Name: JAYANTH BALACHANDRA

POS Tagging Using Bi-LSTM CRF Network

Architecture:

The implemented model consists of 4 layers in the following order.

1. Embedding Layer
2. Bi-LSTM Layer
3. Fully Connected Layer
4. Linear Chain CRF Layer

Embedding Layer: The sparse input matrix containing token ids is first passed into embedding layer. Here, a 50-Dimensional embedding vector is randomly initialized for each token. This layer substitutes each token with its vector representation. The embedding vectors are also trained similar to other parameters using back propagation.

Bi-Directional LSTM Layer: The output of embedding layer is fed into a Bi-Directional RNN Layer, which are good at capturing dependencies based on past and future inputs. LSTM were the used as the choice for RNN's since they deal with vanishing / exploding gradient problem and efficiently capture dependencies over very long input sentences. Number of LSTM units used are 28 and a dropout of 0.5 is added to the forget gate of each LSTM. The outputs of forward and backward LSTM's are concatenated depth wise.

Fully Connected Layer: The Bi-Directional LSTM outputs internal representation which varies based on number of LSTM units. To make it suitable for our sequence labelling task, the LSTM output is passed to a linear fully connected layer which outputs a [batch_size, sequence_length, num_tags] shaped tensor. L1 regularization value of 0.01 is applied to weights for generalization and biases are not used.

Linear Chain CRF Layer: The LSTM predicts each tag only considering the current word, independent of neighboring tag predictions. This might lead to improbable Part of Speech tag sequences if there are dependencies between tags. To overcome this limitation, a Linear Chain Conditional Random Field is used in the final layer. CRF considers both transition probabilities between tags and the word-tag decisions by the Bi-LSTM layer for making its predictions. The transition probabilities are initialized randomly and are updated through backpropagation. The most probable tag sequence is found out using the Viterbi algorithm. More details about the CRF can be found in the reference paper linked below.

Hyper Parameters:

The following hyper parameter values were chosen based on execution environment (CPU), maximum time limit and the data.
Learning Rate: 0.025
Optimizer: Adam
Loss Function: CRF Loss
Batch Size: 55
Total Epochs: 2
Maximum Train Time: 11 minutes

Reference Papers:

1. Improving Neural Sequence Labelling using Additional Linguistic Information