# NLP ML
# LSTM for PoS

Kolmakova Elizaveta
*Innopolis University*
Kazan, Russia
e.kolmakova@innopolis.university

***Index Terms*—LSTM, PoS, long short term memory, part of speech, deep learning.**

## I. Introduction and Motivation

This is a comprehensive report-comparison of different setups of machine learning models: simple LSTM and bidirectional LSTM, different parameters, different amount of available data. The motivation of this work is to investigate whether simply configured LSTM network is capable to get significantly better performance in comparison with other configuration of model in terms of limited data.

The work is constructed from the following sections: section 2 introduces experiment settings, section 3 shows results with limited data and section 4 represents conclusions.

1) Section 1. Introduction and Motivation.
2) Section 2. Experiments settings.
3) Section 3. Results.
4) Section 4. Conclusion.

## II. Experiments settings

The task is to train a neural network to distinguish different part of speech. The given dataset from UD contains sentences/text and tags with its part of speech.

The dataset is downloaded and preprocessed with Torchtext. Then the vocab is build with representation pretrained in-build vector. After that the dataset is split between train, validation and test sets. In the part of experiments with limited data, only train and validation sets are reduced, while the test set is still the same. This would provide with better understanding of performance of the model.

### *A. Comparison*

Technically, LSTM network for this task contains from three main part: LSTM layer with hidden cells and final linear layer.

In order to obtain the best parameters which would provide with the highest performance measured by accuracy, the simple version of GridSearch was applied. GridSearch is simple algorithm of brute force of all parameters. There were 4 parameter's groups to tune: amount of epochs, batch size, amount of layers, configuration of LSTM.

1) Amount of epochs: 2, 5, 10, 20
2) Batch size: 32, 64, 128
3) Layers: 1, 2, 4
4) Configuration of LSTM: simple, bidirectional

The table I represent different configurations for 100% of data to identify initially the scope of parameters which would lead to low performance to omit them:

After exploration of results of GridSearch the following parameters were selected as preferable:

1) amount of layers is better to be moderate - 1 or 2 - due to higher volatility of performance for 4 layers.
2) batch size does not show significant difference, therefore any of batch sizes could be chosen. For the following experiments $batch\_size = 32$ is chosen.
3) for almost every pair of experiments with simple and bidirectional LSTM in comparison, the bidirectional LSTM had a better performance, therefore this setup is preferable to use.
4) for most of cases 5 epochs were enough to get the average performance, while the difference in performance between 10 and 20 epochs was insignificant, therefore the 5 and 10 epochs are preferable.

For the following exploration of limited data there is a necessity to identify 2 configurations: weak model which nevertheless had successful performance for 100% of data and strong model . We assume the first configuration will not manage to cover the difference between parts of speech for severely limited data (10%, 20%), therefore we need 2 different configuration to compare the performance. But due to good pretrained vector for data representation, the model can handle the weak model.

1) Weak configuration: 1 layer, 5 epochs, 32 batch size, simple LSTM
2) Strong cofiguration: 2 layer, 10 epochs, 32 batch size, bidirectional LSTM

## III. Results

The limited data was presented as 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of initial dataset. The table II represent result of training for two configurations: weak and strong. For a visual representation we will assume accuracy below 75% is low (red color), between 75 and 85 % is satisfactory (yellow color) and more 85% is high (green color).

**Exploration of outlines**

To consider tokens which were identified incorrectly, there are a couple of sentences from 10% limited data and strong heightconfiguration of model:

| Val accuracy | Test accuracy, % | Batch size | Layers | Is bidirectional | Number of epochs |
|---|---|---|---|---|---|
| 81.44 | 80.24 | 32 | 1 | False | 2 |
| 82.22 | 81.73 | 32 | 1 | False | 5 |
| 82.97 | 82.13 | 32 | 1 | False | 10 |
| 82.72 | 82,13 | 32 | 1 | False | 20 |
| 83.95 | 83.16 | 32 | 1 | True | 2 |
| 85.75 | 84.77 | 32 | 1 | True | 5 |
| 86.07 | 84.56 | 32 | 1 | True | 10 |
| 86.05 | 84.56 | 32 | 1 | True | 20 |
| 80.82 | 80.30 | 32 | 2 | False | 2 |
| 82.33 | 81.43 | 32 | 2 | False | 5 |
| 83.06 | 82.01 | 32 | 2 | False | 10 |
| 82.65 | 81.70 | 32 | 2 | False | 20 |
| 84.53 | 83.17 | 32 | 2 | True | 2 |
| 85.50 | 84.72 | 32 | 2 | True | 5 |
| 85.70 | 84.63 | 32 | 2 | True | 10 |
| 85.08 | 84.63 | 32 | 2 | True | 20 |
| 76.41 | 74.93 | 32 | 4 | False | 2 |
| 81.56 | 81.17 | 32 | 4 | False | 5 |
| 82.69 | 81.42 | 32 | 4 | False | 10 |
| 83.17 | 81.52 | 32 | 4 | False | 20 |
| 83.16 | 82.49 | 32 | 4 | True | 2 |
| 85.29 | 84.27 | 32 | 4 | True | 5 |
| 85.98 | 84.93 | 32 | 4 | True | 10 |
| 86.09 | 84.93 | 32 | 4 | True | 20 |
| 80.53 | 79.66 | 64 | 1 | False | 2 |
| 82.35 | 81.40 | 64 | 1 | False | 5 |
| 82.77 | 81.69 | 64 | 1 | False | 10 |
| 83.07 | 81.88 | 64 | 1 | False | 20 |
| 83.41 | 82.12 | 64 | 1 | True | 2 |
| 85.46 | 84.56 | 64 | 1 | True | 5 |
| 86.03 | 84.88 | 64 | 1 | True | 10 |
| 85.79 | 84.88 | 64 | 1 | True | 20 |
| 80.09 | 79.59 | 64 | 2 | False | 2 |
| 82.30 | 81.56 | 64 | 2 | False | 5 |
| 82.83 | 82.11 | 64 | 2 | False | 10 |
| 83.14 | 82.05 | 64 | 2 | False | 20 |
| 83.58 | 82.91 | 64 | 2 | True | 2 |
| 85.73 | 84.58 | 64 | 2 | True | 5 |
| 85.70 | 84.87 | 64 | 2 | True | 10 |
| 84.95 | 84.87 | 64 | 2 | True | 20 |
| 71.96 | 70.47 | 64 | 4 | False | 2 |
| 80.28 | 79.70 | 64 | 4 | False | 5 |
| 82.20 | 81.89 | 64 | 4 | False | 10 |
| 83.14 | 81.69 | 64 | 4 | False | 20 |
| 80.87 | 80.37 | 64 | 4 | True | 2 |
| 84.63 | 84.39 | 64 | 4 | True | 5 |
| 86.46 | 84.87 | 64 | 4 | True | 10 |
| 83.80 | 84.87 | 64 | 4 | True | 20 |
| 78.60 | 77.76 | 128 | 1 | False | 2 |
| 81.92 | 81.50 | 128 | 1 | False | 5 |
| 82.50 | 82.01 | 128 | 1 | False | 10 |
| 82.74 | 81.98 | 128 | 1 | False | 20 |
| 84.68 | 83.47 | 128 | 1 | True | 2 |
| 85.91 | 84.32 | 128 | 1 | True | 5 |
| 86.06 | 84.62 | 128 | 1 | True | 10 |
| 86.57 | 84.62 | 128 | 1 | True | 20 |
| 77.57 | 76.92 | 128 | 2 | False | 2 |
| 82.57 | 81.57 | 128 | 2 | False | 5 |
| 83.21 | 82.03 | 128 | 2 | False | 10 |
| 81.27 | 80.34 | 128 | 2 | False | 20 |
| 85.63 | 84.22 | 128 | 2 | True | 2 |
| 85.20 | 84.82 | 128 | 2 | True | 5 |
| 85.52 | 84.82 | 128 | 2 | True | 10 |
| 85.82 | 84.82 | 128 | 2 | True | 20 |
| 47.62 | 47.35 | 128 | 4 | False | 2 |
| 59.88 | 78.46 | 128 | 4 | False | 5 |
| 82.30 | 81.68 | 128 | 4 | False | 10 |
| 83.40 | 81.96 | 128 | 4 | False | 20 |
| 77.89 | 76.07 | 128 | 4 | True | 2 |
| 85.07 | 83.57 | 128 | 4 | True | 5 |
| 86.10 | 84.79 | 128 | 4 | True | 10 |
| 86.27 | 85.22 | 128 | 4 | True | 20 |

TABLE I: Best parameters according to GridSearch

| DataLimitation | First Val Acc, % | Final Val Acc, % | Test Acc, % | overfitted? |
|---|---|---|---|---|
| Weak: 1 layer, 5 epochs, 32 batch size, simple LSTM | | | | |
| 10 | 42.7 | 76.02 | 63.70 | likely possible |
| 20 | 61.37 | 82.37 | 76.98 | possible |
| 30 | 68.29 | 83.25 | 81.21 | |
| 40 | 71.51 | 80.53 | 79.90 | |
| 50 | 71.45 | 79.00 | 81.16 | |
| 60 | 78.08 | 79.81 | 82.40 | |
| 70 | 75.15 | 80.79 | 83.37 | |
| 80 | 77.08 | 85.17 | 85.99 | |
| 90 | 78.97 | 83.30 | 84.76 | |
| 100 | 80.85 | 84.57 | 85.02 | |
| Strong: 2 layer, 10 epochs, 32 batch size, bidirectional LSTM | | | | |
| 10 | 34.29 | 86.66 | 72.13 | likely possible |
| 20 | 68.58 | 88.67 | 81.38 | possible |
| 30 | 73.76 | 88.29 | 84.58 | |
| 40 | 75.54 | 90.35 | 85.42 | possible |
| 50 | 75.42 | 86.47 | 86.53 | |
| 60 | 77.62 | 84.28 | 86.12 | |
| 70 | 78.03 | 84.68 | 87.00 | |
| 80 | 80.93 | 85.71 | 87.71 | |
| 90 | 81.91 | 87.34 | 88.01 | |
| 100 | 82.99 | 88.52 | 88.37 | |

TABLE II: Models performance

1) 'two', 'of', 'them', 'were', 'being', 'run', 'by', '2', 'officials', 'of', 'the', 'ministry', 'of', 'the', 'interior', '!'

The following words were tagged incorrectly: ministry (Actual: PROPN) - predicted NOUN, interior (Actual: PROPN) - predicted NOUN. These two examples indicates quite complicated contextual task, the PROPN could be distinguished from NOUN with "the" word or capital letter, but this network does not analyze the context and initially the whole text were lowered. The rest of words were identified correctly.

2) 'the', 'moi', 'in', 'iraq', 'is', 'equivalent', 'to', 'the', 'us', 'fbi', ',', 'so', 'this', 'would', 'be', 'like', 'having', 'j.', 'edgar', 'hoover', 'unwittingly', 'employ', 'at', 'a', 'high', 'level', 'members', 'of', 'the', 'weathermen', 'bombers', 'back', 'in', 'the', '1960s', '.'

The following words were tagged incorrectly: moi (Actual: PROPN) - predicted NOUN, iraq (Actual: PROPN) - predicted NOUN, equivalent (Actual: ADJ) - predicted VERB, us (Actual: PROPN) - predicted ADJ, fbi (Actual: PROPN) - predicted NOUN, be (Actual: VERB) - predicted AUX, having (Actual: VERB) - predicted ADV, employ (Actual: VERB) - predicted NOUN. Again, the difference between PROPN and NOUN is hard to detect, therefore we can assume the data was not enough to learn these aspects. The rest of misclassifications are evidence of having not enough data to cover the distribution. The rest of words were identified correctly.

## IV. Conclusions

Based on the results the following was concluded:

1) LSTM is a good approach for Part of Speech classification.
2) Having enough data is obligatory condition to obtain good performance. The data itself has tremendous impact on performance, therefore it should be managed properly. If the data is limited, good representation or generation of new data could help. We used pretrained representative vector to get sophisticated level of performance even in terms of severely limited data or model simplicity.
3) It is doubtfulit is possible to get 95-100% accuracy without diverse and tremendously huge dataset. Part of speech of many words depends on its context, while LSTM can not really understand the meaning of text.
4) The stronger configuration of LSTM (more epochs, preference for bidirectional model, moderate amount of layers), the faster convergence. Table II shows better accuracy of advanced model, it reached the appropriate level of accuracy with the same data limitation in most of cases. But significant data reduction should be assessed whether the model were overfitted, because is is highly possible learn only limited features of limited data.

**Link to Google Colab**: https://colab.research.google.com/drive/1FBqfTYR-EQR5z-b8oi52O77s8gXDr4hN?usp=sharing

**Link to GitHub** with code and limited data: https://github.com/Elizaveta55/NLP_ML

**Amount of words** (without tables, head and abstract): 925