POS Tagging

Natural Language Processing Assignment

Francesco Ballerini
francesco.ballerini3@studio.unibo.it

Emmanuele Bollino
emmanuele.bollino@studio.unibo.it

Tommaso Giannuli
tommaso.giannuli@studio.unibo.it

Manuel Mariani
manuel.mariani2@studio.unibo.it

AY 2021–2022

Abstract

We tackle part-of-speech tagging as a sequence labeling task performed by recurrent neural architectures. The models we test—some variations on a simple neural baseline—are all trained on a freely available sample of the Penn Treebank corpus and use GloVe-50 pre-trained embeddings. Our best model reaches an F1-macro score on the test set of 0.8.

1 Task

A *part of speech* (POS) is a category of words or lexical items that have similar grammatical properties and display analogous syntactic behaviour—nouns, verbs, adjectives, adverbs, etc. *Part-of-speech tagging* is the process of assigning to each word or symbol in a text the corresponding part of speech.

We modeled this task by feeding single sentences to a recurrent neural architecture trained to produce in output the corresponding list of POS tags. The dataset on which our models were trained and tested is the 10% sample of the 45 Penn Treebank corpus provided by the NLTK library.

2 Models

We tested the following recurrent neural architectures:

- `BiLSTM-Dense`: our baseline model, consisting of a bidirectional LSTM followed by a dense layer with softmax activation.
- `BiGRU-Dense`: bidirectional GRU followed by a dense layer with softmax activation.
- `2xBiLSTM-Dense`: a stack of two bidirectional LSTMs with a final Dense layer with softmax activation on top.
- `BiLSTM-2xDense`: bidirectional LSTM followed by two dense layers, the first with ReLU and the second with softmax activation. The first dense layer has twice as many neurons as the second one.

Every input, before being fed to a model, goes through a non-trainable embedding layer storing GloVe-50 word embeddings. Out-of-vocabulary tokens are encoded by averaging the embeddings of all their left and right neighors throughout the dataset split they belong to.

All models share the same hyperparameters: 32 memory units for LSTM/GRU layers and 46 units (45 tags + 1 for padding) for the final Dense layers.

Each model was trained with categorical cross-entropy as loss function and the Adam optimizer. Interactive plots of the training history can be found on WandB.

3 Results

Being some parts of speech intrinsically more common than others, our dataset naturally suffers from class imbalance. In an attempt to counteract this effect, we trained our models with both "regular" and weighted cross-entropy as loss function, with each weight being inversely proportional to the frequency of its associated tag. However, as shown in Table 1, taking class imbalance into account did not provide any significant improvement in the F1-macro on the validation set; as a matter of fact, the best-scoring model was trained with no weights in the loss.

When looking at the most misclassified tags in Table 2, some of them are, unsurprisingly, associated to uncommon parts of speech, such as `FW` (foreign word), `LS` (list item marker), `UH` (interjection), and `#` (pound sign). Upon further inspection, we discovered that those tags do not appear in the test set at all, which might also explain why F1 scores on the test set are higher than those on the validation set in Table 1.

More interesting classification mistakes are those involving plural proper nouns (`NNPS`), which are often misclassified as plural common nouns (`NNS`). This is probably due to the fact that GloVe does not contain embeddings for capitalized words, and we therefore had to convert the whole dataset to lowercase before feeding it to the embedding layer; by doing so, however, we lost some information that might have been useful to discriminate between common and proper nouns—especially plural proper nouns, which are rarer than their singular counterparts and therefore harder to learn.

Model	F1-macro (validation)
`BiLSTM-Dense`	0.74
`2xBiLSTM-Dense`	0.73
`BiGRU-Dense`	0.71
`BiLSTM-2xDense`	0.71

(a) Unweighted loss and accuracy

Model	F1-macro (validation)
`BiGRU-Dense`	0.73
`BiLSTM-2xDense`	0.73
`BiLSTM-Dense`	0.69
`2xBiLSTM-Dense`	0.67

(b) Weighted loss and accuracy

Model	F1-macro (test)
`BiLSTM-Dense`	0.80
`2xBiLSTM-Dense`	0.76

(c) Selected models (unweighted)

Table 1: F1-macro scores on validation and test set. Each model was trained for 500 epochs and its weights saved through a callback when maximum validation accuracy was recorded. We then computed the F1-macro score on the test set for the best performing model on the validation set (`BiLSTM-Dense` trained with unweighted loss) and the one immediately below it within the same category.

Tag	Most pred. as
`FW`	`NNP`
`LS`	`CD`
`NNPS`	`NNS`
`PDT`	`DT`
`UH`	`DT`
`#`	`CD`

(a) `BiLSTM-Dense` (validation)

Tag	Most pred. as
`FW`	`NNP`
`LS`	`NN`
`NNPS`	`NNS`
`PDT`	`DT`
`UH`	`DT`
`#`	`#`

(b) `2xBiLSTM-Dense` (validation)

Tag	Most pred. as
`NNPS`	`NNS`
`PDT`	`JJ`
`RBS`	`JJS`

(c) `BiLSTM-Dense` (test)

Tag	Most pred. as
`NNPS`	`NNS`
`PDT`	`JJ`
`RBR`	`JJR`
`RBS`	`JJS`
`WP$`	`JJ`

(d) `2xBiLSTM-Dense` (test)

Table 2: Tags which get correctly predicted in less than 50% of their occurrences and corresponding most frequent labeling mistake.