E1246 - Natural Language Understanding
Assignment 3 : Named Entity Recognizer

Monish Keswani(14698)
monishkumar@iisc.ac.in

Abstract

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

1 Dataset

We are given a training dataset of labeled sentences. The format of each line in the training dataset is token label. There is one token per line followed by a space and its label. Blank lines indicate the end of a sentence. It has a total of 3655 sentences. The sentences are split by blank spaces. The dataset is divided into train-dev-test in ratio 80:10:10. The model is trained on the train set. The hyperparameters are tuned using the dev set. The final testing is done on the test set.

2 Sequence Tagger

I have used CRF approach for Named Entity Recognition

2.1 Condtional Random Fields

Conditional Random Fields (CRFs) are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. A conditional random field (CRF) is a type of discriminative probabilistic model used for the labeling sequential data such as natural language text. Conditionally trained CRFs can easily include large number of arbitrary non independent features. The expressive power of models increased by adding new features that are conjunctions to the original features. When applying CRFs to the named entity recognition problem an observation sequence is the token sequence of a sentence or document of text and state sequence is its corresponding label sequence.

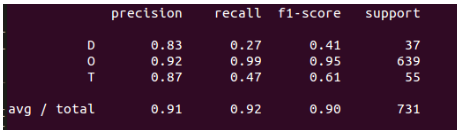
2.2 Feature Learning

The Features that were used during the model training are listed as follows:

1. Word Suffix
2. Word Prefix
3. Word meanings
4. POS tags
5. Neighbourhood Information

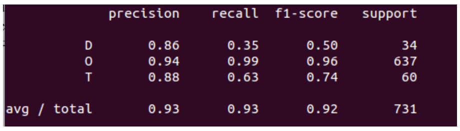
3 Methodology and Experiments

The Algorithm that is used for training 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 a popular algorithm for parameter estimation in machine learning. 10-fold cross validaton is used to tune the hyerparameters. Accuracy, F1-score, precision and recall are used as evaluation measures



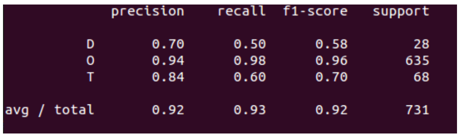
```
             precision    recall  f1-score   support

          D       0.83      0.27      0.41        37
          O       0.92      0.99      0.95       639
          T       0.87      0.47      0.61        55

avg / total       0.91      0.92      0.90       731
```

Figure 1: Without considering any features



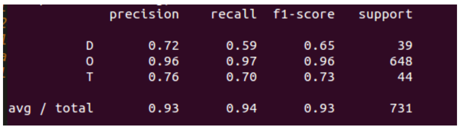
```
             precision    recall  f1-score   support

          D       0.86      0.35      0.50        34
          O       0.94      0.99      0.96       637
          T       0.88      0.63      0.74        60

avg / total       0.93      0.93      0.92       731
```

Figure 2: Conidering Word Suffix, Prefix and meanings



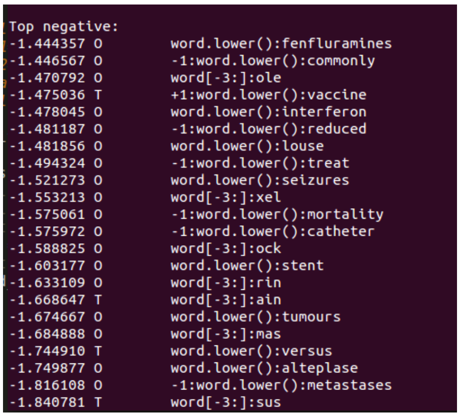
```
             precision    recall  f1-score   support

          D       0.70      0.50      0.58        28
          O       0.94      0.98      0.96       635
          T       0.84      0.60      0.70        68

avg / total       0.92      0.93      0.92       731
```

Figure 3: After POS taggings



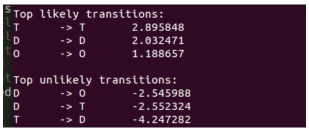
```
             precision    recall  f1-score   support

          D       0.72      0.59      0.65        39
          O       0.96      0.97      0.96       648
          T       0.76      0.70      0.73        44

avg / total       0.93      0.94      0.93       731
```

Figure 4: After Hyperparameter Tuning



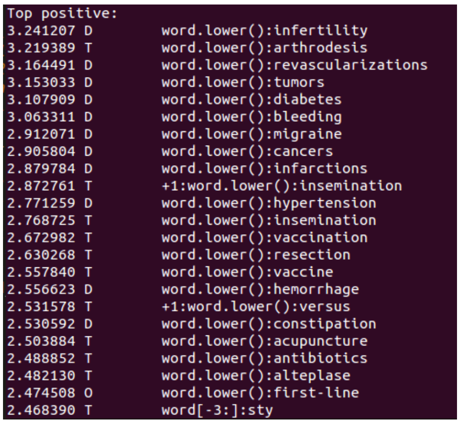
```
Top negative:
-1.444357 O        word.lower():fenfluramines
-1.446567 O        -1:word.lower():commonly
-1.470792 O        word[-3:]:ole
-1.475036 T        +1:word.lower():vaccine
-1.478045 O        word.lower():interferon
-1.481187 O        -1:word.lower():reduced
-1.481856 O        word.lower():louse
-1.494324 O        -1:word.lower():treat
-1.521273 O        word.lower():seizures
-1.553213 O        word[-3:]:xel
-1.575061 O        -1:word.lower():mortality
-1.575972 O        -1:word.lower():catheter
-1.588825 O        word[-3:]:ock
-1.603177 O        word.lower():stent
-1.633109 O        word[-3:]:rin
-1.668647 T        word[-3:]:ain
-1.674667 O        word.lower():tumours
-1.684888 O        word[-3:]:mas
-1.744910 T        word.lower():versus
-1.749877 O        word.lower():alteplase
-1.816108 O        -1:word.lower():metastases
-1.840781 T        word[-3:]:sus
```

Figure 5: Top Negative Features



```
Top likely transitions:
T      -> T       2.895848
D      -> D       2.032471
O      -> O       1.188657

Top unlikely transitions:
D      -> O       -2.545988
D      -> T       -2.552324
T      -> D       -4.247282
```

Figure 6: Top likely and unlikely features



```
Top positive:
3.241207 D        word.lower():infertility
3.219389 T        word.lower():arthrodesis
3.164491 D        word.lower():revascularizations
3.153033 D        word.lower():tumors
3.107909 D        word.lower():diabetes
3.063311 D        word.lower():bleeding
2.912071 D        word.lower():migraine
2.905804 D        word.lower():cancers
2.879784 D        word.lower():infarctions
2.872761 T        +1:word.lower():insemination
2.771259 D        word.lower():hypertension
2.768725 T        word.lower():insemination
2.672982 T        word.lower():vaccination
2.630268 T        word.lower():resection
2.557840 T        word.lower():vaccine
2.556623 D        word.lower():hemorrhage
2.531578 T        +1:word.lower():versus
2.530592 D        word.lower():constipation
2.503884 T        word.lower():acupuncture
2.488852 T        word.lower():antibiotics
2.482130 T        word.lower():alteplase
2.474508 O        word.lower():first-line
2.468390 T        word[-3:]:sty
```

Figure 7: Top Positive Features