CMR
INSTITUTE OF
TECHNOLOGY

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

Second Internal Test- Oct 2019

Sub:	Natural Language Processing							Code:	15CS741
Date:	14/ 10 / 2019	Duration:	90 mins	Max Marks:	50	Sem:	VII	Branch:	ISE (A&B)

Answer Any **FIVE FULL** Questions

		Marks	OBE	
			CO	RBT
1 (a)	Explain the importance of Part of Speech Tagging with an example and explain the categories of part of speech tagging.	[10]	CO3	L2

Part of speech tagging is the process of assigning a part of speech such as noun, verb, adverb, pronoun to each word of a sentence. The input to a tagging algorithm is the sequence of words of a natural language and a specified finite part of speech tag set. The output is a single best part-of-speech tag for each word.

Part-of-Speech tagging is done as a pre-requisite to simplify a lot of different problems such as Text to speech conversion, Word sense disambiguation.

Ex

She saw a **bear**.

Your efforts will **bear** fruit.

Bear is acting as verb in one sentence and other is a verb.

Categories of Part-of-speech tagging:

1. Rule based tagger
2. Stochastic tagger
3. Hybrid tagger

Rule Based Tagger: Rule based tagger uses handcoded rule to assign tags to words. These rules uses a lexicon to obtain a list of candidate tags and then use rules to discard incorrect tags.

Ex: TAGGIT, ENGTWOL

Stochastic Tagger: Data-driven approach in which frequency-based information is automatically deroived from corpus and used to tag words. Stochastic tagger disambiguates words based on probability that a word occur with a particular tag.

Ex: CLAWS(Constituent likelihood automatic word-tagging system), HMM(Hidden Markov Model)

Hybrid Taggers: Hybrid model combines features of both these approaches. They use both rules based system and tagging induced from training corpus to automatically tag the words.

Ex: Tranformation-based Tagger, Brill tagger

2 (a)	Explain the working of Two Level Morphological parser model. Write a simple Finite State Transducer (FST) for mapping Nouns in English language.	[10]	CO3	L2

In two level Morphological parsing model, a word is represented as a correspondence between its lexical level form and its surface level form. The surface level represents the actual spelling of the word while the lexical level represents the concatenation of its constituent morphemes.

Ex:

For example, the surface form 'playing' is represented in th form as play + V + PP as shown in Figure 3.5. The lexical form of the stem 'play' followed by the morphological information which tells us that 'playing' is the present participle form of the

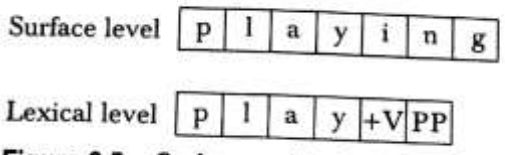


Figure 3.5 Surface and lexical forms of a word

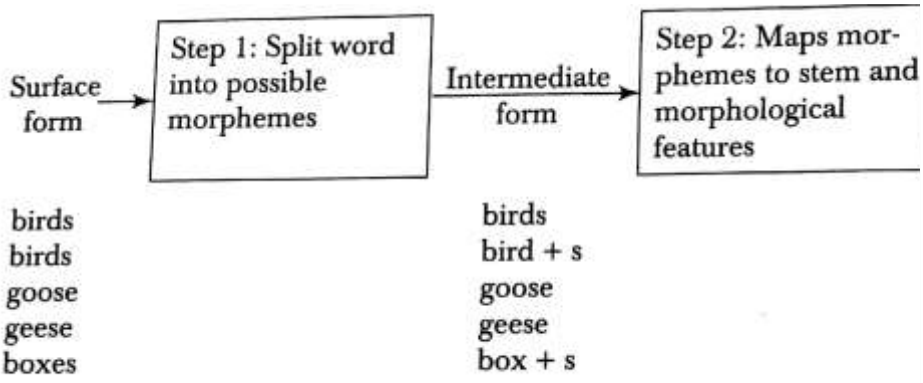
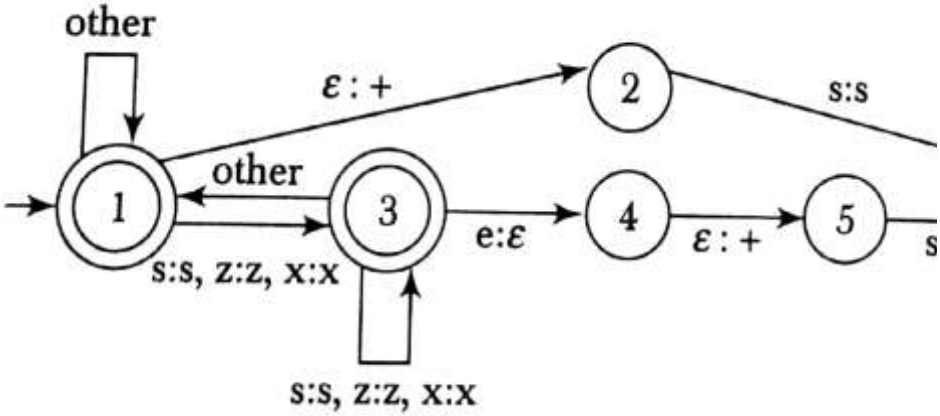


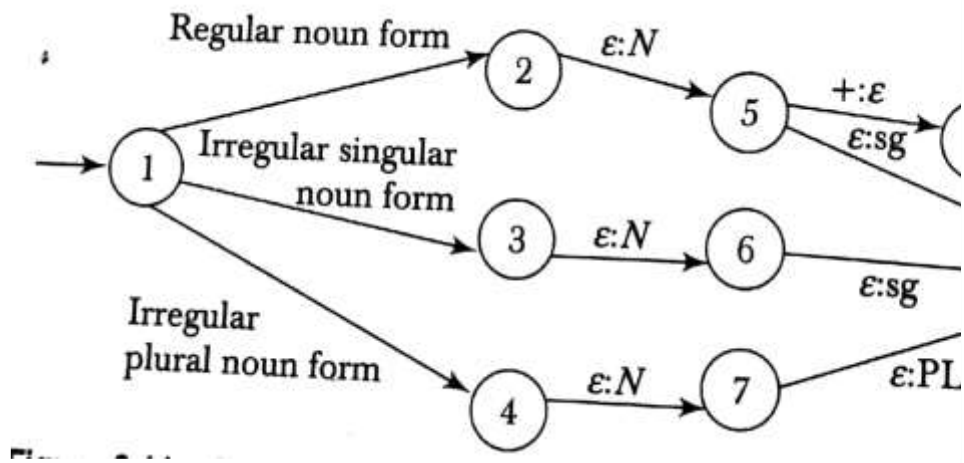
Figure 3.7 Two-step morphological parser

FST for mapping English Noun in two step morphological parsing:

Step 1: Transducer that maps from surface form to intermediate form



Step 2: Transducer that maps from intermediate level to the lexical level.



3 (a)	Write a note on common causes of spelling errors and differentiate between non-word and real-word error.	[10]	CO3	L1

Most common causes of spelling errors are due to single letter misspellings such as :

a. Substitution of a single letter ex: error and errpr
b. Omission of a single letter ex: adjacent and adacent
c. Insertion of a single letter ex: wrrite and write
d. Transposition of two adjacent letter ex: are and aer

Non-word error: When an error results in a word that does not appear in given lexicon or is not a valid orthographic word form. It is a solved problem and most widely used techniques to handle non-word errors are n-gram analysis and dictionary lookup.

Ex: are and aer

Real-word error: An error results in actual word of the language. It may cause local syntactic error, global syntactic error, semantic error or errors at discourse or pragmatic levels.

Ex: Peace and Piece

4 (a)	List the categories of spelling correction algorithms. Write an algorithm for minimum edit distance spelling correction and apply the same to compute the minimum edit distance between words ***tutor and tumour.***	[10]	CO3	L3

Categories of spelling correction algorithm:

a. Minimum edit distance
b. Similarity key techniques
c. N-gram based techniques
d. Neural nets
e. Rule-based techniques

```
Input: Two strings, X and Y
Output: The minimum edit distance between X and Y
m ← length(X)
n ← length(Y)
for i = 0 to m do
    dist[i,0] ← i
for j = 0 to n do
    dist[0,j] ← j
for i = 0 to m do
    for j = 0 to n do
        dist[i,j] = min{ dist[i-1,j] + insert_cost,
                         dist[i-1,j-1] + subst_cost(X_i, Y_j),
                         dist[i,j-1] + delet_cost }
```

Figure 3.13 Minimum edit distance algorithm

5 (a) Write the CFG for the following phrases with an example for each [10]

a. Noun Phrase (NP)
b. Verb Phrase (VP)
c. Preposition Phrase (PP)
d. Adjective Phrase (AP)
e. Adverb phrase (AdvP)

CO2 L2

Noun Phrase:

NP-> (Det) (AP) Nominal (PP)
Nominal-> Noun | Noun Nominal

Ex: A beautiful lake in Kashmir

Verb Phrase:

VP->Verb (NP)(NP) (PP)*
VP-> Verb S

Ex: The boy gave the girl a book with blue cover

Preposition Phrase:

PP->Prep (NP)

Ex: We played volleyball on the beach

Adjective Phrase:

AP-> (Adv) Adj (PP)

Ex: My sister is fond of animals

Adverb Phrase:

AdvP-> (Intens) Adv

Ex: Time passes very quickly

6 (a)	Illustrate the Boolean model of Information retrieval with a suitable example, list its advantages and limitations	[10]	CO5	L2

Boolean model is a classical model of Information retrieval (IR) and the oldest model.

- It is based on Boolean logic and classical set theory. Represents documents as a set of keywords, usually stored in an inverted file.
- Users are required to express their queries as a boolean expression consisting of keywords connected with boolean logical operators (AND, OR, NOT).
- Retrieval is performed based on whether or not document contains the query terms.

Boolean model is given as below: Given a finite set

$T = \{t1, t2, ...,ti,...,tm\}$

of index terms, a finite set

$D = \{d1, d2, ...,dj,...,dn\}$

of documents and a boolean expression in a normal form - representing a query Q as follows:

$$Q=\wedge(\vee\Theta_i), \Theta_i \in\{t_i, \neg t_i\}$$

1. The set R_i of documents are obtained that contain or not term t_i:

$$R_i = \{d_j|\Theta_i \in d_j\}, \Theta_i \in\{t_i, \neg t_i\}$$

Where $\neg t_i \in d_j$ means $t_i \notin d_j$

2. Set operations are used to retrieve documents in response to Q:

$$\cap R_i$$

Advantages: Simple, efficient, easy to implement, performs well in terms of recall and precision if the query is well formulated.

Drawbacks:

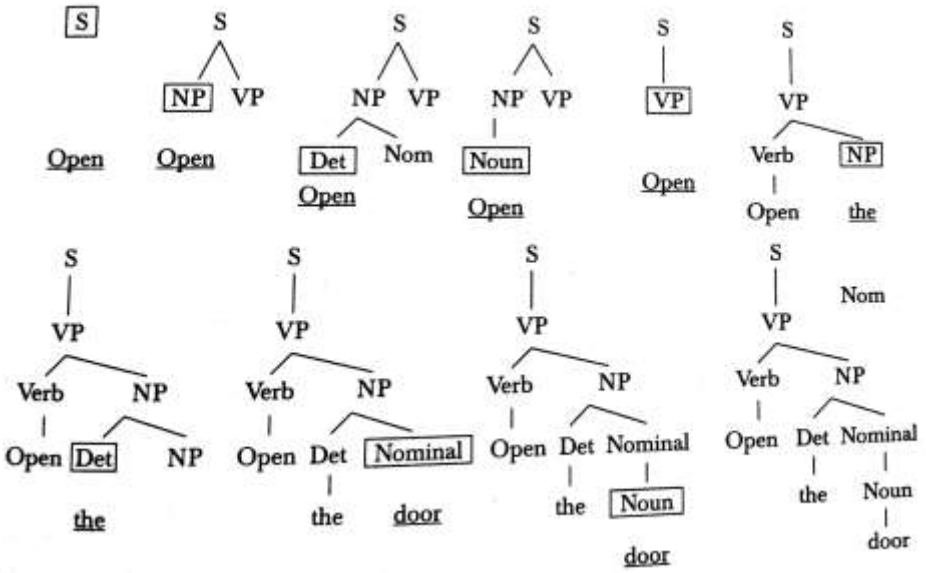
Cannot retrieve documents that are only partly relevant to user query. Boolean system cannot rank the retrieved documents. User need to formulate the query in pure Boolean expression

7 (a)	Give an algorithm for the Simple Top-Down, Depth-First search, Left-Right parser. Parse the sentence 'Open the Door', illustrate the parsing step by step using the same algorithm	[10]	CO3	L3

Algorithm for simple top-down parser:

1. Initialize agenda
2. Pick a state, let it be curr_state, from agenda
3. If (curr_state) represents a successful parse then return parse tree
   else if curr_stat is a POS then
   if category of curr_state is a subset of POS associated with curr_word
   then apply lexical rules to current state
   else reject
   else generate new states by applying grammar rules and push them into a
4. If (agenda is empty) then return failure
   else select a node from agenda for expansion and go to step 3.

Parsing the sentence 'Open the door'

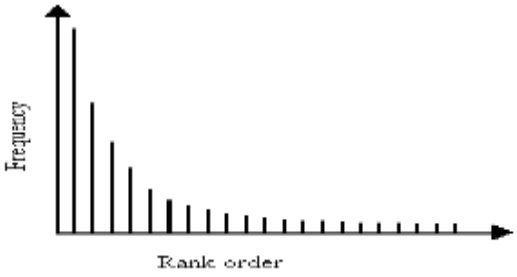


8 (a) Define indexing. Explain how the set of representative keywords are reduced during indexing. [10] | CO5 | L2

- **Indexing :** Transforming text document to some logical representation. Most of the indexing techniques involve identifying good document descriptors, such as keywords or terms, to describe information content of the documents. A good descriptor is one that helps in describing the content of the document and in discriminating the document from other documents in the collection. *Term* can be a single word or it can be *multi-word phrases*.
  Example:
  'Design Features of Information Retrieval systems'
  can be represented by single word terms :
  Design, Features, Information, Retrieval, systems
  or by the set of multi-term words terms:
  Design, Features, Information Retrieval, Information Retrieval systems

- **Reducing set of representative keywords is done using the following techniques:**
    1. Stop word elimination
    2. Stemming
    3. Zipf's law

    1. **Stop word elimination:** Stop words are high frequency words, which have little semantic weight and are thus unlikely to help with retrieval. Such words are commonly used in documents, regardless of topics; and have no topical specificity.
       Example :
       articles ("a", "an" "the") and prepositions (e.g. "in", "of", "for", "at" etc.).
       Advantage: Eliminating these words can result in considerable reduction in number of index terms without losing any significant information.
       Disadvantage: It can sometimes result in elimination of terms useful for searching, for instance the stop word *A* in *Vitamin A*. Some phrases like "*to be or not to be"* consist entirely of stop words.
    2. **Stemming:** Stemming normalizes morphological variants. It removes suffixes from the words to reduce them to some root form.
       e.g. the words *compute*, *computing*, *computes* and *computer* will all be reduced to same word stem ***comput.***
       Most widely used stemmer **Porter Stemmer(1980)**.
       Example:
       The stemmed representation of
       Design Features of Information Retrieval systems
       will be
       {*design, featur, inform, retriev, system*}
    3. **Zipf's law:** Zipf's law states that "Frequency of words multiplied by their ranks in a large corpus is approximately constant", i.e. **Frequency x Rank** $\approx$ $\boldsymbol{Constant}$



High frequency words: Common words with less discriminatory power.
Low frequency words: Less likely to be included in query.
Medium frequency words: Content bearing words that can be used for indexing , extracted using high and low frequency threshold.

5 (a)	Write the CFG for the following phrases with an example for each a. Noun Phrase (NP) b. Verb Phrase (VP) c. Preposition Phrase (PP) d. Adjective Phrase (AP) e. Adverb phrase (AdvP)	[10]	CO2	L2
6 (a)	Illustrate the Boolean model of Information retrieval with a suitable example, list its advantages and limitations.	[10]	CO5	L2
7 (a)	Give an algorithm for the Simple Top-Down, Depth-First search, Left-Right parser. Parse the sentence 'Open the Door', illustrate the parsing step by step using the same algorithm	[10]	CO3	L3
8 (a)	Define indexing. Explain how the set of representative keywords are reduced during indexing.	[10]	CO5	L2