Biomedical Named Entity Recognition at Scale

Jayakrishnan B[1], Priyanka B. G.[2], Gayathri Gutla [3]

***Abstract*— The General Practioners(GP) and the hospitals make use of the Electronic Health Records (EHR) to keep track of the medical history of the patients. EHRs has all the details of the patient like the reason for administering drugs, previous health history of the patient and the outcome of past drugs administeres. EHRs being the largest source of real time patient data in biomedical research, allows for major scientific findings in highly relevant disorders be it cancer, Alzheimer's disease or any other such illness. All these information are crucial as it has a major role to play in deciding the future course of treatment for the patient. All these data in the EHRs are widely available in natural language format, thus making it largely inaccessible for statistical analysis of such data. Unlocking this information will bring a significant advancement in the field of biomedical research and gives a scope to determine and confirm the illness of the patient at an early stage.**

Keywords: **EHR, GP**

I. INTRODUCTION

The General Practioners(GP) and the hospitals make use of the Electronic Health Records (EHR) to keep track of the medical history of the patients. These EHRs have all the detailed information regarding the patients medication, allergies, family history, previous disorders or injuries, treatment of those old traumas, outcome of previous medications etc... All such information cumulatively allows for a major findings in highly relevant and decentralized disorders like cancer and Alzheimer's disease. As majority of the documents are in natural language, i.e., human understandable and readable format, it becomes highly difficult for the computer analytic to understand and analyse the data. Building this gap in the study helps the hospitals and the GPs by reducing their time and manpower in doing this work thus bringing significant advancement to the biomedical research. All these information on the EHG can be extracted if a strict standard method of labelling the different categories/sections of the EHG is followed but, this method is not practically possible due to many reasons variations of text patterns across the globe, time consuming and requires expert to do it meticulously, to name a few.

We have tried to implement Named Entity Recognition (NER) to revolve the issues mentioned above. Named entity recognition (NER) a Natural Language Processor based technique that is widely used to easily identify the key elements in a text, like names of people, places, brands, monetary values, and more. Extracting the main entities in a text helps sort unstructured data and detect important information from different types of EHR, which is crucial when we have to deal with large datasets which do not follow a standard pattern. This technique of NER has gained popularity among the researchers in recent times as its technique of extracting valuable information from biomedical literature has proved reliable.

(Manning et al., 2008)

II. LITERATURE SURVEY

The medical information with good accuracy can be extracted by rule based systems in simpler terms that follow the default or normal speech patterns [1]. These systems don't scale well to complex patterns (e.g. elaborations of symptoms), variations of text patterns (e.g. different slangs of the language used) or improperly framed text (e.g. non globalized terminologies), which are the most commonly seen conditions found in EHRs. It is specifically noticed that the Neural Networks are more trust-able in the unique NLP tasks, where the traditionally used methods like rule-based and other ML-based methods do not succeed [4]. For instance, a combination of the long short-term memory (LSTM) type of recurrent neural networks (RNNs) and a convolutions neural network (CNN) has been successfully applied to set new state-of the-art performance for NER tasks based on CoNLL-2003 and OntoNotes 5.0 data (J. Chiu and Nichols, 2016). It is a very difficult task to get the related data to train for medical purposes and a very few are publicly available datasets exist for medical NLP, and even fewer exist with annotations for NLP tasks (e.g. document classification; or slot filling). The best known of these datasets are MIMIC-III [5] and i2b2 (i2b2, 2018, p. 2).

III. METHODOLOGY

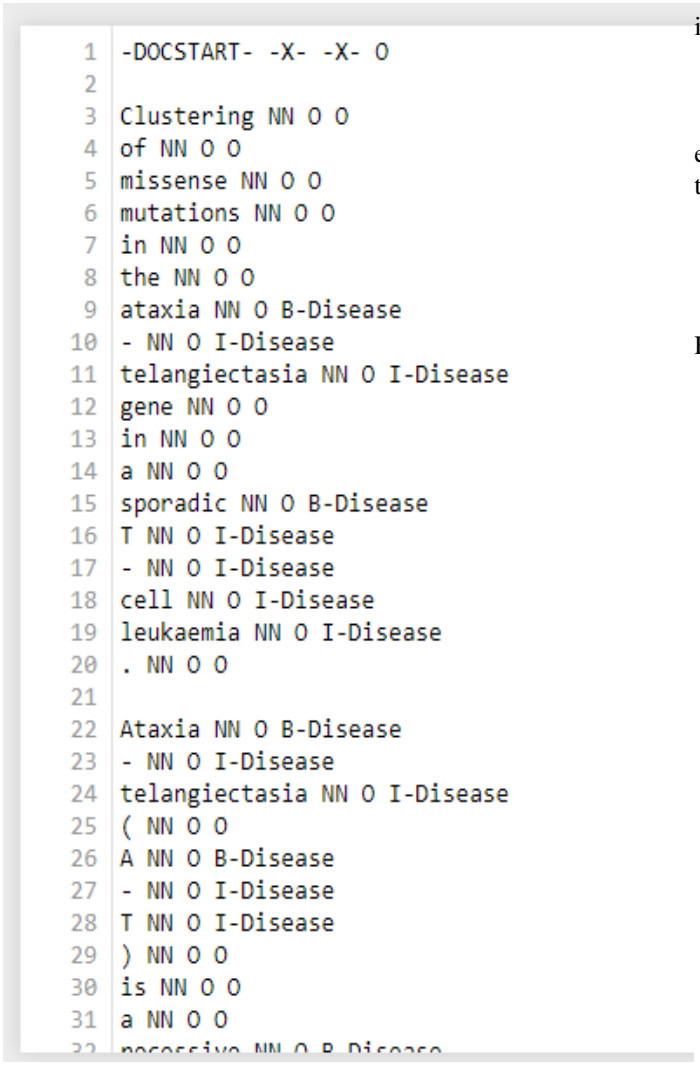
A. Data

The dataset used for this purpose is a medical EHR dataset. The preview of dataset used for the training of the model is pasted below as Fig1.

B. Procedure

The baseline model is based on the state-of-the-art NER architecture. The model has three inputs, namely a character-level, word-level and casing input, each of which encodes a different aspect of the text. The architecture starts processing these three inputs independently, but then merges them to process further. This architecture does so through a number of atomic operations or layers. These layers of operations are explained in the upcoming section of the paper.

We make use of the CNN-BLSTM model, a joint learning of shape, appearance and dynamics done by a deep learning technique. This includes a convolutional neural networks and bidirectional long short term memory(CNN-BLSTM). We create a word-level and character-level embedding and



```
-DOCSTART- -X- -X- O

Clustering NN O O
of NN O O
missense NN O O
mutations NN O O
in NN O O
the NN O O
ataxia NN O B-Disease
- NN O I-Disease
telangiectasia NN O I-Disease
gene NN O O
in NN O O
a NN O O
sporadic NN O B-Disease
T NN O I-Disease
- NN O I-Disease
cell NN O I-Disease
leukaemia NN O I-Disease
. NN O O

Ataxia NN O B-Disease
- NN O I-Disease
telangiectasia NN O I-Disease
( NN O O
A NN O B-Disease
- NN O I-Disease
T NN O I-Disease
) NN O O
is NN O O
a NN O O
```

Fig. 1: Preview of the dataset used

thus the set of unique words and labels from the three datasets are stored. This word embeddings and character embeddings are mapped to a 50 dimensional and 30 dimensional vector embeddings using the Global Vectors for Word representation (GLoVE) embedding model. All these mapped tokens are undergone the process of casing where we consider 8 different types of cases before they are mapped using the GLoVE model. In the next steps, we use the word embeddings from the GLoVE model and use them to generate a dictionary of all possible characters from the available set of words. We tag the datasets with the numerical values and generate a class predictor.

We have defined a class called BuildModel where all the previously defined models are called and we concatenate the processed character-level vector, word-level vector and casing data vector into a single vector. Dropout has been used with a rate of 0.5 to prevent the risks of overfitting in the model.

The parameter we considered for the evaluation purpose if F1 - score which is compared against the values as mentioned in the results sections.

IV. RESULTS

The results are as given below: With the 50d glove embedding and 30 epochs on the medical dataset we got the following distribution:

B-Disease: 3.71
I-Disease: 4.51
O: 91.78

For the general dataset the result is pasted below in fig2.

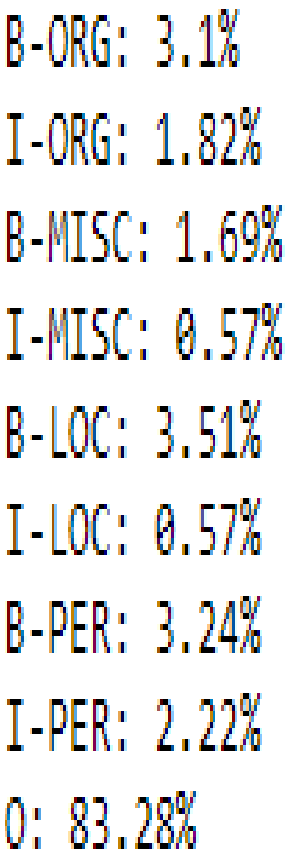


Fig. 2: Result on the normal dataset

We use F1-score as the metric for our model

V. CONCLUSIONS

Despite the increased interest and ground-breaking achievements in NLP research and NER systems, production-ready models and tools in the biomedical area remain sparse. It is one of the most significant barriers for clinical NLP researchers to overcome to start incorporating the most recent algorithms into their workflow as soon as possible

REFERENCES

Manning, C. D., Schütze, H., and Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge university press.

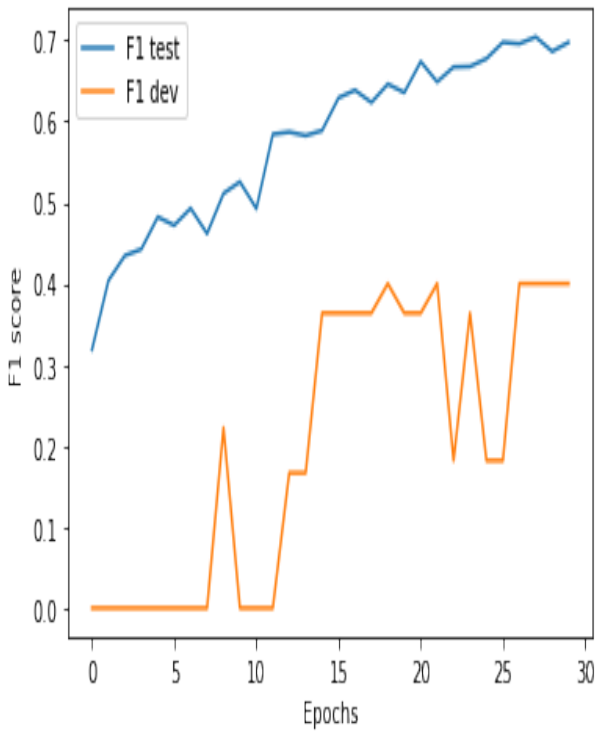


Fig. 3: F1-score

[1] George Karystianis, Kristina Thayer, Mary Wolfe and Guy Tsafnat, Journal of Biomedical Informatics, April 2017

[2] Young-Jin Cha, Wooram Choi and Oral Buyukozturk, conference of Computer-Aided Civil and Infrastructure Engineering, Deep learning based SHMDeveloping an Advanced Hybrid System of Structural Health Monitoring, March 2017

[3] The Medical HSR, Dataset (A. E. W. Johnson et al., 2016) and i2b2 (i2b2, 2018, p. 2).

[4] Erik Cambria, Bebo White, Jumping NLP Curves: A Review of Natural Language Processing Research, IEEE Computational Intelligence Magazine

[5] Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José Millet Roig, Ikaro Silva, Alistair E W JohnsonAn open access database for the evaluation of heart sound algorithms, 21 November 2016