**Medical Entity Recognition**

**Team No 5**

**Geethika Dudyala(20161154)**

**Aditya Saripalli(20173071)**

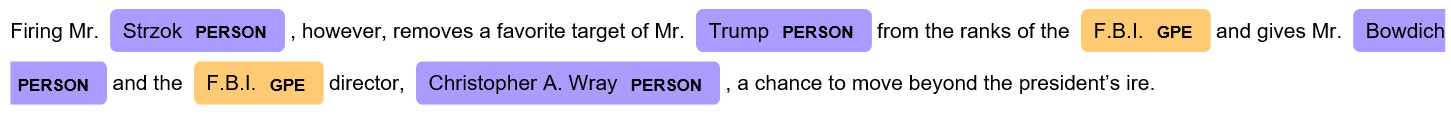
**Aman Bansal(20161008)**

**Tarun Vatwani(2018201075)**

**Abstract**

Medical Entity Recognition(MER) is an application of a very famous problem in the field of Information Extraction(IE) i.e. Named Entity Recognition(NER). In other words it is a domain specific NER in this case the domain is medical field.

NER is a method/model which takes a series of text(sentences or paragraphs) and provides labels to noun phrases present in that piece of text. The image below shows an example of how NER model typically works.



This project aims at parsing named entities and in this project, we have to recognize and classify medical data into the relevant categories, namely drugs, diseases, symptoms, side-effects, treatment, etc. Twitter data will be the input and based on previous medical data from databases and ontologies, relevant medical terms have to be parsed and classified (medical named entities are recognized and classified based on the category they belong to (ex: drug or a disease or cure etc....)

As the name suggests, a Medical Name Entity Recognizer identifies medical entities in text. Medical entities, in the context of our project, are fixed, there are 5 categories as mentioned above. Previously, researchers in the field have used hand crafted features to identify medical entities in medical literature. In this project, we have to extend medical entity recognition on tweets. We would use NLP toolkits designed for processing tweets along with other medical ontologies (or databases) to exploit semantic features for this task.

**Methodologies Tried**

1. **Dataset Preprocessing :** For this phase we have worked with three datasets. We were provided with two datasets i.e. CADEC and Micromed.

We found another dataset [TwiMed](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5438461/) for our training, it is a corpus of text from PubMed and Twitter which is annotated by two pharmacists, comprises semantically correct annotations for a set of drugs, diseases, and symptoms. This corpus contains the annotations for 3144 entities, 2749 relations, and 5003 attributes.

The datasets available to us did not contain all the desired labels. CADEC is annotated with 5 labels (Drug, Disease, Symptom, Finding and Adverse Drug Reactions) whereas TwiMed and Micromed are labeled with Drug/Pharmacologic Substance, Disease and Symptoms. So we have trained with the given labels. We hope to train on desired labels in the next phase with generalized twitter corpus.

We used the following approaches to preprocess our datasets:

* CADEC and TwiMed are annotated using brat annotation tool so to convert it into a dataframe we use the following [script](https://github.com/spyysalo/standoff2conll). The script gives output in ConLL data format from which it is converted into CSV file using a custom script. For twitter part of TwiMed it is again brat annotated .ann file, with named on their tweet IDs, so using their names we fetched out tweets(only 521 were present), and created a .txt file for each tweet
* MicroMed dataset is of .linejson format, with tweet IDs as keys, and ‘type’ contains type of annotation that word belongs to, and an attribute of json type named ‘location’ has two keys ‘start’ and ‘end’, which tells us that from where in text that word is starting and where it is ending. But in this field, data has very discrepancy, as the pointers are very different from the actual occurrences of words, for this we have to go manually file by file, where it occurs, to resolve it, and after doing all this we got only 370 tweets, as many of them were either deleted or account from which they were done is deactivated. After doing all this again we had converted it to ‘brat-flavored standoff’ format, as we had a ready script to get CSV from ‘brat-flavored standoff’ format, and after doing all this preprocessing we got ready to use CSV format of Twitter data.

1. **Model Preparation:** MER can be implemented as a sequence classification task, where every chunk is predicted IOB-style as Drug, Disease, Symptom, Treatment and Test. The IOB format (short for inside, outside, beginning) is a common tagging format for tagging tokens. The B- prefix before a tag indicates that the tag is the beginning of a chunk, and an I- prefix before a tag indicates that the tag is inside a chunk. The B- tag is used only when a tag is followed by a tag of the same type without O tokens between them. An O tag indicates that a token belongs to no chunk.

We approached the problem of identifying the medical entities in the dataset using the following approaches – A naïve max-word-frequency approach, Conditional Random Fields (CRF), Long Short Term Memory (LSTM) and combination of LSTM-CRF. The LSTM-CRF is used by most of the state-of-the-art approaches to named entity recognition.

The LSTM-CRF based approach is mentioned below:

We are given an input sequence x = (x_1,\dots, x_m), i.e. the words of a sentence and a sequence of output states s = (s_1,\dots, s_m), i.e. the named entity tags. In conditional random fields we model the conditional probability of the output state sequence given an input sequence given by:

We did this by defining a feature map that maps an entire input sequence x paired with an entire state sequence s to some d-dimensional feature vector. Then we can model the probability as a log-linear model with the parameter vector w\in\mathbb{R}^d

\[p(s|x; w) = \frac{\exp(w\cdot\Phi(x, s))}{\sum_{s^\prime} \exp(w\cdot\Phi(x, s^\prime))},\]

where s^\prime ranges over all possible output sequences. We can view the expression w\cdot\Phi(x, s) = \text{score}_{crf}(x,s) as a scoring how well the state sequence fits the given input sequence. The idea is now, to replace the linear scoring function by a non-linear neural network. So we define the score where W_{s_{i-1}, s_i} and b are the weight vector and the bias corresponding to the transition from s_{i-1} to s_i, respectively.\[\text{score}_{lstm-crf}(x,s) = \sum_{i=0}^n W_{s_{i-1}, s_i} \cdot \text{LSTM}(x)_i + b_{s_{i-1},s_i},\]

After constructing this score function, we can optimize the conditional probability and propagating back through the network. We are going to use the implementation provided by the [keras-contrib](https://github.com/farizrahman4u/keras-contrib) package that contains useful extensions to the official keras package.

**Findings from the current implementation**

1. The TwiMed, CADEC and MicroMed dataset has 9435 unique sentences with 13204 unique tokens.
2. The Bidirectional LSTM-CRF model achieves a f1-score of 58.1 %
3. From what we observed the it fails to classify certain category words such as symptoms and adverse drug reactions properly as they are somewhat similar

**Code link**

<https://github.com/adisarip/medical_entity_recognition>

**Difference from Original Scope**

Our original scope document contained the following milestones:

1. Studying and Implementing basic sequence to sequence models such as LSTM for NER
2. Study state of the art methods sequential models and try to test them for twitter data set
3. Working on twitter corpus to create a training set for tweets using R1,R2 and R3
4. Final documentation and PPT
5. Preparation of working demo

So far we have completed the implementation of the basic LSTM and Bidirectional LSTM-CRF based models. Also we have finished reading of state of the art models using contextual word embeddings. Now after this we will request the access of general twitter corpus and some helper resources to create and annotate twitter data for our task.

**New Timeline**

The above datasets are very small in size, hence may not be sufficient for training.

We have to increase the size of the dataset that is available for training heuristically.

For this, we would be given three resources,

R1 : A list of hashtags, which are relevant to the medical domain

R2 : A general tweet corpus (about 40-50 GB in size)

R3 : A list of medical terms and their appropriate categories.

1. Using the list of hashtags (R1) obtain a new dataset consisting of a subset of the general tweet corpus (R2) - medical domain related tweets. This new dataset can be used both for training as well as testing purpose. A part of this new dataset can be annotated using R3 and could be used for training purposes. The rest of the dataset can be used for testing purposes. (7-9 days)
2. Incorporating Contextual Word Embedding such as Elmo, GloVe etc. with the model. (7-9 days)
3. Final Documentation and video demo. (3-4 days)