E1 246 Assignment-3

Rahul Chittimalla
Sr no: 06-02-01-10-51-17-1-14585
rahulc@iisc.ac.in

In this assignment, we have developed an NER system for diseases and treatments. There are three possible labels for each token in the dataset D, T & O signifying disease, treatment or other. The dataset is read and formatted such that each list is a sentence with (token, label) pairs of that sentence.

CRF

The algorithm used in this NER system is conditional random field(CRF). We denote $x = (x_1, ..., x_m)$ as the input sequence, i.e. the words of a sentence and $s = (s_1, ..., s_m)$ as the sequence of output states, i.e. the named entity tags. In condional random fields we model the conditional probability

$$p(s_1, ..., s_m | x_1, ..., x_m)$$

We do this by defining a feature map

$$\phi(x_1, ..., x_m, s_1, ..., s_m) \epsilon \Re^d$$

Then we can model the probability as a log-linear model with the parameter vector $w \ \epsilon \ \Re^d$

$$p(s|x : w) = \frac{exp(w \cdot \phi(x, s))}{\sum_{s'} exp(w \cdot \phi(x, s'))}$$

Features

The features played an important role in the sequence tagging. There are number of features possible to be used in the named entity recognition task. Some of the features that are used in this assignment are,

- bias
- whether the word is capitalized
- is it a number
- chunking
- previous word
- type of previous word
- next word
- type of next word, etc.

Methodology and Experiments

Experiments are done on two options:

i train on 70% data and test on remaining

ii 10-fold cross validation

Option 1

In this option, the model is trained on 70% of the data and the remaining data is used for testing. The results for this model are shown in Table 1.

label	precision	recall	f1-score
D	0.885	0.362	0.514
O	0.747	0.985	0.850
T	0.796	0.118	0.206
avg/total	0.775	0.757	0.699

Table 1: Scores for option1

The features tend to play a cruicial role in performance of the model. This model is tested on two sets of features

i contains a label as feature of both previous and next word

ii doesn't contain the label of previous and next word

The difference is scores is captured in Figure 1.

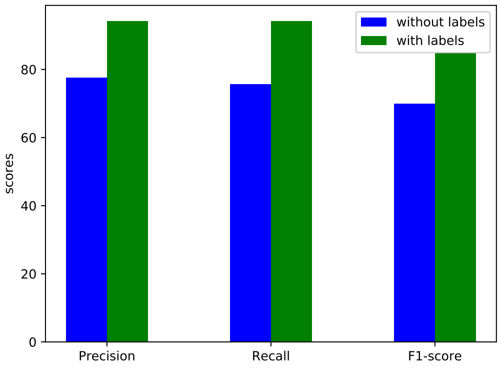

Figure 1: Indicates the difference in scores

We can conclude by looking at the increase in scores with the addition of that extra feature that the feature at consideration is very important and a deciding factor.

Option 2

A 10-fold cross validation is done in this option. The results for this model are shown in Table 2.

label	precision	recall	f1-score
D	0.77	0.62	0.69
O	0.93	0.97	0.95
T	0.68	0.42	0.52
avg/total	0.90	0.91	0.91

Table 2: Scores for option2

Figure 2 shows the difference in scores and the importance of the feature which denotes the labels of previous and next words of the token being considered.

Tables 1 & 2 conclude that option2 i.e. 10-fold cross validation performs way better than option1 even in comparision with addition/removal of the feature in discussion.

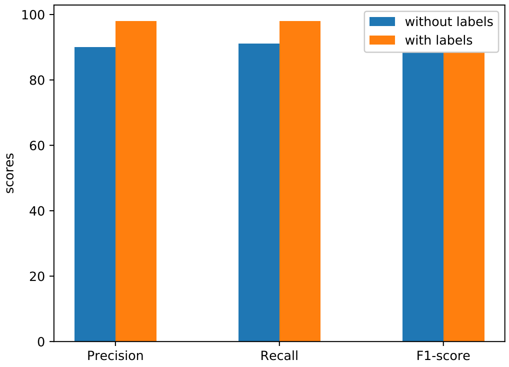

Figure 2: Indicates the difference in scores