Assignment 3 Report

Rupali Dakhane
Computer Science and Automation
Indian Institute Of Science
rupalid@iisc.ac.in

Abstract

Build a Named Entity Recognition system for diseases and treatments. The task involves writing a sequence tagger that labels the given sentences in a tokenized test file.

1 Preprocessing

I have divided the datasets in train ,dev and test in the ratio of 7:2:1.

2 Implementation

The below two models have been used as sequence tagger for the given dataset.

2.1 Sequence Tagger based on Conditional Random Field model

The following additional features were added to the dataset to improve tagging accuracy :

- **Part of Speech Tags:** Pos tags are added as additional feature using nltk tool.Some of the Pos tags are- 'NN', 'NNP', 'JJ', 'CD', 'DT' etc.

- **IsUppercase:** The feature value is True if the word is upper-cased, otherwise the feature value is False.

- **IsTitle:** The feature value is True if the word is title-cased, otherwise the feature value is False.

- **Isdigit:** The feature value is True if the current token is a digit, otherwise the feature value is False.

2.2 Deep Sequence tagging Model

This Deep sequence tagging model uses:

- **Embedding:** The model uses word embeddings. The sentences are mapped to a sequences of numbers and then padded. These sequences are given as inputs to the embedding layer of deep model. The embeddings are also trained while training the model.

- **Layers:** The model consists of one Bidirectional LSTM hidden layer with 75 memory cells and one CRF layer. The model is a hybrid of deep sequence tagging and CRF based sequence tagging model

- **Number of Epochs:** The model is run for 10 epochs until the accuracy increases.

2.3 Deep Sequence tagging Model using character embeddings

Deep sequence tagging model uses:

- **Embedding:** The model uses character embeddings. The sentences are mapped to a sequences of numbers each number representing a character mapped using a dictionary of characters and then padded. These sequences are given as inputs to the embedding layer of deep model. The embeddings are also trained while training the model.

- **Layers:** The model consists of one Bidirectional LSTM hidden layer with 75 memory cells along with a dropout layer.

- **Number of Epochs:** The model is run for 30 epochs until the accuracy increases.

3 Result

3.1 Results obtained for CRF based model

Accuracy achieved : 92.56

Labels	precision	Recall	F1-score
T	0.81	0.67	0.73
D	0.71	0.56	0.63
O	0.94	0.97	0.96

3.2 Results obtained for Deep Sequence tagger model

Accuracy achieved : 91.20

Labels	precision	Recall	F1-score
T	0.74	0.60	0.66
D	0.60	0.53	0.56
O	0.94	0.97	0.96

3.3 Results obtained for CRF based model

Accuracy achieved : 91.59

Labels	precision	Recall	F1-score
T	0.75	0.57	0.65
D	0.69	0.59	0.63
O	0.94	0.97	0.95

4 Accuracy/Measures

Accuracy in percentage is used as the measure for this task.

5 Github Link

https://github.com/Rupali0408/NLU_Assignment3