Output of the model

10 Model Evaluation

The effectiveness of a model can be assessed in a number of ways. Measures taken into account when estimating the recommendation prediction include accuracy, precision, recall, and F-score. The model demonstrates high precision, with scores of 0.98 on the training set and 0.95 on both the validation and test sets, indicating its ability to make accurate positive predictions Table 4.Additionally, the model exhibits robust overall performance, as evidenced by its high F1 scores, which measure the balance between precision and recall. The F1 score ranges from 0.93 to 0.97 across the three datasets.

Although the model's recall is slightly lower than precision, it remains high, with values ranging from 0.91 to 0.96. The loss is relatively low, with a value of 0.11 on the training set and 0.48 on both the validation and test sets. Finally, the model

maintains strong accuracy on all three datasets, ranging from 92.89% to 96.94%. Overall, the model demonstrates reliable performance, suggesting its effectiveness in making accurate predictions.

Table 3 represents the performance metrics of various text analysis methodologies, including openNLP, a BERT model, BERT-CRF, Bilstm CRF, and this proposed model. Overall, "The proposed model" demonstrates the highest precision (0.98), indicating its ability to make accurate positive predictions, while maintaining a relatively high F1 score of 0.97, which balances precision and recall.

Table 3 : Comparison of Different Models

Methodologies	**Precision**	**F1 score**	**Recall**	**Accuracy in %**
openNLP	0.95	0.95	0.95	76.99
BERT model	0.96	0.96	0.95	82.41
BERT-CRF	0.96	0.96	0.96	82.02
Bilstm CRF	0.96	0.96	0.96	94.51
The proposed model	0.98	0.97	0.96	96.94

Table 4. Comparison of testing scores on split dataset

	Precision	**F1 score**	**Recall**	**Loss**	**Accuracy (%)**
Training set	0.98	0.97	0.95	0.11	96.94
Validation set	0.95	0.93	0.92	0.48	92.89
Test Set	0.95	0.93	0.91	0.48	92.89

11 Conclusion

In conclusion, in contrast to conventional methods, our methodology prioritizes complex understanding of details as well as model performance. The outcomes highlight the model's outstanding performance. This demonstrates the superiority of a Bi-LSTM-based RNN over existing techniques in the area of biomedical named entity recognition and demonstrates its potential to make a significant contribution to the field. Furthermore, investigating methods to handle noisy and incomplete data, a common challenge in real-world scenarios, would be crucial for practical applicability. Lastly, focusing on model interpretability and visualization techniques can enhance our understanding of the model's decision-making process, fostering trust and comprehension in critical biomedical applications. In future, various other new models could be built and compared to with other models hence forming a hybrid model.

12 Reference

[1] Grishman, R.; and Sundheim, B. (1996). Message understanding conference6: A brief history. Proceedings of the 16th Conference on Computational Linguistics, Volume 1. Pennsylvania, USA, 466-471.

[2] Yoon, W., So, C., Lee, J. *et al.* CollaboNet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics* **20** (Suppl 10), 249 (2019).

[3] Chiu, J.P.C.; and Nichols, E. (2016). Named entity recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4: 357-370.
[4] Ramachandran, R., Arutchelvan, K. Named entity recognition on bio-medical literature documents using hybrid-based approach. J Ambient Intell Human Comput (2021).
[5] Huang, H.; Wang, H.; and Jin, D. (2018). A low-cost named entity recognition research based on active learning. Scientific Programming, Special Issue, Volume 2018, Article ID 1890683.
[6] Sang, E.F.T.K.; and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-Independent named entity recognition. Proceedings of the Seven
[7] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, Future Healthcare J. 6 (2) (2019) 94.
[8] E. Rencis, Natural language-based knowledge extraction in healthcare domain.Proceedings of the 2019 3rd International Conference on Information System and Data Mining, 2019, pp. 138–142.
[9] Y. Tong, F. Zhuang, D. Wang, H. Ying and B. Wang, "Improving Biomedical Named Entity Recognition with a Unified Multi-Task MRC Framework," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8332-8336, doi: 10.1109/ICASSP43922.2022.9746482.
[10] C. Pearson, N. Seliya and R. Dave, "Named Entity Recognition in Unstructured Medical Text Documents," 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 2021, pp. 1-6, doi: 10.1109/ICECET52533.2021.9698694.
[11] Kunli Zhang, Chenghao Zhang, Yajuan Ye, Hongying Zan, and Xiaomei Liu. 2022. Named Entity Recognition in Electronic Medical Records Based on Transfer Learning. In Proceedings of the 2022 International Conference on Intelligent Medicine and Health (ICIMH '22). Association for Computing Machinery, New York, NY, USA, 91–98.
[12] M. S. Ullah Miah, J. Sulaiman, T. B. Sarwar, S. S. Islam, M. Rahman and M. S. Haque, "Medical Named Entity Recognition (MedNER): A Deep Learning Model for Recognizing Medical Entities (Drug, Disease) from Scientific Texts," IEEE EUROCON 2023 - 20th International Conference on Smart Technologies, Torino, Italy, 2023, pp. 158-162, oi:10.1109/EUROCON56442.2023.10199075.
[13] Eric Brill. 1992. A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, USA, 152–155.
[14] D. Frye, P.D. Zelazo and T. Palfai, "Theory of mind and rule-based reasoning", *Cogn. Dev.*, vol. 10, no. 4, pp. 483-527, 1995.
[15] X. Li, C. Fu, R. Zhong, D. Zhong, T. He and X. Jiang, "Bacterial Named Entity Recognition Based on Language Model," 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 2019, pp. 2715-2721, doi: 10.1109/BIBM47256.2019.8983133.