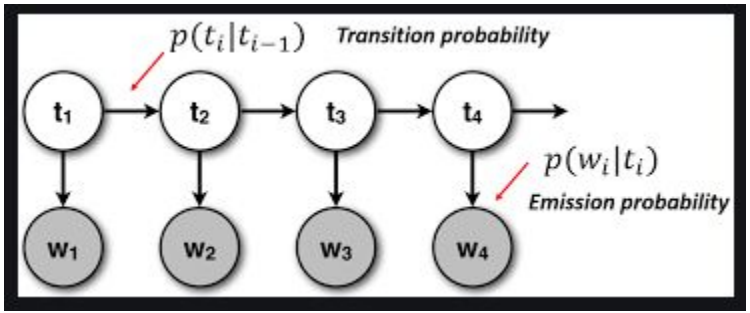
CS 626
Assignment 1- POS Tagging

Team Members:
Saurabh Parekh - 170100016
Inderjeet Nair - 170020013
Manas Jain - 170040068

Hidden Markov Model

Model:



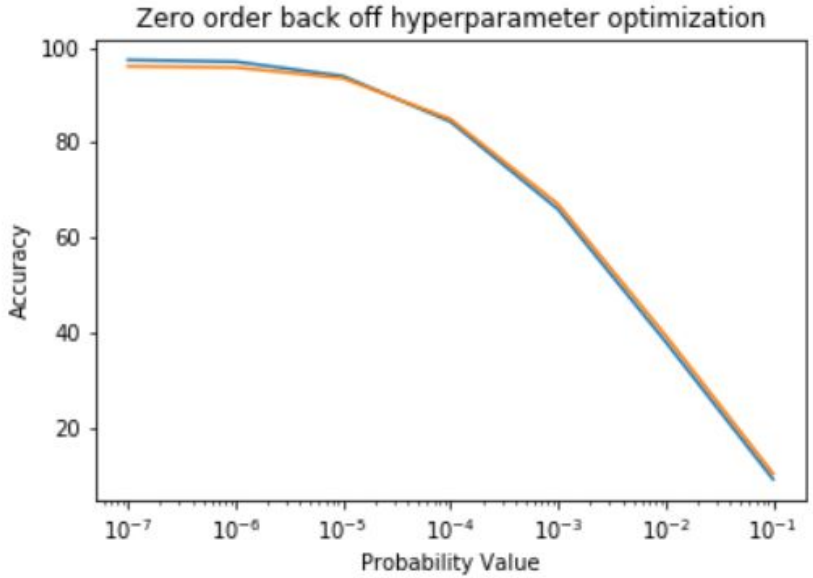
Steps:

1) Computation of Transition Probability Matrix
2) Computation of Emission Probability Matrix
3) Inference using Vitterbi Algorithm

http://people.cs.georgetown.edu/nschneid/cosc572/f16/11_hmm_slides.pdf

HMM: Hyperparameter Optimization

Unknown Transition probabilities were assumed to be a constant quantity λ



Hidden Markov Model

Accuracy:

Train	97.35%
Test	96.01%

Per-PoS Statistics:

Tag	Precision	Recall	F1-Score
NUM	0.926384	0.917104	0.92172
.	0.99631	0.999133	0.997719
VERB	0.971693	0.953625	0.962574
ADP	0.950455	0.967099	0.958705
ADV	0.908005	0.893846	0.90087
PRT	0.904712	0.903014	0.903862
DET	0.966233	0.985922	0.975978
X	0.733918	0.54329	0.624378
PRON	0.962992	0.984757	0.973753
ADJ	0.917711	0.915266	0.916487
NOUN	0.962	0.95565	0.958815
CONJ	0.992979	0.993552	0.993266

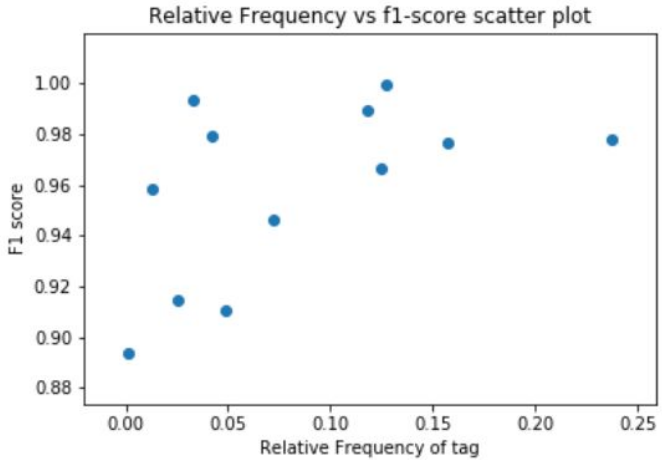
HMM Confusion Matrix

	NUM	.	VERB	ADP	ADV	PRT	DET	X	PRON	ADJ	NOUN	CONJ
NUM	13641	93	105	93	26	20	71	7	12	48	576	33
.	4	147437	95	22	14	4	0	36	0	55	316	0
VERB	43	0	174275	108	272	49	3	47	2	652	3901	0
ADP	37	0	900	140003	1777	2234	684	49	412	174	1020	11
ADV	0	0	115	1921	50269	271	99	5	3	2384	172	123
PRT	2	0	17	1748	760	26936	1	7	6	197	99	0
DET	320	0	421	296	266	20	135090	82	292	1107	1840	77
X	3	35	5	5	0	3	3	753	0	4	214	1
PRON	34	0	20	295	45	22	1032	11	48582	15	393	0
ADJ	62	0	690	80	2279	31	1	38	1	76627	3689	0
NOUN	728	0	6107	51	443	239	4	347	24	2458	263337	1
CONJ	0	0	0	144	88	0	31	4	0	0	1	37905

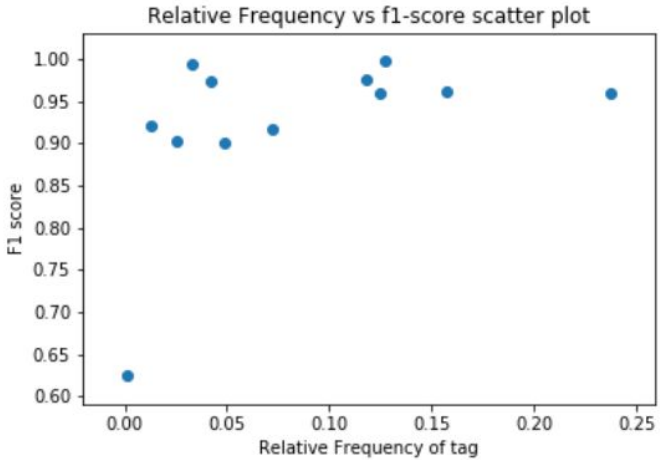
HMM Heat Map of Confusion matrix

HMM: Variation of Per-POS Accuracy with relative frequency

We noted that if the relative frequency of occurrence of a tag is very small, then the performance with respect to that Tag is very Poor.



(a) Train Set



(b) Test Set

HMM: Error analysis

Low Precision and Recall Associated with ‘ADV’:

The model is confused about the predictions associated with tags ‘ADV’

Perhaps the data is confusing.

Quantizing the confusing aspect about a tag ‘S’: H(p(. | S)) [Entropy of transition probability distribution from tag S]

Some entropy values:

TAG	F1-Score	H(p(. | TAG))
ADV	0.900	2.10
PRON	0.973	1.14
ADJ	**0.916**	**1.28**
CONJ	**0.993**	**2.04**

HMM: Entropy Theory Modified

Combining Entropy (H(p(.|s))) and relative frequency occurrence (f)

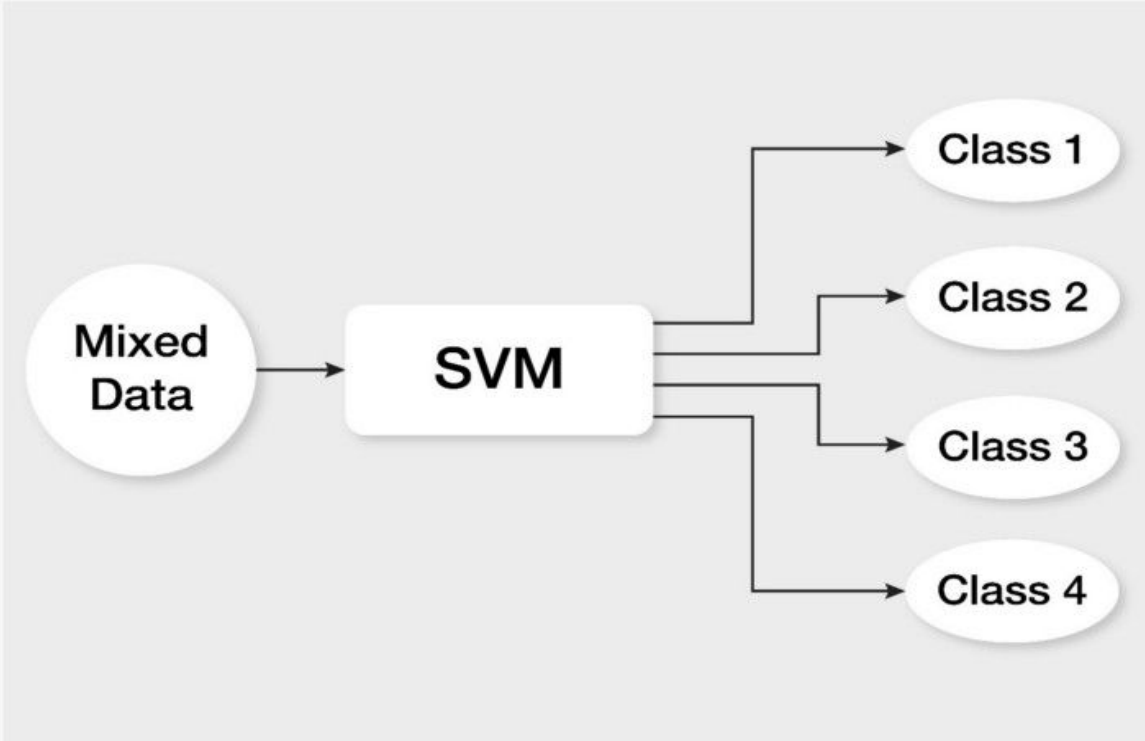
Define M = log(H(p(.|s)) / f)

Interpretation of M being high and low

M range	Average F1-Score
0.5 - 1	0.967
1 - 1.5	0.962
1.5 - 2	0.933
2 - 2.5	0.922
3 - 3.5	0.624

Support Vector Machine

Model Overview



https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

Support Vector Machine

Accuracy:

Train	83.36%
Test	83.25%

Per-PoS Statistics:

Tag	Precision	Recall	F1-Score
NOUN	0.800860	0.816398	0.808555
VERB	0.775557	0.756405	0.765861
ADP	0.846252	0.898056	0.871385
.	0.996299	0.999946	0.998119
DET	0.936171	0.967771	0.951708
ADJ	0.595159	0.545419	0.569204
ADV	0.622987	0.646117	0.634342
CONJ	0.947711	0.983408	0.96523
PRON	0.935717	0.935035	0.935376
PRT	0.711897	0.448154	0.550044
NUM	0.919363	0.932096	0.925686
X	0.494881	0.313853	0.384106

SVM Confusion Matrix

	NOUN	VERB	ADP	.	DET	ADJ	ADV	CONJ	PRON	PRT	NUM	X
NOUN	224965	21701	2019	0	456	23220	4337	72	360	2666	425	683
VERB	18200	138233	4693	0	806	6034	6399	86	339	3233	132	82
ADP	3779	4664	130008	4	1233	1902	2889	69	840	8204	14	22
.	87	189	140	147557	44	4	50	1	0	5	3	25
DET	1002	900	2053	1	132603	1412	1394	58	1472	482	247	20
ADJ	19266	6032	949	0	500	45663	3444	67	19	560	149	75
ADV	4531	8120	2447	0	520	4935	36337	227	99	1093	9	9
CONJ	364	570	174	1	393	143	375	37518	27	11	7	5
PRON	615	751	699	2	446	42	398	12	46129	178	13	13
PRT	1444	1511	1558	0	15	197	592	34	40	13368	9	10
NUM	988	55	18	0	3	114	16	5	0	10	13864	7
X	317	24	8	0	0	55	8	2	9	19	2	435

SVM Heat Map

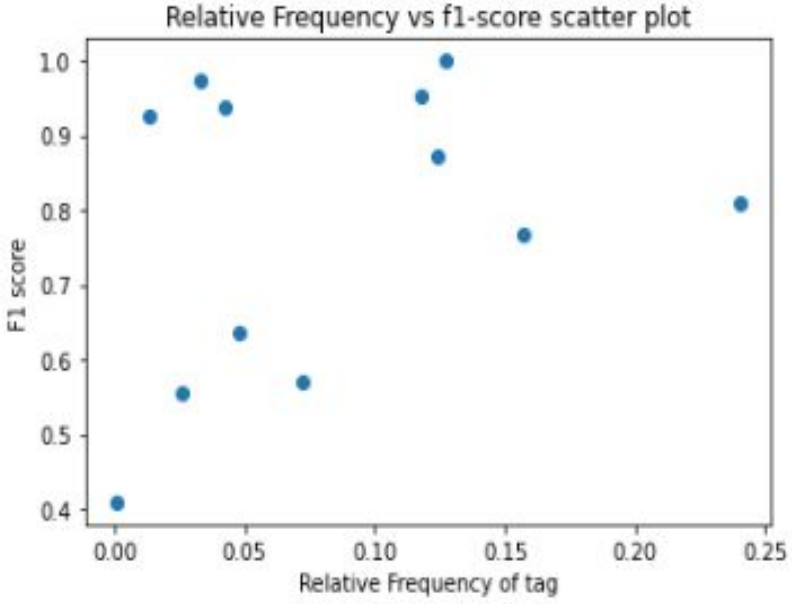

NOUN 2.2e+05 2.2e+04 2e+03 0 4.6e+02 2.3e+04 4.3e+03 72 3.6e+02 2.7e+03 4.2e+02 6.8e+02
VERB 1.8e+04 1.4e+05 4.7e+03 0 8.1e+02 6e+03 6.4e+03 86 3.4e+02 3.2e+03 1.3e+02 82
ADP 3.8e+03 4.7e+03 1.3e+05 4 1.2e+03 1.9e+03 2.9e+03 69 8.4e+02 8.2e+03 14 22
. 87 1.9e+02 1.4e+02 1.5e+05 44 4 50 1 0 5 3 25
DET 1e+03 9e+02 2.1e+03 1 1.3e+05 1.4e+03 1.4e+03 58 1.5e+03 4.8e+02 2.5e+02 20
ADJ 1.9e+04 6e+03 9.5e+02 0 5e+02 4.6e+04 3.4e+03 67 19 5.6e+02 1.5e+02 75
ADV 4.5e+03 8.1e+03 2.4e+03 0 5.2e+02 4.9e+03 3.6e+04 2.3e+02 99 1.1e+03 9 9
CONJ 3.6e+02 5.7e+02 1.7e+02 1 3.9e+02 1.4e+02 3.8e+02 3.8e+04 27 11 7 5
PRON 6.2e+02 7.5e+02 7e+02 2 4.5e+02 42 4e+02 12 4.6e+04 1.8e+02 13 13
PRT 1.4e+03 1.5e+03 1.6e+03 0 15 2e+02 5.9e+02 34 40 1.3e+04 9 10
NUM 9.9e+02 55 18 0 3 1.1e+02 16 5 0 10 1.4e+04 7
X 3.2e+02 24 8 0 0 55 8 2 9 19 2 4.4e+02
NOUN VERB ADP . DET ADJ ADV CONJ PRON PRT NUM X
200000
175000
150000
125000
100000
75000
50000
25000
0

SVM Feature Engineering

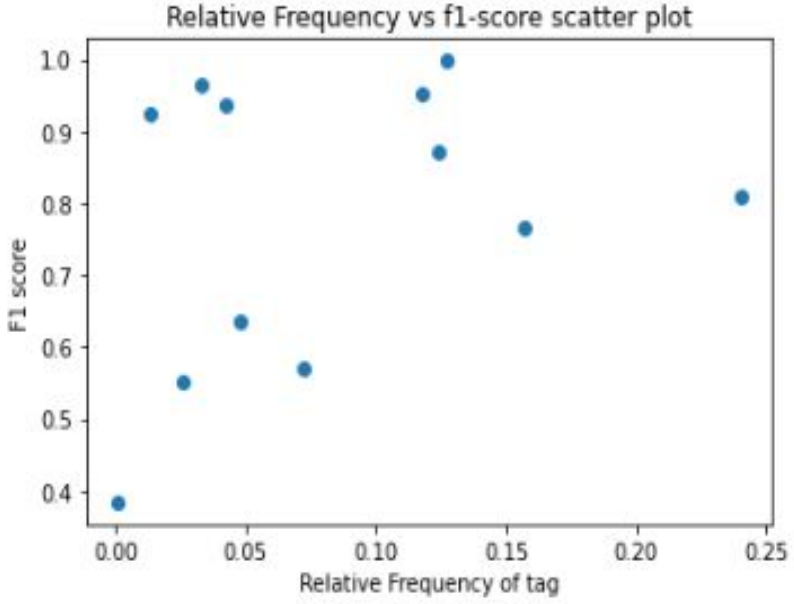
Features Selected	Accuracy
Word length, capitalisation, upper-case, lower-case, isNumeric	40%
Prefix and suffix for nouns, verbs, adjectives and adverbs	55%
Word stems using PorterStemmer	60%
Tag of Previous Word	65%
Pre-trained Glove word-embeddings (1 lakh 50-dimensional vectors)	83%
Pre-trained word2vec embeddings (10 crore 300-dimensional vectors)	90%
Including the features of previous 3 words and following 3 words	95%

SVM: Variation of Per-POS Accuracy with relative frequency

It can be observed that the distribution of F1-score for training set and test set is almost same.



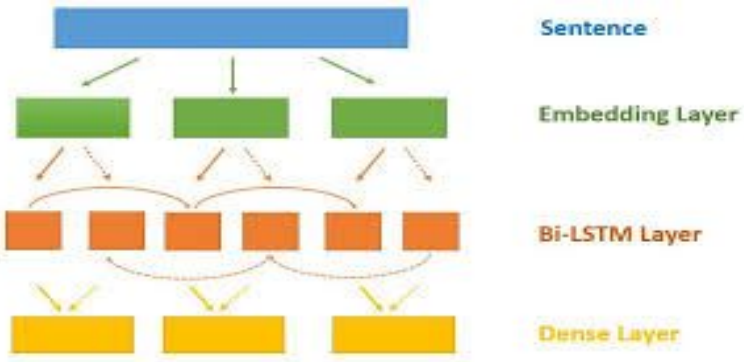
(a) Train Set



(b) Test Set

Bidirectional Long Short term Memory (Baseline)

Model overview



reference

Steps :-

1. Preprocessing of the corpus (Tokenization of words & padding)
2. Extracted embeddings of all vocabulary words using Word2Vec
3. Defined the model architecture adding embedding, BiLSTM & Time Distributed Dense layer
4. Trained the model using 5 fold cross validation

Bi-LSTM (Baseline) - Architecture

Embedding dimension - 300 (Word2vec)

Max sequence length used for padding - 180

Vocabulary length - 49816

BiLSTM cells - 32

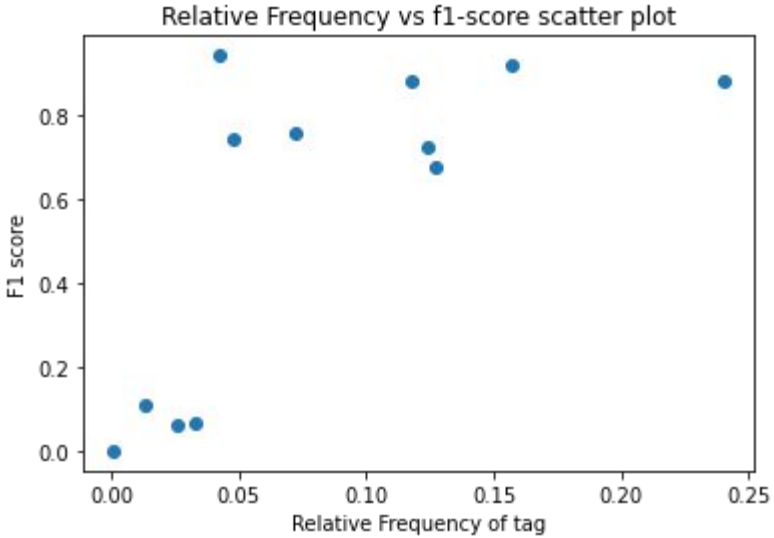
Final Dense layer neurons - 13

Model architecture summary (baseline)

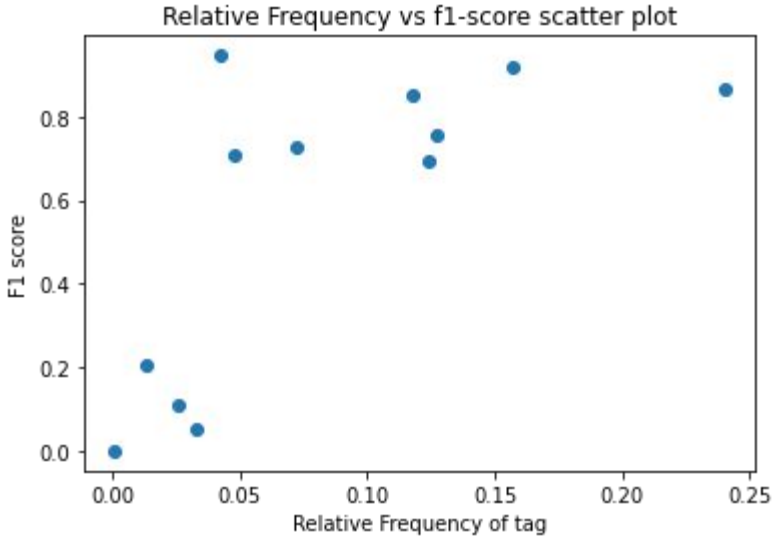
```
Layer (type)                 Output Shape              Param #
=================================================================

embedding (Embedding)        (None, 180, 300)          14944800
_________________________________________________________________

bidirectional (Bidirectional (None, 180, 64)           85248
_________________________________________________________________

time_distributed (TimeDistri (None, 180, 13)           845
=================================================================

Total params: 15,030,893

Trainable params: 86,093

Non-trainable params: 14,944,800
```

Bi-LSTM (baseline) Heat map of confusion matrix

Bi-LSTM (Baseline): Variation of Per-POS Accuracy with relative frequency



Train set



Test set

Here we can see the dependence of data on accuracy on a deep learning model. The labels having less examples in training set have low F1 scores

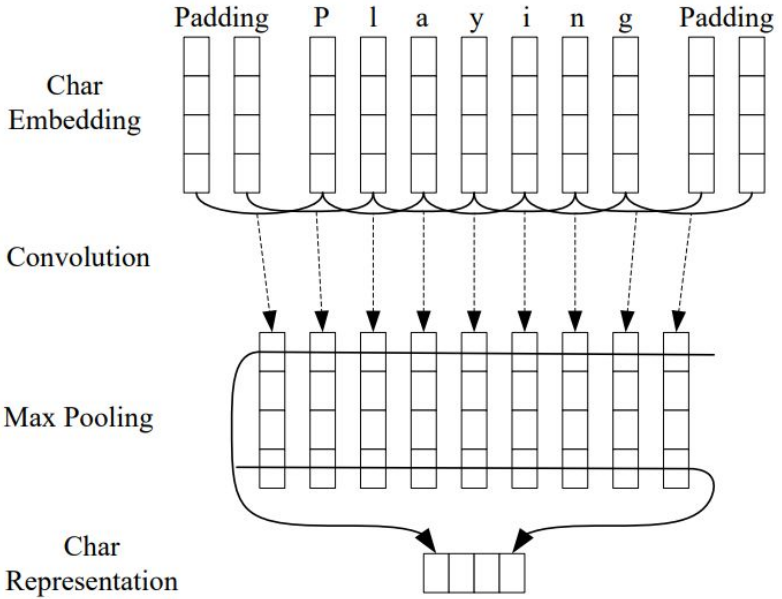
CNN Bi-LSTM

Modifications over the baseline in upcoming slides

Combining Bi-LSTM and CNN

Baseline Bi-LSTM does not takes into account character level information

Word Embedding construction:

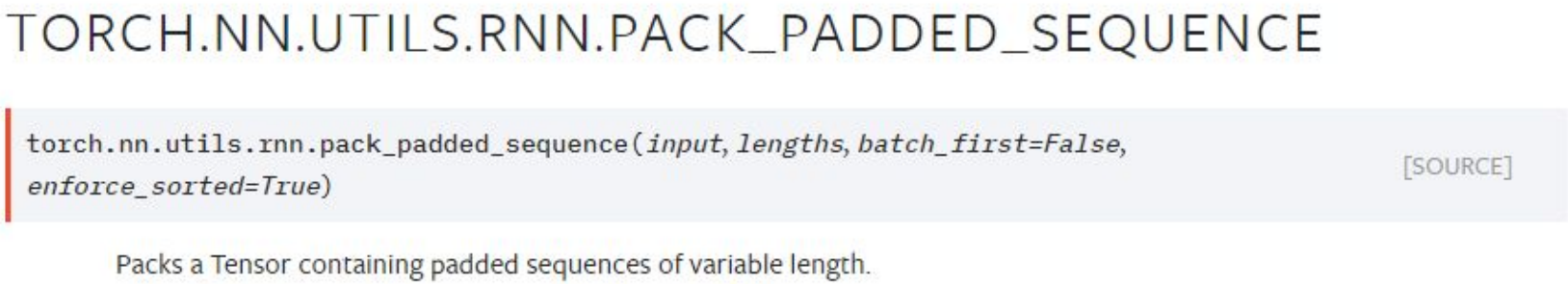


\+ Word embedding

Avoiding Training on <PAD> tags

In the baseline model, sentences were extrapolated to a fixed number of tokens by padding with <PAD> tokens. The RNN model trained on this becomes biased towards the <PAD> tags

In the CNN based model, we used this without compromising the speed:



Improvements over baseline

Each of the models were trained for same of epoches = 3

Model	Accuracy Test(%)	Accuracy Train(%)
Bi-LSTM Baseline	79.38	78.39
Bi-LSTM-CNN	87.15	87.19

Tag	P(Baseline)	R(Baseline)	F1(Baseline)	P	R	F1
NOUN	0.872538	0.859111	0.865772	0.8467462957	0.8749519157	0.8606180673
VERB	0.925502	0.914175	0.919804	0.8455539818	0.859124487	0.8522852188
.	0.612080	0.981999	0.754118	0.992072625	0.9990241588	0.9955362569
ADP	0.691465	0.697326	0.694383	0.9009276846	0.9043007336	0.9026110579
DET	0.916463	0.795872	0.851921	0.9399036033	0.9507075661	0.9452747149
ADJ	0.747382	0.712002	0.729263	0.704576745	0.605152829	0.6510910633
ADV	0.686423	0.731869	0.708418	0.6817838382	0.6738917833	0.677814839
PRON	0.959272	0.933033	0.945971	0.8935800928	0.9287915028	0.9108456248
CONJ	1.000000	0.027419	0.053375	0.9679943101	0.9631988677	0.965590635
PRT	0.855478	0.058729	0.109913	0.7969689423	0.7510141138	0.7733093997
NUM	1.000000	0.112994	0.203046	0.8789575866	0.7913809332	0.8328734168
X	NaN	0.000000	NaN	0.6857142857	0.0173160173	0.0337790289