Assignment 1

Alessandro Folloni, Gabriele Fossi, Daniele Napolitano and **Marco Solimé**
Master's Degree in Artificial Intelligence, University of Bologna
{ alessandro.folloni2, gabriele.fossi, daniele.napolitano4, marco.solime }@studio.unibo.it

Abstract

This paper addresses the POS Tagging problem using recurrent neural networks. The methodology involves preprocessing steps such as tokenization, encoding of words and labels, and embedding creation using GloVe pre-trained word vectors. Three models are constructed and evaluated: a baseline bidirectional LSTM, a bidirectional LSTM extended model, and a model incorporating a dense layer. The analysis reveals that there is not a substantial difference between the performance of three models, except for the effectiveness of the last aforementioned model to correctly identify some of the tags which have a very low frequency in the training data.

1 Introduction

This work deals with the problem of POS Tagging, the process of assigning the part of speech tag to words. A possible method to tackle this kind of problem is to resort to Rule Based Approaches: together with context information, hand written rules are applied in order to assign tags to words (Kumawat and Jain, 2015). However, this technique has some limitations. Rules have to be constructed manually and there is the need of an expert in the language that is being tagged. Our approach overcomes this limitations and consists in trying different Recurrent Neural Networks (RNN) that in the end achieve a better tagging (Chiche and Yitagesu, 2022). In this paper we compare three different neural networks: (1) the Baseline model which consists in a Bidirectional LSTM with a Dense layer on top, (2) the Baseline model with the addition of a LSTM layer and (3) the Baseline model with the addition of a Dense layer. All the models include pre-initialized GloVe embedding layers, and are evaluated using macro-F1 score, precision and recall. The results show a similar performance of the three models, with the Baseline model with the Dense layer achieving a slightly higher score. Indeed, this model achieves a higher F1 score on some tags which have a low frequency in the training data. Moreover, the frequency of a tag highly affects the score of the corresponding tag: tags with a low frequency have very low scores.

2 System Description

Our investigation involves three distinct models aimed at analyzing various behaviors and performances within simple architectures:

- **Baseline**: This comprises an Embedding layer, a bidirectional LSTM layer, and a Dense layer. The Embedding layer is configured with essential parameters to accurately accommodate the vocabulary size, embedding vector size, and embedding matrix. We determined a maximum sequence length of 300 after analyzing the sequences to encompass the majority effectively.

- **Baseline + LSTM**: This model extends the Baseline by adding an extra bidirectional LSTM layer.

- **Baseline + Dense**: Here, an additional Dense layer is incorporated into the Baseline model.

To evaluate these models, we employed three distinct metrics—F1 score, precision, and recall. For streamlined monitoring of their performance during training, we designed a dedicated class.

Our workflow follows a simple yet effective pipeline: model definition, training, weight preservation, and comprehensive result analysis.

Model	Validation set	Test set
Baseline	0.587	0.572
Baseline + LSTM layer	0.563	0.559
Baseline + Dense layer	0.617	0.601

Table 1: Macro-F1 scores of each model.

3 Experimental setup and results

To ensure robust tokenization of our dataset, we harnessed the GloVe vocabulary as a repository of pre-existing knowledge. This strategic approach resulted in our vocabulary consisting of the amalgamation of GloVe tokens and training tokens. This hybrid vocabulary proved instrumental in curtailing out-of-vocabulary (OOV) tokens in the validation and test sets, boasting a mere 0.57% occurrence. This reduction significantly enhanced our models' capacity for generalization.

The Embedding layer served as a pivotal component, initialized with GloVe embedding vectors of length 50. Although we experimented with other dimensions (100, 200), these alterations failed to yield substantial improvements. On the contrary, they substantially inflated the number of parameters. Consequently, the embedding matrix adheres to a predefined structure: row 0 reserved for the PAD token (zero-initialized), row 1 designated for the UNK token (a static embedding), rows 2-400002 populated by GloVe embeddings, and the residual rows allocated for tokens present in the training set but absent in GloVe (randomly initialized). Notably, during our experiments, the trainable parameter was set to True, enabling the model to glean insights from the training data.

Across all models, the output space's dimensionality for each layer remains fixed at 64, barring the final layer, whose dimensionality aligns with the number of possible tags + 1. This ultimate layer employs the softmax activation function. For the training regimen, we adopted the Adam optimizer and leveraged categorical cross-entropy as the preferred loss function, tailored for our multi-class classification problem. Each model underwent rigorous training for 100 epochs, utilizing a batch size set to 32.

For reference, Table 1 elucidates the macro-F1 scores attained by each model on both the validation and test sets.

4 Discussion

The results show that there is not a huge difference between the three models in terms of performance, even though the Baseline model with the addition of a Dense layer has a bit higher F1 score. An interesting result regards the performance of the models on each tag related to its frequency (Figures 1 and 2).

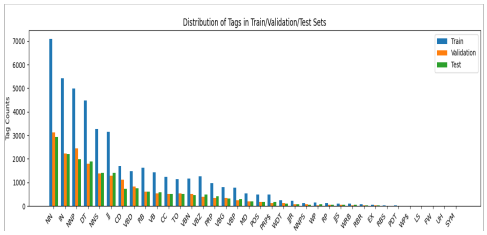

Figure 1: Frequency of the tags in training, validation and test sets.

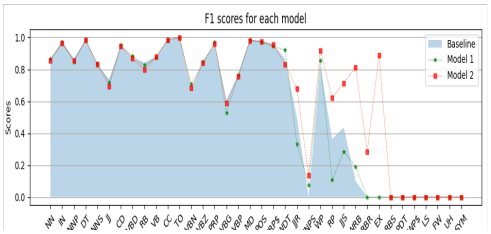

Figure 2: F1 scores of the three models for each tag

It is possible to notice that the three models have almost the same F1-scores corresponding to tags with a high frequency in the training data. When it comes to tags with a low frequency, the F1 scores are very low, equal to 0 in some cases. In those labels, the Baseline model with a LSTM layer (Model 1) performs very similar to the Baseline. However, the Baseline model with a Dense layer (Model 2) has a quite high F1 score in some of these labels, compared to the other two models. This explains why the macro-F1 score of the last model is slightly higher.

5 Conclusion

In this work we compared three different models to solve the POS tagging problem. The results show that overall the three models have a similar performance. However the Baseline model with the Dense layer outperforms the other two in terms of F1 scores on some tags which have a very low frequency in the training data. Indeed, as expected, the frequency of the tags plays a crucial role on the performance of the model for that specific label. Very frequent tags have very high F1 scores, while infrequent tags have low scores, suggesting that the models struggle to learn them. Some of them are even equal to 0.

References

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*.

Deepika Kumawat and Vinesh Jain. 2015. Pos tagging approaches: A comparison. *International Journal of Computer Applications (0975-8887)*.