Assignment-3: Named Entity Recognition

Sonu Dixit
sonudixit2k@gmail.com

Abstract

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories. Here we have to classify every word in three given categories(’D’: disease, ’T’:treatment,’O’:other.)

1 Introduction

In this assignment, the input is a sequence, and we have to tag each token in the sequence as (’O’,’D’,’T’.) Input file is ’ner.txt’, each line contains a word and its tag. First i have divided this file in three parts for training, validation and testing. For completing this task, i have used CRF model, and biLSTM model combined with CRF. For CRF i have used pycrfsuite, and for biLSTM i have used keras.

Here, for comparing the result, accuracy is not a good measure. Since, ’O’ category tokens are in very large amount, simply tagging all tokens as ’O’ would result in very high accuracy.

2 Models:

First model i used is simple crf model, using pycrfsuite library. In this model, first thing i tried was just giving each words independently as input. And then i modified it to give each sentence as input. That is a sentence is being considered as a sequence and this sequence is being given as input. Just using simple CRF, although i was getting good accuracy(about 0.94), Precision and recall was poor around 0.50.

Then i used a biLSTM CRF model. Its a CRF on top of biLSTM. In this model, our input size must be fixed. Here i searched for a maximum length sentence, and then made each sentence of that length by padding required number of ”unk” tokens at the end of those sentences. For these ”unk” tokens , i have tagged it as ”O”, since the file has more more than 90 percent tags as ’O’. Here the maximum length sentence was 110, so i converted all sentences to 110 length.

Figure 1: model for CRF

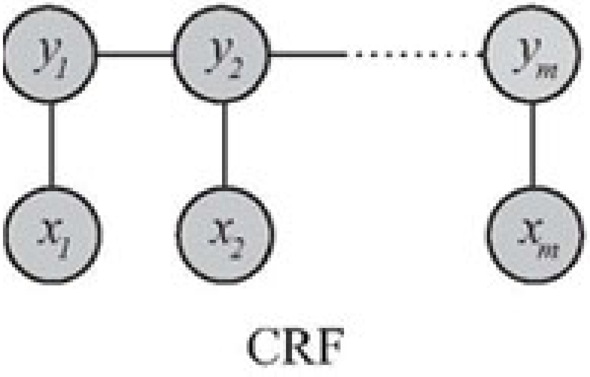
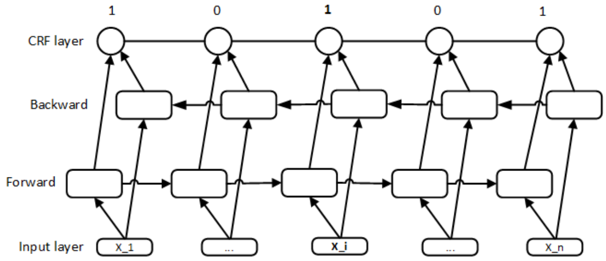


Figure 2: BiLSTM CRF model

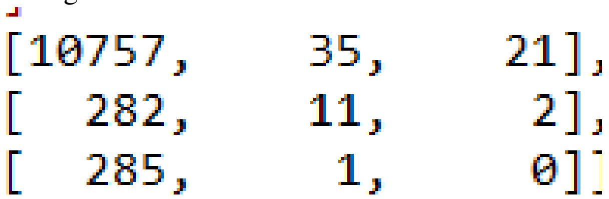


3 Results

Using the preparefiles.py i have divided the input file in three parts train.txt, valid.txt, test.txt. Result for crf model is: Accuracy =94 percent confusion matrix: label = O, D,T

for biLSTM CRF model i have used batch size of 30, and trained for 7 epochs. 20 percent vali-

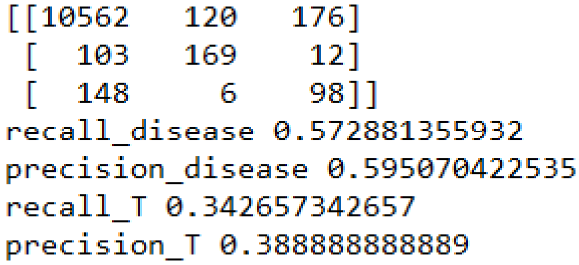
Figure 3: confusion matrix for CRF model



```
[10757,    35,    21],
[   282,    11,     2],
[   285,     1,     0]]
```

dation set has been used. Confusion matrix is as follows:

Figure 4: confusion matrix for biLSTM CRF model



```
[[10562   120   176]
 [  103   169    12]
 [  148     6    98]]
recall_disease 0.572881355932
precision_disease 0.595070422535
recall_T 0.342657342657
precision_T 0.388888888889
```