Natural Language Processing:

Assignment 5: Qu' bopbe' paqvam

Jordan Boyd-Graber

Submitted by Reed Anderson

1 Tagging and Tag Sets (10 points)

1.1 When taggers go bad (5 points)

Consider the following sentences: British Left Waffles on Falkland Islands

1. British Left Waffles on Falkland Islands

- British/NOUN Left/NOUN Waffles/VERB on/PREPOSITION Falkland/NOUN Islands/NOUN
- British/NOUN Left/VERB Waffles/NOUN on/PREPOSITION Falkland/NOUN Islands/NOUN

1.2 Exploring the tag set (5 points)

There are 265 distinct words in the Brown Corpus having exactly four possible tags (assuming nothing is done to normalize the word forms).

1. Create a table with the integers 1...10 in one column, and the number of distinct words in the corpus having $\{1, \ldots, 10\}$ distinct tags.

Tag Count	Words with Tag Count
1	40235
2	7229
3	1594
4	463
5	176
6	75
7	23
8	10
9	5
10	2

2. For the word with the greatest number of distinct tags, print out sentences from the corpus containing the word, one for each possible tag.

```
TAG: (’to’, ’TO’)
[(’The’, ’AT’), (’September-October’, ’NP’), (’term’, ’NN’), (’
    jury’, ’NN’), (’had’, ’HVD’), (’been’, ’BEN’),q (’charged’,
     ’VBN’), (’by’, ’IN’), (’Fulton’, ’NP-TL’), (’Superior’, ’
    JJ-TL’), (’Court’, ’NN-TL’), (’Judge’, ’NN-TL’), (’Durwood’
    , ’NP’), (’Pye’, ’NP’), (’to’, ’TO’), (’investigate’, ’VB’)
    , (’reports’, ’NNS’), (’of’, ’IN’), (’possible’, ’JJ’), (’
    ‘‘’, ’‘‘’), (’irregularities’, ’NNS’), ("’’", "’’"), (’in’,
     ’IN’), (’the’, ’AT’), (’hard-fought’, ’JJ’), (’primary’, ’
    NN’), (’which’, ’WDT’), (’was’, ’BEDZ’), (’won’, ’VBN’), (’
    by’, ’IN’), (’Mayor-nominate’, ’NN-TL’), (’Ivan’, ’NP’), (’
    Allen’, ’NP’), (’Jr.’, ’NP’), (’.’, ’.’)]

TAG: (’to’, ’IN’)
[(’It’, ’PPS’), (’recommended’, ’VBD’), (’that’, ’CS’), (’
    Fulton’, ’NP’), (’legislators’, ’NNS’), (’act’, ’VB’), (’‘‘
    ’, ’‘‘’), (’to’, ’TO’), (’have’, ’HV’), (’these’, ’DTS’), (
    ’laws’, ’NNS’), (’studied’, ’VBN’), (’and’, ’CC’), (’
    revised’, ’VBN’), (’to’, ’IN’), (’the’, ’AT’), (’end’, ’NN’
    ), (’of’, ’IN’), (’modernizing’, ’VBG’), (’and’, ’CC’), (’
    improving’, ’VBG’), (’them’, ’PPO’), ("’’", "’’"), (’.’, ’.
    ’)]

TAG: (’to’, ’IN-HL’)
[(’Cost’, ’NN-HL’), (’up’, ’RP-HL’), (’to’, ’IN-HL’), (’$37’, ’
    NNS-HL’), (’a’, ’AT-HL’), (’year’, ’NN-HL’)]

TAG: (’to’, ’TO-HL’)
[(’Three’, ’CD-HL’), (’groups’, ’NNS-HL’), (’to’, ’TO-HL’), (’
    meet’, ’VB-HL’)]

TAG: (’to’, ’IN-TL’)
[(’On’, ’IN’), (’the’, ’AT’), (’clock’, ’NN’), (’given’, ’VBN’)
    , (’him’, ’PPO’), (’was’, ’BEDZ’), (’the’, ’AT’), (’
    inscription’, ’NN’), (’,’, ’,’), (’‘‘’, ’‘‘’), (’For’, ’IN-
    TL’), (’Outstanding’, ’JJ-TL’), (’Contribution’, ’NN-TL’),
    (’to’, ’IN-TL’), (’Billiken’, ’NP-TL’), (’Basketball’, ’NN-
    TL’), (’,’, ’,’), (’1960-61’, ’CD’), ("’’", "’’"), (’.’, ’.
    ’)]

TAG: (’to’, ’TO-NC’)
[(’Or’, ’CC’), (’an’, ’AT’), (’‘‘’, ’‘‘’), (’I’, ’PPSS-NC’), (’
    want’, ’VB-NC’), (’to’, ’TO-NC’), (’go’, ’VB-NC’), (’home’,
     ’NR-NC’), ("’’", "’’"), (’,’, ’,’), (’or’, ’CC’), (’
    whatever’, ’WDT’), (’--’, ’--’), (’but’, ’CC’), (’a’, ’AT’)
    , (’nonverbal’, ’JJ’), (’one’, ’CD’), (’which’, ’WDT’), (’
    reveals’, ’VBZ’), (’itself’, ’PPL’), (’,’, ’,’), (’
    gradually’, ’RB’), (’,’, ’,’), (’as’, ’CS’), (’the’, ’AT’),
     (’condensed’, ’VBN’), (’expression’, ’NN’), (’of’, ’IN’),
    (’more’, ’AP’), (’than’, ’IN’), (’one’, ’CD’), (’latent’, ’
    JJ’), (’meaning’, ’NN’), (’.’, ’.’)]

TAG: (’to’, ’IN-NC’)
[(’When’, ’WRB’), (’go’, ’VB-NC’), (’represents’, ’VBZ’), (’
    itself’, ’PPL’), (’and’, ’CC’), (’a’, ’AT’), (’complement’,
     ’NN’), (’(’, ’(’), (’being’, ’BEG’), (’equivalent’, ’JJ’),
     (’,’, ’,’), (’say’, ’UH’), (’,’, ’,’), (’to’, ’IN’), (’go’
    , ’VB-NC’), (’to’, ’IN-NC’), (’Martinique’, ’NP-NC’), (’)’,
     ’)’), (’in’, ’IN’), (’which’, ’WDT-NC’), (’boat’, ’NN-NC’)
    , (’did’, ’DOD-NC’), (’Jack’, ’NP-NC’), (’go’, ’VB-NC’), (’
    on’, ’IN-NC’), (’?’, ’.-NC’), (’?’, ’.-NC’)]

TAG: (’to’, ’NIL’)
[(’As’, ’CS’), (’the’, ’AT’), (’field’, ’NN’), (’on’, ’IN’), (’
    which’, ’WDT’), (’my’, ’PP$’), (’tent’, ’NN’), (’was’, ’
    BEDZ’), (’pitched’, ’VBN’), (’was’, ’BEDZ’), (’a’, ’AT’), (
    ’favorite’, ’JJ’), (’natural’, ’JJ’), (’playground’, ’NN’),
     (’for’, ’IN’), (’the’, ’AT’), (’kids’, ’NNS’), (’of’, ’IN’
    ), (’the’, ’AT’), (’neighborhood’, ’NIL’), (’,’, ’,’), (’I’
    , ’NIL’), (’had’, ’NIL’), (’made’, ’NIL’), (’many’, ’NIL’),
     (’friends’, ’NIL’), (’among’, ’NIL’), (’them’, ’NIL’), (’,
    ’, ’,’), (’taking’, ’NIL’), (’part’, ’NIL’), (’in’, ’NIL’),
     (’their’, ’NIL’), (’after-school’, ’NIL’), (’games’, ’NIL’
    ), (’and’, ’NIL’), (’trying’, ’NIL’), (’desperately’, ’NIL’
    ), (’to’, ’NIL’), (’translate’, ’NIL’), ("Grimm’s", ’NIL’),
     (’Fairy’, ’NIL’), (’Tales’, ’NIL’), (’into’, ’NIL’), (’an’
    , ’NIL’), (’understandable’, ’JJ’), (’French’, ’NP’), (’as’
    , ’CS’), (’we’, ’PPSS’), (’gathered’, ’VBD’), (’around’, ’
    IN’), (’the’, ’AT’), (’fire’, ’NN’), (’in’, ’IN’), (’front’
    , ’NN’), (’of’, ’IN’), (’the’, ’AT’), (’tent’, ’NN’), (’.’,
     ’.’)]

TAG: (’to’, ’NPS’)
[(’Also’, ’RB’), (’noted’, ’VBN’), (’are’, ’BER’), (’the’, ’AT’
    ), (’marriages’, ’NNS’), (’of’, ’IN’), (’Elizabeth’, ’NP’),
     (’Browning’, ’NP’), (’,’, ’,’), (’daughter’, ’NN’), (’of’,
     ’IN’), (’the’, ’AT’), (’George’, ’NP’), (’L.’, ’NP’), (’
    Brownings', 'NPS'), (',', ','), ('to', 'NPS'), ('Austin', '
    NP'), ('C.', 'NP'), ('Smith', 'NP'), ('Jr.', 'NP'), (';', '
    .'), (';', '.')]

TAG: ('to', 'QL')
[('He', 'PPS'), ('suggested', 'VBD'), ('offering', 'VBG'), ('
    half', 'NN'), ('to', 'IN'), ('Sir', 'NP'), \\('Edward', 'NP
    '), (',', ','), ('fearing', 'VBG'), ('lest', 'CS'), ('``',
    '``'), ('he', 'PPS'), ('shall', 'MD'), ('thinke', 'VB'), ('
    it', 'PPO'), ('to', 'QL'), ('good', 'JJ'), ('for', 'IN'), (
    'us', 'PPO'), ('and', 'CC'), ('procure', 'VB'), ('it', 'PPO
    '), ('for', 'IN'), ('himselfe', 'PPL'), (',', ','), ('as',
    'CS'), ('he', 'PPS'), ('served', 'VBD'), ('us', 'PPO'), ('
    the', 'AT'), ('last', 'AP'), ('time', 'NN'), ("''", "''"),
    ('.', '.')]
```

2 Viterbi Algorithm (30 Points)

2.1 Emission Probability (10 points)

The below table represents β as the emission probabilities, where $\beta_{noun,tera'ngan}$ is approximately equal to 0.44, or (4.1/9.3).

	NOUN	VERB	CONJ	PRO
'e	0.1	0.1	0.1	1.1
'eg	0.1	0.1	1.1	0.1
ghaH	0.1	0.1	0.1	1.1
ja'chuqmeH	0.1	1.1	0.1	0.1
legh	0.1	0.1	0.1	0.1
neH	0.1	1.1	0.1	0.1
pa'Daq	1.1	0.1	0.1	0.1
puq	2.1	0.1	0.1	0.1
qIp	0.1	2.1	0.1	0.1
rojHom	1.1	0.1	0.1	0.1
taH	0.1	1.1	0.1	0.1
tera'ngan	4.1	0.1	0.1	0.1
yaS	0.1	0.1	0.1	0.1
SUM	**9.3**	**6.3**	**2.3**	**3.3**

2.2 Start and Transition Probability (5 points)

The below table represents θ as the transition probabilities, where $\theta_{n,v}$ is approximately equal to 0.48, or (3.1/6.4).

	NOUN	VERB	CONJ	PRO	**SUM**
START					
N	0.1	3.1	1.1	2.1	**6.4**
V	5.1	0.1	0.1	0.1	**5.4**
CONJ	1.1	0.1	0.1	0.1	**3.4**
PRO	0.1	1.1	0.1	0.1	**1.3**

The below table represents π as the initial probabilities, where π_n is approximately equal to 0.62, or (2.1/3.4).

N	2.1
V	1.1
CONJ	0.1
PRO	0.1
SUM	**3.4**

2.3 Viterbi Decoding (15 points)

Now consider the following sentence: “tera’ngan legh yaS”.

1. Compute the probability of the sequence NOUN, VERB, NOUN.

   For “tera’ngan/NOUN legh/VERB yaS/NOUN”:

$$\pi_n \beta_{n,tera'ngan}\,\theta_{n,v}\,\beta_{v,legh}\,\theta_{v,n}\,\beta_{n,yaS} =$$
$$(2.1/3.4)(4.1/9.3) * (3.1/6.4)(0.1/6.3) * (5.1/5.4)(0.1/9.3) = 2.1e - 5$$

2. Create the decoding matrix of this sentence $\ln \delta_n(z)$ (word positions are columns, rows are parts of speech). Only provide log probabilities, and only use base 2.

POS	$n = 1$	$n = 2$	$n = 3$
$z =$N	-1.9	-14.2	-15.5
$z =$V	-7.6	-8.9	-14.7
$z =$CONJ	-9.6	-9.0	-16.7
$z =$PRO	-10.1	-8.5	-15.9

Evaluation of n = 1 :

$$\begin{aligned}
\delta_1(k) &= \pi_k \beta_{k,x_i} \\
\delta_1(n) &= \pi_n \beta_{n,tera'ngan} \\
&= lg((2.1/3.4)(4.1/9.3)) \\
&= -1.9
\end{aligned}$$

$$\begin{aligned}
\delta_1(v) &= \pi_v \beta_{v,tera'ngan} \\
&= -7.6
\end{aligned}$$

$$\begin{aligned}
\delta_1(c) &= \pi_c \beta_{c,tera'ngan} \\
&= -9.6
\end{aligned}$$

$$\begin{aligned}
\delta_1(p) &= \pi_p \beta_{p,tera'ngan} \\
&= -10.1
\end{aligned}$$

Evaluation of n = 2 :

$$\begin{aligned}
\delta_n(k) &= \max_j(\delta_{n-1}(j)\theta_{j,k})\beta_{k,x_n} \\
\delta_2(n) &= (-7.6 + lg(\theta_{v,n})) + \beta_{n,legh} \\
&= -7.7 + lg(0.1/9.3) \\
&= -14.2
\end{aligned}$$

$$\begin{aligned}
\delta_2(v) &= (-1.9 + lg(\theta_{n,v})) + \beta_{v,legh} \\
&= -2.9 + lg(0.1/6.3) \\
&= -8.9
\end{aligned}$$

$$\begin{aligned}
\delta_2(c) &= (-1.9 + lg(\theta_{n,c})) + \beta_{c,legh} \\
&= -4.4 + lg(0.1/2.3) \\
&= -9
\end{aligned}$$

$$\begin{aligned}
\delta_2(p) &= (-1.9 + lg(\theta_{n,p})) + \beta_{p,legh} \\
&= -3.5 + lg(0.1/3.3) \\
&= -8.5
\end{aligned}$$

Evaluation of n = 3 :

$$\delta_n(k) = \max_j(\delta_{n-1}(j)\theta_{j,k})\beta_{k,x_n}$$

$$\begin{aligned}\delta_3(n) &= (-8.9 + lg(\theta_{v,n})) + \beta_{n,yaS} \\ &= -8.9 + lg(0.1/9.3) \\ &= -15.5\end{aligned}$$

$$\begin{aligned}\delta_3(v) &= (-8.5 + lg(\theta_{p,v})) + \beta_{v,yaS} \\ &= -8.7 + lg(0.1/6.3) \\ &= -14.7\end{aligned}$$

$$\begin{aligned}\delta_3(c) &= (-8.5 + lg(\theta_{p,c})) + \beta_{c,yaS} \\ &= -12.2 + lg(0.1/2.3) \\ &= -16.7\end{aligned}$$

$$\begin{aligned}\delta_3(p) &= (-8.5 + lg(\theta_{p,p})) + \beta_{p,yaS} \\ &= -12.2 + lg(0.1/3.3) \\ &= -15.9\end{aligned}$$

3. What is the most likely sequence of parts of speech?

   - NOUN PRO VERB

4. Let's compare this to the probability of your previous answer.

   (a) How does this compare to the sequence NOUN, VERB, NOUN?

   - When we reconstruct the sequence with viterbi decoding, we see a $\log_2$ probability of -14.7 for VERB at Position 3 and -8.5 PRO at Position 2, making them the most probabile sequence. For NOUN at Position 3 (-15.5), and VERB at Position 2 (-8.9), these are the second most probable outcomes.

   (b) Which is more plausible linguistically?

   - The NVN sequence seems the most probable for known words; however, given transition probablities and the limited number of PRO and CONJ POS seen so far (i.e., considering the probability that a PRO will most often preceed a VERB and that the occurances of PRO following NOUN are not dramatically less probable than a VERB following a NOUN) the NPV sequence seems plausible for new data.

(c) Does an HMM model encode the intuition that you used to answer the previous question?

- Yes. One of the highest probabilities supporting the NPV model is the transition probability from PRO to VERB (0.84), and the comparatively high transition probabilties from NOUN to both PRO and VERB; and, because the training data has such few CONJ and PRO, new words such as legh and yaS thus receive high probabilities when evaulated as unknowns for those POS.

5. (For fun, not for credit) What do you think this sentence means? What word is the subject of the sentence?

N	PRO	V
tera'ngan	legh	yaS
human	she	eat

She is eating the human