Assignment 1

Lorenzo Cassano, Jacopo D'Abramo, Kilian Tiziano Le Creurer and **Francesco Pivi**
Master's Degree in Artificial Intelligence, University of Bologna
{ lorenzo.cassano2, jacopo.dabramo, kilian.lecreurer, francesco.pivi }@studio.unibo.it

Abstract

This study focused on the employment of three distinct recurrent neural network variations for solving the Part-of-Speech Tagging task, alongside different dimensions of GloVe embeddings. The aim was to identify the model that demonstrates the highest macro-F1 score, crucial for accurate tagging. Our research highlighted that Recurrent Networks are well-suited for this specific classification task.

1 Introduction

Part-of-Speech tagging employs various approaches: rule-based methods rely on predefined linguistic rules for transparency but may struggle with unknown patterns (Pham, 2020). Statistical models adapt well to language variations by learning from data, yet they require substantial annotated information (Silfverberg and Lindén, 2011). Finally, the deep learning approach offers high accuracy but demands significant computational resources (Chiche and Yitagesu, 2022).
The employed method involves constructing three variations of the Bi-LSTM + Dense Layer, incorporating GloVe embeddings (Tifrea et al., 2018) to tackle the task within the Penn Treebank corpus (Taylor et al., 2003). To further explore potential enhancements, a series of experiments were conducted, evaluating diverse modifications within the architecture as depicted in Figure 1. Each model underwent meticulous comparison within a validation subset, and the most robust model has been assessed on the test set. The evaluation criterion utilized the F1-macro metric, without punctuation. The results emphasized the performance based on macro F1 scores on the validation set. Through rigorous validation, our top-performing model achieved a notable 79% on the validation set and 81% on the test set.

2 System description

The dataset comprises pre-tokenized text paragraphs, consisting of 45 tags with uneven distribution across the training, validation, and test sets. Notably, sentence statistics reveal significant variations in maximum sentence length among these sets.
Preprocessing involved several key steps: converting the text to lowercase, segmenting sentences based on the line space, and padding sentences to match the maximum sentence length. Additionally, GloVe embeddings were generated while masking the padding to prevent its contribution to the loss function. Out-of-vocabulary (OOV) words were addressed by employing a random uniform distribution method, totaling around 350 such instances. The structural layout of the model is visually represented in Figure 1. A random search was performed to find the best hyperparameters tuning for each model in order to identify the optimal configuration.

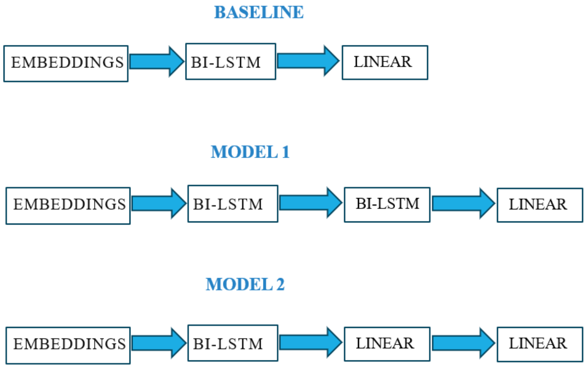

Figure 1: Model architectures: linear is the dense layer

3 Experimental setup and results

The dataset is divided as follows: documents 1-100 constitute the training set, 101-150 are allocated for validation, and 151-199 serve as the test set.
The models are trained using the AdamW optimizer

for 50 epochs. Following this, a random search for hyperparameters tuning was conducted to identify the top configurations for each model (Bergstra and Bengio, 2012).
In particular, the hyperparameters used during the random search are:

Drop.	Hidd. dim	Emb dim	Smooth.	LR	LSTM dim
0.0	32	100	0.0	0.01	32
0.1	64	200	0.05	0.005	64
0.2	128	300	0.1	0.001	128

Table 1: Hyperparameters grid for random search

Various learning rate schedulers were tested, and ultimately, the OneCycle scheduler emerged as the most effective configuration. Table 2 displays the results of the models on validation set and the result of the most robust model on test set. Notably, the aggregation excludes the F1-score of punctuation.

Validation Results			
Model	Precision	Recall	F1
Baseline	0.810	**0.781**	0.777
Model1	**0.842**	0.769	**0.787**
Model2	0.809	0.775	0.783
Test Result			
Model2	0.831	0.814	**0.805**

Table 2: Best results of the model on validation and test set

The optimal model was determined by calculating the mean F1 score across three different seeds. The model with the highest mean was selected to ensure the extraction of the most robust model. Subsequently, from the choices of baseline, model1, and model2, the model exhibiting the best performance on its respective seed was employed for the test set.

4 Discussion

The results underscore Model 2's effectiveness as the most robust model, exhibiting the highest performance across three different seeds on the validation set on average, with a macro-F1 mean of 0.780. It is worth mentioning that, all the approaches selected yield similar results. Additionally, it's notable that the test set consistently outperforms the validation set. This difference could be attributed to:

- The validation set contains twice as many sentences as the test set
- The different distributions of classes across the training, test, and validation data.

From the error analysis, several observations arise:

- Some classes, such as Foreign Words (FW) and Interjections (UH), Symbol (SYM) have almost negligible support, leading to the network's struggle in learning their role due to their low occurences in the training set. However, classes like "Preposition or subordinating conjunction" (IN) are correctly classified, mainly due to their clear and unique contextual cues.
- Inherently ambiguous classes, exemplified by the Proper Noun in Plural Form (NNPS), often get confused with Singular Proper Nouns (NNP) or Plural Nouns (NNS)
- Predeterminers (PTD), which usually precede Determiners (DT), should be easily detected, yet they are frequently predicted as DT or Adjectives (JJ), possibly due to the smaller class size of PDT and contextual similarity.

5 Conclusion

In this task we presented a solution to the POS tagging problem with an RNN architecture and GloVe word embeddings, trying different combinations of layers and hyper-parameters.
Ultimately, we achieved highly favorable outcomes on the validation set using the Model2, and we obtained even higher results on the test set.
Nevertheless, all models continue to grapple with class imbalance and inherent ambiguities, which might be mitigated by: considering larger datasets to rectify these issues, employing more complex models, incorporating rule-based corrections post-prediction to disambiguate certain classes like PDT, and leveraging sub-word encodings to enhance the prediction accuracy of tokens intrinsically tied to their word form.

References

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):1–25.

Bao Pham. 2020. Parts of speech tagging: Rule-based.

Miikka Silfverberg and Krister Lindén. 2011. Combining statistical models for pos tagging using finite-state calculus. In *Proceedings of the 18th Conference of Computational Linguistics NODALIDA 2011*. Northern European Association for Language Technology.

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The penn treebank: an overview. *Treebanks: Building and using parsed corpora*, pages 5–22.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*.