CMR
INSTITUTE OF
TECHNOLOGY

USN | | | | | | | | | | |

First Internal Test- Sep 2019

Sub:	Natural Language Processing							Code:	15CS741
Date:	21/ 09 / 2019	Duration:	90 mins	Max Marks:	50	Sem:	VII	Branch:	ISE (A&B)

Answer Any **FIVE FULL** Questions

		Marks	OBE	
			CO	RBT
1 (a)	Define Natural Language processing. What are its purposes? List and explain different levels of processing involved in it.	[2+2+6]	CO1	L1

It is the **understanding and generation** of Natural languages by generating computational models of natural languages.

Purposes:

a. To develop automated tools for language processing
b. To gain a better understanding of human communication

Different Levels of processing:

a. Lexical Analysis:
   It involves identifying and analyzing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of text into its small units.
b. Syntactic Analysis:
   It involves analysis of words in the sentence for grammar and arranging words in a manner that shows the relationship among the words. It creates an appropriate order words and symbols.
c. Semantic Analysis: Establishes the exact meaning of the sentence by linking or mapping one interface to another. Established the relations ship between words, phrases, signs and symbols.
d. Pragmatic Analysis: It involves the understating the sentence or other text in the context of overall world knowledge. It established the actual meaning of the sentence.

2 (a)	"Processing Indian languages using NLP is far more challenging", justify the statement.	[10]	CO1	L2

Processing Indian languages are far more challeing as Indian languages differes from English in more than many ways:

Differences between Indian languages and English

I. Indic scripts have a non-linear structure.
II. Unlike English, Indian languages have SOV(subject-object-Verb) as default sentence structure.
III. Indian languages have a free word order i.e words within sentence can be freely moved without changing the meaning of the sentence.

IV. Rich set of morphological variants as languages have evolved over centuries.
V. Indian Languages uses post-positions case markers instead of pre-positions.
VI. Indian languages makes extensive and productive use of complex predicates.
VII. Indian languages use verb complexes consisting of sequence of verbs. Auxilary verbs provide information about tense, aspect , modality.

3 (a)	Define morphology. Explain the 3 ways of word formation and what are the information sources used in morphological parsing?	[1+6+3]	CO3	L2

Morphology is a sub-discipline of linguistics. It studies word structure and the formation of words from smaller units known as morphemes. The goal of the Morphology parsing is to discover the morphemes that build a given word. There are two broad classes of morphemes called stems and affixes.

3 ways of word formation:

a. Inflection
b. Derivation
c. Compounding

Inflection: In inflection root word is combines with a grammatical morpheme to yield a word of the same class as that of original stem.
Ex: Egg and Eggs, sing and singing
Derivation: In derivation root word is combined with a grammatical morpheme to yield a word belonging to a different class.
Ex: Compute and Computation
Compounding: Compounding is the process of merging two or more words to form a new word.
Ex: Desktop, Overlook
Information source used in morphological parser:

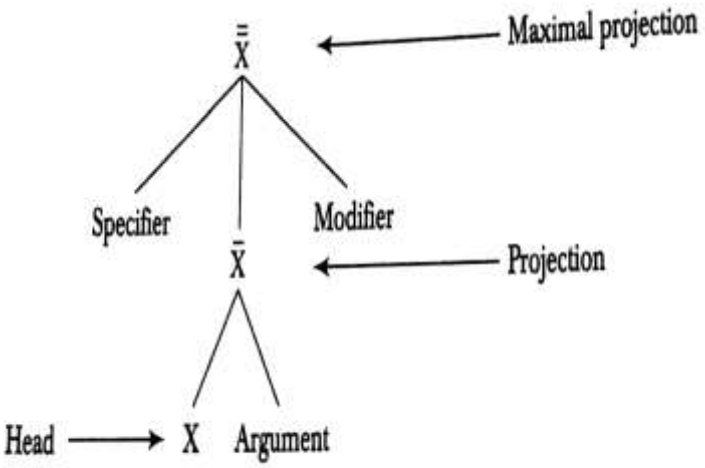
a. **A Lexicon:** A lexicon is a list of stems and affixes together with basic information about them.
b. **Morphotactics:** It deals with the ordering of the morphemes. It describes the way morphemes are arranged or touch each other.
Ex: Rest-less-ness not rest-ness-less
c. **Orthographic rules:** Spelling rules that specify the changes that occur when two given morphemes combine. Ex: y->ier i.e easy+ ier= easier

4 (a)	What is X-Bar theory? Write the general phrase and sentential structure of X-bar theory, apply the same for the following sentences: **"ate the food in a dhaba"**	[2+2+6]	CO2	L3

X-bar theory makes the claim that every single phrase in every single sentence in the mental grammar of every single human language, has the same core organization.X-Bar theory is one of the central concept in GB theory. Instead of defining several phrase structures and the sentence structure with separate

set of rules , X-bar theory defines them both as maximal projections of some head X. In this manner entities become independent of language. Here's a tree diagram that shows the general phrase structure of X-bar theory. Noun Phrase(NP) , Verb Phrase (VP), Adjective Phrase(AP), and Prepositional Phrase(PP) are maximal projection of Noun(N), Verb (V), Adjective (A) and Preposition(P) respectively.

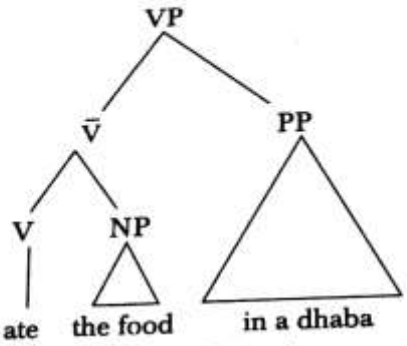
The projection is at two levels – first projection of head at the semi-phrasal level denoted by X' and then the second maximal projection at the phrasal level denoted by X''.



General Phrase Structure

2. VP: ate the food in a dhaba

$[_{VP}[_{\bar{V}}[_{V}\ ate][_{NP}\ the\ food]][_{PP}\ in\ a\ dhaba]]$



5 (a) Define Information retrieval and discuss major issues involved in it. [10] CO1 L2

Information retrieval is the organization, storage, retrieval and evaluation of information relevant to a query. Its application can be found in database management systems, bibliographic text retrieval system, question answering system and also in search engines.

Major Issues in IR:

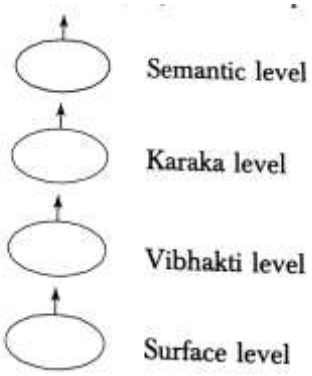
1. **Representation of the document.**

Generally implemented using keywords. Retrieval issues due to Polysemy, Homonymy, Synonymy. Keyword based IR ignores semantics and contextual information of the data.

2. **Vagueness and inaccuracy of the user's queries.**
   Feedback mechanism to modify or expand query.
3. **Measure of similarity:**
   Selection of appropriate similarity measure is a crucial issue in the design of IR system.
4. **Performance of the IR system:**
   Effectiveness of the IR
5. **Degree of relevance:**
   Binary function: Either information is related or not (0 or 1).
   Continuous function: Percentage or degree of relevance is continuous .

6 (a)	What is Government and Binding Theory and explain its components and organization.	[10]	CO2	L2
7 (a)	Explain Karaka Theory of Paninian Grammar (PG). Identify different karaka's in the following Hindi sentence: ***Maa ne bachchi ko aangan me haath se roti khilati thi***	[5+5]	CO2	L3

Levels of Paninian Grammar:



Karaka literally means CASE, these case relations are based on the way the word group participates in the activity denoted by the verb group. Karaka relations are assigned based on the roles players by various participants in main activity.

Various karaka's are (case marker in hindi)

1. Karta (Subject) case marker: 'ne' or Φ
2. Karma (Object) case marker: 'ko' or Φ
3. Karana (instrument) case marker: 'dwara' or 'se'
4. Sampradana (Beneficiary) case marker: 'ko' or 'ko liye'
5. Apadana (Seperation) case marker: part that serves as separation

6. Adhikaran (Locus) case marker: (support in space or time)
7. Sambandh (Relation)
8. Tadarthya (Purpose)

Problem: Identify the Karakas in the following sentence:

"Maan bachchi ko aangan mein haath se rotii khilaati hei"

Karta: maan
Karma: Rotii
Karana: haath
Sampradana: bachchi
Adhikarana: aangan

8 (a) What is statistical language modeling? Explain n-Gram modeling. Find the probability of the third sentence in the corpus given below using bi-Gram modeling: [2+3+5] CO2 L3

The Arabian Knights
These are the fairy tales of the east
The stories of the Arabian Knights are translated in many languages

n-Gram modelling:

n-Gram predicts the probability of a word by considering all the previous words by the conditional probability given previous n-1 words.

$$P(w_i/h_i) \approx P(w_i/w_{i-n+1} \cdots w_{i-1})$$

It makes use of the markov model, if the model limits the previous words to one only then it is known as bi-gram model. Proability of a sentence is the product of bi-gram probability of all words in it, which is given as below:

$$P(s) \approx \prod_{i=1}^{n} P(w_i/w_{i-1})$$

Training set:

The Arabian Knights
These are the fairy tales of the east
The stories of the Arabian knights are translated in many languages

Bi-gram model:

$P(\text{the}/<s>) = 0.67$	$P(\text{Arabian/the}) = 0.4$	$P(\text{knights /Arabian}) = 1.0$
$P(\text{are/these}) = 1.0$	$P(\text{the/are}) = 0.5$	$P(\text{fairy/the}) = 0.2$
$P(\text{tales/fairy}) = 1.0$	$P(\text{of/tales}) = 1.0$	$P(\text{the/of}) = 1.0$
$P(\text{east/the}) = 0.2$	$P(\text{stories/the}) = 0.2$	$P(\text{of/stories}) = 1.0$
$P(\text{are/knights}) = 1.0$	$P(\text{translated/are}) = 0.5$	$P(\text{in /translated}) = 1.0$
$P(\text{many/in}) = 1.0$		
$P(\text{languages/many}) = 1.0$		

Probability of Third sentence:

0.67*0.2*1.0*1.0*0.4*1.0*1.0*0.5*1.0*1.0*1.0= **0.0268**