C11 - 017
Group 1

THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

# NLP written Assignment

1] Book the flight

S → NP/VP

S → VP

NP → Det/N

NP → Det/Adj/N

VP → V

VP → V/NP

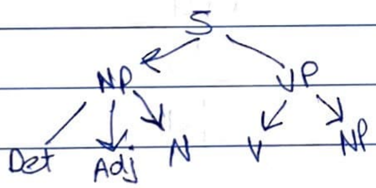
Ans

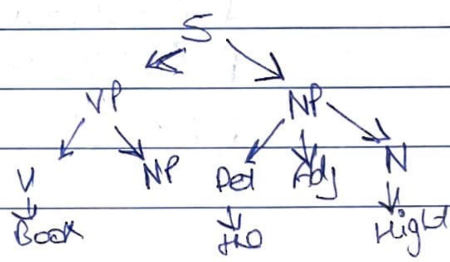
Det → the

Vb → Book

N → flight

Rearranging




2) Given Corpus

<S> I am Henry </S>

<S> I like college </S>

<S> Do Henry like college </S>

<S> Do I like Henry </S>

<S> Do I like college </S>

<S> I do like Henry </S>

1) I like College

2) Do I like Henry

| Word | Frequency |
|---|---|
| <s> | 7 |
| </s> | 7 |
| I | 6 |
| am | 2 |
| Henry | 5 |
| like | 5 |
| college | 3 |
| do | 4 |

Next word prediction probability $W_{i-1} = do$

| Next word | Probability Next word |
|---|---|
| P (</s / do) | 0/4 |
| P (< I >/ do) | 2/4 |
| P (<am>/ do) | 0/4 |
| P (< Henry / do) | 1/4 |
| P (<like >/ do) | 1/4 |
| P (<college >/ do) | 0/4 |
| P (do / do) | 0/4 |

Next word prediction Probability $W_{i-1}$ = Henry

| Next word | Probability next word = N/D |
|---|---|
| P (</s>/ Henry) | 3/5 |
| P (<I>/ Henry) | 1/5 |
| P (<am> Henry) | 0 |
| P (< Henry> Henry) | 0 |
| P (< like > Henry) | 1/5 |
| P (<college> Henry) | 0 |
| P (< DO > Henry) | 0 |

<15) Do I Like

use Trigram

P <I like> = 5    Next word $w_{i-2}$ = I and $w_{i-1}$ = like

| Next word | Probability Next word = count ($w_{i-2} w_{i-1} w_i$) |
|---|---|
| P (<15> I like) | 0/3 |
| P (<I> I like) | 0/3 |
| P (<am> / I like) | 0/3 |
| P (<Henry> I like) | 1/3 |
| P (<like / like) | 0/3 |
| P (<college / I like) | 2/3 |
| P (do / I like) | 0/3 |

4) <S>> Do I like college?

use forgram

$w_{i-3}$ = I    $w_{i-2}$ = like    $w_{i-1}$ = college

| Next word | Probability = $\frac{\text{Count}(w_{i-3} w_{i-2} w_{i-1} w_i)}{\text{Count}(w_{i-3}, w_{i-2}, w_{i-1})}$ |
|---|---|
| P (<15> / I like college) | 2/2 |
| P (<I> / I like college) | 0/2 |
| P (<am> / I like college) | 0/2 |
| P (<Henry> / I like college) | 0/2 |
| P (< like> / I like college) | 0/2 |
| P (< College> / I like college) | 0/2 |
| P (<do / I like college) | 0/2 |

1) <s> I like college </s>

$= P(I \mid <s>) \times P(like \mid I) \times P(college \mid like) \times P(</s> \mid college)$

$= \frac{3}{7} \times \frac{3}{6} \times \frac{3}{5} \times \frac{3}{3} - \frac{9}{70} = 0.13$

2) <s> Do I like Henry </s>

$P(do \mid <s>) \times P(I \mid do) \times P(like \mid I) \times P(Henry \mid like)$
$\times P(</s> \mid Henry)$

$= \frac{3}{7} \times \frac{2}{4} \times \frac{3}{6} \times \frac{2}{5} \times \frac{3}{5} = \frac{9}{350} = 0.0257$

Q3

<s> the student pass the test </s>
<s> the student wait for the pass </s>
<s> teachers test students </s>

Bigrams are
(the | <s>), (the | student), (student pass), (pass the),
(the | test), (test </s>, (<s>, the), (the | student), (student | wait)
(wait | for) (for line), (the pass) (pass | </s>), (<s> teachers)
(teachers | test) (test students), (students | </s>)

$P(the \mid student) = P(student \mid the) \mid P(student)$
$= 0$

$P(the \mid <s>) = P(<s> \mid the) \mid P(<s>)$
$= 2 \mid 3$

$P(student \mid the) = P(the \mid student) \mid P(the)$
$= 2 \mid 4$

$P(<\!/s\!>|\,test) = P(test\,|<\!s\!>)/P(test)$
$= 1/2$

$P(student\,|\,test)$ count (Test, student) / count (test)
$= 1/2$

Spelling correction using N-gram is a simple & effective approach to address misspelled words in a text. N-grams are used to identify & suggest correction for misspelled words by comparing them to correctly spelled words for language model

1) Building an n gram model :- first we need to arrage a corpus of text with correctly spelled words to build an n-gram language model.

2) Tokenization into N-germ :- The test in the corpus is tokenized into N-grams (typically bigrams or trigram)

3) Counting N-grams :- For each N-gram, you count the number of times it appears in the corpus

4) Spelling correction :- To correct a missspelled word, you tokenize the words into N-gram & compare these N-grams to N-grams in your language model & calculate the likelihood probability for each N-gram from misspelled, you consider possible correct alternative

Q4

<s> the/DT students/NN pass/V the/DT test/NN </s>
<s> the/DT students/NN wait/V for/P the/DT pass/NN </s>
<s> the/DT students/NN </s>
<s> teachers/N test/V students/NN </s>

Emission probability

| Words | Noun | Verb | Model | Preposition | DT |
|---|---|---|---|---|---|
| The | 0 | 0 | 4/4=1 | 0 | |
| Student | 3/5=0.6 | 0 | 0 | 0 | |
| Pass | 0 | 1/3=0.3 | 0 | 0 | |
| wait | 0 | 1/3=0.3 | 0 | 0 | |
| for | 0 | 0 | 0 | 1/1=1 | |
| teachers | 1/5=0.2 | 0 | 0 | 0 | |
| test | 1/5=0.2 | 1/3=0.3 | 0 | 0 | |

| | N | V | M | P | </s> |
|---|---|---|---|---|---|
| <s> | 1 | 0 | 2 | 0 | 0 |
| N | 0 | 3 | 0 | 0 | 3 |
| V | 1 | 0 | 1 | 1 | 0 |
| M | 4 | 0 | 0 | 0 | 0 |
| P | 0 | 0 | 1 | 0 | 0 |

| | N | V | M | P | </s> |
|---|---|---|---|---|---|
| <s> | 1/3=0.3 | 0 | 2/3=0.6 | 0 | 0 |
| N | 0 | 3/6=0.5 | 0 | 0 | 3/6=0.5 |
| V | 1/3=0.3 | 0 | 1/3=0.3 | 1/3=0.3 | 0 |
| M | ~~0~~1 | 0 | 0 | 0 | 0 |
| P | 0 | 0 | 1 | 0 | 0 |

5

In the two sentences different meanings arise from different type ambigity

a) Time flies like an arrow.
The ambiguity in this sentence is an example of structural ambiguity. It can be represented in two different ways based on structural grouping of data

i) Time flies [i.e insects known as flies that exist over time] are similar to an arrow. In this like, an arrow is a similar, indicating comparison between time flies & arrow.

ii) Time (the concept of time passes quickly in a manner similar to an arrow" like an arrow is a simple indicating comparision between speed of time & arr

b) "He crushed the key to my heart
i) Semantic ambigity the word crushed is the source of semantic ambiguity. It can be interpreted in two different ways:
- literally as in physically crushing into pie
- figuratively emotional impact (crushed) it is emotional

6

a) Important to Bill

→ Important is an Adjective

→ to Bill → prepositional phrase

∴ The entire phrase is therefore an adjective phrase

b) "looked up the tree"

"Looked" is a verb.

"up the tree" is a prepositional phrase

∴ the entire segment is a verb phrase.

7. For the given corpus

N : Noun [Martin, Justin, will, spot, Pat]

M : Model [I, we, will]

V : Verb [warm, spot, Pat]

Statement : Justin will spot will.

Training corpus.

<S> Martin Justin can watch will <E>

<S> Spot will watch Martin <E>

<S> will Justin Spot Martin <E>

<S> Martin will spot <E>

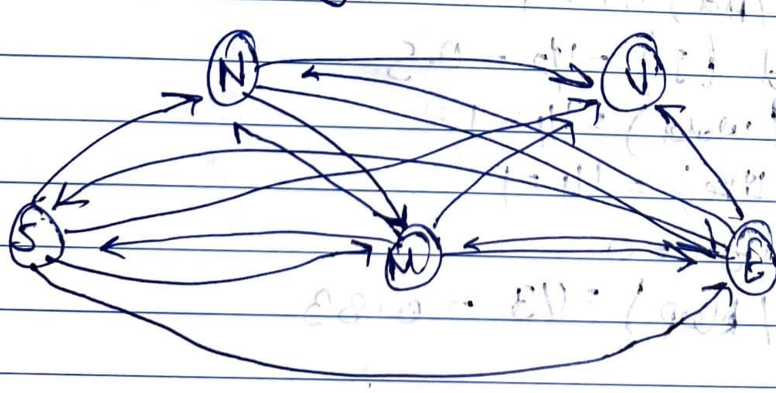
Emission Probability Matrix

| | N | M | V |
|---|---|---|---|
| Martin | 4/10 | 0 | 0 |
| Justin | 2/10 | 0 | 0 |
| can | 0 | 1/3 | 0 |
| watch | 0 | 0 | 2/4 |
| will | 2/10 | 2/3 | 0 |
| spot | 2/10 | 0 | 1/4 |
| pat | 0 | 0 | 1/4 |

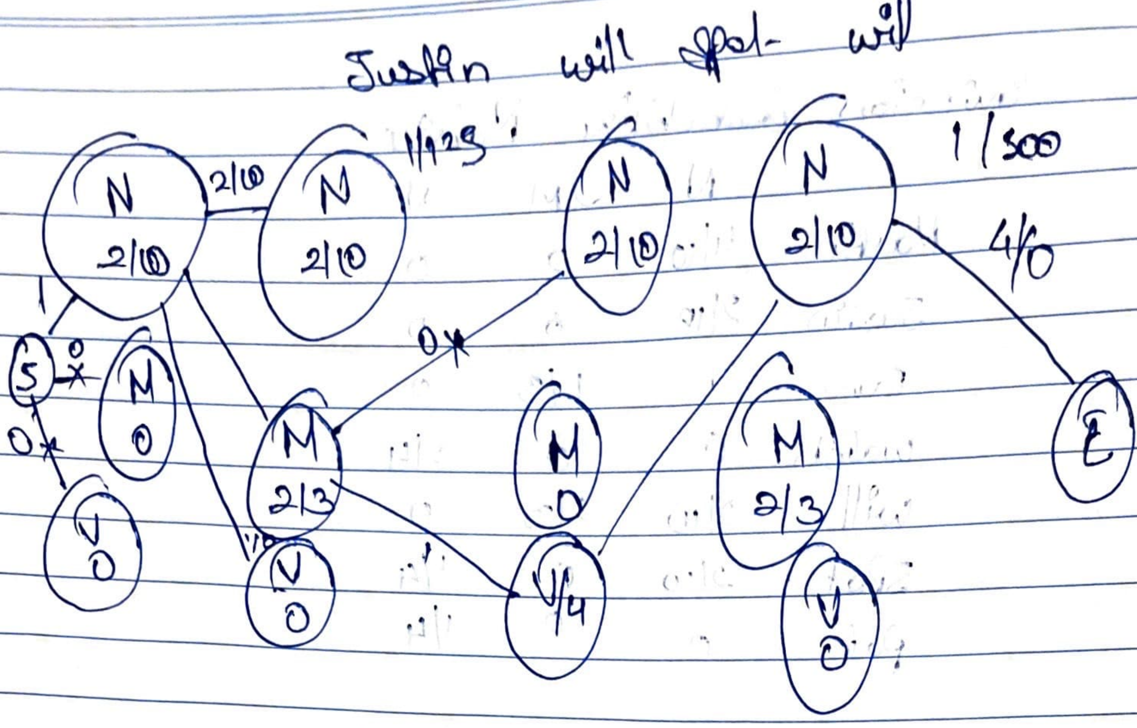
Transition probability Matrix

| | N | M | V | E |
|---|---|---|---|---|
| S | 4/4 | 0 | 0 | 0 |
| M | 2/10 | 3/10 | 1/10 | 4/10 |
| N | 0 | 0 | 5/3 | 0 |
| E | 4/4 | 0 | 0 | 0 |

Transition Probability




Martin Justin can watch will spot pat




S → N → M → V → N → E

Q8 <s> I tell you to sleep & rest </s>
<s> I would like to sleep for an hour </s>
<s> sleep helps one to relax </s>

P( I | <s> ) = count ( <s> I ) / count <s> = 2/3 = 0.66
P ( tell / I ) = 1/2 = 0.5
P ( you / tell ) = 1/1 = 1
P ( to / you ) 1/1 = 1
P ( would / I ) = 1/2 = 0.5
P ( like / would ) = 1/1 = 1
P ( to / like ) = 1/1 = 1
P ( sleep / to ) = 0
P ( for / sleep ) = 1/3 = 0.33

TSEC ENGINEERING COLLEGE

P( an | fox ) = 1/1 = 1
P ( hour | an ) = 1/1 = 1
P ( </s> | hour ) = 1/1 = 1
P ( helps | sleep ) = 1/3 = 0.3
P ( one | helps ) = 1/1 = 1
P ( to | one ) = 1/1 = 1
P ( relax | to ) = 1/3 = 0.33
P ( </s> | relax ) = 1/1 = 1

Highest => P( sleep | to ) = 0.26
∴ word is sleep

Q9) Emission probability:

| Words | Noun | Model | Verb |
|---|---|---|---|
| Martin | 4/9 | 0 | 0 |
| Justin | 2/9 | 0 | 0 |
| an | 0 | 1/4 | 0 |
| watch | 0 | 0 | 2/4 |
| with | 1/9 | 3/4 | 0 |
| spot | 2/9 | 0 | 1/4 |
| pat | 0 | 0 | 1/4 |

Q11)
<S> Mary Jane can see will <S>
<S> Spot will see Marry </S>
<S> will Jane Spot Marry </S>
<S> Marry will pat Spot </S>

Emission probability Matrix

| | N | M | V |
|---|---|---|---|
| Marry | 4/9 | 0 | 0 |
| Jane | 2/9 | 0 | 0 |
| Can | 0 | 1/4 | 0 |
| See | 0 | 0 | 2/4 |
| will | 1/9 | 3/4 | 0 |
| Spot | 2/9 | 0 | 1/4 |
| Pat | 0 | 0 | 1/4 |

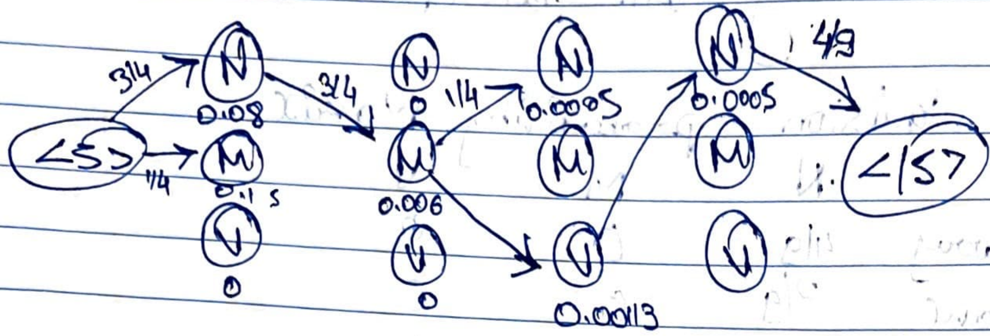
Transmission probability Matrix

| | N | M | V | </S> |
|---|---|---|---|---|
| <S> | 3/4 | 1/4 | 0 | 0 |
| N | 1/9 | 3/9 | 1/9 | 4/9 |
| M | 1/4 | 0 | 3/4 | 0 |
| V | 4/4 | 0 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| | $v_{11}$ N | $v_{21}$ N | $v_{31}$ N | $v_{41}$ N | |
| <S> | $v_{12}$ M | $v_{22}$ M | $v_{32}$ M | $v_{42}$ M | <E> |
| | $v_{13}$ V | $v_{23}$ N | $v_{33}$ N | $v_{43}$ V | |
| | will | can | spot | Marry | |

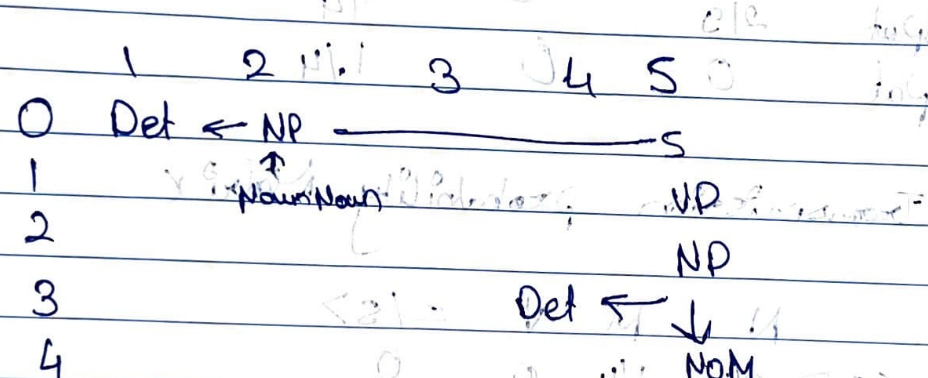
Q 12)

Using the transmission and emission probabilities from the previous numerals.

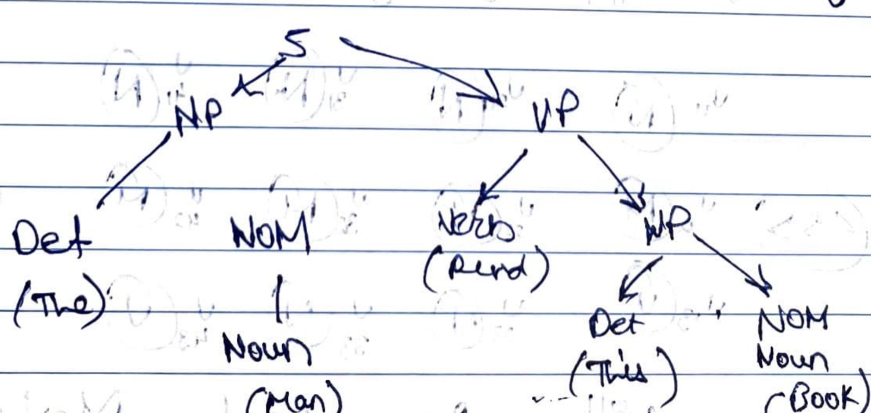



Q 13)

0 The 1 man 2 read 3 this 4 books 5




This can be converted to tree as follow.



THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Q14] i) Write, right → They are homophones, which means they sound the same but have unrelated meaning & spelling

ii) Big, large → These words are synonyms. They both mean having a greater size.

iii) Dark, light → These words are antonymic. Dark is opposite of light.

iv) Car, vehicle → These words have hypernym, homonym relationship. A car is a specific type of vehicle.

Q15)
→ Ans is c
(19|20)\d\d ?5 . [0[1-9]|(0-2?)-([0-2][1-9]|3[0-1])
This regular exp is similar is but further specifies that days can range form 0-31, which words all positive vehicles.

Q16) Holonym → Hyponym

"[ function " [W,] is a holonym of the wheels in sentence:
Hyponym - Hyponym

In second sentence 'drivers car" [W2] is a hyponym of companies.

(A)

15/10/24