Assignment-3: Name Entity Recognition using Sequence Model

Sqn Ldr Abhimanyu Vatta
M.Tech. First Year, CSA, IISc
SR No: 15251
abhimanyuv@iisc.ac.in, v.manyu@gmail.com

1 Problem Statement

The third and final assignemnt in Natural Language Understanding course deals with implementing a Name-Entity Recognition Model for a system (dataset) of diseases and treatments using Sequence Model. A set of tokenized sentences and there corresponding labels serve as the input and output for the task to be carried out. Dataset as described has been provided along with problem statement and the division of the data is as given below:

- Training : 80%
- Held-out : 10%
- Testing : 10%

2 Model Description

Development of NER using Sequence Model is based on paper written by Xuezhe Ma, Eduard Hovy(**?**), wherein, NER task makes use of word embeddings from GloVe and Character embeddings extracted from training dataset.

2.1 Pre-processing of Data

As compared to previous assignments, preprocessing required is minimal and the following tasks were carried out -

- Token(label) extraction
- Making use of Pre-trained Glove model embeddings downloaded from Stanford
- **Sequence Length** - Maximum length sentence is detected for padding all other sentences to make them of equal length before feeding to neural network (Bi-LSTM in this case). The reason for doing so is just intuitive and may not have any theoritical corelation.

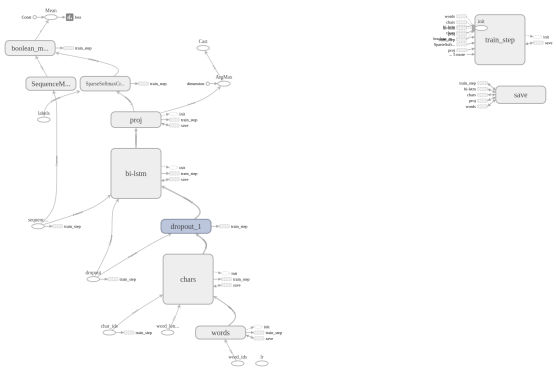

Figure 1: NER Word Bi-LSTM Tensorflow Model

2.2 Bi-LSTM Model Generation

Chris Olah Blog (Olah, 2015) and Goldberg Book(Goldberg, S, 2017) were referred for understanding of the LSTM model, and Tensorflow was used to develop a Bi-LSTM Sequence model for Name-Entity Recognition configured with CPU Id.

2.2.1 LSTM Cell Type

Cell Type of contrib.rnn.LSTMCell is used in this model.

2.2.2 Hidden Layers

Two Hidden layers one each for Character level and Word level are used.

2.2.3 Dropouts

Dropout is used to drop the weights randomly while entering into new Epoch. It helps gaining better training accuracy.

2.2.4 Gradient Optimizer

Adagrad Optimizer is used in this model. Adam Optimizer was also tested but Adagrad provided with better results.

2.3 Learning Rate Decay

It helps to avoid missing minima as the learning progress through iterations and epochs, resulting in model learning progressively slow down by a fixed factor of 0.95 in this implementation.

2.4 Parameters

Following parameters were finalised and used to get results-

- Droupout - 0.5
- Batch Size - 20
- Hidden Layer Size (Word) - 128
- Learning Rate - 0.001
- Learning Rate Decay - 0.95

3 Result

Accuracy (based on sklearn python package) and **multi-class weighted F1** score are used to evaluate the test data. High accuracy may be attributed to observation that data is unambiguous and fixed results as per disease are well tagged. Best Score with different parameters as given in sub-section above, the final results on test data are as follows -

- Accuracy - 94.69
- F1 Score - 94.81

4 Usage

Two python files $Main.py$ and $Model.py$ along with training, testing and held out data are uploaded in clserv at home1/e1-246-30/Assignment_3. The parameters have been kept fixed but can be changed in Main.py file. The usage of this Sequence Model interface is **python3 Main.py**. The program runs through all iterations to train itself in approximately 10 min (30 epochs, 150 steps each) and then given the final result.

References

Guillaume Genthial. 2017. Sequence tagging with tensorflow. https://guillaumegenthial.github.io/sequence-tagging-with-tensorflow.html.

Goldberg, S. 2017. *Neural Network Methods for Natural Language Processing*. Morgan Claypool Publishers.

Chris Olah. 2015. Understanding lstm networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Tobias Sterbak. 2017. Guide to sequence tagging with neural networks in python. https://www.depends-on-the-definition.com/guide-sequence-tagging-neural-networks-python/

Eduard Hovy Xuezhe Ma. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. https://arxiv.org/abs/1603.01354v5.