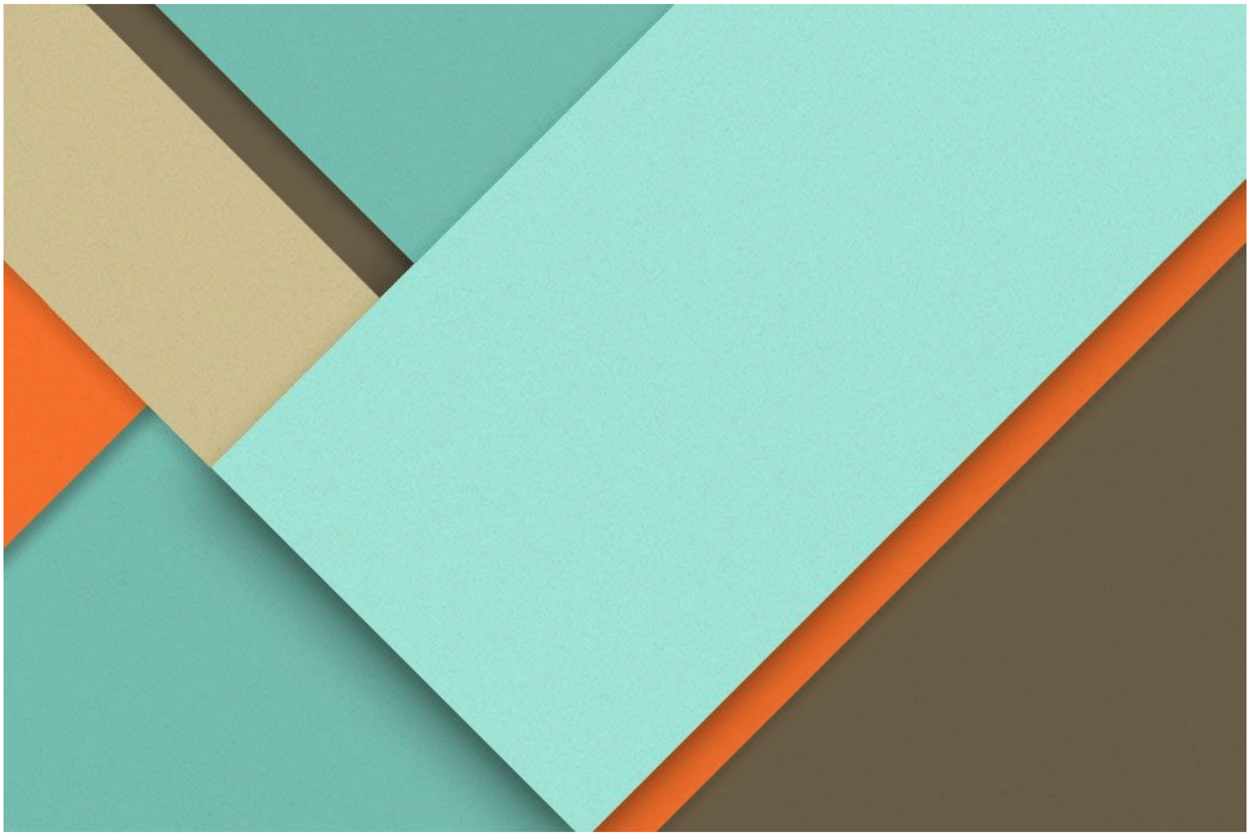

CS 626 Assignment 1

Part of Speech Tagging using SVM, HMM and Bi-LSTM

20.09.2020

Deepti Mittal 193050025

Ankita Singh 19305R002

Overview :

The process of marking up the words in text(corpus) corresponding to a particular Part of Speech on the basis of its definition and context.

Example: Everything is all about Money.

Tagged Sentence: Everything_NN is_VBZ all_DT about_IN money_NN ._.

Corpus Insights :

We used the 'Brown' corpus and universal tag-set for our experiments .

List of universal tags { DET, NOUN, ADJ, VERB, ADP, ., ADV, CONJ, PRT, PRON, NUM, X}

Total number of sentences considered : 57340

Total number of tagged words considered : 1161192

Total tagged words in each category (considering Universal tags) :

TAG	COUNT OF TAGS
DET (DETERMINER)	137019
NOUN	275558
ADJ (ADJECTIVE)	83721
VERB	182750
ADP (ADPOSITION)	144766
'.' (SYMBOL)	147565
ADV (ADVERB)	56239
CONJ (CONJUNCTION)	38151
PRT (PARTICLE)	29829
PRON (PRONOUN)	49334
NUM (NUMERAL)	14874
X (OTHER)	1386

We notice that the most number of POS tags in the corpus belongs to NOUNS. The least is X which means the tag is unidentified. Some sentences are observed to be very long , even ranging to 180 words.

We observed some in-correctly tagged sentences in brown corpus itself, especially with respect to NOUN- (ADV/ADJ/VERB) pairs.

Confusion Matrix and Accuracy values:

Metrics:

1. **Precision**: ratio of true positives to the sum of true positives and false positives.
2. **Recall** : ratio of true positives to the sum of true positives and false negatives.
3. **F1_score** : 2*((precision*recall)/(precision+recall))
4. **Overall Accuracy** = ratio of sum of true positives for all tags to the total tags.

Confusion Matrix :

The confusion matrix for POS tagging contains a set of rows and columns equal to the number of tags in the classification set. The topmost column names represent the true tags and the left hand side column names represent the predicted tags. Thus , the diagonal elements signify the true positives.

We calculate overall accuracy for POS tag classification as:

$$\frac{\text{\# Tags classified correctly (sum of diagonal elements of matrix)}}{\text{\# Total Tags in the corpus (sum of all the elements in the matrix)}}$$

Here are the Confusion matrix and per-tag metrics obtained from the aforementioned three techniques:

HMM:

Confusion Matrix :

	.	<EOS>	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	^
.	<147206>	.	.	20	.	.	.	1	.	.	.	.	20	.
<EOS>	.	<57340>	.	.	.	.	.	.	.	.	.	.	.	.
ADJ	.	.	<75965>	60	3433	.	.	2372	3	1	370	737	20	.
ADP	.	.	143	<129586>	4486	3	2186	116	.	1797	978	43	27	.
ADV	.	.	3407	733	<46176>	135	71	208	.	3	401	84	4	.
CONJ	.	.	.	148	46	<37901>	61	1	.	.	.	.	4	.
DET	.	.	.	1	336	109	<134589>	2	.	989	2	.	18	.
NOUN	.	.	3698	138	675	.	.	<262537>	217	15	199	8647	101	.
NUM	.	.	2	1	.	.	2	481	<14652>	.	.	.	5	.
PRON	.	.	.	.	.	.	17	19	.	<46066>	1	.	6	.
PRT	.	.	15	12955	960	.	1	33	.	7	<27764>	6	4	.
VERB	.	.	430	1086	12	2	.	9723	.	.	38	<173049>	37	.
X	359	.	61	38	115	1	92	65	2	456	76	184	<1140>	.
^	.	.	.	.	.	.	.	.	.	.	.	.	.	<57340>

(row = reference; col = test)

OVERALL ACCURACY : 94.94 %

Ignore <EOS> and ^ tags

Per -POS accuracy :

TAG	PRECISION	RECALL	F1_SCORE
^	1.0	1.0	1.0
DET	0.99	0.98	0.99
NOUN	0.95	0.95	0.95
VERB	0.94	0.95	0.94
ADP	0.93	0.9	0.91
ADJ	0.92	0.91	0.91
PRT	0.67	0.93	0.78
.	1.0	1.0	1.0
<EOS>	1.0	1.0	1.0
NUM	0.97	0.99	0.98
CONJ	0.99	0.99	0.99
ADV	0.9	0.82	0.86
PRON	1.0	0.93	0.97
X	0.44	0.83	0.57

Overall Accuracy : 94.94%

SVM:

Confusion Matrix:

	.	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X
.	<116249>	.	.	.	.	.	3	4	.	.	.	.
ADJ	.	<49302>	78	1745	2	5	10831	54	.	24	1951	14
ADP	14	166	<108783>	921	71	251	290	99	172	1335	332	1
ADV	.	4680	1075	<35727>	63	211	2325	47	35	204	445	2
CONJ	.	50	83	174	<29720>	43	44	1	.	3	10	.
DET	.	110	314	154	9	<105173>	129	4	252	.	9	4
NOUN	.	7353	120	924	5	94	<193845>	553	23	46	2926	85
NUM	.	211	4	35	4	5	674	<9670>	.	.	80	1
PRON	.	73	229	69	2	624	625	1	<39148>	6	73	3
PRT	.	206	1333	679	.	28	663	2	22	<20656>	207	2
VERB	.	2280	275	321	11	31	5709	16	18	120	<135005>	11
X	23	32	27	10	7	17	824	13	6	3	41	<94>

(row = reference; col = test)

Per-POS Accuracy:

TAG	PRECISION	RECALL	F1_SCORE
NOUN	0.9	0.94	0.92
.	1.0	1.0	1.0
CONJ	0.99	0.99	0.99
PRON	0.99	0.96	0.97
VERB	0.96	0.94	0.95
DET	0.99	0.99	0.99
ADV	0.88	0.8	0.84
NUM	0.92	0.91	0.91
ADP	0.97	0.97	0.97
ADJ	0.76	0.77	0.77
PRT	0.92	0.87	0.89
X	0.43	0.09	0.14

Overall Accuracy : 93.71%

Bi-LSTM:

Confusion Matrix :

	.	ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PAD	PRON	PRT	VERB	X
.	<147562>	.	20	.	.	.	1	.	.	.	.	.	24
ADJ	.	<76237>	46	1671	.	1	2453	101	.	2	106	507	29
ADP	.	69	<141939>	1170	8	588	59	3	.	296	1834	127	27
ADV	.	1708	802	<51785>	64	88	185	10	.	11	269	139	44
CONJ	.	.	149	121	<38040>	58	2	.	.	.	.	1	5
DET	.	.	275	169	33	<135957>	108	1	.	509	2	.	19
NOUN	.	4889	48	670	2	12	<268700>	871	77	28	221	4850	756
NUM	.	7	1	3	1	5	349	<13742>	.	.	2	4	31
PAD	.	3	.	1	.	.	52	2	<9068192>	.	.	8	4
PRON	.	.	42	2	1	300	22	1	.	<48484>	4	1	9
PRT	.	49	1350	361	1	1	37	4	.	1	<27298>	5	5
VERB	.	754	93	284	.	8	3544	138	3	3	89	<177105>	117
X	.	5	.	2	.	1	45	1	.	.	3	2	<316>

(row = reference; col = test)

Ignore the 'PAD' tag used for padding.

Per-POS Accuracy:

TAG	PRECISION	RECALL	F1_SCORE
DET	0.99	0.99	0.99
NOUN	0.96	0.98	0.97
ADJ	0.94	0.91	0.92
VERB	0.97	0.97	0.97
ADP	0.97	0.98	0.98
.	1.0	1.0	1.0
PAD	1.0	1.0	1.0
PRT	0.94	0.91	0.93
NUM	0.97	0.92	0.95
ADV	0.94	0.92	0.93
PRON	0.99	0.98	0.99
CONJ	0.99	1.0	0.99
X	0.86	0.23	0.36

Overall Accuracy : 97.07%

Error analysis :

HMM :

1. In preprocessing , conversion of all the words into lower case gave an accuracy of 93.56% whereas without the same, it gave 94.25% overall accuracy.
2. To increase the accuracy further, we performed lower case preprocessing on only **non-NOUN** tagged words. This increased the test accuracy to 94.94%.
3. We used a smoothing factor of 0.00001 for tackling cases when the test set encountered unknown words.
4. We also added beginning(^) and end of sentence(<EOS>) tags for ease of processing.
5. A significant number of words are classified as verbs when nouns and vice versa. We notice the same trend when it comes to particles and adpositions. This is because our HMM model takes bigrams and hence is not able to capture long distance dependencies.

SVM:

1. We used word morphological features in categorisation of POS tags.
2. Feature Set includes following features:
    - Length of the word.
    - Word vector
    - Prev word vector
    - Next word vector
    - Length of the previous word
    - Length of the next word
    - No. of words in a sentence
    - Suffix of word.
    - Suffix of previous word
    - Suffix of next word.
    - Is first letter capital?
    - Are all letters capital?
    - All letters small?
    - Similarity measure of word from previous word
    - Similarity measures of word from next word.
    - Whether the word is the first word in a sentence or not?
    - Whether the word is the last word in the sentence or not?

3. Words are represented using a vector of dimension 50.

4. We have trained the model on 200k words (around 20 percent of the training data) which took 16 hrs for training on the local system. We used an rbf kernel with parameters of gamma = 0.001 and C=1000.

5. Accuracy of 91.33% was achieved when window size was increased to 5 without involving similarity measures. So, we can say similarity measures play an important role in better classification.

6. While training the model with the partial data, when we took all the tags classes of the same size, accuracy decreased.

7. We also tried using stochastic gradient descent SVM classifiers (linear) for training our model. Although the training time reduced significantly, we could only get accuracies close to 85% ,which is much worse if we compare it with our current model.

8. Adjectives and adverbs seem to be misclassified more often in SVM, particle classification accuracy significantly improved if compared to HMM.

Bi-LSTM :

1. The metrics when Bi-LSTM was trained by converting all the words to lower was giving better accuracy.
2. Time taken to train the model for 4 epochs was around 9 hours on whole data.
3. The precision and recall for X tag is quite less.
4. We also added the ‘PAD’ tag for the ease of processing sentences of different length.
5. Embeddings taken are of size 128.
6. Accuracy when all words were converted to lower gives better results in case of Bi_LSTM but not in SVM and HMM.

General Comments :

1. PRT-ADP misclassification reduces significantly as compared to HMM as word dependencies are captured.
2. Recall for X tag is quite less as compared to other tags in both SVM and bi-LSTM. On the other hand HMM does a better job in the recall rate for X .
3. HMM training and testing was comparatively faster than other models. SVM and LSTM took hours to train whereas in HMM the results were obtained in a matter of a few seconds.
4. Few tags like determiners,conjunctions and pronouns are very accurately classified by all the models owing to less ambiguity in occurances.
5. Some common misclassifications that are common in all the models are NOUN-VERB ,PRT-ADP,ADV-ADJ and NOUN-ADJ
6. Comparison table for tags:

TAG	HMM	SVM	Bi-LSTM
ADV	0.90	0.88	0.94
DET	0.99	0.99	0.99
ADJ	0.92	0.82	0.94
NOUN	0.95	0.90	0.96
ADP	0.67	0.97	0.97
VERB	0.94	0.96	0.97
CONJ	0.99	0.99	0.99
PRON	1.0	0.99	0.99
NUM	0.97	0.97	0.97
PRT	0.67	0.94	0.94
X	0.44	0.43	0.86
.	1.0	1.0	1.0
OVERALL	94.94%	93.71%	97.07%