Assignment-Discussion POS tagging using (a) EnCo-DeCo, (b) FFNN-BP

Akshat Gautam, 190110004
Anish Satpati, 190110007
Jaideep Chawla, 190110030
Raghav Gupta, 190040083

10-03-2023

Problem Statement: **Part 1**

- Objective:Given a sequence of words, produce the POS tag sequence
- Dataset: Universal Tag Set

  NOUN, VERB, ADJ, ADV,PRON, DET, ADP, NUM, CONJ, PRT, . , X
- Technique used: RNN and Encoder-Decoder LSTM
- Results: RNN gives a better accuracy of around 99%

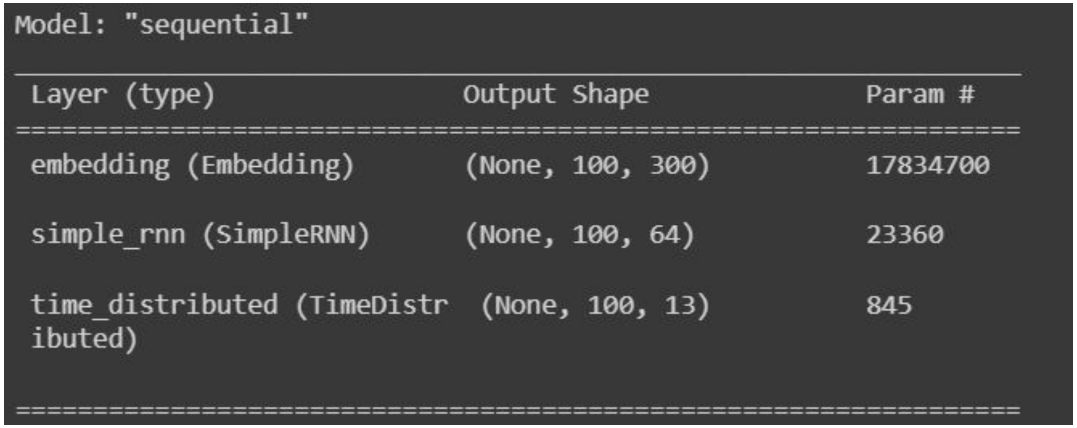
Data Processing Info (Pre-processing)

- Corpus has universal tagset and consists of treebank, brown and conll corpora
  - Total number of tagged sentences: 72202
  - Vocabulary size: 59448
  - Total number of tags: 12
- Lower cased all the sentences
- Tokenized the words and tags
- Set max sequence length to 100

Experimental Setup

- Library used: keras for all models
- Embedding dimension=300, Epochs=10

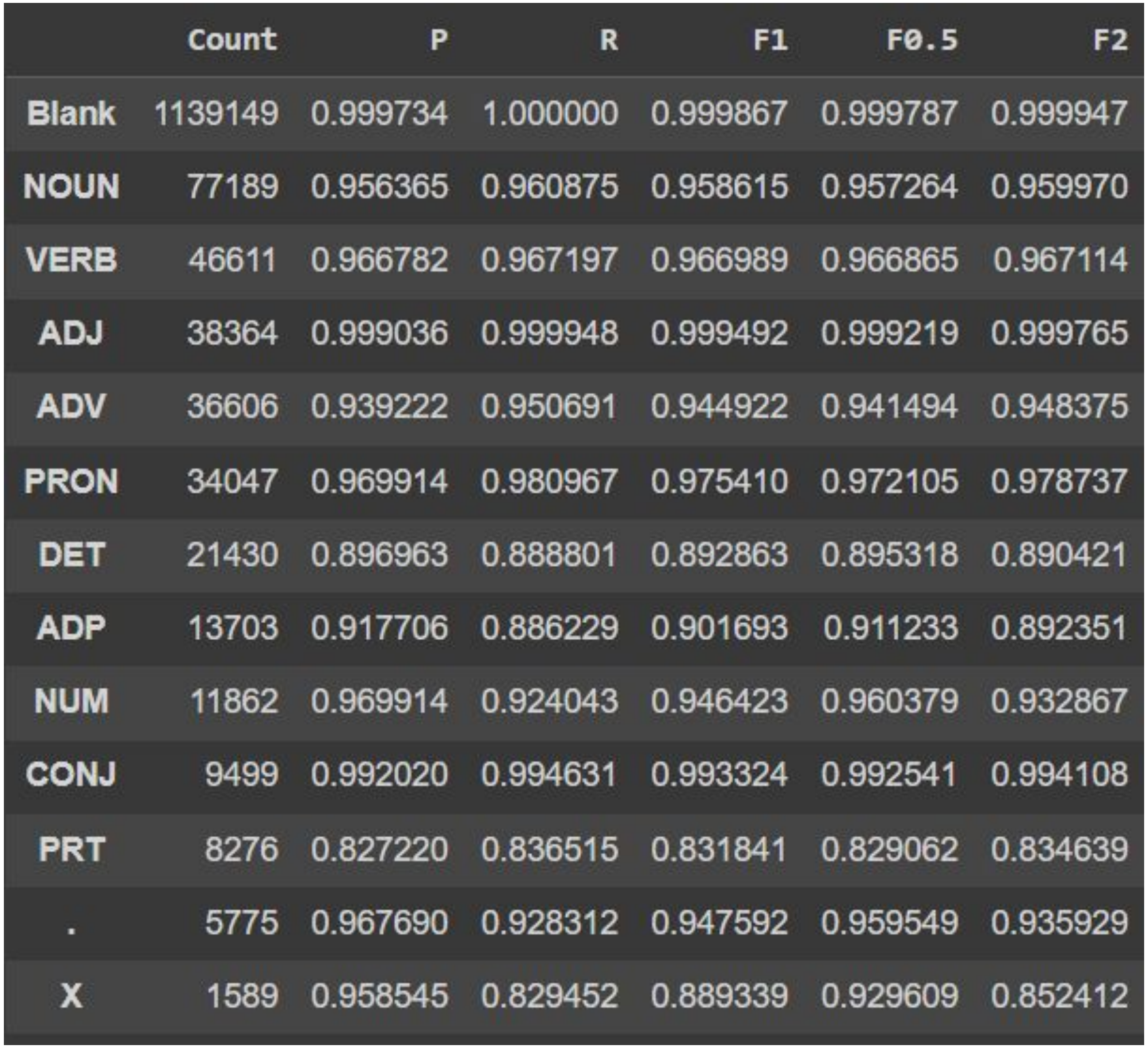
RNN Network:

- 

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 100, 300)          17834700

 simple_rnn (SimpleRNN)      (None, 100, 64)           23360

 time_distributed (TimeDistr  (None, 100, 13)          845
 ibuted)

=================================================================
```

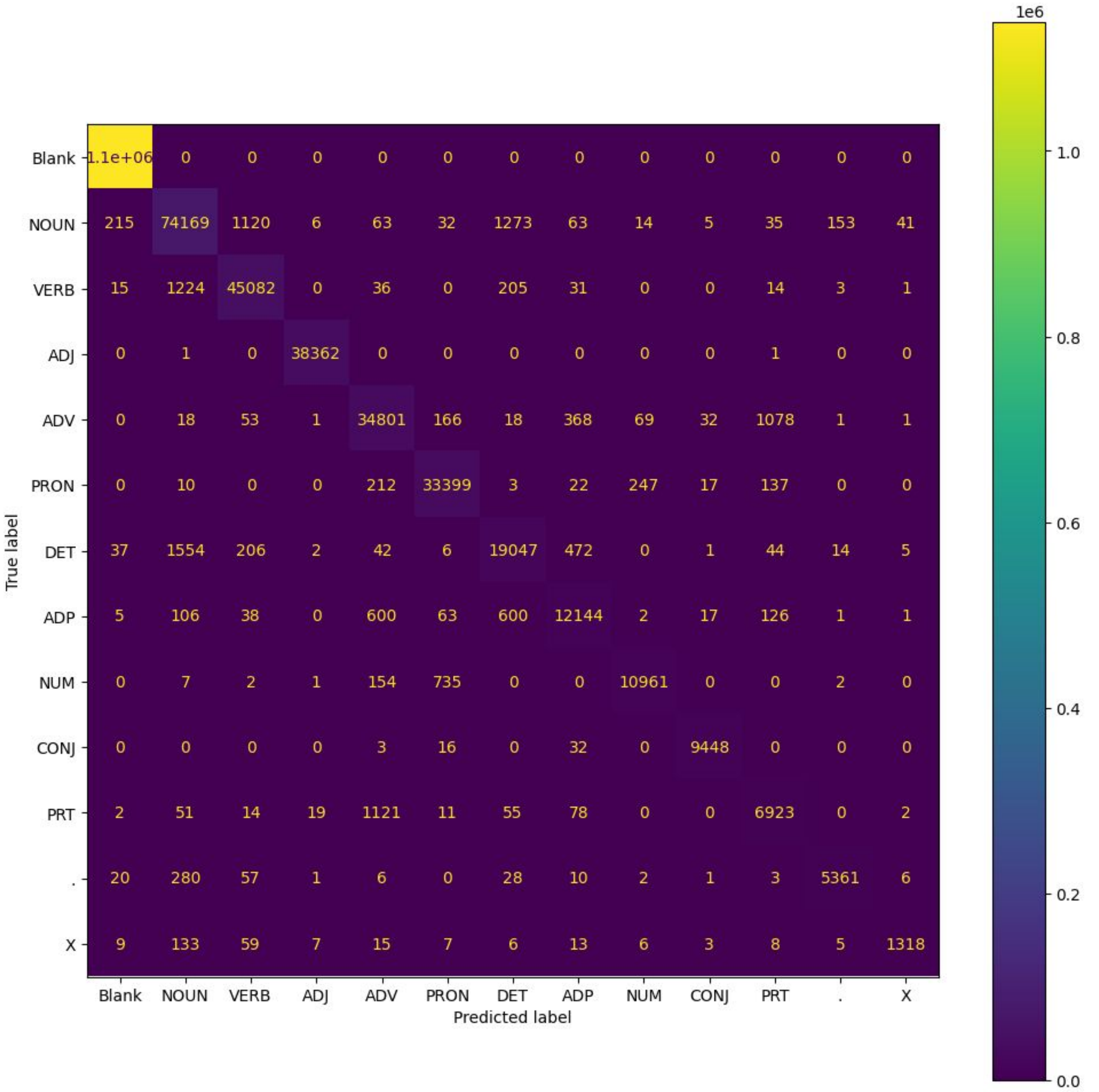
- Initialized weights of the RNN network with word2vec embeddings (gensim)

Overall performance (Part 1)



	Count	P	R	F1	F0.5	F2
Blank	1139149	0.999734	1.000000	0.999867	0.999787	0.999947
NOUN	77189	0.956365	0.960875	0.958615	0.957264	0.959970
VERB	46611	0.966782	0.967197	0.966989	0.966865	0.967114
ADJ	38364	0.999036	0.999948	0.999492	0.999219	0.999765
ADV	36606	0.939222	0.950691	0.944922	0.941494	0.948375
PRON	34047	0.969914	0.980967	0.975410	0.972105	0.978737
DET	21430	0.896963	0.888801	0.892863	0.895318	0.890421
ADP	13703	0.917706	0.886229	0.901693	0.911233	0.892351
NUM	11862	0.969914	0.924043	0.946423	0.960379	0.932867
CONJ	9499	0.992020	0.994631	0.993324	0.992541	0.994108
PRT	8276	0.827220	0.836515	0.831841	0.829062	0.834639
.	5775	0.967690	0.928312	0.947592	0.959549	0.935929
X	1589	0.958545	0.829452	0.889339	0.929609	0.852412

Confusion Matrix (Part 1)



	Blank	NOUN	VERB	ADJ	ADV	PRON	DET	ADP	NUM	CONJ	PRT	.	X
Blank	1.1e+06	0	0	0	0	0	0	0	0	0	0	0	0
NOUN	215	74169	1120	6	63	32	1273	63	14	5	35	153	41
VERB	15	1224	45082	0	36	0	205	31	0	0	14	3	1
ADJ	0	1	0	38362	0	0	0	0	0	0	1	0	0
ADV	0	18	53	1	34801	166	18	368	69	32	1078	1	1
PRON	0	10	0	0	212	33399	3	22	247	17	137	0	0
DET	37	1554	206	2	42	6	19047	472	0	1	44	14	5
ADP	5	106	38	0	600	63	600	12144	2	17	126	1	1
NUM	0	7	2	1	154	735	0	0	10961	0	0	2	0
CONJ	0	0	0	0	3	16	0	32	0	9448	0	0	0
PRT	2	51	14	19	1121	11	55	78	0	0	6923	0	2
.	20	280	57	1	6	0	28	10	2	1	3	5361	6
X	9	133	59	7	15	7	6	13	6	3	8	5	1318

Interpretation of confusion (error analysis)

- Maximal confusions:
  - ADV with PRT
  - DET with NOUN
  - PRT with ADV