E1246 - Natural Language Understanding
Assignment3 : Named Entity Recognizer

Shivank Gupta(14638)
shivankgupta@iisc.ac.in

Abstract

In this assignment we have build a **NER**(Named Entity Recognizer) which classify words into some predefined categories.The input to the model is the tokenized sentence and we have to predict the label for each word in the sentence.

1 Dataset

Training dataset that we have contain token label. There is 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.

- **Preprocessing** I have read the sentences as list of (word,tag) therefore having list of sentences and within that list of (word,tag) pair.we have 3655 sentences containing 11311 different tokens with 3 different tags. I have taken length of sentences to be 100 if the actual length is shorter than 100 than padding is done at the end.

- **Data split** I have divided data into 80-20 percent as train and test respectively and further taken 10% of training set as validation set.

2 Implementation Details

I have used combinatory approach combining a bidirectional LSTM model and a CRF model.

- **Bidirectional LSTM** The idea of Bidirectional Recurrent Neural Networks (RNNs) is straightforward. It involves duplicating the first recurrent layer in the network so that there are now two layers side-by-side, then providing the input sequence as-is as input to the first layer and providing a reversed copy of the input sequence to the second.

- **Condtional Random Fields(CRF)** Conditional Random Fields (CRFs) are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to other designated input nodes. A conditional random field (CRF) is a type of discriminative probabilistic model used for the labeling sequential data such as natural language text. The expressive power of models increased by adding new features that are conjunctions to the original features. When applying CRFs to the named entity recognition problem an observation sequence is the token sequence of a sentence or document of text and state sequence is its corresponding label sequence.

- **A CRF and Bi-LSTM**: In this Assignment I have combined Bi-LSTM network and a CRF network to form a Bi-LSTM-CRF model, In this network past input features are handle efficiently by Bi-LSTM whose output is passed to CRF for predicting the sequence of next token. Bidirectional LSTM layer considers the previous input features and obtain sentence level tag information from the CRF layer.

  I have used keras package to achieve this goal.

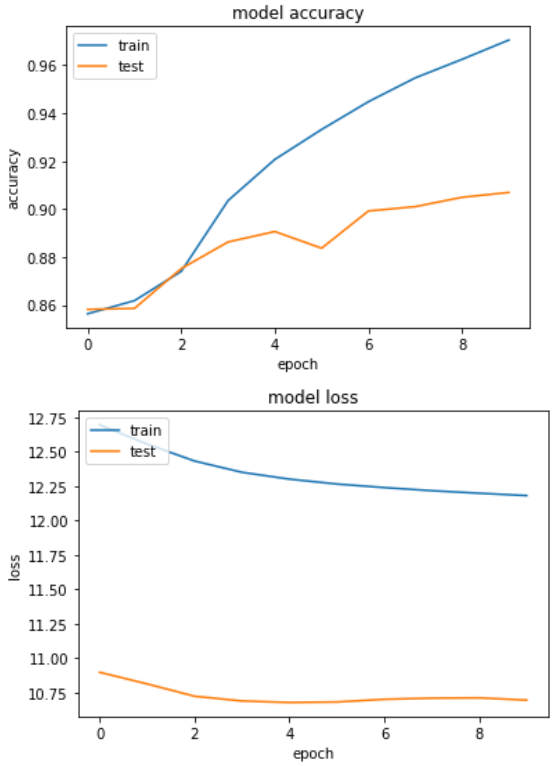
3 Training procedure

I have taken maximum length of each sentence to 100. Training is done in batches with batch size of 64 i.e. at a time 64 sentences are feeded into the network. I have trained my model for 10 epochs and obtain reasonable accuracy within that.

4 Evaluation Metric

Evaluation metrics used for this task are Accuracy, F1 Measure, Precision and Recall.

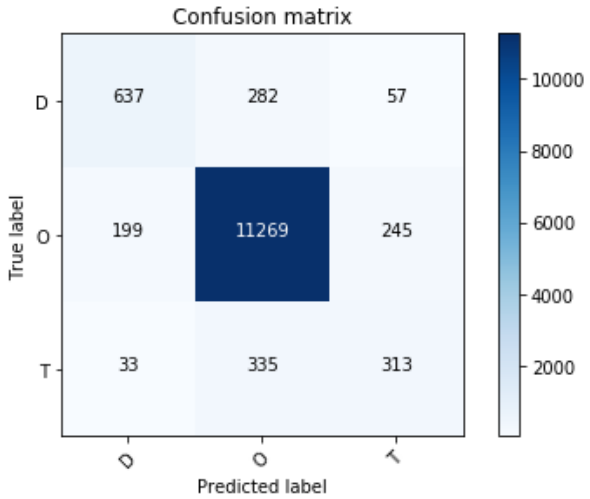
5 Results



Training and Test accuracy graph follows the noraml trend that is should follow and same with the the training and test loss graphs i.e. it should decrease with each iteration.

	precision	recall	f1-score	support
D	0.73	0.65	0.69	976
O	0.95	0.96	0.96	11713
T	0.51	0.46	0.48	681
avg / total	0.91	0.91	0.91	13370

Classification report is on 20% data that we have kept as testing.



Above figure show the confusion matrix of different labels that we have obtain on 20% of the whole data i.e. test data.

6 Github Link

https://github.com/shibu38/Nlu_Assignment_3