E-ISSN 2581-7957
P-ISSN 2277-3916

CVR Journal of Science and Technology, Volume 21, December 2021
DOB: 10.32377/cvrjst2118

Attention-Based Multitask Model for Name Entity Recognition and Intent Analysis of Online Medical Questions

N. Yogitha[1] and R. Seetharamaiah[2]
[1]PG Scholar, CVR College of Engineering/ IT Department (Data Sciences), Hyderabad, India
Email: yogithanirmalkar123@gmail.com
[2]Professor, CVR College of Engineering/ IT Department, Hyderabad, India
Email: seethasyam03@gmail.com

Abstract: **In recent years, various methods of deep learning have been explored and are applied for solving Named Entity Recognition task which are considered as one of the prior task of Natural Language Processing.**

Named Entity Recognition has been recognized as a sub task of Information Extraction, where it recognizes and classifies proper noun into their given pre-defined categories like persons, location, organization, date and time etc. The survey mainly makes a focus on different methods of deep learning, which have been approached for NER task to work well along with relative multitask learning techniques to opt a novel model based on Neural Network architecture by performing sequence tagging and text classification for witnessing Named Entity Recognition task and Intent Analysis task for online medical questions.

Both the attention mechanism and multitask learning have been improved the performance of their respective tasks. This method has achieved superior performance in both Name Entity Recognition (NER) and Intent Analysis when compared with other methods. The present method is considered as light-weighted solution that can be suitable on every small server for its deployment. By making use of both the tasks a simple question-answering system has been developed.

Index Terms: **Name Entity Recognition, Intent Analysis, Attention mechanism, Multitask neural network, Sequence tagging, Text Classification.**

I. Introduction

Internet Technology has been developed rapidly with many medical related health-services from many Smartphone applications and online websites. This health related online website provides services to the people by sending a relative answer to their respective questions. There are several medical services which are helping people through online in obtaining some knowledge regarding the health issues and other relative aspects. These online medical service systems make use of medical expertise to build proper medical knowledge bases, thereafter, applying natural language understanding technologies to generate related answers to questions thereby obtaining the better quality of generated answers for the user's queries.

The two main tasks performed in the method are Name Entity Recognition (NER) [1] and text classification task conducted on Intent Analysis separately.

A. Name Entity Recognition (NER):

The Name Entity Recognition is considered as a form of Natural Language Processing where its main task is identifying and classifying entities. The NER is also referred as an entity extraction where it seeks to extract the named entities from a given unstructured text into some pre-defined categories like person's name, medical codes, locations, percentages, organizations, time expressions and many more. With the help of NER it is feasible to extract key information, to understand given text and further use it by collecting the important information thereafter storing it in a database. In the online medical question-answering services it helps the system to rightly provide the services to its users. For example, consider that a person asks a question like: "Can arrhythmia causes severe headache?". With the help of NER we can understand that the given question contains "arrhythmia" and "headache" as a disease name (or) disease word.

The working principle of NER model Fig. 1, is mainly carried out in two phases i.e. 1) Detecting the named entity from given text. 2) Categorizing the entities.

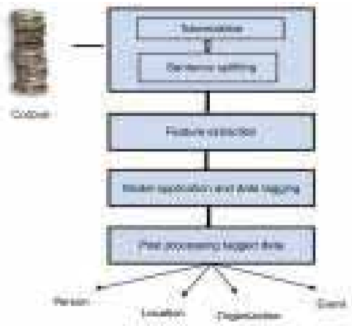

Figure 1. NER Pipeline

Here the first phase involves in detecting the string of words from the given corpus that form entities. During the first phase each word in the corpus is represented as token where the splitting of a sentence, phrase or paragraph has been processed into smaller units which helps in understanding the context and also helps to elucidate the meaning of the text in the corpus by analyzing the words given as sequence. This method is called Tokenization.

The second phase is Categorization phase where creation of entity classifying takes place.

However the problem with the working process of Natural Language Understanding is that, the algorithms of Machine Learning cannot work properly with any raw text. So to make this language processing work properly there is a need of some feature extraction methods that can convert the

Received on 24-07-2021, Revised on 21-08-2021, Accepted on 26-09-2021.

text into vectors. The resultant outcomes of these feature extractions are processed for data tagging, where the conversion of sentences to forms takes place where the forms consist of words and tuples. The list of elements for each tuple is in the form of (word, tags) where the tag is considered as POS tags which resolves ambiguities and classifies the given words based on their Parts-of-Speech types. Thereafter the named entities are created based on the text data, which classifies and recognize the words individually to perform good results on trained data for a given task of entity recognition.

B. Intent Analysis:

Intent Analysis [2] is also called as Intent Recognition, is a text classification technique which is a form of NLP. The task of intent classification is to understand the user's correct goal for which the system should leverage its intent detector to classify the user's statement into any one of the predefined classes which is known as intent.

This intent classification identifies the intentions behind any given statement, and displays it according to its predefined labels by classifying them as the intent of the given query for its easy understanding. For online Question&Answer medical services the intent classification helps the user's to correctly identify the intentions behind the medical query that has been asked. For example, if any query is given as "Can arrhythmia cause headache?" the intent classification for the asked array is given as "Syndrome". This will be helpful for the user to have an easy classification for a large form of text.

The working principal of Intent classification or Intent analysis Fig. 2 is carried out by providing the training data, where the training data is considered as a representative sample of raw text data which is labeled manually as user wants to design their respective model, to work automatically.

Once the labeling of data is done, the user can feed the data into the suitable Machine Learning (ML) model. Once the user is done with training the model the next process is Validation Process where another sample of raw data is used to validate the required model. This Validation Process is used to check whether the model performs well on the given type of data during its process of production. If the raw unlabeled data that is given to model perform well and obtain an accurate output, then it considers that the model is ready for production with good accuracy.

The method introduces a multi-task model [3] for sequence tagging and text classification where both intent analysis task and NER task can be trained. The model uses text data directly from online user questions that are considered as input and generate good results compared to other baseline methods.

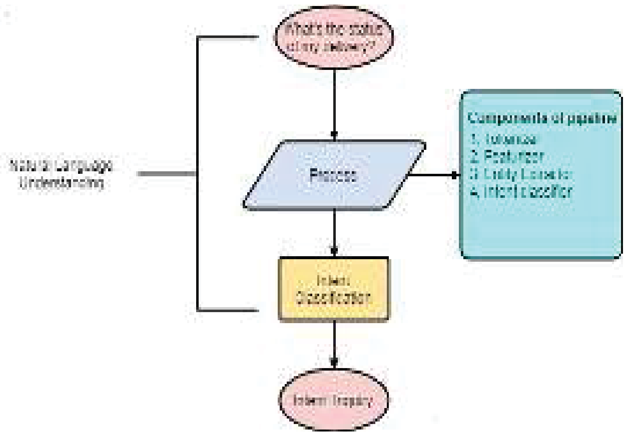

Figure 2. Intent Analysis

II. Literature Survey

I . Attention based multi task model:

The attention based multi task model shares a word embedding layer and the Bi-LSTM layer to generate the NER tags and intent labels for given text. Word embedding [4] represents the words in an array in the form of continuous vectors. So, in the word embedding each token is transformed into a vector where we obtain some sequence of vectors that are considered as the inputs to the Bi-LSTM model. The type of word embeddings used here is word2vector representation to generate pre-trained word vectors. Here the corpus for word2vector pre-training model is collected from online health communities. Later the sequences of word vectors are sent as inputs to the Bi-LSTM layer.

The Bi-LSTM RNN architecture is a Deep Neural Network model Fig. 3, where it tends to learn bidirectional long term dependencies between the sequential data. It is considered as sequence parsing model where it consists of two LSTM's where one LSTM is input that is formed in forward direction and other in backward direction making use of Bi-LSTM models may effectively increase the amount of information available for a network. The usage of Bi-LSTM can be carried out in 2 ways for running the given input i.e. one from past to future and other from future to past. It stores the information for the future use and tries to predict the next words in the given sequential data.

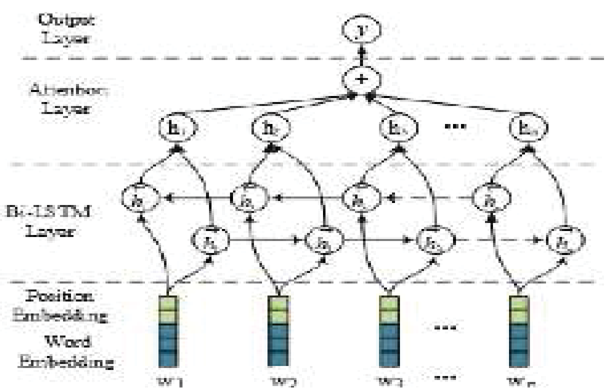

Figure 3. Bi-LSTM model with attention mechanism

E-ISSN 2581-7957 CVR Journal of Science and Technology, Volume 21, December 2021
P-ISSN 2277-3916 DOB: 10.32377/cvrjst2118

Here when the word2vector output is given as input to the Bi-LSTM layer, its function is carried out in both forward and backward direction i.e. from past to future and future to past. The output of Bi-LSTM layer has some weights given using an attention mechanism, where the main function of attention mechanism [5] is to implement the action of selectively concentrating on most relevant information from the available corpus while ignoring the other.

The Bi-LSTM is used to generate a sequence of annotations i.e. h1, h2, h3 ...hnx for each given input sentence. The h1, h2 ...hnx are considered as vectors where it undergoes with some weights for given set of input annotations. Later on, the attention weights are being obtained to calculate with the Bi-LSTM output.

For NER model it contains attention layer for extracting several components of relevant information from a text sentence. Here the sequence of Bi-LSTM output is concatenated with attention layer and further it is given as input to the CRF layer for generating entity labels. For classification task also we make use of same attention layer as of for NER with some different annotations and annotation matrix is used for classifying different labels from the given corpus Fig. 4.

The model applied here is binary entropy as loss function and Adam optimizer. Both the tasks have their respective learning rates with better accuracies, differ from other baseline models.

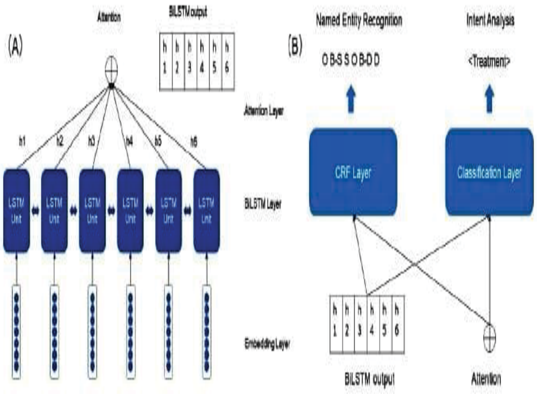

Figure 4. Bi-LSTM model with attention mechanism

The data that is made to use in the model is collected from health community website with around 60,000 user questions with their suitable answers and are stored in the MySQL database. For Intent Analysis the dataset contains 12,525 user questions with 7 types of intent labels for those questions. The intent labels are given as Diagnosis, Syndrome, Treatment, Reason, Concept, Complication and Prognosis. Thus, one user question may contain 0-7 different types of intent labels. For NER around 6,199 user questions were used with labels of 6 types of entities like Disease name, treatment, method, checkup body etc.

After training the datasets on both intent analysis and NER individually, the performance of both trained NER task and Intent Analysis task are calculated separately, and further compared with other models.

For NER task on test dataset the model used F1 score, precision and recall for evaluating the model's result and compared with SSVM, CRF, LSTM-CRF, LSTM-CRF-attention [6] BERT [7] and Multi-task attention models where the BERT model has higher F1 score of 0.8142 and recall of 0.8364 than others.

The second highest result was on model Multi-task attention with F1 score of 0.8135 and recall of 0.8168. The use of attention mechanism helps the model to focus on certain components of the sentences that components can be one or group of characters that are related to word representation, whereas for Intent analysis the model used accuracy, micro averaged F1 score and macro averaged F1 score for evaluation and is compared with CNN model [8], SVM model, Random Forest, BERT, Multi-task attention and other baseline models.

The results are given as; the present applied models obtained better score than CNN and SVM models with high performance of accuracy, Micro F1 score and Macro F1 score with 0.7588, 0.8652, and 0.6888 respectively.

The advantage of using attention mechanism with Bi-LSTM layer during training gave additional improvement for the model, to obtain good performance than other baseline methods and the overfitting of the model can be prevented by using multi-task learning model. The attention based multitask learning models have been generated by giving promising results for both the NER and Intent analysis tasks. It also states that the result suggests that both Attention mechanism and Multi-task learning can improve the Natural Language Understanding for given medical text by working together.

II . Multitask Bi-directional RNN model:

The authors Shanta Chowdhury, Xiangfang Li, Yi Guan and Xishuang Dong proposed a multitask bi-directional RNN model for extracting entity terms from Electronic Medical Records [9] on Chinese text. The model has been divided into two parts.

The first one is Shared layer and the other is Task specific layer. The shared layer is defined as a process of allowing a layer to link with other layers, so that the changes made in the original layer are produced on the shared layer. This sharing layer makes changes on multiple copies by applying changes on any of single linked layer. The task specific layer is defined as a process of performing multiple tasks that are given in a specific way by generalizing the domain specified information.

Many applications which use this task specific layer are given as Machine Learning, Natural Language Processing, Speech recognition, Computer vision.

Here for the proposed model the Fig. 5, gives the vector representation for each given word; the word is obtained as the concatenation of word embedding and character embedding. Later in Fig. 6, the bi-directional RNN [10] model is used in order to extract context information for any given sentence.

Later on, the given layers shared with two differently specified task layers i.e. NER task layer and other is POS tagging layer. Both the task layers are trained in an alternative manner where the knowledge learned from one task layer namely NER task has been enhanced by the knowledge gained from POS tagging task layer.

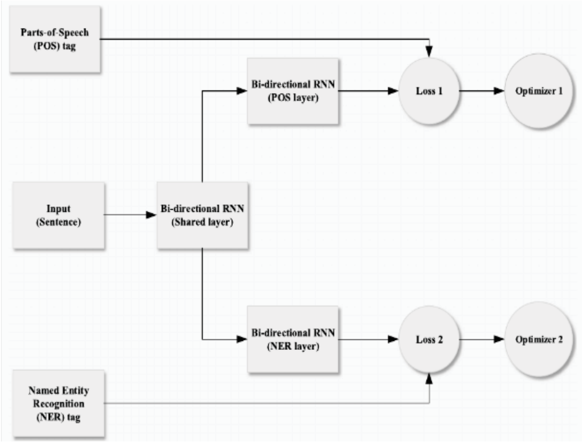

Figure 5. Framework of multitask bi-directional RNN model for NER

The framework proposed here is bi-directional RNN model that is used to exploit the past and the future context, since the RNN model is used to capture previous words and character information in a sequential manner and store in its memory for its further work as it has input layers, output layers, and hidden layers. So here in bi-directional RNN the forward hidden states compute with forward hidden sequence and the backward hidden states computes with backward hidden sequence.

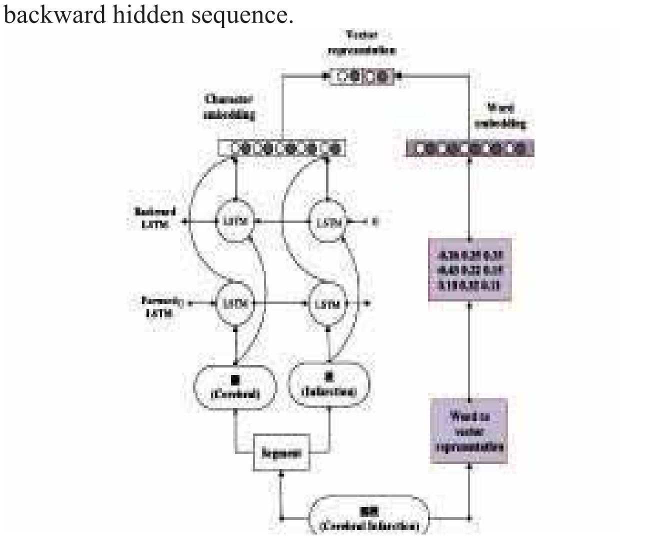

Figure 6. Vector representation as concatenation of word and character embedding.

Here the two bi-directional RNN models are used where one is used to extract features from word embedding by achieving word2vector representation and the other is character level feature extraction and represents it in character embedding. Later both the character embedding, and word embedding are concatenated to represent the words individually in a vector representation. Another bi-directional RNN is used for extracting context information from given text sentence.

Thereafter the outputs of words representation are shared by two different bi-directional RNN for both POS [11] and NER tasks separately. These two layers are trained alternatively, so that the performance of both tasks.

The model uses the EMR dataset which is collected from the department of second affiliated hospital available at Harbin Medical University by discarding the personal information of patients. The labeled corpus for the experiment consists of 500 discharge summaries and around 492 progress notes that were manually created. The data consists of 55,485 sentences that were written in Chinese language. The 5 types of entity labels are begin categorized as disease, treatment, test, symptom, disease group and for Part-of-Speech tagging the categorized entity types are labeled in BIO format where B stands for Beginning of the medical entity type, I stand for inside of the medical entity type and O is given as outstand of the entity type.

Thereafter the comparison results of accuracy on progress notes and discharge summaries are also improved from 5.66% to 9.41% points compared to CRF model and other baseline models for extracting features on NER task by introducing the prior-knowledge. Finally, it is observed that the best accuracy is given as 89.20% points in the test term and lowest performance is given as 36.00% points in recognizing disease terms on discharge summaries and similar results on progress notes.

The future enhancement for the given model can be designed by planning a joint loss function and joint optimizer that can reduce the training time and can later results in improving the performance accuracy on real datasets in better ways for its future research program.

III. Multiclass Classification Model using Deep Learning for NER:

Lijun Qian, Xishuang dong, Yi Guan, Jinfeng yang, Qiubin Yu and Lei Huang from computer science department, Prairie. View A&M university, Houston, USA proposed a model where a multiclass classification method which is based on Convolution Neural Network (CNN) of Deep Learning is used for NER by considering a constructed corpus of annotations which is obtained from Chinese EMR's where the corpus consists of 992 clinical notes in which it contain both Discharge summaries and Progress notes.

Later both the notes were tagged with certain entities and some baseline methods are also used to evaluate its performance. The method consists of two phases. The first phase is given as preprocessing and word embedding phase and the second phase is given as construction of multiclass classification based on CNN [12].

The proposed model constructs a CNN based multiclass classification method for extracting Named entities from the EMR's system Fig. 7. As mentioned earlier about two model's phases, it is given that in the first phase, the EMR's are gone through pre-processing for feature extraction, selection and representation for given words. Here the sentences are being extracted from EMR's with the text extractor tool where the same sentences from the text are been removed. This extractor extracts only the contents which are described by medical Natural Language from both Progress notes and Discharge summaries. These contents record the information which is important for the patients collected from its medical records like Treatment, Symptoms, and Tests etc.

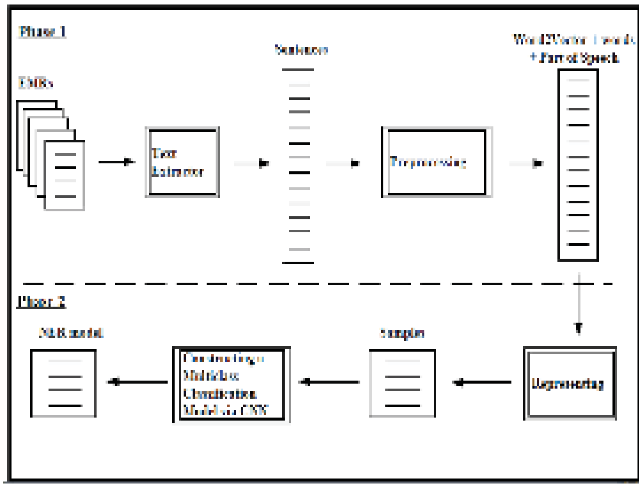

Figure 7. Framework of CNN based Multitask model

The second phase initially uses the word vectors to represent the samples, which are represented as pair of labels and group of word vectors Fig. 8. Then these samples are divided into some subsets with the one-vs-one strategy where the model to train on their subsets and are combined into a pair of labels. Then the CNN model is used to train the binary classifier [13] on each given subset to conduct multiclass classification model and for the prediction process the labels of test samples are generated in terms of voting to predict the results which are conducted by those classifiers.

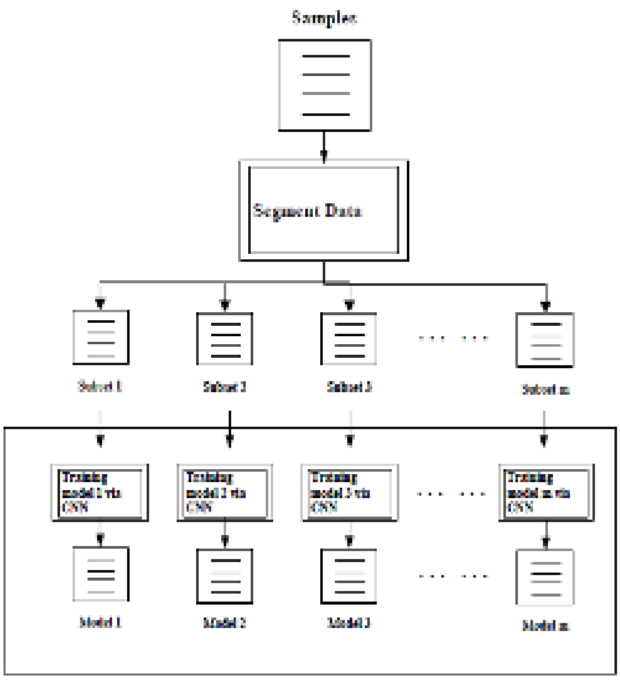

Figure 8. Framework of CNN model

The dataset collection consists of 992 EMR's taken from different departments from the second affiliated hospital, Harbin university of China, where the personal information of patients were removed from the EMR's. For the dataset collection two Chinese doctors were asked to label the records individually and then the results of labeling were merged only if both the doctors have same opinion, then the labels are been fixed otherwise some discussions were conducted till the agreement on the labels are reached by both the doctors.

To obtain more consistency on labeled corpus without any errors they are made to be corrected using some standard guidelines and later on a second round of labeling is been performed by considering two doctors to give a list of relative labels on given EMR dataset. Here if both of them have same opinions then those respective labels are considered as the final list of labels to annotate. There are 5 categories of Named entities that are annotated in the experiment and are given as Symptom, Test, Treatment, Disease name and Disease group.

It is given that the micro average F score of the CRF based method is considered as more than the SVM based method by 2.63% and also 1.87% for NER model based on discharge summaries and progress notes (Table1).

The micro F, micro P and micro R values are also evaluated on the models like Maximum Entropy, SVM and CRF on which the results shown that all the values are same since the testing samples of this model assigned with NER model on both discharge summaries and progress notes on proposed model are less than those of CRF by 1.51% and 3.80% respectively.

TABLE I.
RESULTS ON BASELINE METHODS

	Discharge Summary			**Progress Note**		
Algorithm	**Micro P**	**Micro R**	**Micro F**	**Micro P**	**Micro R**	**Micro F**
NB	78.07	77.91	77.99	79.42	79.37	79.40
ME	88.81	88.81	88.81	91.45	91.45	91.45
SVM	90.52	90.52	90.52	93.07	93.06	93.06
CRF	93.15	93.15	93.15	94.93	94.92	94.93
Our Model	**88.64**	**88.65**	**88.64**	**91.13**	**91.14**	**91.13**

Since the above results are not sufficient and do not show any effectiveness in the method, the accuracy of NER is investigated where the results of overall accuracies of NER using the proposed model shown lower results than CRF based method. It is given that if an entity which consists of one or more than one word, then the entity type for those words cannot simultaneously are made to be tested to their correct labels. The comparison results on recognition of different types of entities shown that the recognition accuracy on test entity type is considered as higher and the disease group entities as least. The performance results on NER model between CRF and proposed method are given as 3.77% on discharge summaries and 3.39% on progress notes (Table2).

TABLE II.
COMPARISON RESULTS WITH SAME FEATURES

	Discharge Summary			**Progress Note**		
Algorithm	**Micro P**	**Micro R**	**Micro F**	**Micro P**	**Micro R**	**Micro F**
CRF	92.41	92.41	92.41	94.52	94.53	94.52
Our Model	88.64	88.65	**88.64**	91.13	91.14	**91.13**

The method constructed on corpus of 992 Chinese EMR's with manually tagged entities of 5 categories. Then the performance evaluation is conducted on two methods for recognizing medical named entities from the corpus.

Later by obtaining the experimental results it is shown that the method achieved micro F score values of 88.64% and 91.13% on discharge summaries and progress notes respectively and for CRF method it is shown as 92.41% and 94.52% respectively on NER model.

These results have shown the effectiveness of proposed method for extracting named entities on Chinese EMR's. As future enhancement the method may build a dependency parser system for extracting the dependency syntactic relationship for improving NER performance on model and by integrating it with POS which can further improve its performance in its best ways.

IV. Named Entity Recognition based on Deep Learning Pretraining:

The Electronic Medical Records record different types of Symptoms, Disease lists and Test results that are taken by patients from their admission process to hospitalization, and based on their medical results proper disease diagnosis, other treatment methods are recommended by the medical professionals. But as this EMR's are not fully structured data. So, in order to convert the unstructured data into the structured form that can be used for Natural Language Processing technology for information extraction and text mining.

The basic task of Information Extraction [14] Text Mining [15] is, to recognizing the types of medical entities which include Disease, Tests, Symptoms and Operations etc based on patient's medical records. Though the method is used for recognizing the named entities on clinical EMR's it undergoes with some difficulties such as: 1. The clinical entities consist of large text data with various types with certain unregistered words such as unregistered drugs lists, disease group lists etc, which make it difficult to understand them and to construct a comprehensive dictionary on clinical text where it obtains different entity dictionaries. 2. These clinical entities are generally divided into complex and simple type of entities with some complex structures, with large no. of variables with lots of nesting, acronyms etc for clinical entities.

By considering the above problems and other deep learning method issues, the model proposed a method for understanding NER on Chinese EMR text based on pre-training, where it has word embedding pre-training model and other is fine-tuning of entity identification model which is pre-trained by having some relevant corpus. The method makes use of Transformers and Bi-LSTM models for feature extraction to recognizing the clinical named entities on Chinese EMR's.

The model makes use of two specific ways for implementing the pre-training mode of deep learning. In the first practice, the input has been initialized by the corpus of EMR's by embedding them and in the second, the entity recognition is pre-trained by the same field of corpus for fine-tuning.

The model performs a fine-tuning mode by pre-training the model and adding the functions of character embedding to Chinese EMR's. As the corpus of Chinese EMR is difficult to annotate based on their existing research method, the model is used to fine-tune the recognition task which is based on the model of clinical entity recognition by training a medical dataset Fig. 9.

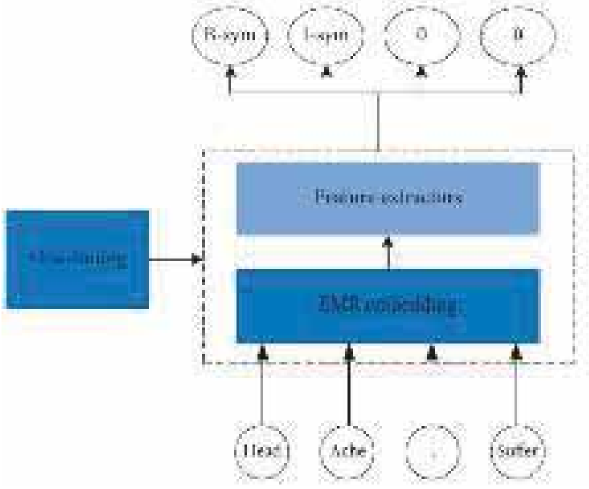

Figure 9. Pipeline model for Fine-tuning

Later the model uses Bi-LSTM model as a feature extractor with CRF for performing sequence annotation. This Bi-LSTM is used to convert the given input sequence through the embedding layer into the inputs formed by vector sequences with two LSTM networks, forming the forward and reverse with two hidden layers and the outputs formed are contacted with softmax layer for classification task. Since the LSTM model learns only the context relation of some features whereas not learning the context relation of tags, which may produce the outputs in wrong sequence tagging form. For this reason, the softmax layer is replaced with CRF layer where it works for sequence annotation and the Bi-LSTM works in producing automatic feature selection Fig. 10, which helps in improving the recognition performance on Chinese EMR's clinical named entity recognition model which is based on deep learning pre-training model.

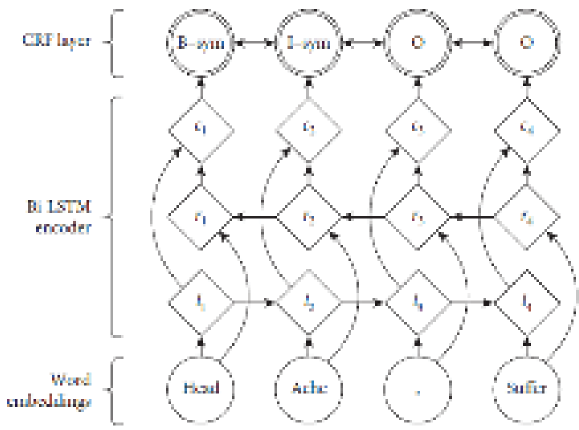

Figure 10. Bi-LSTM CRF model

The data consists of 1064 respiratory records and around 30,262 unrestricted records were collected from a website (https://www.iiyi.com). According to the annotation specification, 200 of 1064 respiratory department of EMR's manually created annotations with semantic types of English text indicating four categories of medical entities such as Disease, Drug, Operation and Symptoms.

The skip-gram model for word2vector used to adopt the word embedding for EMR's from 30,262 sets of EMR's which are considered as unmarked dataset and in addition to the first dataset the model used universal word embedding consisting of 268G of news corpus as second data. Later for the sequential annotation task for entity recognition, it is composed of two parts namely 1. The entity category tag and, 2. Location in the entity tag along with BIO representation which is used to represent the entity categories along with position of entity where BIO stands for B is beginning of entity, I is inside of entity and O is outside (not in) entity and the labeled corpus consists of 9 types of labels and 4 types of entities for its recognition in the given EMR's.

Here a pre-trained method is used on Chinese EMR's Name Entity Recognition model keeping in view of the Electronic Medical Records language features with many unclear entity boundaries, missing entities, annotation corpus difficulty etc. The pre-training method is classified into two steps 1. The first is to adopt the same set of corpus for pre-training word embedding along with Bi-LSTM and Transformer is used to identify Medical entities on Chinese EMR's, which acts as feature extractor and the second is fine-tuning the NER pre-training it with other relevant corpus having annotations for improving the effectiveness of recognition on Chinese medical entities, with results of Macro P having 75.06%, Macro R with 76.40% and Macro F with 75.72%. By considering the above performance results it can be concluded that the Chinese clinical named entity recognition model can improve its recognition performance effectively based on some deep learning pre-training methods.

V. Recognizing Named Entities by combining multi-task Bi-directional LSTM RNN with deep transfer learning:

Research Scholars' Shanta Chowdhury, Lijun Qian YiGuan, X. Dong, J. Yang and Q. Yu., from department of Electrical and Computer Engineering from praineview A&M University, Texas, United States of America, proposed a new model in order to overcome the liabilities i.e. combining the deep transfer Bi-directional RNN model with multitask Bi-directional RNN model [16] on Chinese EMR's in order to extract medical named entity terms, for both multitask deep learning and deep transfer learning which shows the potentials to strengthen Named Entity Recognition performance.

Here to build the proposed model, two steps were carried out i.e. In the first step, the model has obtained the general knowledge for Named Entity Recognition by considering some specific list of domains, by training Bi-directional RNN on Chinese corpus, and the second is to transfer the obtained general knowledge to construct multitask Bi-directional RNN model on Chinese EMR corpus. The proposed method applies a Bi-directional RNN model for extracting medical entity terms from EMR, where it is divided into two phases 1. Extracting domain knowledge phase 2. Multitask learning phase Fig. 11.

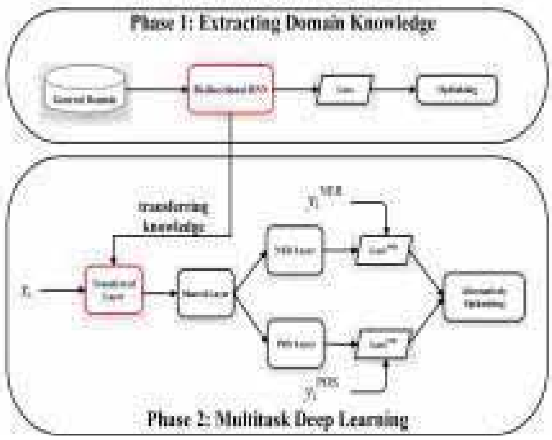

Figure 11. Framework of NER using Bi-directional RNN model

During the first phase, a Bi-directional LSTM RNN model is trained in a general domain, where it selects the optimal hyper-parameters like batch size, learning rate etc, which can obtain highest accuracy rate on extracting named entities from the specific domain. Later on by assuming that the obtained knowledge can uplift the performance of NER in a specific domain and then after transfer the knowledge for performing the complete extraction of NER on Chinese EMR. In the second phase, the knowledge has been transferred to the multitasking deep learning model by initializing the transferred layers as the suitable knowledge that could be activated for improving the accuracies of Named Entity Recognition on Chinese EMR's. In the next step the multitask Bi-directional LSTM RNN model is fine-tuned on transferred layer on the given Chinese corpus of EMR, and the output of this transferred layer is opted as input to the shared layer for extracting more accurate relationship between the words, later on these relations are shared with two different task layers which are given as Named Entity Recognition task layer and other is Parts-of - Speech tagging. Both the task layers are being trained alternatively in such a way that the knowledge which is learned from NER task layer can be upgraded by the knowledge obtained from Parts-of-Speech tagging.

In this model, the vector representations of each given word from both the phrases are concatenated by word embedding and the character embedding. Subsequently the shared layer undergoes two consecutive parts, during the first part, each of the given word is represented into a form of vector. This vector is built as a concatenation of both character embedding and word embedding. Here the word embedding is obtained by word to vector representation and the Bi-directional RNN with LSTM cell is used for extracting the features of characters and represent it as character embedding. Both the word and character are combined together to exhibit the vector representation for each word. This vector representation is then adapted as an input component for both transferred layer and shared layer.

Thereafter the output i.e. contextual word representation from both the layers are shared with two different Bi-

CVR Journal of Science and Technology, Volume 21, December 2021
DOB: 10.32377/cvrjst2118

directional RNN with LSTM cell model for performing two different tasks such as Parts-of-Speech tagging and Named Entity Recognition. Both the tasks are trained alternatively in order to improve the performance of NER task Fig. 12.

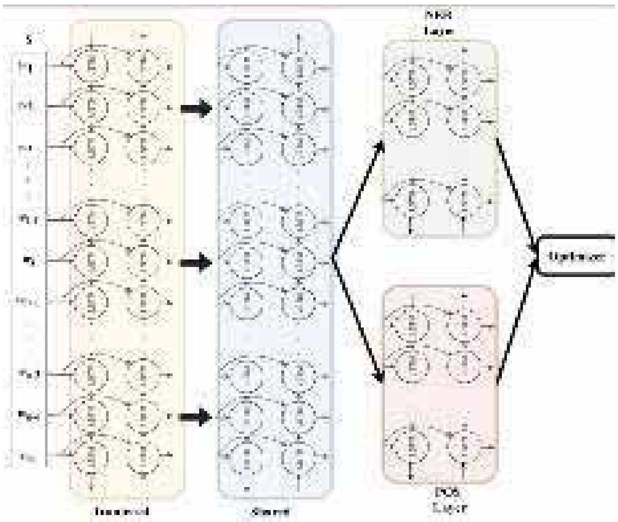

Figure 12. Architecture of transfer and shared layers for NER and POS

The datasets of EMR are collected from the department of Second Affiliated Hospital from the Harbin University, where the patient's personal information has been dispensed. Then the labeled corpus consists of 550 discharge summaries and the progress notes of 492 datasets which are manually created. There was total 55,485 sentences in EMR data written in Chinese language, and the labels were categorized into 5 entity types. The 5 types of entities are disease group, test, treatment, symptoms and disease.

For obtaining the results on the model, different metrics such as macro average F, micro average F score and accuracy have been used. The macro average is used to calculate precision, recall, and F-score on each class independently, whereas the micro average is used for aggregating all the classes for computing the average metrics. These three metrics are compared with classifiers such as Support Vector Machine (SVM), Naive Bayes (NB), Maximum Entropy (ME), Conditional Random Field, Multitask Bi-directional RNN, Transfer Bi-directional RNN [17] etc are used for resolving NER where Bi-directional RNN is considered as a baseline model and MBRNN is considered as state-of-art model where the micro F value of proposed model is improved with 2.55% points, from BRNN and CNN model it has improved its performance value to 4.85% points respectively on discharge summaries and, on progress notes the micro average F score value is been improved on proposed model by 2.23% points and 4.08% points when compare to other baseline models.

Later on the overall accuracy is also checked on both discharge summaries and progress notes where it is observed that the accuracy on discharge summaries is improved by 1.71% points but on progress notes it has decreased by 5.78% points compared to other state-of-art models and the best accuracy model is considered as 90.84% points on test term of named entity and lowest accuracy is recorded on disease group with 60.00% points on proposed model in case of discharge summaries.

III. CONCLUSIONS

By studying all the above models, the classification of Named Entities has been performed effectively on Chinese Electronic Medical Record (EMR's). Each of the models has their own deep learning techniques with some specific tasks which are meant to be carried out in such a way that the Named Entity Recognition tasks are correctly performed on the given words or sentences from its datasets. Many relative annotations/labels were created for NER task where the classification of words on given datasets works based on the listed entity types such as: Disease, Test, Treatment, Medication etc. For analyzing the performance rate and accuracy rate on each model different evaluation metrics such as macro average F score, micro average F score, precision & recall have been evaluated to retrieve the best performance results. Each of the individual models obtained best results based on their individual features on the given tasks. So, by considering all the above models it can be concluded that the Named Entity Recognition task on mentioned deep learning models opt the better results in terms of performance and accuracies on the clinical EMR datasets, but the best performance rate is obtained from TBRNN model for classifying the named entities on clinical corpus more effectively by creating the entity labels & by understanding the datasets which would help in improving its performance and accuracies results, and are used for its future enhancements.

REFERENCES

[1] Aixin Sun, Jianglei Han, Chenliang li, Jing Li, et al., "A survey on deep learning for Named Entity Recognition, IEEE Transaction on Knowledge and Data Engineering, (2020).
[2] B Kasthuriarachchy, M Chetty, D Walls, G Karmakar, "Pre-trained language models with limited data for Intent Classification", IEEE International Joint Conference on Neural Networks (NN), (2020).
[3] Xishuang Dong, Lijun Qian, Jinfeng Yang, Lei Huang, Yi Guan, "A Multiclass Classification Method based on Deep Learning for NER in EMR", IEEE (2016).
[4] Habibi M, Weber L, Neves M, et al., "Deep learning with Word Embeddings to improve Biomedical NER, Bioinformatics, (2017).
[5] Zuko-hregoric A, Bachrach Y, Maksak B, et al., "Neural named entity recognition (NER) using self-attention mechanism", IEEE (2017).
[6] J Xie, C Luo, Y Jin, D Wu, W Guo, R Wang, "LSTM-CRF Neural Network with Gated Self-Attention for Chinese NER", IEEE Access (2019).
[7] Ming-Wei Chang, Kristina Toutanova, Keton Lee, Jacob Devlin, "BERT: Pre-training of deep Bidirectional-Transformers for language understanding", arXiv: 1810.04805v2 [cs.CL], 18 (Oct 2018).
[8] C Mao, L Yao, Y Luo, "Clinical text classification with rule-based features and knowledge-guided convolution neural networks", BMC Med. Information. Decision. Making, (2019).
[9] Licu, Shis, Hao Z, et al., "Improved deep belief network model & its application in Named Entity Recognition (NER) of Chinese electronic medical record (EMR's)", IEEE Big Data Analysis (2018).
[10] Qian Y, Zhao H, Wang P, et al., "A unified tagging solution: Bi-directional LSTM recurrent neural network with word embedding", arXiv: 1511.00215v1 [cs.CL], (Nov 2015).

[11] Xiaoling Cai, Shoubin Dong, Jinlong Hu, "A deep learning model incorporating Parts-of-Speech and self-matching attention for named entity recognition (NER) of Chinese electronic medical record (EMR's)", BMC medical informatics in (2019).
[12] S Reheman, A Farroq, M Awais, S Anwar, "A deep Convolution Neural Network based multi-class classification for Alzheimer's Disease using MRI", IEEE International Conference on Imaging System & Technique,(2017).
[13] S Abdhulla, A Alashoor, "An Artificial Neural Network for the Binary Classification of Network Traffic", International Journal of Advanced Computer Science and Applications, (2020)
[14] Rashmi Agarwal, Simran Kaur, "A Detailed Analysis of core NLP for Information Extraction", International Journal of Machine Learning & Network Collaborative Engineering, (2018).
[15] W Sun, Z Cui, Y Li, F Liu, et al., "Data Processing & Text Mining Technologies on electronic medical record (EMR's): A Review", Journal of Healthcare Engineering, (2018).
[16] S. Chowdhury, X. Dong, L. Qian, X. Li, J. Yang, Q. Yu, Y. Guan, "A multitask Bi-directional RNN model for NER on Chinese EMR's", BMC bioinformatics 19, (2018).
[17] J Yang, X Dong, S Chowdhury, L Qian, Q Yu, Y Guan, "Transfer bi-directional LSTM RNN for named entity recognition on Chinese Electronic Medical Records (EMR's)", IEEE 19th International Conference on e-Health Networking, Applications and Service in (2017).