Assignment 2

Neha Iyer
SR No. 14942

1 Task 1

To build an NER system for diseases and treatments.
Build a sequence tagger that labels the given sentences in a tokenized test file.

2 Datasets used

The input 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.he format of each line in the training dataset is token label. The input file has one token per line followed by a space and its label. Blank lines indicate the end of a sentence. It has a total of 3655 sentences.

3 Data Split

The dataset was divided into train dev and test sets in following manner:
The test set was created by taking 20% of the input file and dev set was created by taking 10% of the input file. The model was trained on the remaining 70% of the input file.

4 Data Preprocessing

Following preprocessing steps were performed in order to make it suitable before feeding it to the training model:
1. The training and test data is converted to tab-separated format. The format is one word per line, separate column for token and label, empty line between sentences.
2. Removal of numeric characters
3. Transform the text case to lower case

5 Approach followed

The approach followed to develop token-level language model is as follows:
1. A bidirectional LSTM is implemented sequence tagging.
2. Pre trained Glove word embeddings were used with vocab size as 6 billion and embedding size of 300.
3. Maximum vocab size used is 7500.
4. CRF is used as the output layer.
5. Learnin rate used is 1.0.
6. Optimizer used is adadelta.
7. Maximum number of epochs set is 10.
8. Batch size used is 64.

6 Evaluation metric

Evaluation Metric used for computing the loss is precision, recall and accuracy.

7 Results

Please note that the bidirectional LSTM model was implemented using Theano library and all the following results were obtained purely by executing the code without GPU support.
The following test results were recorded on the test set by averaging the results of multiple iterations.
Test label: T

- Test accuracy: 77%
- Test precision: 76%
- Test recall: 55%
- Test F1 measure: 70%

Test label: T

- Test accuracy: 77%

Token	Gold Label	Predicted Label
advanced	D	D
lung	D	D
cancer	D	D
methods	O	O
author	O	O
use	O	O
treatment	T	T
modalities	T	O

Table 1: Tagged output

- Test precision: 78%
- Test recall: 17%
- Test F1 measure: 28%

Test label: O

- Test accuracy: 77%
- Test precision: 77%
- Test recall: 96%
- Test F1 measure: 85%

The evaluation metrics for label O is better than other two labels. The poor test score for label D may be attributed to the skewed label distribution in given data set.

8 Sample Output

The sample tagged test data output is as shown in Table 1