Branch	Test Date	Semester	Div.	Roll No.	Student's Signature
CSE(N)		VII			

IA Test No.	Subject
	Natural Language Processing (NLP)

Junior Supervisor's full signature with date :	Question No.	1	2	3	Total 20	Examiners Signature	Student's Sign After receiving the assessed answer sheet
	Marks obtained						

Q.1 (a) List the phases of NLP

Ans (1) Morphological phase
(2) Syntax analysis phase
(3) Semantic analysis phase
(4) Pragmatic phase
(5) Discourse phase.

Q.1 (b) What is referential ambiguity in natural language processing? Give one example

Ans Either a text is mentioned as an entity (something/someone) and then refers to it again, possibly in a different sentence, using another word. Pronoun causes ambiguity when it is not clear which noun it is refering to.

Eg - The boy told to his father about the theft. He was very upset. He is referentially ambiguous because it can refer to both the boy and the father.

Q.1 (c) List the applications of NLP

Ans (1) Auto correction (2) Online chatbot (3) Auto complete
(4) Sentiment Analysis (5) Spam Mail filtering
(6) Recomendation engine (7) Information Retrieval.

Q1(d) Preprocessing steps

Ans Removing HTML tags, Removal of white space, Removal of accented character removal, expand contraction, removal of special character, convert all words into lower case, removal of punctuation mark, removal of words containing digits, removal stop w

Q1(e) Stemming

1) It is an elementary rule based process for removing inflectional word forms from a token & output are the stem.
2) Laughing → laugh
   studies → studie

Lemmitization

It is a systematic step by steps process for removing inflectional forms of a word.

2) Running → Run
   Studies → Study

Q1(f) Inflection Morphology

1) It is a morphological process that adapts existing words so that they function efficiently
2) Mostly added as suffix
3) Cannot change part of speech
4) They are more regular
5) Eg - cats → cat

Derivational Morphology

1) It is concerned with the way morphemes are connected to existing lexical form as affix.
2) Both as prefix, suffix
3) Can change POS.
4) They are less regular
5) Danger - Dangerous
   Legal → Legalize

Q1(g) Apply porter algorithm on — Writing, King

Ans As per Porter's Algorithm

(*vowel*) ing → ϕ

(*vowel*) ed → ϕ

∴ Writing → write

King → King

If the word contains vowel then suffix -ing will be removed.

Q5(b) Define Minimum Edit distance used in Spelling correction

Ans Minimum edit distance is the minimum number of insertion, deletion, transposition & substitution reqd to tranform the misspelled word into a valid word.

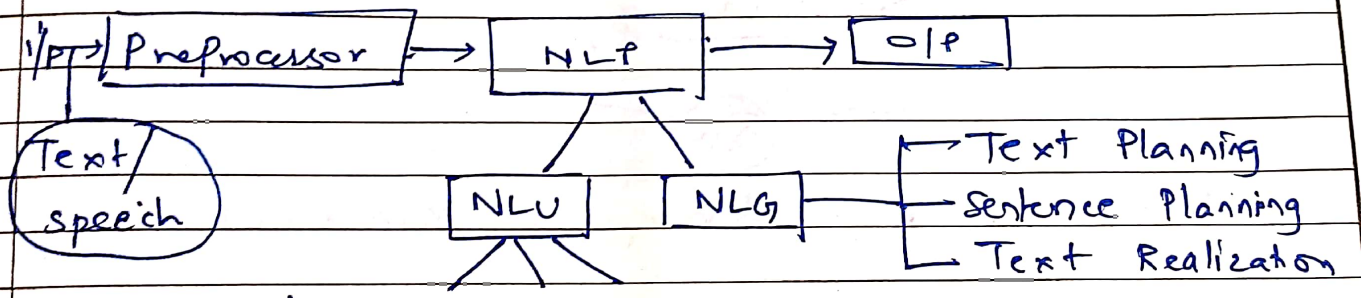
Eg INTENTION EXECUTION

INTENTION
EXECUTION
↑ ↑ ↑ ↑ ↑↑
I S S D I

insertion (I) = 2
Substitution (S) = 2
Deletion (D) = 1
(5)
Minimum edit distance

Q6(a) Explain Natural language Generator component of NLP generic block diagram

Ans Generic NLP system



NLG - Natural Language Generator

It is the proces of producing meaningful phrases & sentences in the form of natural language from some internal representation. It involves -

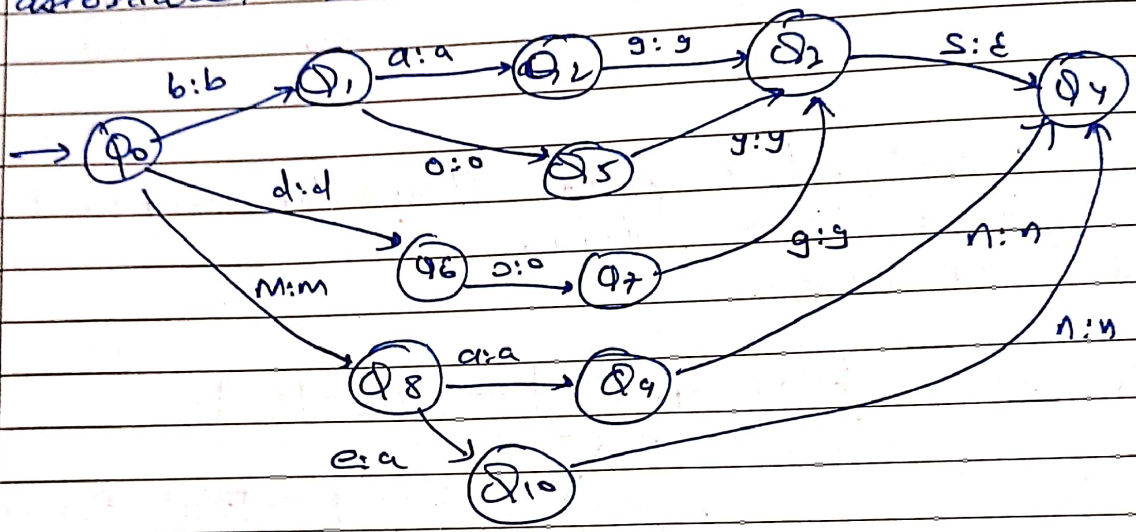
1) Text Planning - It includes retrieving the relevant content from the knowledge base.
2) Sentence planning - It includes choosing required words, forming meaningful phrase & setting the tone of sentence.

3) Text Realization – It is mapping sentence plan into sentence structure.

Q2(b) Explain FST in detail

Ans Finite state Transducer are finite state machine to check a string. It is used to build morphological analyzer. It translates strings from one language to another language.

– Like finite state automata, but here each edge is associated with two strings.



Take men as i/p & recognize Man +PL+N.

2-level Morphology

Cat → Cats (regular word +N+PL)

fox → foxes (irregular word + N+PL)

Given i/p cats we get o/p cat +N+PL

2 level morphology [lexical level
surface level

lexical level – morphemes & feature

surface level – Actual spelling.

lexical { | c | a | t | +N | +PL | | }

Surface { | c | a | t | s | }

Q2(c) for a corpus, MLE for bigram "battery life" is 0.27 frequency of "battery" is 800. After applying laplace smoothening the MLE for "battery life" becomes 0.025. what is the vocabulary size of the corpus?

Soln

$$P_{MLE}(\text{life} \mid \text{battery}) = f(\text{battery life}) / f(\text{battery})$$

$$0.27 = f(\text{battery life}) / 800$$

$$f(\text{battery life}) = 800 \times 0.27 = 216$$

with laplace smoothening

$$P_{MLE}(\text{life} \mid \text{battery}) = \{ f(\text{battery life}) + 1 \,/\, f(\text{battery}) + V \}$$

$$0.025 = \frac{217}{800+V}$$

$$V = \frac{197}{0.025} = 7880$$

Size of vocabulary = 7880

Q3(a) find the probability based on given corpus.

① Michael & zack played at the playground = S1

$P(\text{Michael} \mid <S7) = \frac{1}{5}$

$P(\text{and} \mid \text{Michael}) = \frac{2}{2}$

$P(\text{zack} \mid \text{and}) = \frac{3}{3}$

$P(\text{playground} \mid \text{the}) = \frac{2}{5}$

$P(\text{played} \mid \text{zack}) = \frac{1}{3}$

$P(\text{at} \mid \text{played}) = \frac{1}{1}$

$P(\text{the} \mid \text{at}) = \frac{2}{2}$

$P(</s> \mid \text{playground}) = \frac{1}{2}$

$P(S_1) = \frac{1}{5} \times 1 \times 1 \times \frac{2}{5} \times \frac{1}{3} \times 1 \times 1 \times \frac{1}{2} = \frac{1}{75}$

(2) S_2 = Bob went to the school

$P(S_2) = ?$ ∴ one probability is 0

∴ without laplace smoothing

$P(\text{Bob} | \langle s \rangle) = \frac{2}{5}$

$P(S_2) = 0$

$P(\text{went} | \text{Bob}) = \frac{0}{2}$

(3) The school was huge = S_3

$P(\text{The} | \langle s \rangle) = \frac{2}{5}$

$P(\text{school} | \text{The}) = \frac{2}{3}$

$P(\text{was} | \text{school}) = \frac{2}{3}$

$P(\text{huge} | \text{was}) = \frac{1}{2}$

$P(\langle /s \rangle | \text{huge}) = \frac{1}{1}$

∴ $P(S_3) = \frac{2}{5} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{2} \times 1$

$= \frac{2}{45}$

(4) Zack went to the playground = S_4

$P(\text{Zack} | \langle s \rangle) = \frac{1}{8} 0$ ∴ $P(S_4) = 0$

$P(\text{went} | \text{Zack}) = $ ____

∴ Highest probability is for sentence "The school was Huge"

Perplexity $= (P(s))^{-1/N}$

$= \left(\frac{2}{45}\right)^{-1/6} = \left(\frac{45}{2}\right)^{1/6} = 1.495$

Q3(b) Write a note on Language model & evaluation metrics for language model

Ans Language modelling is the way of determining the probability of any sequence of words. Language modelling is used in wide variety of application such as speech recognition, spam filtering etc.

- There are two types of language modelling -

1) Statistical Language modelling

- Development of probabilistic model that are able to predict the next word in the sequence given the word that precedes.
Eg - N-gram model

2) Neutral Language Modelling

- Neutral network methods are achieving better results than classical methods both on standalone language model & when it is incorporated into large model on challenging task like speech recognition & machine translation.

Eg - Word embedding.

3) N gram Language Model

- It can be defined as the contiguous sequence of items from a given sample of text or speech.

- It predict the probability of a given N-gram within any sequence of words in the language.

Matrics for language Model

1) Entropy - It is a measure of the amount of information conveyed by Claude Shannon.

$$H(p) = \mathcal{E}_x p(x) \cdot (-\log(p(x)))$$

H(p) is always greater than equal to 0.

2) Cross Entropy - It measures the ability of the trained model to represent test data.

$$H(p) = \sum_{i=1}^{n} \frac{1}{n} (-\log_2 (p(w_i | w_i^{i-1})))$$

The cross entropy is always greater than or equal to entropy i.e the model uncertainity can be no less than the true uncertainity.

3) Perplexity - It is a measure of how good a probability distribution predicts a sample. It can be understood as a measure of uncertainity The perplexity can be calculated by cross entropy to the exponent of 2.

Lower the perplexity = better the model