CS563-NLP

Assignment 1: POS Tagging using HMMs

Group Name: **1801cs31_1801cs33**
Students:

Names	**Roll No.**	**Batch**
M **Maheeth Reddy**	1801CS31	B.Tech.
Nischal A	1801CS33	B.Tech.

Solution:

We have implemented two strategies for POS tagging

1. Replacing unknown words with the most commonly occurring POS tag
   ➔ *python trigram_tagger.py*
2. Using a _RARE_ token for all words which occur less than 5 and model the trigram transition probabilities distribution, and emission probabilities using the _RARE_ token.
   ➔ *python trigram_tagger_rare.py*

Approach	Tags Used	Overall Accuracy	% Improvement
1. Replacing unkown word with most frequent POS tag	36 Tags	73%	12%
	4 Tags	85%	
2. Modeling _RARE_ probability distribution	36 Tags	80%	5%
	4 Tags	85%	

We further explain the results in the next few pages.

Approach 1: Replacing unkown words with most commonly occurring POS tag

1. Results using 36 POS Tags

	precision	recall	f1-score	support
#	1.00	1.00	1.00	3
''	0.00	0.00	0.00	3
-LRB-	1.00	1.00	1.00	23
-RRB-	1.00	0.92	0.96	26
:	1.00	0.98	0.99	42
CC	1.00	0.96	0.98	471
CD	0.99	0.43	0.60	729
DT	0.99	0.97	0.98	1660
EX	1.00	0.62	0.76	13
IN	0.97	0.96	0.96	2014
JJ	0.91	0.49	0.63	1177
JJR	0.83	0.65	0.73	77
JJS	0.86	0.59	0.70	41
LS	0.00	0.00	0.00	4
MD	1.00	0.98	0.99	173
NN	0.39	0.98	0.56	2618
NNP	0.98	0.43	0.60	1971
NNPS	0.56	0.09	0.15	57
NNS	0.97	0.44	0.60	1146
PDT	0.40	0.40	0.40	5
PRP	1.00	0.92	0.96	331
PRP$	0.99	0.99	0.99	137
RB	0.85	0.52	0.65	449
RBR	0.47	0.42	0.44	19
RBS	0.60	0.43	0.50	7
RP	0.60	0.50	0.55	52
SYM	0.00	0.00	0.00	1
TO	1.00	0.96	0.98	439
VB	0.96	0.58	0.72	496
VBD	0.98	0.69	0.81	653
VBG	0.94	0.32	0.48	275
VBN	0.87	0.47	0.61	445
VBP	0.94	0.67	0.78	227
VBZ	0.99	0.71	0.83	375
WDT	0.89	0.79	0.84	85
WP	0.97	0.84	0.90	45
WP$	1.00	1.00	1.00	1
WRB	1.00	0.84	0.91	31
accuracy			0.73	16321
macro avg	0.81	0.65	0.70	16321
weighted avg	0.87	0.73	0.74	16321

Overall Accuracy: 73%

2. Results using only 4 Tags

	precision	recall	f1-score	support
A	0.95	0.52	0.68	1801
N	0.73	0.99	0.84	6306
O	0.99	0.91	0.95	5743
V	0.96	0.60	0.74	2471
accuracy			0.85	16321
macro avg	0.91	0.75	0.80	16321
weighted avg	0.88	0.85	0.84	16321

Overall Accuracy is 85%

Approach 2: Modeling _RARE_ words probability distribution: In this approach we term all the words less than 5 in number as _RARE_ observation during the train time. We replace all unknown words with _RARE_ token during the test time.

1. Results using 36 POS Tags

	precision	recall	f1-score	support
#	1.00	1.00	1.00	3
''	0.00	0.00	0.00	3
-LRB-	1.00	1.00	1.00	23
-RRB-	1.00	0.92	0.96	26
:	1.00	0.98	0.99	42
CC	0.98	0.96	0.97	471
CD	0.72	0.70	0.71	729
DT	0.99	0.97	0.98	1660
EX	0.89	0.62	0.73	13
FW	0.00	0.00	0.00	0
IN	0.96	0.96	0.96	2014
JJ	0.64	0.65	0.65	1177
JJR	0.78	0.65	0.71	77
JJS	0.83	0.59	0.69	41
LS	0.00	0.00	0.00	4
MD	1.00	0.99	1.00	173
NN	0.77	0.73	0.75	2618
NNP	0.64	0.88	0.74	1971
NNPS	0.31	0.09	0.14	57
NNS	0.70	0.63	0.66	1146
PDT	0.40	0.40	0.40	5
PRP	0.99	0.92	0.95	331
PRP$	0.99	0.99	0.99	137
RB	0.73	0.60	0.66	449
RBR	0.47	0.42	0.44	19
RBS	0.60	0.43	0.50	7
RP	0.60	0.50	0.55	52
SYM	0.00	0.00	0.00	1
TO	1.00	0.96	0.98	439
VB	0.89	0.84	0.86	496
VBD	0.82	0.75	0.78	653
VBG	0.57	0.41	0.48	275
VBN	0.64	0.69	0.66	445
VBP	0.81	0.70	0.75	227
VBZ	0.89	0.74	0.81	375
WDT	0.89	0.79	0.84	85
WP	0.97	0.84	0.90	45
WP$	1.00	1.00	1.00	1
WRB	0.96	0.84	0.90	31
accuracy			0.80	16321
macro avg	0.73	0.67	0.69	16321
weighted avg	0.81	0.80	0.80	16321

Overall accuracy is 80%

2. Using 4 POS tags

	precision	recall	f1-score	support
A	0.93	0.53	0.67	1801
N	0.74	0.98	0.84	6306
O	0.99	0.91	0.95	5743
V	0.91	0.63	0.75	2471
accuracy			0.85	16321
macro avg	0.89	0.76	0.80	16321
weighted avg	0.88	0.85	0.85	16321

Overall accuracy is 85%

Observation: The overall POS tagging accuracy for 4 POS tag setting is better than the 36 POS tag setting. For Approach 1, we get 12% improvement and for approach 2 we get 5% improvement.

The reason for this improvement is that, in case of the 4 tags POS tagging, ambiguities are lesser as compared to 36 tags. For the 36 tags setting more transition probabilities and emission probabilities are calculated. This increases the scope of error as at each step there are more choices to take than in the 4-step setting. Due to this the rate of committing error is higher in the 36-tag setting. Also, since number of tags are higher for the 36 tags setting, more amount of data will be required to correctly capture the right transition and emission probability distributions for it to perform as good as the 4-tag setting.