Assignment Report

Praveen Gupta
Computer Science and Automation
Indian Institute Of Science
praveengupta@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.

- **Task:** You need to write a sequence tagger that labels the given sentences in a tokenized test file. The tokenized test file follows the same format as training except that it does not have the final label in the input. Your output should label the test file in the same format as the training data.

1 Implementation

I have implemented using the 'pycrfsuite' library in python. I have added following features :-

- I added the Last 4 character of word. I tried with last 2 and 3 word also but 4 give best results.
- I added the length of word because the most of disease and treatment name are of larger length.
- I added the starting two character but F1 score decreases so i removed it.
- I used the nltk pos tagger to tag each word and added it. This gives good improvement in F1 score
- I tried with the IsUpper flag variable but F1 decreases so i removed it.
- I used the IsTitle flag variable.
- I used the IsDigit flag variable, but it also remove the F1 score so i removed it.
- I used the start 3 letter of the word and it slightly increase the F1.
- I have added the word by removing all of its vowel.
- I added the same set of features for the previous word also.
- I added these for the next word also.
- I added these features for the n-2 and n+2 word but slightly decrease the F1 score so i remove them.

2 Result

Train:Test = 8:2

2.1 Before using POS

Features	**Recall Score**			
	T	**D**	**O**	**Avg.**
POS Tag	0.63	0.54	0.91	0.89

Features	**F1 Score**			
	T	**D**	**O**	**Avg.**
POS Tag	0.69	0.60	0.92	0.88

Features	**Precision Score**			
	T	**D**	**O**	**Avg.**
POS Tag	0.74	0.67	0.90	0.89

2.2 After using POS for current, previous and next word

.

Features	**Recall Score**			
	T	**D**	**O**	**Avg.**
POS Tag	0.69	0.60	0.97	0.93

Features	**F1 Score**			
	T	**D**	**O**	**Avg.**
POS Tag	0.74	0.66	0.96	0.93

Features	Precision Score			
	T	D	O	Avg.
POS Tag	0.79	0.72	0.95	0.93

.

The score reported are after 300 iterations.

3 Acuuracy/Measures

3.1 Task 1

F1 score, Recall score and Precision score are used as the measurement metrics.

4 Observation

After adding the POS tags for the word and adding POS tag for the previous and the next word the accuracy and F1 score has increased by a good margin.

5 Link to Github

https://github.com/praveengupta18/NLU-Assignment3.git