CS571 Nature Language Processing
Homework 3 Report: Word Segmentation

Zhexiong Liu

1 Problem Description

This project aims to implement a named entity problem based on deep neural networks.

2 Modeling

2.1 Data Preprocessing

'fasttext-50-180614.bin' is used to embed each words in the sentences into 50-dimensional vectors. The character-level embedding are also implemented by using a CNN models. Furthermore, casing information is considered in the data preprocessing stage. For the convenience, the labels in training, developing, and test dataset are encoded into numbers, which increases the model efficiency.

2.2 Model

As can be seen in Figure 1, CNN is used to extract character-level representation of a given word. Then the character-level representation vector is concatenated with the word embedding vector to feed into the BLSTM network. Finally, the output vectors of BLSTM are fed to the CRF layer to jointly decode the best label sequence. Furthermore, dropout layers are applied on both the input and output vectors of BLSTM.

3 Experiments

In the experiment part, two models are implemented in Keras framework. As for the training, the epoch iteration is set to 80, and the trained embedding, the character-level embedding is padded to 52 dimension, the drop out rate is 0.5, and index parameters is stored in a pickle file, which will be used in evaluation process. As for the results, the model archived 92% F-1 score on the developing dataset.

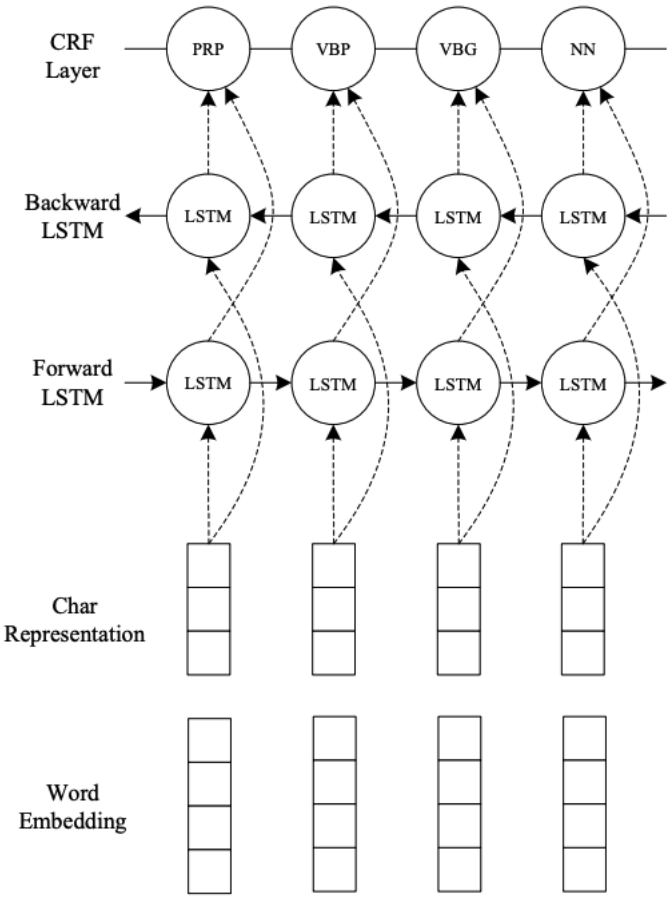

Figure 1: Model

4 Problems

Since the batch function reordered the input labels and return a predicted label sequence that has different orders from the input labels, we need to reorder the input label to match the corresponding sequence of the predicted labels in evaluation process. But, this modification will not effect the performance of the model.