NLP ETE

Sol 1:- (a) CKY Parsing Table:-

0 I 1 eat 2 sushi 3 with 4 chopstics 5 with 6 you 7

	1	2	3	4	5	6	7
0	NP	.	S		S	.	S
1		Verb	VP		VP	.	VP
2			NP	.	NP	.	NP
3				Prep	PP	.	PP
4					NP	.	NP
5						Prep	PP
6							NP

Sol 1 (b) CNF:- $\begin{cases} A \to \epsilon \\ A \to AA \\ A \to a \end{cases}$

Given, $S' \to S$ # add new production

$S \to ASB$

$A \to aASA \mid a \mid \epsilon$

$B \to SbS \mid A \mid bb$

(1) Remove ϵ productions

$S' \to S$

$S \to ASB \mid SB \mid AS \mid S$

$A \to aASA \mid a \mid aS \mid aAS \mid aSA$

$B \to SbS \mid bb \mid A$

(2) Remove unit products:

$S' \rightarrow ASB \mid SB \mid AS$

$S \rightarrow ASB \mid SB \mid AS$

$A \rightarrow aASA \mid a \mid aS \mid aAS \mid aSA$

$B \rightarrow SbS \mid bb \mid aASA \mid a \mid aS \mid aAS \mid aSA$

(3) Remove useless productions.

$S' \rightarrow XB \mid SB \mid AS$

$S \rightarrow XB \mid SB \mid AS$

$A \rightarrow ZA \mid a \mid YS \mid YX \mid UA$

$B \rightarrow MS \mid TT \mid ZA \mid a \mid XS \mid YX \mid UA$.

$X \rightarrow AS$

$Y \rightarrow a$

$Z \rightarrow YX$

$U \rightarrow YS$

$T \rightarrow b$

$M \rightarrow ST$

Sol 3(a):- Mean Reciprocal Rank Method:

Q No:-	1	2	3	4	5	6	7
Correct Ans No :	4	3	6	2	8	1	2

$$\text{Mean Reciprocal Rank} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

$$MRR = \frac{1}{7}\left[\frac{1}{4} + \frac{1}{3} + \frac{1}{6} + \frac{1}{2} + \frac{1}{8} + \frac{1}{1} + \frac{1}{2}\right] \Rightarrow \frac{1}{7}\left[\frac{69}{24}\right] = \frac{23}{56}$$

$\Rightarrow$ Ans

Sol 3(b) :- Rouge-2 score:

Human	System
Water Spinach ✓	Water spinach ✓
Spinach is ✓	Spinach is ✓
is a ✓	is a ✓
a commonly	a leaf
commonly eaten ✓	leaf vegetable ✓
eaten leaf	vegetable commonly
leaf vegetable ✓	commonly eaten ✓
vegetable of	eaten in
of Asia ✓	in tropical
	tropical areas
	areas of
	of Asia ✓

(Human Bigrams) = 9

(System Bigrams) = 12

(common bigrams) = 6

Recall = $\frac{6}{9}$ Precision = $\frac{6}{12}$.

$$f_1 \text{ score} = \frac{2PR}{R+P} = \frac{2 \times \frac{6}{12} \times \frac{6}{9}}{\frac{6}{9} + \frac{6}{12}}$$

Rouge 1 :- Pre Re f_1

$\frac{10}{13}$ $\frac{10}{10}$ $= \frac{2 \times \frac{10}{10} \times \frac{10}{13}}{1 + \frac{10}{13}} = \frac{4}{7}$.

$\Rightarrow \frac{20}{23}$

Sol 4:- Cosine Similarity:

	Doc1	Doc2	Doc 3	Doc4	Q
new	1	0	0	1	0
home	1	1	1	1	1
sales	1	1	1	1	1
top	1	0	0	0	0
forcasts	1	0	0	0	0
rise	0	1	0	1	0
in	0	1	2	0	0
july	0	1	1	1	0
increase	0	0	1	0	0
is	0	0	0	0	1
very	0	0	0	0	1
bad	0	0	0	0	1

Similarity Matrix

	D_1	D_2	D_3	D_4	Q
D_1	1	0.4	0.316	0.6	0.4
D_2		1	0.79	0.8	0.4
D_3			1	0.474	0.316
D_4				1	0.4
Q					1

Using Cosine Similarity = $\frac{D_1 \cdot D_2}{|D_1|\,|D_2|}$

⇒ formula, fill similarity matrix.

$\lambda = 0.3$

MMR = $\arg\max [\lambda \, Sim_1(D_i, Q) - (1-\lambda) \, Sim_2(D_i, D_j)]$

First iteration:-

S is empty set.

∵ max pairwise similarity within S = 0

So,

MMR = arg max (Sim (d_i, q)).

∵ d_1 has max similarity, pick any one d_2, d_4

S = {d_1}.

Second Iteration:-

find max distance of an element in S to given d_i,

sim (d_1, d_i)

for d_2,

sim (d_1, d_2) = 0.4

sim (d_2, q) = 0.4

MMR = $0.4 \times 0.3 - 0.6 \times 0.3$

= $0.12 - 0.18 = -0.06$

Similarly for d_3, d_4,

for d_3,

$$Sim(d_1, d_3) = 0.316$$

$$Sim(d_3, q) = 0.316$$

$$MMR = 0.3 \times 0.316 - 0.7 \times 0.316$$

$$= -0.1264$$

for d_4,

$$Sim(d_1, d_4) = 0.6$$

$$Sim(d_4, q) = 0.4$$

$$MMR = 0.3 \times 0.4 - 0.7 \times 0.6$$

$$= -0.3$$

Here, for d_2, max MMR is there,

So, $S = \{d_1, d_2\}$.

Third Iteration :-

$$\max \begin{cases} Sim(d_i, d_i) & \text{for 1st part} \\ Sim(d_i, d_2) & \text{for 2nd part} \end{cases}$$

$$\max \underbrace{(Sim(d_1, d_3)}_{0.316}, \underbrace{sim(d_2, d_3)}_{0.79}) = 0.79$$

$$sim(d_3, q) = 0.316$$

$$MMR = 0.3 \times 0.316 - 0.7 \times 0.79$$

$$= -0.4582$$

for d_4,

$$\max \{ \underset{0.6}{sim(d_1, d_4)} ; \underset{0.8}{sim(d_2, d_4)} \} = 0.8$$

$$Sim(d_4, q) = 0.4$$

$$MMR = 0.3 \times 0.4 - 0.7 \times 0.8$$

$$= -0.44.$$

MMR for d_4 is max &,

$$S = \{ d_1, d_2, d_4 \}.$$

These are 3 sentences in summary set.

Sol 6:- Hidden Markov Model → HMM.

· Initial Probability :-

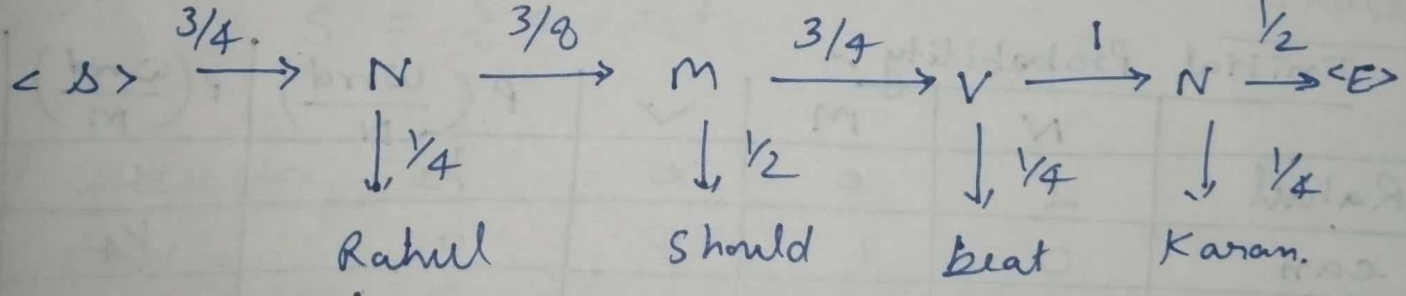
	N	M	V	$P(\frac{word}{N})$	$P(\frac{word}{M})$	$P(\frac{word}{V})$
Rahul	2	0	0	1/4	0	0
can	0	1	0	0	1/4	0
play	0	0	2	0	0	1/2
chess	2	0	0	1/4	0	0
Karan	2	0	0	1/4	0	0
Should	0	2	0	0	1/2	0
Clean	0	0	1	0	0	1/4
table	1	0	0	1/8	0	0
will	0	1	0	0	1/4	0
Shayam	1	0	0	1/8	0	0
beat	0	0	1	0	0	1/4
	8	4	4			

Using HMM,

P(N M V N / Rahul should beat Karan).

	N	M	V	<E>
<S>	3/4	1/4	0	0
N	0	3/8	1/8	4/8
M	1/4	0	3/4	0
V	4/4	0	0	0

"Rahul should beat Karan"



$$P = \frac{3}{4} \times \frac{1}{4} \times \frac{3}{8} \times \frac{1}{2} \times \frac{3}{4} \times \frac{1}{4} \times 1 \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{2}$$

$$= \frac{27}{2^{15}} = 0.00082357.$$

Sol 5(b):-

For, Sim_{path} :

$Path_{len}$ (Hill, shore)

$\Rightarrow 1+3 \Rightarrow 4$

$Sim_{path} = \frac{1}{4} = 0.25$

for Sim_{zensik} :

$-\log P(LCS(Hill, shore))$

$-\log_2 P(geo\text{-}form)$

$-\log_2(0.00176) = 9.15$

for Sim_{Lin} :

$$\frac{2 \log P(LCS(Hill, shore))}{\log P(hill) + \log P(shore)}$$

$$\Rightarrow \frac{2 \log P(geoform)}{\log P(hill) + \log P(shore)}$$

$$\Rightarrow \frac{2 \ln(0.00176)}{\ln(0.0000187) + \ln(0.0000836)}$$

$\Rightarrow 0.6259.$

Sol 5 (a):-

Transition Probability :-

	jouer	croquet	la	grillon.
play	1/3	1/3	1/3	1/3
cricket	1/3	1/3	1/3	1/3
the	1/3	1/3	1/3	1/3
team	1/3	1/3	1/3	1/3

Alignment Probability :-

Play | jouer : $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

cricket | croquet

play — cricket
jouer — croquet : $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

the | la . cricket | grillon : $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

the — cricket
la — grillon : $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

cricket | croquet . team | equipe : $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

cricket — team
croquet — equipe : $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

$$\text{Normalization} = \frac{\frac{1}{9}}{\frac{1}{9} + \frac{1}{9}} = \frac{1}{2}$$

Weighted translation count →		jouer	croquet	la grille	epuis
	play	1/2	1/2	0	0
	the	0	0	1/2	1/2
	Cricket	1/2 × 1/3	1/3 (1/2 + 1/2)	1/2 × 1/3	1/2 × 1/3
	team	0	1/2	0	0

Normalizing

~~Alignment~~

~~Normalized Foo~~ :-

Alignment Probability :-

$\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$

$\frac{1}{2} \times \frac{1}{6} : \frac{1}{12}$

$\frac{1}{2} \times \frac{1}{6} \times \frac{1}{12}$

$\frac{1}{2} \times \frac{1}{6} \times \frac{1}{12}$

$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$

$\frac{1}{8} \times \frac{1}{2} = \frac{1}{12}$

Sol 7:-

To handle co-references in English Text, we can propose a novel methodology that combines traditional rule based approaches with machine learning techniques. The methodology involves the following steps:

1) Text Preprocessing :- Text is tokenized and parts of speech (POS) tagging and Named entity Recognition (NER) are performed to identify intities and their types.

2) Coreference Resolution:- Initially, a rule based approach is applied to resolve simple coreferences based on exact match of words or phrases. For ex:- in the sentence "John went to the market. He bought some fruits", the rule-based system can resolve "He" to refer to "John".

3) Feature Extraction:- Features such as mention headword, syntactic distance between mentions and semantic similarity between mentions are extracted to represent mentions and their contexts.

4. Machine learning models:- A machine learning model, such as neural network or a conditional random field (CRF) is trained on dataset annotated with coreferences. The model uses the extracted features to learn patterns in coreference resolution.

5) Evaluation:- The methodology is evaluated using standard coreference resolution evaluation metrics such as MUC, B. CUBED, and CEAF. These metrics measures the precision, recall, and F1 score of the resolved coreferences compared to a gold standard.

Ex:- "Tom visited Mary at her house. He brought her a gift

Using our methodology, the system first identifies 'Tom' & "Mary' as entities and resolves the pronouns "He" and 'Her' to refer to "Tom" and mary" respectively. The rule based approach handles the simple coreferences, while the machine learning model improves the resolution by considering contextual features.

Evaluation:- methodology performance is evaluated on a dataset of similar contexts annotated with coreferences

Sol:- PPMI

context / Term	Auto	comp	Money	House	Politics	Σ
1 car	10	8	1	0	3	22
2. Auto	5	1	0	0	1	7
3. Insu	1	0	4	3	0	8
4. Window	1	2	1	2	1	7
5. Comp	3	2	1	0	0	6
6. Tech	1	3	1	1	1	7
Σ	21	16	8	6	6	57

N = 57

P context / P Term	21/57	16/57	8/57	6/57	6/57
22/57	10/57	8/57	1/57	0	3/57
7/57	5/57	1/57	0	0	1/57
8/57	1/57	0	4/57	3/57	0
7/57	1/57	2/57	1/57	2/57	1/57
6/57	3/57	2/57	1/57	0	0
7/57	1/57	3/57	1/57	1/57	1/57

$$PPMI(w_1, w_2) = \max\left(\log_2 \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)}, 0\right)$$

$$PPMI(w, c) = \max\left(\log_2 \frac{p(w, c)}{p(w)\, p(c)}, 0\right)$$

PPMI :-	Auto	Comp	Money	Sports	Politics
car					
auto					
insu					
win					
com					
Tea					

LESK'S ALGORITHM

Sense Bag: *contains the words in the definition of a candidate sense of the ambiguous word.*

Context Bag: *contains the words in the definition of each sense of each context word.*

E.g. "On burning ***coal*** we get ***ash***."

Ash

- Sense 1
  Trees of the olive family with pinnate leaves, thin furrowed bark and gray branches.
- Sense 2
  The ***solid*** residue left when ***combustible*** material is thoroughly ***burn***ed or oxidized.
- Sense 3
  To convert into ash

Coal

- Sense 1
  A piece of glowing carbon or ***burn***t wood.
- Sense 2
  charcoal.
- Sense 3
  A black ***solid combustible*** substance formed by the partial decomposition of vegetable matter without free access to air and under the influence of moisture and often increased pressure and temperature that is widely used as a fuel for ***burn***ing

In this case Sense 2 of ash would be the winner sense.

Summary: thesaurus-based similarity

$$\text{sim}_{\text{path}}(c_1, c_2) = \frac{1}{pathlen(c_1, c_2)}$$

$$\text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(LCS(c_1, c_2)) \quad \text{sim}_{\text{lin}}(c_1, c_2) = \frac{2\log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

$$\text{sim}_{\text{jiangconrath}}(c_1, c_2) = \frac{1}{\log P(c_1) + \log P(c_2) - 2\log P(LCS(c_1, c_2))}$$

$$\text{sim}_{eLesk}(c_1, c_2) = \sum_{r,q \in RELS} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))$$