University of St Andrews

Practical 1

POS Tagging and Smoothing

Student ID: 120022067

University of St Andrews

CS5012 Language and Computation

Contents

1 Introduction

The aim of this practical is to develop a first-order HMM (Hidden Markov Model) for POS (part of speech) tagging in Python. The corpora used in the experiment are Conll2000 (default and universal tagset), Treebank (default and universal tagset), and Brown (universal tagset).

1.1 Features Implemented

1. Training of the HMM using relative frequency estimation, with handling of unknown words and Laplace smoothing.
2. Computation of the most probable sequence of tags for a given (untagged) sentence, according to the trained HMM.
3. Accuracy measurement of the tagged sentence.
4. Good-Turing smoothing

2 Design

2.1 Training and Testing Data

When the HMM is initialised, the training and testing sentences are created. Training data consists of 90% of tagged sentences in the corpus, while the testing data has a maximum of 500 untagged sentences from the corpus. The number of testing sentences was capped at 500 to make sure that the POS tagging program does not take too long to run.

2.2 Transition Probability Table

According to Jurafsky & Martin (2009), The transition probability of a tag with Laplace smoothing calculated by:

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_i)} \tag{1}$$

2.2.1 Laplace Smoothing

Transition probability table is stored in a dictionary. The code used to create transition probability table can be found in Listing 1.

```
for row in transition_prob.keys():
  for col in transition_prob[row].keys():
    transition_prob[row][col] = (1.0 * transition_count[col][row] + 1) / \
      (self.tags_dist[col] + len(self.tags))
return transition_prob
```

Listing 1: The code used to create transition probability table with Lapplace smoothing.

2.2.2 Good-Turing Smoothing

The code used to create transition probability table can be found in Listing 2.

```
for row in transition_prob.keys():
for col in transition_prob[row].keys():
  if transition_count[row][col] == 0:
    transition_prob[row][col] = (self.get_nc(1, linreg) * 1.0) / N
  else:
    c_star = (transition_count[row][col] + 1) * \
    (self.get_nc(transition_count[row][col] + 1, \
    linreg)/self.get_nc(transition_count[row][col], linreg))
    transition_prob[row][col] = c_star / N
return transition_prob
```

Listing 2: The code used to create transition probability table with Good-Turing smoothing.

2.3 Emission Probability Table

According to Jurafsky & Martin (2009), The emission probability of a word given a tag can be calculated by:

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)} \tag{2}$$

The code used to create transition probability table can be found in Listing 2.

```
for row in emission_prob.keys():
  for col in emission_prob[row].keys():
      emission_prob[row][col] = (1.0 * emission_count[row][col]) / self.
   tags_dist[row]
return emission_prob
```

Listing 3: The code used to create emission probability table.

2.4 Testing

A maximum of 500 sentences are used to test the POS tagging with Viterbi algorithm. For corpora with universal tagset, the confusion matrix will be printed at the end of the test. Otherwise, a three columns summary will be printed. This is because the confusion matrix won't fit in the terminal if there are too many tags.

3 Results

Corpus	Tagset	Smoothing	Accuracy
conll2000	default	laplace	88.63%
conll2000	default	good-turing	84.99%
conll2000	universal	laplace	90.06%
conll2000	universal	good-turing	86.05%
treebank	default	laplace	86.56%
treebank	default	good-turing	83.51%
treebank	universal	laplace	89.27%
treebank	universal	good-turing	85.26%
brown	universal	laplace	88.56%
brown	universal	good-turing	83.97%

Table 1: Experiment results

3.1 Conll2000

```
Tag:   DT        Correct:  996/1005      Accuracy: 99.10
Tag:   NN        Correct: 1541/1636      Accuracy: 94.19
Tag:  VBZ        Correct:  227/ 253      Accuracy: 89.72
Tag:   IN        Correct: 1234/1245      Accuracy: 99.12
Tag:  NNS        Correct:  651/ 781      Accuracy: 83.35
Tag:  VBD        Correct:  319/ 349      Accuracy: 91.40
Tag:   RB        Correct:  310/ 361      Accuracy: 85.87
Tag:  VBN        Correct:  234/ 287      Accuracy: 81.53
Tag:  NNP        Correct:  999/1122      Accuracy: 89.04
Tag:    .        Correct:   42/ 486      Accuracy: 8.64
Tag:  PRP        Correct:  202/ 202      Accuracy: 100.00
Tag:   MD        Correct:  126/ 127      Accuracy: 99.21
Tag:   VB        Correct:  335/ 376      Accuracy: 89.10
Tag:   JJ        Correct:  688/ 768      Accuracy: 89.58
Tag:  VBG        Correct:  130/ 195      Accuracy: 66.67
Tag:   CD        Correct:  373/ 427      Accuracy: 87.35
Tag:    ,        Correct:  525/ 525      Accuracy: 100.00
Tag:    :        Correct:   52/  67      Accuracy: 77.61
Tag:   ``        Correct:  100/ 100      Accuracy: 100.00
Tag:   CC        Correct:  298/ 298      Accuracy: 100.00
Tag: PRP$        Correct:   67/  67      Accuracy: 100.00
Tag:   ''        Correct:   61/  97      Accuracy: 62.89
Tag:    $        Correct:   84/  84      Accuracy: 100.00
Tag:   TO        Correct:  322/ 322      Accuracy: 100.00
Tag:    (        Correct:   14/  14      Accuracy: 100.00
Tag:    )        Correct:   14/  14      Accuracy: 100.00
Tag:   EX        Correct:   14/  18      Accuracy: 77.78
Tag:  POS        Correct:   90/  96      Accuracy: 93.75
Tag:  WDT        Correct:   39/  43      Accuracy: 90.70
Tag:   WP        Correct:   35/  35      Accuracy: 100.00
Tag:  VBP        Correct:  132/ 157      Accuracy: 84.08
Tag:  JJS        Correct:   12/  14      Accuracy: 85.71
Tag:  JJR        Correct:   46/  48      Accuracy: 95.83
Tag:  WRB        Correct:   28/  28      Accuracy: 100.00
Tag:  RBR        Correct:   11/  15      Accuracy: 73.33
Tag:  RBS        Correct:   20/  21      Accuracy: 95.24
Tag: NNPS        Correct:   21/  38      Accuracy: 55.26
Tag:   RP        Correct:    0/   5      Accuracy: 0.00
Tag:  PDT        Correct:    1/   1      Accuracy: 100.00
Tag:  WP$        Correct:    1/   1      Accuracy: 100.00
Tag:   FW        Correct:    1/   1      Accuracy: 100.00

overall accuracy, correct: 10395/11729 percentage: 88.63
```

Figure 1: Experiment result of conll2000 corpus with Laplace smoothing

```
Tag:   DT        Correct:  981/1005      Accuracy: 97.61
Tag:   NN        Correct: 1367/1636      Accuracy: 83.56
Tag:  VBZ        Correct:  212/ 253      Accuracy: 83.79
Tag:   IN        Correct: 1156/1245      Accuracy: 92.85
Tag:  NNS        Correct:  643/ 781      Accuracy: 82.33
Tag:  VBD        Correct:  269/ 349      Accuracy: 77.08
Tag:   RB        Correct:  302/ 361      Accuracy: 83.66
Tag:  VBN        Correct:  182/ 287      Accuracy: 63.41
Tag:  NNP        Correct:  991/1122      Accuracy: 88.32
Tag:    .        Correct:  486/ 486      Accuracy: 100.00
Tag:  PRP        Correct:  202/ 202      Accuracy: 100.00
Tag:   MD        Correct:  114/ 127      Accuracy: 89.76
Tag:   VB        Correct:  166/ 376      Accuracy: 44.15
Tag:   JJ        Correct:  521/ 768      Accuracy: 67.84
Tag:  VBG        Correct:  119/ 195      Accuracy: 61.03
Tag:   CD        Correct:  340/ 427      Accuracy: 79.63
Tag:    ,        Correct:  525/ 525      Accuracy: 100.00
Tag:    :        Correct:   52/  67      Accuracy: 77.61
Tag:   ``        Correct:  100/ 100      Accuracy: 100.00
Tag:   CC        Correct:  298/ 298      Accuracy: 100.00
Tag: PRP$        Correct:   67/  67      Accuracy: 100.00
Tag:   ''        Correct:   59/  97      Accuracy: 60.82
Tag:    $        Correct:   84/  84      Accuracy: 100.00
Tag:   TO        Correct:  322/ 322      Accuracy: 100.00
Tag:    (        Correct:   14/  14      Accuracy: 100.00
Tag:    )        Correct:   14/  14      Accuracy: 100.00
Tag:   EX        Correct:   10/  18      Accuracy: 55.56
Tag:  POS        Correct:   95/  96      Accuracy: 98.96
Tag:  WDT        Correct:   21/  43      Accuracy: 48.84
Tag:   WP        Correct:   35/  35      Accuracy: 100.00
Tag:  VBP        Correct:   80/ 157      Accuracy: 50.96
Tag:  JJS        Correct:   12/  14      Accuracy: 85.71
Tag:  JJR        Correct:   43/  48      Accuracy: 89.58
Tag:  WRB        Correct:   28/  28      Accuracy: 100.00
Tag:  RBR        Correct:   10/  15      Accuracy: 66.67
Tag:  RBS        Correct:   20/  21      Accuracy: 95.24
Tag: NNPS        Correct:   25/  38      Accuracy: 65.79
Tag:   RP        Correct:    0/   5      Accuracy: 0.00
Tag:  PDT        Correct:    1/   1      Accuracy: 100.00
Tag:  WP$        Correct:    1/   1      Accuracy: 100.00
Tag:   FW        Correct:    1/   1      Accuracy: 100.00

overall accuracy, correct: 9968/11729 percentage: 84.99
```

Figure 2: Experiment result of conll2000 corpus with Good-Turing smoothing

3.2 Conll2000 (Universal Tagset)

	DET	NOUN	VERB	ADP	ADV	.	PRON	ADJ	NUM	CONJ	PRT	X	Total	Accuracy
DET	1038	0	0	26	3	0	0	0	0	0	0	0	1067	97.28
NOUN	0	3527	27	0	1	0	0	17	0	0	0	0	3577	98.60
VERB	0	229	1508	0	0	0	0	6	0	0	1	0	1744	86.47
ADP	17	4	0	1223	0	0	0	0	0	0	0	0	1245	98.23
ADV	4	22	2	22	356	0	0	19	0	0	0	0	425	83.76
.	0	0	0	0	0	892	0	0	0	0	1	0	1387	64.31
PRON	0	0	0	0	0	0	305	0	0	0	0	0	305	100.00
ADJ	0	179	9	0	8	0	0	634	0	0	0	0	830	76.39
NUM	0	60	4	0	0	0	0	0	363	0	0	0	427	85.01
CONJ	0	0	0	0	0	0	0	0	0	298	0	0	298	100.00
PRT	0	0	0	2	3	0	0	0	0	0	418	0	423	98.82
X	0	0	0	0	0	0	0	0	0	0	0	1	1	100.00

overall accuracy, correct: 10563/11729 percentage: 90.06

Figure 3: Experiment result of conll2000 corpus (universal tagset) with Laplace smoothing

	DET	NOUN	VERB	ADP	ADV	.	PRON	ADJ	NUM	CONJ	PRT	X	Total	Accuracy
DET	1032	0	0	31	4	0	0	0	0	0	0	0	1067	96.72
NOUN	0	3306	246	1	0	0	0	13	6	0	0	0	3577	92.42
VERB	0	271	1465	0	0	0	0	6	0	0	2	0	1744	84.00
ADP	33	13	3	1184	8	0	0	1	0	0	2	0	1245	95.10
ADV	5	20	2	22	350	0	0	25	1	0	0	0	425	82.35
.	0	0	0	0	0	892	0	0	0	0	1	0	1387	64.31
PRON	0	0	0	0	0	0	305	0	0	0	0	0	305	100.00
ADJ	0	217	73	0	31	0	0	509	0	0	0	0	830	61.33
NUM	0	81	9	0	0	0	1	0	336	0	0	0	427	78.69
CONJ	0	0	0	0	0	0	0	0	0	298	0	0	298	100.00
PRT	0	0	3	2	2	0	0	1	0	0	415	0	423	98.11
X	0	0	0	0	0	0	0	0	0	0	0	1	1	100.00

overall accuracy, correct: 10093/11729 percentage: 86.05

Figure 4: Experiment result of conll2000 corpus (universal tagset) with Good-Turing smoothing

3.3 Treebank

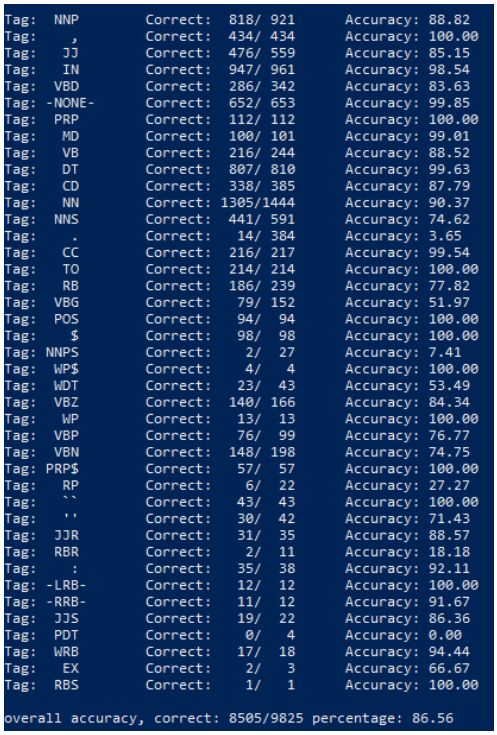

Figure 5: Experiment result of treebank corpus with Laplace smoothing

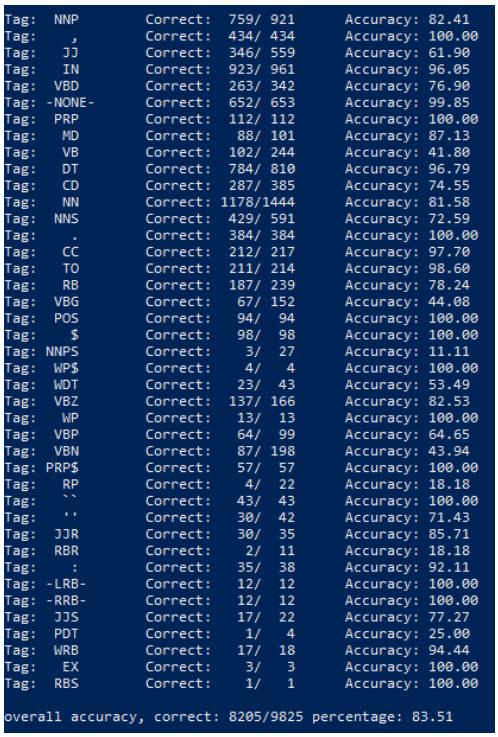

Figure 6: Experiment result of treebank corpus with Good-Turing smoothing

3.4 Treebank (Universal Tagset)

	NOUN	.	ADJ	ADP	VERB	X	PRON	DET	NUM	CONJ	PRT	ADV	Total	Accuracy
NOUN	2925	0	18	0	33	0	0	3	0	0	0	0	2983	98.06
.	0	677	0	0	0	0	0	0	0	0	0	0	1063	63.69
ADJ	151	0	435	3	8	0	0	0	0	0	0	19	616	70.62
ADP	2	0	0	936	2	0	0	14	0	0	0	7	961	97.40
VERB	191	0	15	0	1096	0	0	0	0	0	0	0	1302	84.18
X	1	0	0	0	0	652	0	0	0	0	0	0	653	99.85
PRON	0	0	0	0	0	0	186	0	0	0	0	0	186	100.00
DET	0	0	2	20	0	0	0	837	0	0	0	1	860	97.33
NUM	58	0	2	0	33	0	0	0	291	0	0	0	385	75.58
CONJ	2	0	0	0	0	0	0	0	0	215	0	0	217	99.08
PRT	2	1	0	13	0	0	0	0	0	0	308	6	330	93.33
ADV	18	0	13	18	5	0	0	2	0	0	0	213	269	79.18

overall accuracy, correct: 8771/9825 percentage: 89.27

Figure 7: Experiment result of treebank corpus (universal tagset) with Laplace smoothing

	NOUN	.	ADJ	ADP	VERB	X	PRON	DET	NUM	CONJ	PRT	ADV	Total	Accuracy
NOUN	2703	0	11	0	260	0	0	3	2	0	0	0	2983	90.61
.	0	677	0	0	0	0	0	0	0	0	0	0	1063	63.69
ADJ	149	0	351	6	73	0	0	0	0	0	0	37	616	56.98
ADP	4	0	1	926	1	0	0	14	0	0	4	11	961	96.36
VERB	217	0	7	0	1077	0	0	0	0	0	1	0	1302	82.72
X	1	0	0	0	0	652	0	0	0	0	0	0	653	99.85
PRON	0	0	0	0	0	0	186	0	0	0	0	0	186	100.00
DET	10	0	2	23	0	1	0	823	0	0	0	1	860	95.70
NUM	125	0	0	0	3	3	0	0	253	0	0	0	385	65.71
CONJ	4	0	3	0	0	0	0	0	0	210	0	0	217	96.77
PRT	2	1	1	6	3	0	0	0	0	0	313	4	330	94.85
ADV	18	0	21	13	4	0	0	2	0	0	4	206	269	76.58

overall accuracy, correct: 8377/9825 percentage: 85.26

Figure 8: Experiment result of treebank corpus (universal tagset) with Good-Turing smoothing

3.5 Brown (Universal Tagset)

	PRON	VERB	ADV	NUM	NOUN	ADJ	.	DET	PRT	ADP	CONJ	X	Total	Accuracy
PRON	541	0	0	0	0	0	0	21	0	19	0	0	581	93.12
VERB	0	1403	0	0	107	2	0	0	1	4	0	0	1517	92.49
ADV	0	3	366	0	22	18	0	2	3	10	2	0	426	85.92
NUM	0	0	0	32	6	0	0	0	0	0	0	0	38	84.21
NOUN	0	15	0	0	1663	8	0	0	0	0	0	0	1686	98.64
ADJ	0	3	13	0	66	399	0	0	0	0	0	0	481	82.95
.	0	0	0	0	0	0	833	0	0	0	0	0	1333	62.49
DET	9	0	0	0	0	0	0	960	0	14	0	0	983	97.66
PRT	0	2	11	0	20	3	0	0	200	76	0	0	312	64.10
ADP	0	0	7	0	1	0	0	1	8	838	2	0	857	97.78
CONJ	0	0	0	0	0	0	0	0	0	0	352	0	352	100.00
X	0	0	0	0	1	0	0	0	0	0	0	0	1	0.00

overall accuracy, correct: 7587/8567 percentage: 88.56

Figure 9: Experiment result of brown corpus (universal tagset) with Laplace smoothing

	PRON	VERB	ADV	NUM	NOUN	ADJ	.	DET	PRT	ADP	CONJ	X	Total	Accuracy
PRON	552	0	0	0	0	0	0	10	0	19	0	0	581	95.01
VERB	0	1362	0	0	144	6	0	0	1	4	0	0	1517	89.78
ADV	0	3	363	0	19	19	0	1	6	12	3	0	426	85.21
NUM	0	1	0	24	13	0	0	0	0	0	0	0	38	63.16
NOUN	0	117	4	1	1554	6	0	0	0	4	0	0	1686	92.17
ADJ	0	49	48	0	81	302	0	0	0	1	0	0	481	62.79
.	0	0	0	0	0	0	833	0	0	0	0	0	1333	62.49
DET	41	0	3	0	2	0	0	912	0	25	0	0	983	92.78
PRT	0	2	13	0	20	2	0	0	200	75	0	0	312	64.10
ADP	6	9	44	0	4	4	0	0	43	744	3	0	857	86.81
CONJ	0	0	3	0	0	0	0	1	0	0	348	0	352	98.86
X	0	0	0	0	1	0	0	0	0	0	0	0	1	0.00

overall accuracy, correct: 7194/8567 percentage: 83.97

Figure 10: Experiment result of brown corpus (universal tagset) with Good-Turing smoothing

4 Evaluation

Overall, POS tagging program's performance is at satisfactory level. The accuracy varies from 83% to 90%. It appears that the model achieves higher accuracy with smaller size tagsets. Due to time constraints, it is not possible to explore other smoothing techniques that may help improve the accuracy even further.

Bibliography

Jurafsky, D. & Martin, J. H. (2009), *Speech and language processing*, Vol. 2, Pearson.