Page No.
Date

Q1 What is stemming. Explain Porter's algorithm in detail.

Stemming

- It is a process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as lemma
- Stemming is a part of linguistic studies in morphology and artificial intelligence, information retrieval and extraction. Stemming and AI knowledge extract meaningfull info from vast source like big data or the internet :: additional forms of a word related to a subject may need to be searched to get best results.

Porters's stemming Algorithm

Consonant: a letter other than A, E, I, O, U and if preceded by consonant

Vowel - Only other letter.

eg - Troubles
C V C V C.

The rules are of the form:

(condition) $S_1 \rightarrow S_2$

Teacher's Sign.: _____

Page No.
Date

Step1: SSES → SS caresses → caress
IES → I ponies → poni
SS → SS caress → caress
S → ϵ cats → cat

Step 2a: (m>1) EED → EE agreed → agree
(*V*) ED → E plastered → plaster
(*V*) ING → ϵ motoring → motor

step 2b: AT → ATE
BL → BLE

(*d & !(*L or *S or *Z)) → single letter
(m=1 & *o) → E

Step 3 (*V*) Y → I

Step 4- (m>0) ATIONAL → ATE
(m>0) IZATION → IZE
(m>0) BILITI → BLE

step 5: (m>0) ICATE → IC
(m>0) FUL → ϵ
(m>0) NESS → ϵ

step 6: (m>0) ANCE → ϵ
(m>0) ENT → ϵ
(m>0) NE → ϵ

step 7: (m>1) E → ϵ
(m=1 & !*o) NESS → ϵ

Teacher's Sign.:

2) Explain ambiguity in NLP with example.
- Ambiguity can be referred as the ability of having more than one meaning or being understand in more than one way.
- Natural languages are ambigous, so ~~continous~~ computers are not able to understand language the way people do.
- NLP is concerned with the development of computational model of aspects of human language processing.
- Ambiguity could be lexical, syntactic, semantic pragmatic etc

eg
I saw a man on a hill with telescope

Alternate meaning
1. There is a man on a hill, and I'm watching him with my telescope
2) There's a man and he's on a hill that also has a telescope on it

or what is NLP ? Describe various stages involved in NLP process with suitable

NLP :
It is a subfield of linguistics, CS, info engg and AI concerned with the interaction b/w computers and human languages in particular how to program computers to process and analyze large amount of

Teacher's Sign.: _____

natural language data

Challenges in NLP involve speech recognition, natural language understanding and natural language generation

Steps in NLP

1. 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 chunks of text into paragraphs, sentences and words.

2. Syntactic Analysis

It involves analysis of words in the sentence for grammar and arranging words in a manner that shows relationships.

3) Semantic Analysis

It draws the exact meaning or the dictionary meaning or the text. The text is checked for meaningfulness. It is done by mapping syntactic structures and objects in the task domain. The semantic analyzer disregards sentence such as "hot" ice-cream.

Teacher's Sign.:

4. Discourse Integration:
The meaning of any sentence depends upon the meaning of sentence just before it. In addition, it also brings about the meaning of immediately succeeding sentence.

5. Pragmatic Analysis:
During this, what was said is re-interpreted on what it actually meant. It involves deriving those aspects of ~~the~~ language which require real world knowledge

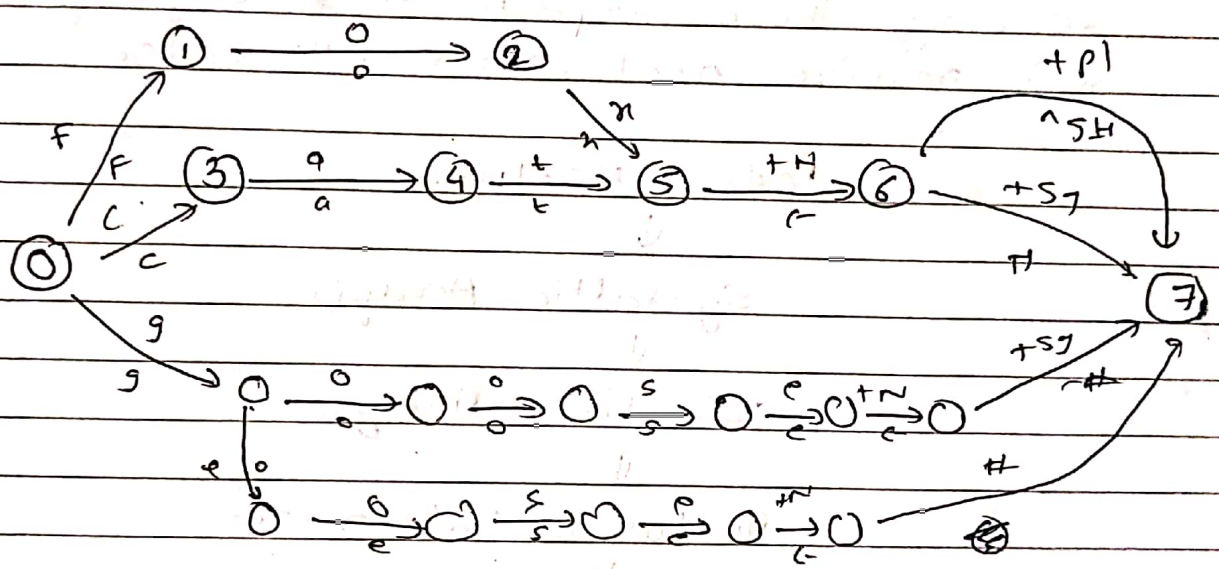
lexical Analysis
↓
Syntaltic Analysis
↓
Semantic Analysis
↓
Disclosure Integration
↓
Pragmatic Analysis

Q3
Ans Finite state Transducers
⇒ Fs Automata
- An FSP represents a set of strings
  eg work, walks, walked
- A recognizer function
  recognize (str) → true or ~~false~~ false.

Teacher's Sign.: _____

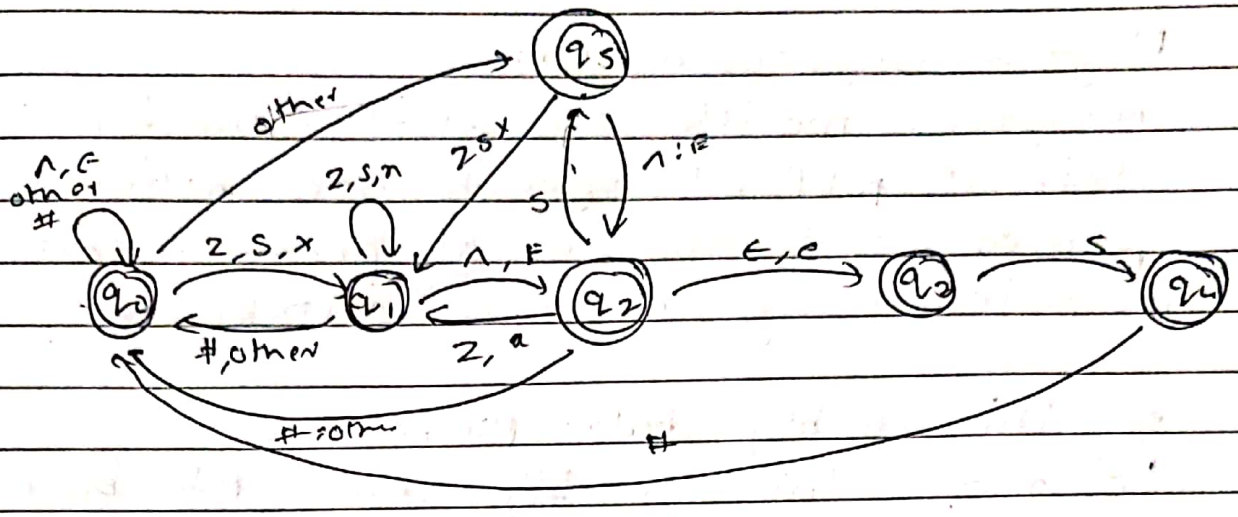
- FS transducer.

: An FST represents a set of pairs of strings (think of as ip, op, pairs)
Can return multiple answers if ambiguity
eg If you don't have pos-tagged input
"walk" could be the verb. They walk to the store versus the noun. I took a walk.



lexical	f	o	x	+N	+PI		
Intermediate	f	o	x	^	s	#	

Teacher's Sign.: _____



lexical	f	0	x	+N	+P		
Intermediate	f	0	x	^	s	#	
Surface		f	0	x	E	s	

04 what is language model.
write a note on n-gram language model

language modelling is one of the most imp parts of modern NLP. There are many sorts of application for language modelling like: Machine translation, spell correction, speech recognition, summarization, Question-Answering, sentiment analysis etc. Each of those tasks require use of language model.

Teacher's Sign.: _____

N-gram language model:

N-gram models are widely used in statistical NLP. In speech recognition, phonemes and sequences of phonemes are modelled using n-gram distribution

eg This is big data AI book.

Unigram : This, is, big, data, AI, book = 6

Bigram : This is, is big, big data, data AI.

Trigram : This is big, is big data, big data AI book = 4

The chain rule probability is :

$$P(w_1 \ldots w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1 w_2) \ldots P(w_n|w_1 w_2 \ldots w_{n-1})$$

The maximum likelihood estimate

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

eg <s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$P(I|<s>) = \frac{2}{3}$ $P(Sam|am) = \frac{1}{2}$

$P(Sam|<s>) = 1/3$ $P(do|I) = 1/3$

$P(am|I) = 2/3$ $P(</s>|Sam) = 1/2$

Teacher's Sign.: _____

Page No.
Date

or Explain various approaches to perform Part of speech (Pos).

→ Part of speech tagging may be defined as the process of assigning one of the POS to given word. In a simp- words. Pos tagging is a task of labelling each word in a sentence with its appropriate Pos.

Most of Pos tagging falls under Rule Bas, stochasti- and transformation based tagging.

* Rule Based Pos tagging

One of the oldest techniques of tagging is r rule based Pos tagging. Rule Based tagge use dictionary or lexicon for getting possible tags for tagging each word. If the word has more than one possible tag, then rule-based taggers use hand-written rules to identify the correct tag. Disambiguou can also be performe in this by analyzing the linguistic feature of a word along with its preceding as well as foll words

Rules may be either
- context : Pattern Rules
- Regex into FSA

Teacher's Sign.: _____

Page No.
Date

Stochastic Pos Tagging
- The model that includes frequency or probability can be called stochastic

The simplest stochastic trigger applies the foll approaches for POS tagging.

Word frequency approach :
In this the stochastic tagger disambiguates the words based on the prob that word occurs with a particular tag.
The main issue is that it may yield inadmissible seqence of tags

Tag seqence probabilities
It is another approach where tagger calculates the prob of a given seqence of tags occuring. It is also called n-gram approach.

* Transformation Based Tagging :
It is also called as Brill tagging. It is an instance of transition Based learning which is a a rule based algo for automatic tagging of POS. TBL allows to have linguistic knowledge in a readable form, transform one state to another using transition rules. If we see similarity blw rule based and translation tagger then like rule-based, it is also based on rules

Teacher's Sign.: _____

which specify what tags to be needed to assigned specific words. It is machine learning technique in which rules are automatically included from data.

Teacher's Sign.: _____