NATURAL LANGUAGE PROCESSING - ASSIGNMENT #1

Lorenzo Pratesi, Martina Rossini, Riccardo Foschi, Vairo Di Pasquale
{lorenzo.pratesi2, martina.rossini3, riccardo.foschi4, vairo.dipasquale}@studio.unibo.it
Artificial Intelligence Master's Degree

Alma Mater Studiorum

ABSTRACT

In this work, after analyzing and preparing the given corpus, we developed a series of neural architectures according to the assignment specification. After that, we selected the two best models based on the validation F1 score and evaluated them on the test set. Finally, after tuning some hyperparameters and adding regularization to the baseline model, we evaluated it on the test set and we obtained an accuracy of 92.89% and an F1 score of 85.8%.

1 Introduction

Part-Of-Speech (POS) tagging is the task of labeling each word in a sentence using the appropriate word class. Having this information is a useful feature in many NLP applications, ranging from information extraction to NER and co-reference resolution. The approach we take when solving this problem is based on neural architectures, in particular on Recurrent Neural Networks. Indeed, none of the models we explored use more complex architectures based on attention and transformers. Section 2 shows how we pre-process the initial corpus (a dependency-parsed version of a sample of the Penn Treebank, which is available here) and how we create our embedding matrix starting from the pre-trained GloVe embeddings. Section 3 describes in more detail the models we tried and how we trained them, while in Section 4 we show our results and briefly analyze them.

2 Data Preparation

We first split each document in subsentences, in order to reduce the length of each sample in the datasets. Later, we realized that many tokens were OOV words only because they contained particular symbols, so we chose to apply a light pre-processing for each of them. Next, we created a custom vocabulary containing all the tokens associated with the GloVe embedding of dimension 50 and all the other tokens we found in the train set. For the creation of the embedding matrix, we have concatenated the index of each word that belongs to the GloVe's vocabulary to the corresponding pre-trained embedding. Instead, for the other words, we decided to proceed like this:

- If the word is a compound word, where each sub-word is separated by the "-" character we check if their associated embeddings exist. If so, we use the their mean as a new embedding, otherwise we simply use a random embedding for each non existing word for the final computation.
- If the word is not a compound word, we use a random vector as new embedding.

We also update vocabulary and embedding matrix for the other dataset splits in a separated manner. Next, we transformed each sentence in sequences of identifiers, with respect to the vocabulary. Finally, we zero-padded each sequence with a fixed maximum length (each dataset has its own), in order to be able to create batches during training.

3 Models' Description

As required by the assignment, we defined a baseline model with a Bidirectional LSTM before a Fully Connected layer as classificator. For the embedding implementation, we used the Keras Embedding Layer. We set the LSTM layer's hidden state dimension to 100, while the FC layer - with a softmax activation - has an output dimension equal to the number of POS-tags. Each LSTM layer outputs an hidden state for every embedding received as input and, before passing them to the classifier, they get concatenated. We also defined the three extensions required:

1. Additional BiLSTM: same hyperparameters as the other BiLSTM layer; gets as input the output of the previous BiLSTM.
2. BiGRU instead of BiLSTM: we used the same hidden state dimension as before.
3. Additional FC layer before the baseline's FC layer: output dimension of 100 with a ReLu activation function.

We trained each type of model with a categorical cross-entropy loss and using the Adam optimizer with the default starting learning rate. We defined a batch size of 8 and a maximum number of epochs of 100. We also implemented a custom early stopping criterion based on the F1 score on the validation set, without considering batched data.

Finally, as extra work, we tried to optimize the baseline model by changing some hyperparameters like the embeddings' dimension and the use of class weights. We also added some Dropout layers as a form of regularization. The results of these choices are shown below.

4 Result / Error Analysis

The results we obtained for validation accuracy and F1 score on the four originally required models are shown in Table 1. As we can see, all models achieve similar performances; however, models "Baseline" and "2x FC" obtain higher values for the macro-F1 score, which was our target metric. Thus, we also evaluated these two models on the test set and we show the results in Table 2. Finally, notice that we also tried tuning some hyperparameters and employing some forms of regularization, in order to improve performance of our overall best model (i.e.: "Baseline"), resulting in an F1 score of 85.8% on the test set.
As we can clearly see from the confusion matrix displayed in Figure 1, classes with indexes 35, 39, 41 and 42 (respectively NNPS, PDT, FW and UH) are where the majority of our errors are located. This is not surprising, as in our small training corpus classes are quite unevenly distributed and the aforementioned labels appear with a very low frequency.
Moreover, we can see that class NNPS is often missclassified as being of index 4 (NNS) or 1 (NNP), which could be explained by several factors: first, note that proper names, both plural and singular, were probably not part of the pre-trained GloVe model, thus their embeddings were randomly generated drawing from a uniform distribution. It is then reasonable to think that many of the embeddings for the words in these two classes would be similar. Moreover, from a morphological point of view, it is plausible that some proper nouns in plural are mistaken for simple plural nouns, especially when - as in our case - we transformed all the text to lowercase. Similarly, it is also morphologically reasonable that the pre-determiners (class PDT), which are very much a minority class, are mistaken for determiners and adjectives (respectively, with index 8 and 5). Other noticeable errors have to do with foreign words and interjections, which again are minority classes and most probably not included in the GloVe vocabulary.

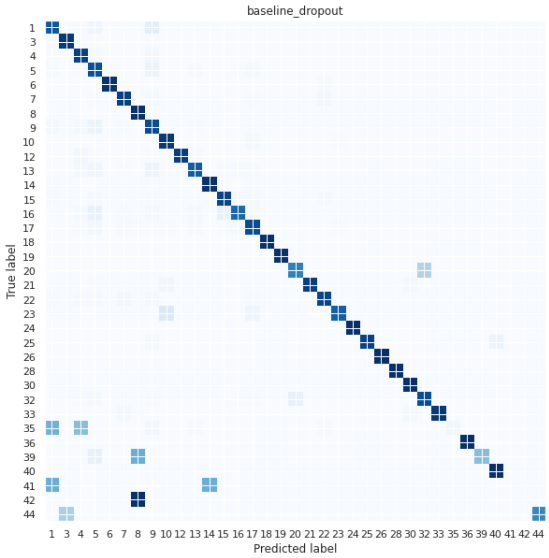

Figure 1: Final Model's Confusion Matrix

	Baseline	2x BiLSTM	BiGRU	2x FC
val acc	87.39%	88.17%	87.32%	87.77%
val F1	76.44%	74.86%	74.23%	75.02%

Table 1: Models evaluation on the validation set

	Baseline	2x FC	Final Model
test F1	79.7%	77.7%	85.8%

Table 2: Best models evaluation on the test set

5 Future work

Our work could be further improved by considering more advanced neural architectures that include an attention mechanism, like transoformers. We could also try to obtain more training data to reduce the problems related to overfitting and to class imbalance. Finally, a further point of possible exploration regards the way we deal with Out-Of-Vocabulary words, for which we could try to employ some more advanced strategies based on string similarity or which use subword information (i.e.: similarly to what is done in fastText).