CMU CS 11-737 - Multilingual NLP: Assignment 1

Uygar Kurt
kurt.uygar@icloud.com

Abstract

This is the report for 1st assignment of CMU CS 11-737 - Multilingual NLP [1]. We well be working with part of speech tagging for multiple languages, and investigating the challenges when the available labeled data is scarce.

1 Baseline Multilingual Results

Results of POS sequence tagging using the base model on 8 different languages: English(en), Czech(cs), Spanish(es), Arabic(ar), Afrikaans(af), Lithuanian(lt), Armenian(ar) and Tamil(ta). Configuration used for these experiments is as follows: embedding_dim : 100, hidden_dim : 128, n_layers:2, bidirectional : true, dropout : 0.25, batch_size : 128,max_epoch" : 10, min_freq : 2

2 Analysis Cases

2.1 Effects of Training Size and Language Family on Performance

From the Table 2 it can be observed that the decisive factor on the results is the data-set size. Data-sets with high number of training examples such as Arabic, Czech and English performs much better compared to data-sets with few number of training examples such as Tamil and Armenian.

Data-set size is not the only factor. Even though Czech has the highest number of training examples its results are not the highest. All languages used for the experiment are from different language families except English and Afrikaans which belong to Germanic language family. It can be observed that English data-set size is ten times of Afrikaans. Yet their accuracy is different only by small decimals.

2.2 Effects of Hyperparametes on Performance

See Table 2 for an example of a table and its caption.

Due to limited access to computation resources only a subset of hyper-parameters have been tuned. With the increase minimum frequency we observe that scores have decreased. Reverse would be expected actually because with increased minimum frequency the rare words supposed to be eliminated. We observe that larger batch size had a small improvement. Minimum number of epochs didn't have any effect which was expected.

3 Additions and Possible Improvements

Scripts.ipynb file contains an additional easy to use inference function. It loads one of the pre-trained models and loads vocabulary. Models can be enhanced such as adding CNN or CRF layer. These options will be tried later on.

[1] http://phontron.com/class/multiling2022/index.html

English(en)	**Czech(cz)**
Unique tokens in TEXT vocabulary: 9863	Unique tokens in TEXT vocabulary: 41251
Unique tokens in UD$_TAGvocabulary$: 19	Unique tokens in UD$_TAGvocabulary$: 19
Number of training examples: 12543	Number of training examples: 41559
Number of validation examples: 2002	Number of validation examples: 9270
Number of testing examples: 2077	Number of testing examples: 10148
Epoch: 10 | Epoch Time: 1m 49s	Epoch: 10 | Epoch Time: 5m 24s
Train Loss: 0.122 | Train Acc: 96.13%	Train Loss: 0.042 | Train Acc: 98.61%
Val. Loss: 0.291 | Val. Acc: 91.35%	Val. Loss: 0.194 | Val. Acc: 94.36%
Test Loss: 0.266 | Test Acc: 91.58%	Test Loss: 0.186 | Test Acc: 94.05%

Spanish(es)	**Arabic(ar)**
Unique tokens in TEXT vocabulary: 18727	Unique tokens in TEXT vocabulary: 15889
Unique tokens in UD$_TAGvocabulary$: 18	Unique tokens in UD$_TAGvocabulary$: 18
Number of training examples: 14187	Number of training examples: 6174
Number of validation examples: 1552	Number of validation examples: 786
Number of testing examples: 274	Number of testing examples: 704
Epoch: 10 | Epoch Time: 2m 40s	Epoch: 10 | Epoch Time: 2m 6s
Train Loss: 0.109 | Train Acc: 96.50%	Train Loss: 0.069 | Train Acc: 97.72%
Val. Loss: 0.180 | Val. Acc: 94.35%	Val. Loss: 0.174 | Val. Acc: 94.13%
Test Loss: 0.278 | Test Acc: 93.30%	Test Loss: 0.168 | Test Acc: 94.24%

Afrikaans(af)	**Lithuanian(lt)**
Unique tokens in TEXT vocabulary: 2368	Unique tokens in TEXT vocabulary: 4547
Unique tokens in UD$_TAGvocabulary$: 18	Unique tokens in UD$_TAGvocabulary$: 19
Number of training examples: 1315	Number of training examples: 2341
Number of validation examples: 194	Number of validation examples: 617
Number of testing examples: 425	Number of testing examples: 684
Epoch: 10 | Epoch Time: 0m 13s	Epoch: 10 | Epoch Time: 0m 30s
Train Loss: 0.339 | Train Acc: 90.73%	Train Loss: 0.400 | Train Acc: 81.17%
Val. Loss: 0.402 | Val. Acc: 88.47%	Val. Loss: 0.608 | Val. Acc: 79.33%
Test Loss: 0.388 | Test Acc: 88.53%	Test Loss: 0.721 | Test Acc: 75.65%

Armenian(hy)	**Tamil(ta)**
Unique tokens in TEXT vocabulary: 3897	Unique tokens in TEXT vocabulary: 926
Unique tokens in UD$_TAGvocabulary$: 19	Unique tokens in UD$_TAGvocabulary$: 15
Number of training examples: 1975	Number of training examples: 400
Number of validation examples: 249	Number of validation examples: 80
Number of testing examples: 278	Number of testing examples: 120
Epoch: 10 | Epoch Time: 0m 23s	Epoch: 10 | Epoch Time: 0m 2s
Train Loss: 0.481 | Train Acc: 84.99%	Train Loss: 1.830 | Train Acc: 39.90%
Val. Loss: 0.518 | Val. Acc: 83.69%	Val. Loss: 1.770 | Val. Acc: 41.73%
Test Loss: 0.620 | Test Acc: 80.02%	Test Loss: 1.862 | Test Acc: 39.84%

Table 1: Baseline model experiment results.

Hyper Parameter	Test Score
Min Freq. 2	91.58%
Min Freq. 8	89.30%
Batch Size 128	91.58%
Batch Size 16	91.50%
Epochs 10	91.58
Epochs 15	91.58

Table 2: Examples of different hyper-parameters experimented on the base model.