Biomedical Named Entity Recognition

Jayakrishnan B (181IT120)

Priyanka B. G. (181IT135)

Gayathri Gutla (181IT216)

Introduction

- Electronic health records (EHRs) are the databases used by general practitioners (GPs) to store the medical history of patients.
- Most of the information held in these EHRs is in the form of natural language.
- Therefore it is largely inaccessible for statistical analysis.
- Unlocking this information can bring a significant advancement to biomedical research.

Literature Survey

- Rule-based systems can extract medical information with good accuracy in simpler situations, such as information that follows regular speech patterns (Karystianis et al., 2018).
- Neural networks are being especially successful in complex NLP tasks (Young et al., 2017), where more traditional rule-based and other ML-based methods fail (Cambria & White, 2014).

Contd..

- For instance, a combination of the long short-term memory (LSTM) type of recurrent neural networks (RNNs) and a convolutional 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 & Nichols, 2016).
- Amounts of annotated text, which is especially troublesome for their application to EHRs. Namely, only a few 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 (A. E. W. Johnson et al., 2016) and i2b2 (i2b2, 2018, p. 2).

Problem Statement

To apply the concept of NER natural language processing on the Electronic Health Records in order to make the process of extracting the required information from the unstructured, huge, unorganized medial history of a patient in a more faster and efficient way to reduce the effort of the medical official doing the work.

Dataset

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

EU NNP B-NP B-ORG
rejects VBZ B-VP O
German JJ B-NP B-MISC
call NN I-NP O
to TO B-VP O
boycott VB I-VP O
British JJ B-NP B-MISC
lamb NN I-NP O
. . O O

Peter NNP B-NP B-PER
Blackburn NNP I-NP I-PER

BRUSSELS NNP B-NP B-LOC
1996-08-22 CD I-NP O

The DT B-NP O
European NNP I-NP B-ORG
Commission NNP I-NP I-ORG
said VBD B-VP O
on IN B-PP O
Thursday NNP B-NP O
it PRP B-NP O
disagreed VBD B-VP O
with IN B-PP O
German JJ B-NP B-MISC
advice NN I-NP O
to TO B-PP O
consumers NNS B-NP O
```

```
-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
```

Methodology

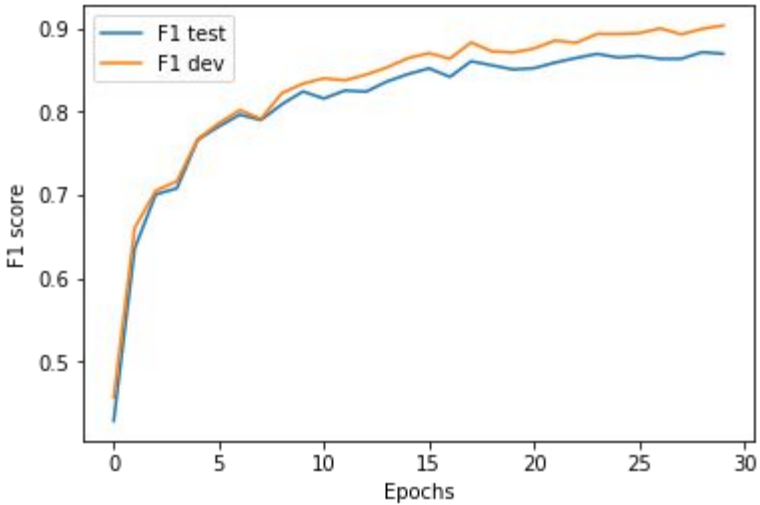
- The model has three inputs character-level, word-level and casing input.
- The architecture starts processing these three inputs independently and then merges them to process further.
- We create a word-level and character-level embedding and the set of unique words and labels from the three datasets are stored.
- These word and character embeddings are mapped to a 50 dimensional and 30 dimensional vector embeddings using the Global Vectors for Word representation.
- All these mapped tokens are undergone the process of casing where we consider 6 different types of cases before they are mapped using the GLoVE model.
- 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 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.

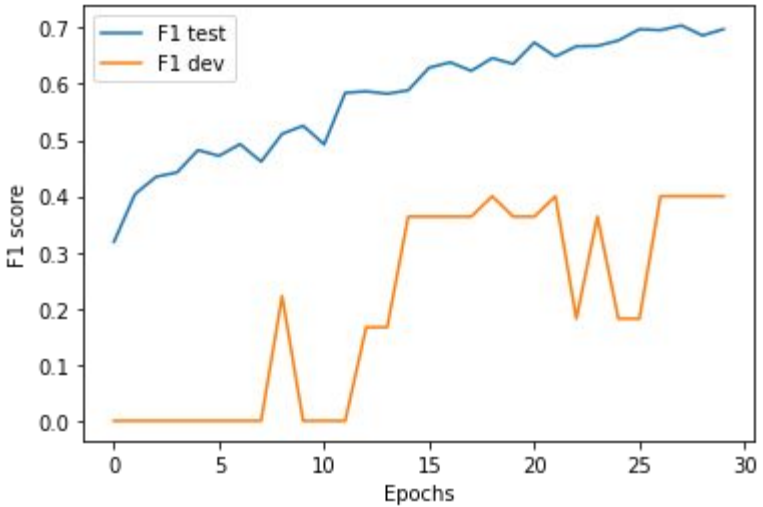
Contd.

- We tag the datasets with the numerical values and generate a class predictor.
- Build Model defines all the models that 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.

Results

We are using F1-score as a metric for our model and the graph plotted over 30 epochs is pasted below. The first figure is the graph over medical data and the second one over non- medical data





References

[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 SHM Developing 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

Thank You