Natural Language Processing

Assignment 1 Report

Anna Fabris
anna.fabris@studio.unibo.it

Miki Mizutani
miki.mizutani@studio.unibo.it

Leonidas Gee
leonidas.gee@studio.unibo.it

Academic Year: 2021 / 2022

Abstract

A part-of-speech tagging task was solved using pre-trained GloVe embeddings alongside Recurrent Neural Networks (RNNs). Four different RNN architectures were optimised using random search and the two best models were evaluated using the F1-macro metric on the test set. Final results show scores of 84.2% and 84.0% for the first (bidirectional LSTM with a dense layer) and second (bidirectional GRU with a dense layer) best-performing models respectively.

Description

Part-of-Speech (POS) tagging can be viewed as a multiclass classification task. One approach to solving this is to use pre-trained GloVe embeddings as inputs to RNNs. For this project, the Penn Treebank dataset is used, which consists of documents and the corresponding POS tag for each token.

To begin, each document is divided into sentences along with their corresponding list of POS tags. Each sentence is tokenised and lower cased. The transformed dataset is subsequently divided for training, validation, and testing.

Next, GloVe embeddings are downloaded and used to build an embedding dictionary that will convert words into vectors. To account for Out-of-Vocabulary (OOV) words, a random embedding is assigned to words not found in the embedding dictionary. The POS tags are also encoded using integer values from 1 to the number of unique POS tags. The value 0 is reserved for the padding tags. The embeddings and labels are then padded to a maximum length of 250 as RNNs require inputs of fixed length. Finally, the class labels are one-hot encoded to match the output of the model.

After the data preparation phase, 4 RNN architectures are prepared for training and validation. Random search is used to optimise the hyperparameters of the RNN architectures. A custom accuracy metric that ignores padding is employed and maximised by the random search during training to obtain the optimised model for all the architectures. Finally, the two best model architectures are selected for final evaluation.

In the final evaluation phase, an F1-macro score is used to evaluate the performance of each model on the test set. The punctuations are removed from both the true labels and the predictions before evaluation.

Models

A baseline model is established with 3 additional variations of the given architecture. The baseline model is a bidirectional LSTM with a dense layer. The 3 variations are the following: bidirectional GRU with a dense layer, bidirectional LSTM with an additional LSTM layer and a dense layer, and a bidirectional LSTM with two dense layers. The best hyperparameters were chosen with random search with trials of 3 and epochs of 2. The best model of each architecture is then trained from scratch using 10 epochs.

Tables of Findings

Bidirectional LSTM with a dense layer	Bidirectional GRU with a dense layer	Bidirectional LSTM with an additional LSTM layer	Bidirectional LSTM with an additional dense layer
0.9270	0.9239	0.8551	0.9029

Table 1: Validation accuracy of models excluding padding after training.

The models with the highest validation accuracy were the bidirectional LSTM and the bidirectional GRU. WIth the two best-performing models, the F1-macro scores for both models were then calculated.

Error Analysis

The confusion matrix was plotted and the F1-macro score was calculated.

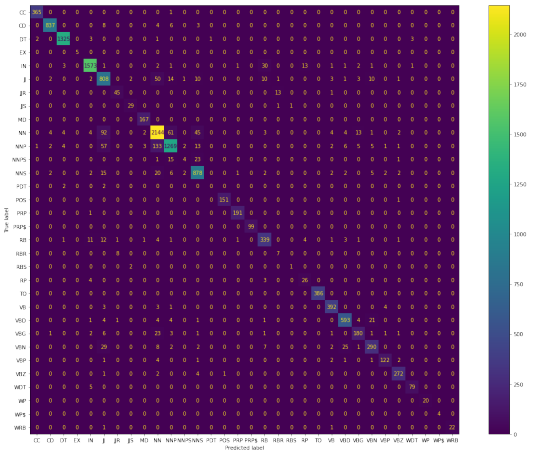

Figure 1: Model 1

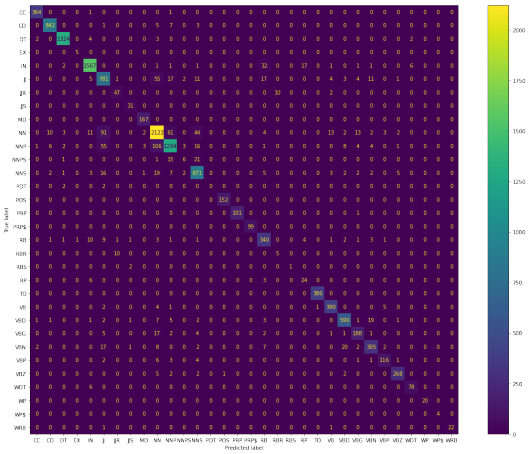

Figure 2: Model 2

Model 1

The Model 1 is the Bidirectional LSTM with a dense layer model.

Most errors of this model are from the *NN* and *NNP* labels which were incorrectly switched or predicted as the *JJ* label. The confusion between *NN* and *NNP* likely comes from having lower-cased all the words. The confusion with *JJ* likely comes from when a noun is used to describe another noun (ex: history teacher).

The model has low performance with less common classes like *NNPS*, *PDT* and *RBR*. This is probably caused by the lack of enough examples in the training data. If necessary, many approaches could be tried to solve this problem: using more data, resampling the training data, merging the minority classes with other classes (ex: *NNPS* could be merged with *NNP*).
F1-macro score = 0.8422.

Model 2

The Model 2 is the Bidirectional GRU with a dense layer.

Most of the errors of model 2 are very similar to the ones of model 1.
F1-macro score = 0.8396.

Conclusion

Through this project, we were able to prove the effectiveness of GloVe embeddings with RNNs, and in particular, LSTMs with dense layers in POS tagging. The models were able to accurately classify tokens into POS tags with a high degree of accuracy.