Home Exercise 08

Viet-Hoa Nguyen
Martikelnummer: 2581721

20-00-0947 Deep Learning für Natural Language Processing
Technische Universität Darmstadt

June 16, 2020

1 Mandatory Paper

- Due to the non-deterministic nature of the process of training neural network, the performance of a system between different single runs can be greatly varied. Evaluating a system with score distribution reduces the risk of rejecting good approaches.

2 Sequence Tagging with RNN

2.1 Download Pretrained Embeddings

2.2 LSTM and Bi-LSTM Model

See code implementations

2.3 Intermediate Result

- F1 score on the test set trained by LSTM: 0.38539
- F1 score on the test set trained by BiLSTM: 0.51267

2.4 F1 Model Checkpointer

- F1 score is a better metric when there are imbalanced classes as in our case. The nouns phrases are predominantly in the datasets. Accuracy for unlabeled dataset takes into account the correct classification for the most populous class members than others minor classes.
- See code implementation. The macro F1-score on the test set does not change in comparison from the results in 2.3.

2.5 Hyperparameter Optimization

Three best hyperparameter sets trained with BiLSTM in increasing order:

- F1 score on test set: **0.54217**
  dropout 1: 0,2
  dropout 2: 0.5
  hidden units: 100
  batch size = 10

- F1 score on test set: **0.60626**
  dropout 1: 0,1
  dropout 2: 0.3
  hidden units: 150
  batch size = 10

- F1 score on test set: **0.6125**
  dropout 1: 0,1
  dropout 2: 0.3
  hidden units: 200
  batch size = 50

2.6 Error Analysis

By observing the tagged dataset and calculating the accuracy of the tags by group, it is easy to recognized that the following patterns. The trivial tags such as ")", "(", "w-questions" or "TO" tags tend to get the highest accuracy ¿85%. The tags with the lowest accuracy ¡ 50% are "JJ", "RB", "VGB", "RP". A possible explaination for this can be as follows: the tags with higher scores are filler word such as "a", "the", "what", "where" that exists very often in the training dataset, while tags with lower scores occur not as often.