NLU-Assignment3 Report

Akash Mishra
SR No- 14421
akashmishra@iisc.ac.in

1 Introduction

In Assignment-3, we need to develope an NER system for diseases and treatment. by classifying words into some predefined categories. I included multiple feature for every training instance. the labels in input are D,T and O which represents disease, treatment and other. After Shuffling the given training data, I have used 80% of the data for training and rest 20% for testing the accuracy of the model. I also add and removed different subset of features and added the results for same.

2 Implementation Details

2.1 Model:

I have used Mallet which is a java based package for statistical natural language processing. It has several inbuilt tools for document classification with help of which it can convert texts into features and it also has tools for sequential tagging. After that with the help of an algorithm it can measure the performance of classifier.

2.2 Features:

For features , I append semantic labels of each tokens as a feature. If the instance is made up of only numerical values then I have added another feature Digits for it. For the identification of stopping words, I used another feature stopwords. For the identification of parts of speech of every instances, I have used another feature named POS and to know if an instance is starting with a capital letter, I have used another feature Capital.

2.3 Tagging:

I have used mallets SimpleTagger tool for sequence tagging

3 Results

I splitted my dataset into a ratio of 80-20% for training and testing the model respectivelly.

1.) No features added
Train accuracy: 0.8784
Test accuracy: 0.8652

2.) Stopword
Train accuracy: 0.8893
Test accuracy: 0.8776

3.) POS
Train accuracy: 0.8718
Test accuracy: 0.8762

4.) Semantic
Train accuracy: 0.8875
Test accuracy: 0.8763

5.) StopWord + POS
Train accuracy: 0.8876
Test accuracy: 0.0.8759

6.) StopWord + Semantic
Train accuracy: 0.8838
Test accuracy: 0.8745

7.) POS + Semantic
Train accuracy: 0.8753
Test accuracy: 0.0.8712

8.) Capital + Digit
Train accuracy: 0.8856
Test accuracy: 0.0.8779

9.) POS + Semantic + Capital + Digit
Train accuracy: 0.8871
Test accuracy: 0.0.8745

10.) All features included
Train accuracy: 0.8856
Test accuracy: 0.8732

4 Conclusion

We can observe that the StopWord feature changes the accuracy significantly for both trainig and testing dataset. When we does not include StopWord, accuracy is always less than the accuracy with using it.

Github Link:
https://github.com/AkashMishra2k16/Assignment3/tree/master