E1246 - Natural Language Understanding
Assignment-3 : Build an NER system

Divya Agarwal(14722)
divyaagarwal@iisc.ac.in

Abstract

The goal of the assignment is to build an NER system for diseases and treatments. The input of the code will be a set of tokenized sentences and the output will be a label for each token in the sentence. Labels can be D, T or O signifying disease, treatment or other.

1 Conditional Random Fields

CRF is the discriminative equivalent of the Generative HMM model. Therefore it is clear that through CRF we are only trying to model the probability distribution.In the Graphical model the edges are undirected for CRFs and each node in the graph denotes a random variable. Basically we call it a random field because the state of each node(Random Variable) depends on all the neighbors in the graph. It is through this manner we incorporate context in tag prediction.

2 Feature Used

These are the features which i used for experiments.

- **POS Tag**
- **Word Suffix**
- **Word special character**
- **Word is digit or not**
- **Neighbourhood Information**

3 Method and Experiments

The Algorithm that is used for training our model is LBFGS which is an optimization algorithm in the family of quasi-Newton methods that approximates the BroydenFletcherGoldfarbShanno(BFGS) algorithm using a limited amount of computer memory. It is one of the popular algorithm for parameter estimation in machine learning.For tuning hyper parameter, i used 10-fold cross validation. Evaluation metrics are Accuracy, F1-Measure, precision and recall.

- **Without Using extra feature**

	D	O	T	Avg
precision	0.83	0.92	0.87	0.87
recall	0.27	0.99	0.47	0.57
f1-measure	0.41	0.95	0.61	0.65

- **Using POS tag, suffix**

	D	O	T	Avg
precision	0.84	0.95	0.87	0.88
recall	0.56	0.99	0.59	0.71
f1-measure	0.56	0.96	0.74	0.75

- **Using Special character, Neighbourhood Information,POS Tag**

	D	O	T	Avg
precision	0.83	0.96	0.92	0.90
recall	0.62	0.99	0.63	0.76
f1-measure	0.58	0.96	0.82	0.77

- **Using POS tag, suffix, Special character, Neighbourhood Information**

	D	O	T	Avg
precision	0.90	0.98	0.95	0.94
recall	0.70	0.99	0.68	0.79
f1-measure	0.58	0.97	0.85	0.79

4 Git Hub Link

https://github.com/Divyagarwalfeb/NLU-project.git