DHANALAKSHMI SRINIVASAN ENGINEERING COLLEGE

Department of Artificial Intelligence and Data Science

Question Bank

U20AI703 – NATURAL LANGUAGE PROCESSING

IV Year / VII Semester

REGULATION-2020

UNIT I OVERVIEW AND LANGUAGE MODELLING 9

Overview: Origins and challenges of NLP Language and Grammar-Processing Indian Languages- NLP Applications Information Retrieval. Language Modeling: Various Grammar - based Language Models, Statistical Language Model.

Part – A

1. Define NLP. (R)
   NLP stands for Natural Language Processing, which is a part of Computer Science, Human language, and Artificial Intelligence. It is the technology that is used by machines to understand, analyse, manipulate, and interpret spoken and written human's languages.

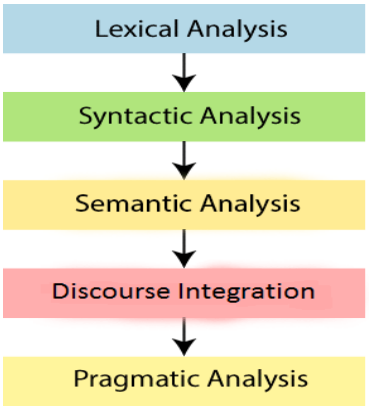
2. List out the driving force of NLP. (U)
   - ✔ Virtual assistants
   - ✔ Speech recognition
   - ✔ Sentiment analysis
   - ✔ Automatic text summarization
   - ✔ Machine translation and much more

3. Mention the components of NLP. (R)
   There are two components of NLP,
   a. Natural Language Understanding (NLU)
   b. Natural Language Generation (NLG).

4. Compare NLU and NLG. (AN)
5.

NLU	NLG
NLU is the process of reading and interpreting language.	NLG is the process of writing or generating language.
It produces non-linguistic outputs from natural language inputs.	It produces constructing natural language outputs from non-linguistic inputs.

6. Illustrate the phases of NLP. (U)



7. What are the applications of NLP? (U)
   - ✔ Question Answering
   - ✔ Spam Detection
   - ✔ Sentiment Analysis
   - ✔ Machine Translation
   - ✔ Spelling Correction
   - ✔ Speech Recognition
   - ✔ Chatbot
   - ✔ Information Extraction
   - ✔ NLU

8. Outline the operations involved in Morphological Analysis. (U)
   - ✔ Tokenization
   - ✔ Stop-word removal
   - ✔ Stemming
   - ✔ N-gram models

9. What is meant by Tokenization? Give example. (U)
   Tokenization refers to the process of converting a sequence of text into smaller parts, known as tokens.
   Example. John ate the pizza!
   We extract tokens or the words separated by spaces.
   John, ate, the, pizza, !

10. How will you remove unwanted tokens from the input sentence? (U)
    Stop word removal: It is a preprocessing step in NLP that involves removing common, non-meaningful words like ―the‖ and ―and‖ from text data.
    Example. John ate the pizza!
    We extract tokens or the words separated by spaces.
    John, ate, the, pizza, !

For the above example, from generated tokens I don‘t want unwanted tokens like punctuations, articles irrelevant, prepositions, meaningful words.

After removing stopwords, I get the words,

John, ate, pizza

11. Define the term Stemming. ®

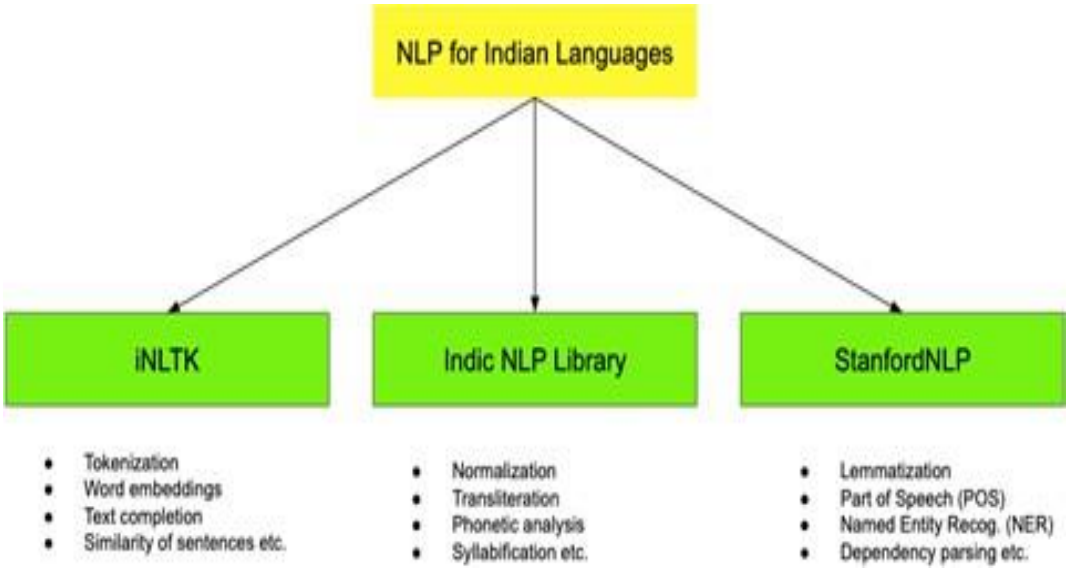
It is process of reducing words into its base form (root form/stem form)

Example.

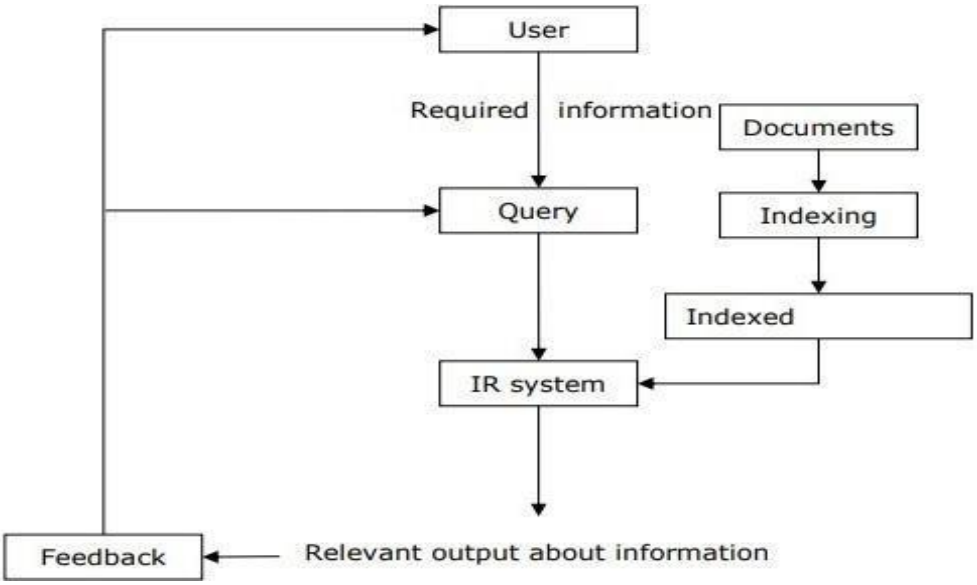
John / John
Ate / eat
Pizza / pizza
Car, cars / car
Run, ran, running / run
Stemmer, stemming, stemmed / stem

12. Analyse the various challenges involved in developing NLP applications. (AN)
    1. Contextual words and phrases and homonyms
    2. Synonyms
    3. Irony and sarcasm
    4. Ambiguity
    5. Errors in text or speech
    6. Colloquialisms and slang
    7. Domain-specific language
    8. Low-resource languages
    9. Lack of research and development

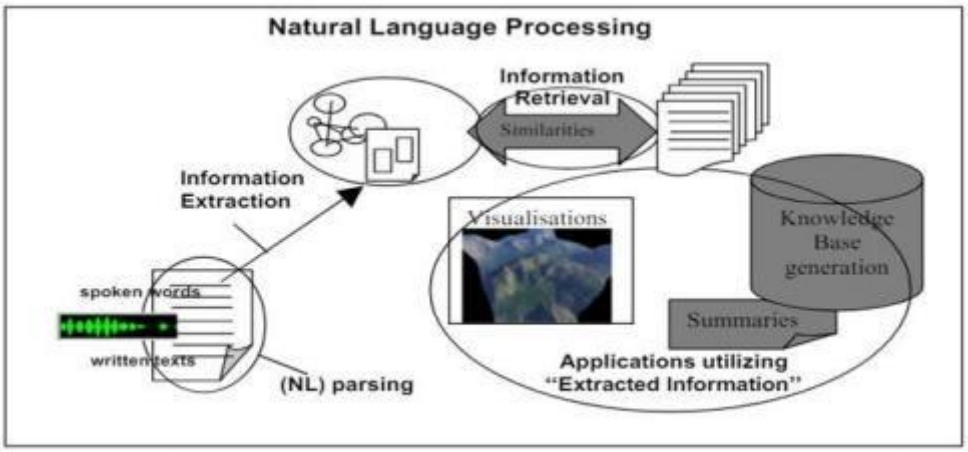
13. Mention the libraries used for processing Indian languages. (R)



14. Sketch out the process of information retrieval. (U)



15. Sketch the various processes related with Information Retrieval. (U)



Natural Language Processing, NL parsing, Information Extraction and Information Retrieval

16. What are the basic components of IR model? (U)

    IR model is basically a pattern that defines the above-mentioned aspects of retrieval procedure and consists of the following −

    - ✔ A model for documents.
    - ✔ A model for queries.
    - ✔ A matching function that compares queries to documents.

17. What do you mean by Language Modelling? (U)

    Language Modelling (LM) is the technique for utilization of different factual and probabilistic procedures to decide the occurrence of a given group of words happening in a corpus / sentence. These models help to examine collections of text information to give a premise to their base and forecast a prediction.

18. List out the capabilities of Language Models. (U)
    - ✔ Content generation
    - ✔ Part-of-Speech (PoS) tagging
    - ✔ Question Answering
    - ✔ Text Summarization
    - ✔ Sentiment Analysis
    - ✔ Conversational AI
    - ✔ Machine translation
    - ✔ Code completion

19. What are the tasks that cannot be performed by Language Models? (AN)
    Language Models can‘t perform tasks that involve
    - ✔ Common-sense knowledge,
    - ✔ Understanding abstract concepts, and
    - ✔ Making inferences based on incomplete information.
    - ✔ They also lack the ability to understand the world as humans do, and they can't make decisions or take actions in the physical world.

20. What is statistical language modeling in NLP? (U)
    Statistical Language Modeling, or Language Modeling and LM for short, is the development of probabilistic models that can predict the next word in the sequence given the words that precede it.

21. Categorize Language Model. (U)
    1. Grammar-based models
        - Rule-based models
            - Context-Free Grammars (CFG)
            - Phrase Structure Grammar (PSG)
            - Transformational Grammar
        - Dependency models
    1. Statistical Language Models
        - N-gram Model
        - Neural Network-Based Model

i.

Part – B

1. Summarize the phases of Natural Language Processing. (U)
2. Explain the various applications of NLP with relevant examples. (U)
3. Apply the various steps of NLP to process the below sentence: ‒London is the capital and most populous city of England and the United Kingdom.‖ (A)

4. Dissect the various challenges of NLP with proper justification. (AN)
5. Explain in detail about the text processing of Indian languages using Python. (U)
6. Explain NLP applications Information Retrieval in detail with relevant sketch. (U)
7. Explain the various types of Information Retrieval Model. (U)
8. Explain the different types of language modelling. (U)
9. Detail about the applications and drawbacks of Statistical Language Modelling. (AN)

UNIT II WORD LEVEL ANALYSIS AND MORPHOLOGY 9

Unsmoothed N-grams, Evaluating N-grams, Smoothing, Interpolation and Backoff – Word Classes, Part-of-Speech Tagging,Rule-based, Stochastic and Transformation-based tagging, Issues in PoS tagging – Hidden Markov and Maximum Entropy models- Morphological analysis and generation using Finite State Automata and Finite State transducer

Part – A

1. What Are N-Grams(ngrams)? (U)
   N-grams are continuous sequences of words or symbols, or tokens in a document. In technical terms, they can be defined as the neighboring sequences of items in a document. They come into play when we deal with text data in NLP (Natural Language Processing) tasks. They have a wide range of applications, like language models, semantic features, spelling correction, machine translation, text mining, etc.

2. Give some examples of N-Grams. (U)
   Let‘s understand n-grams practically with the help of the following sample sentence:
   ―I reside in Bengaluru‖.

SL.No.	Type of n-gram	Generated n-grams
1	Unigram	[―I‖,‖reside‖,‖in‖,‖Bengaluru‖]
2	Bigram	[―I reside‖,‖reside in‖,‖in Bengaluru‖]
3	Trigram	[―I reside in‖, ―reside in Bengaluru‖]

3. Classify N-grams. (AN)
   N-grams are classified into different types depending on the value that n takes.
   - When n=1, it is said to be a unigram.
   - When n=2, it is said to be a bigram.

- When n=3, it is said to be a trigram.
- When n=4, it is said to be a 4-gram, and so on.

4. What are n-grams used for in NLP? (U)
   N-grams are used in the various use cases of NLP, such as spelling correction, machine translation, language models, semantic feature extraction, etc.

5. What is the difference between n-grams and bigrams? (U)
   The ‗n' in n-grams refers to the no. of sequences of tokens. Hence, when the value of n=2, it's known as bigrams.

6. What are the advantages and disadvantages of using n-grams in NLP? (U)
   Here are the advantages and disadvantages of n-grams in NLP.
   Pros
   The concept of n-grams is simple and easy to use yet powerful. Hence, it can be used to build a variety of applications in NLP, like language models, spelling correctors, etc.
   Cons
   N-grams cannot deal Out Of Vocabulary (OOV) words. It works well with the words present in the training set. In the case of an Out Of Vocabulary (OOV) word, n-grams fail to tackle it.

   Another serious concern about n-grams is that it deals with large sparsity.

7. Define Chain Rule. (R)
   Chain Rule calculates the joint probability of a sequence by using the conditional probability of a word given the previous words. It allows us to decompose the probability of a sequence of words.
   It is given as:
   $P(X_1,X_2,.....X_n) = P(X_1)\ P(X_2|X_1)\ P(X_3|X_1X_2)\ ….. p(X_n|X_1X_2…X_{n-1})$

   The chain rule to compute the joint probability of words in sentence is:

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

8. Define Markov Model. (R)
   Markov Models are the class of probabilistic models that assume we can predict the probability of some future unit without looking too far into the past.
   Generalized equation for N-gram approximation to the conditional probability of the next word in a sequence is given as,

$$\mathbf{P(w_n|w_1,w_2,....w_{n-1})} \approx \mathbf{P(w_n|w_{n-k}....w_{n-1})}$$

$$\mathbf{P(w_1,w_2,....w_k)} \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

9. Demonstrate the use of MLE. (U)

   MLE is a method that determines values for the parameters of the model that maximizes the likelihood of the data. It is the simplest method to estimate parameters. 21ML1601 – NLP Unit – 2 III Year / VI Semester AI&DS 5 It produces the MLE estimate for the parameters of an N-gram model by normalizing counts from a corpus so that they lie between 0 and 1. It is done using the following formula: P(wn|wn-1) = $\boldsymbol{C(wn-1wn)}$ /$\boldsymbol{C(wn-1)}$

10. Compare the two types of N-gram Evaluation Models. (AN)

    There are two types of Evaluation:

    1. Extrinsic Evaluation
    2. Intrinsic Evaluation

    1. Extrinsic Evaluation:

    The best way to evaluate a model is to check how well it is predicted in end-to-end application testing. This approach is known as extrinsic evaluation, but it is time-consuming and expensive.

    2. Intrinsic Evaluation:

    It is an alternative approach is to define a suitable metric and evaluate it regardless of the application called intrinsic evaluation. This does not guarantee the performance of the application. However, this is a quick first step in verifying algorithmic performance.

11. Define the term Perplexity. (R)

    The perplexity (sometimes called PP for short) of a language model in a test sentence is the inverse probability of the test sentence, normalized by the number of words.

    If the model gives the

    highest P(sentence), then Lower value of perplexity -> Better Model.

    Lowest P(sentence), the Higher value of perplexity -> Confused for prediction

12. What is meant by Smoothing? (U)

    Smoothing is the process of flattening a probability distribution implied by a language model so that all reasonable word sequences can occur with some probability.

13. Why do we need smoothing? (AN)

✔ In a language model, we use parameter estimation (MLE) on training data. We can‘t actually evaluate our MLE models on unseen test data because both are likely to contain words/n-grams that these models assign zero probability to.

- ✔ Relative frequency estimation assigns all probability mass to events in the training corpus. But we need to reserve some probability mass to events that don‘t occur (unseen events) in the training data.
- ✔ Smoothing mechanisms are used in language models to address the issue of data sparsity and improve the generalization of the model.
- ✔ Data sparsity refers to the problem of encountering unseen or infrequent n-grams in the training data, which can lead to zero probabilities or unreliable probability estimates.

14. List out the various smoothing techniques. (R)
    1. Add-1 (Laplace) Smoothing
    2. Good Turing Discounting
    3. Back off Smoothing

15. Compare Open and Closed word classes. (AN)

Aspect	Open Word Classes	Closed Word Classes
Definition	Categories accepting new members, content words	Fixed membership, structural words
Characteristics	More varied, represent concrete concepts	Limited members, grammatical functions
Examples	Nouns: "dog," "house"	Determiners: "the," "a"
	Verbs: "run," "eat"	Pronouns: "he," "she"
	Adjectives: "big," "happy"	Prepositions: "in," "on"
	Adverbs: "quickly," "happily"	Conjunctions: "and," "but"
		Articles: "a," "an"
		Auxiliary Verbs: "is," "have"
Membership	Dynamic, potential for new additions	Fixed, rarely changes

16. Define Part of Speech. (R)
    Part of Speech is the classification of words based on their role in the sentence.

17. What is meant by Parts-Of-Speech tagging? (U)
    Part-of-speech (POS) tagging is the process of labeling words in a text with their corresponding parts of speech in natural language processing (NLP). It helps algorithms understand the grammatical structure and meaning of a text.

18. List out the basic part of speech tags used for NLP. (U)

[1] Noun (N) — A noun is the name of a person, place, thing, or idea. - John, London, Table, Teacher, Pen, City, Happiness, Hope

[2] Pronoun(PRO) — A pronoun is a word used in place of a noun. - I, We, They, You, He, She, It, Me, Us, Them, Him, Her, This, That

[3] Verb (V) — A verb expresses action or being. - Read, Eat, Go, Speak, Run, Play, Live, Have, Like, Are, Is

[4] Adverb(ADV) — An adverb modifies or describes a verb, an adjective, or another adverb. - Slowly, Quietly, Very, Always, Never, Too, Well, Tomorrow

[5] Adjective(ADJ) — An adjective modifies or describes a noun or pronoun. - Big, Happy, Green, Young, Fun, Crazy, Three

[6] Preposition (P) — A preposition is a word placed before a noun or pronoun to form a phrase modifying another word in the sentence. - At, On, In, From, With, Near, Between, About, Under

[7] Conjunction (CON) — A conjunction joins words, phrases, or clauses. - And, Or,But, Because, So, Yet, Unless, Since, If

[8] Interjection (INT) — An interjection is a word used to express emotion. - Ouch! Wow! Great! Help! Oh! Hey! Hi!

[9] Determiner or Article (DT) - A grammatical marker of definiteness (the) or indefiniteness (a, an). These are not always considered POS but are often included in POS tagging libraries.

19. Mention the use of PoS. (U)
    - ✔ To understand the grammatical structure of a sentence.
    - ✔ To disambiguate words with multiple meanings.
    - ✔ To improve the accuracy of NLP tasks.
    - ✔ To facilitate research in linguistics.

20. List out the applications of POS Tagging. (U)
    - Information extraction.
    - Named entity recognition.
    - Text classification.
    - Machine translation.
    - Natural language generation.

21. Categorize PoS Tagging. (U)
    1. Rule-based PoS Tagging
    2. Stochastic PoS Tagging
    3. Transformation-based PoS Tagging

22. Define Markov Chain. (R)

    A Markov Chain is a special case of a weighted automation in which the input sequence uniquely determines which states the automation will go through. The transition assumes that the probability of moving to the next state is solely dependent on the current state.

23. State ―Markov Assumption‖. (R)

For the bigram model, we can approximate the probability of a word given all the previous words $\mathbf{P(w_n|w_1,w_2,....w_{n-1})}$ by using only the preceding word $\mathbf{P(w_n|w_{n-1})}$

ie. $\mathbf{P(w_n|w_1,w_2,....w_{n-1}) \approx P(w_n|w_{n-1})}$

This assumption that the probability of a word depends only on the previous word is called Markov Assumption.

24. Give the formal definition of Markov Chain. (R)

First, let's be more formal. We'll view a Markov chain as a kind of probabilistic **graphical model**; a way of representing probabilistic assumptions in a graph. A Markov chain is specified by the following components:

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{01} a_{02} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
q_0, q_F	a special **start state** and **end (final) state** which are not associated with observations.

25. Summarize about HMM. (U)

HMM is a probabilistic model that consists of a sequence of hidden states, each of which generates an observation. The goal of HMM is to estimate the sequence of hidden states based on the sequence of observations.

Let's begin with a formal definition of a Hidden Markov Model, focusing on how it differs from a Markov chain. An **HMM** is specified by the following components:

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, \ldots, v_V$.
$B = b_i(o_t)$	a sequence of **observation likelihoods:**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i.
q_0, q_F	a special **start state** and **end (final) state** which are not associated with observations, together with transition probabilities $a_{01} a_{02} .. a_{0n}$ out of the start state and $a_{1F} a_{2F} ... a_{nF}$ into the end state.

26. What is the difference between Markov Chain and HMM? (AN)

Markov chain is useful when we need to compute a probability for a sequence of events that we can observe in the world.

HMM can compute the probability for both observed events (like words that we see in the input) and hidden events (like POS tags).

27. What is the significance of Maximum Entropy Models in NLP? (U)
    We can see any problems in natural language processing as linguistic classification problems in which linguistic contexts are used to predict linguistic classes. Maximum entropy models are a clean way to combine various pieces of contextual evidence to estimate the probability of a particular linguistic class occurring with a specific linguistic context.
    Maximum entropy classification is a method that generalizes logistic regression to multiclass problems. The Maximum Entropy model is a type of log-linear model.

28. Give the formal definition of FST. (R)
    A Finite State Transducer (FST) is a finite state machine with two tapes: an input tape and an output tape, with finite number of states.

$Q = \{q_0, q_1, \ldots, q_{N-1}\}$, is finite set of states,
Σ is finite alphabet of complex symbols, and
$\Sigma \subseteq I \times O$, where I is set of input alphabet, and O is set of out alphabet
q_0 is start state, and
$\delta : Q \times \Sigma \rightarrow Q$, for example, $\delta(q', i : o) = q_j$ is transition function.

The input (I) and output (O) symbols, both include the empty string symbol ε. For $\Sigma = \{a, b, !\}$, corresponding to the sheep-talk language discussed earlier, the FST has $i : o$ set as, {*a:a, b:b, !:!, a:!, a:ε, ε:!*}.

The FSTs are useful for variety of applications:

- *Word inflections:*[1] For example, finding the plural of the words, cat → cats, dog → dogs, goose → geese, etc.
- *Morphological parsing:* Extracting the properties of a word, e.g., cats → cat + [nouns] + [plural].
- Simple word *translations:* For example, US English to UK English.
- Simple *commands* to computer.

29. What is meant by Morphology? Mention the types of Morphology. (U)
    Morphology is the study of how the words are constructed. Construction of English language words through attachment of prefixes and suffixes (both together called affix) are called concatenation morphology.
    Types of Morphology:

1. **Inflection:** - Inflectional morphology forms words using the same group word stem.

Table 10.2: Inflectional Morphology

Type	Regular nouns	Irregular nouns
Singular	cat, thrush	mouse, ox
Plural	cats, thrushes	mice, oxen

2. **Derivation:** - Derivations morphology produces a words of different stem, for example computerization (noun) from computerize (verb) – the words belong to different groups.

Table 10.3: Derivational Morphology

Suffix	Base verb/adjective	Derived Noun
-ation	computerize (V)	Computerization
-ee	appoint (V)	appointee
-er	kill (V)	killer
-ness	fuzzy (A)	fuzziness

30. Compare Finite State Automata and Finite State Transducers. (AN)

Aspect	Finite State Automata (FSA)	Finite State Transducers (FST)
Purpose	Recognize or accept strings	Translate or transduce input sequences into output sequences
Description	Finite set of states, transitions, start state, and accepting states	Finite set of states, transitions, input/output alphabets
Functionality	Processes input symbols, accepts or rejects strings	Processes input symbols, generates output symbols based on transitions
Output	Does not produce output symbols	Produces output symbols during processing
Applications	Recognizing strings in regular languages	Natural language processing tasks like morphological analysis, machine translation
Typical Use Cases	Text search algorithms, regular language recognition	Spelling correction, machine translation, speech recognition
Examples	Aho-Corasick algorithm, regex matching	Phonological rule systems, spell checkers

Part – B

1. How N-Grams are evaluated? Explain with relevant example. (U)
2. Make use of N-gram model for predicting the next word. Explain the same in detail with relevant example. (U)
3. What is the most probable next word predicted by the unigram model for the following word sequence? (A)
   <s>I am Henry</s>
   <s>I like college</s>
   <s>Do Henry like college</s>
   <s>Henry I am</s>
   <s>Do I like Henry</s>
   <s>Do I like college</s>
   <s>I do like Henry</s>
4. How the N-gram language models are evaluated? Explain with example. (A)
   Consider the following corpus of 3 sentences:
   There is a big garden

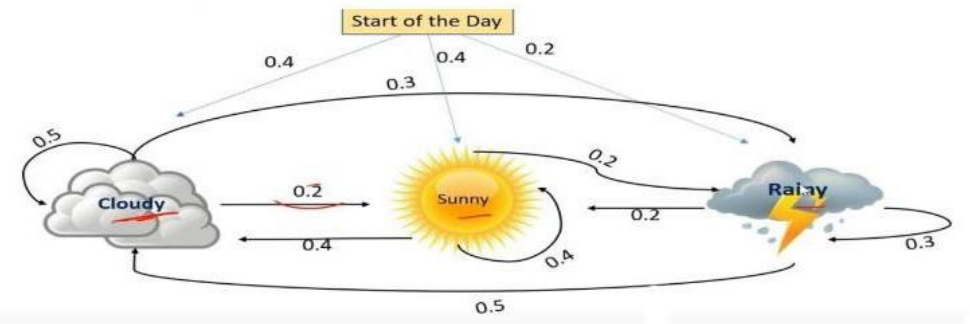
Children play in a garden

They play inside beautiful garden

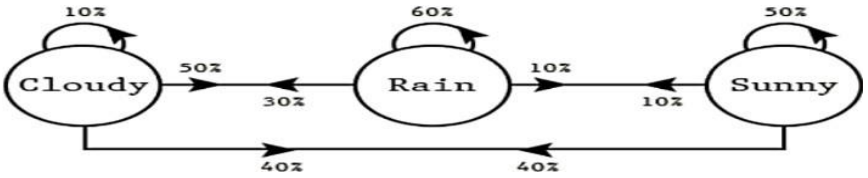
i. Calculate P(They play in a big garden) assuming bi-gram language model.

ii. Calculate the perplexity of They play in a big garden.

5. Why do we need smoothing in language models? Explain the following smoothing techniques: Add-1 Smoothing, and Interpolation. (AN)
6. Elaborate on Parts-of-Speech Tagging. (U)
7. Explain PoS Tagging. Apply PoS Tagging to tokenize the following sentence:
   —Apple is planning to buy Indian startup for $1 billion‖ (A)
8. Explain the steps involved in PoS Tagging. Apply PoS Tagging to tokenize the following sentence: (A)
   —Taj Mahal is the most beautiful historical remark of India"
9. Explain in detail about Hidden Markov and Maximum Entropy models. (U)
10. Explain in detail about Hidden Markov Model. Consider the given HMM and estimate the probability of the observation: (A)
    Cloudy ✇Sunny ✇Cloudy ✇Rainy



11. Explain in detail about Markov Chain. Consider the given Markov Chain and estimate the probability that Wednesday will be cloudy if today (Monday) is sunny. (A)



12. Explain the process of Morphological Analysis and Generation using FSA. (U)

UNIT III SYNTACTIC ANALYSIS 9

Context-Free Grammars, Grammar rules for English, Treebanks, Normal Forms for grammar – Dependency Grammar – Syntactic Parsing, Ambiguity, Dynamic Programming parsing – Shallow parsing – Probabilistic CFG, Probabilistic CYK, Probabilistic Lexicalized CFGs – Feature structures, Unification of feature structures.

Part – A

1. Give the formal definition of CFG. (U)

   Context Free Grammar (CFG) - Formal Definition

   Context-free grammar G is a 4-tuple.

   G = (V, T, S, P)

   These parameters are as follows;

- **V** – ***Set of variables*** (also called as **Non-terminal symbols**)
- **T** – ***Set of terminal symbols*** (**lexicon**)
- The symbols that refer to words in a language are called **terminal symbols.**
- **Lexicon** is a set of rules that introduce these symbols.
- **S – Designated start symbol (**one of the non-terminals, S Є V**)**
- **P – Set of productions (**also called as rules**).**
- Each rule in P is of the form A → s, where
- A is a non-terminal (variable) symbol.
- Each rule can have only one non-terminal symbol on the left hand side of the rule.
- s is a sequence of terminals and non-terminals. It is from (T U V)*, infinite set of strings.
- **A grammar G generates a language L.**

2. Define the term Corpus. (R)

   A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting. Its plural is corpora. They can be derived in different ways like text that was originally electronic, transcripts of spoken language and optical character recognition, etc.

3. What is Treebank? (U)

   Treebank is a corpus in which each sentence is annotated with a parse tree. It represents syntactic and semantic relations of words in a sentence. Treebanks are created by

   - Parsing texts using parsers
   - Human annotations

4. List the types of Treebanks. (U)

   Semantic and Syntactic Treebanks are the two most common types of Treebanks in linguistics.

   1. Semantic Treebanks: -
      - These Treebanks use a formal representation of sentence‘s semantic structure.
      - They vary in the depth of their semantic representation. Robot Commands Treebank, Geoquery, Groningen Meaning Bank, RoboCup Corpus are some of the examples of Semantic Treebanks.
   2. Syntactic Treebanks: -
      - Opposite to the semantic Treebanks, inputs to the Syntactic Treebank systems are expressions of the formal language obtained from the conversion of parsed Treebank data.
      - The outputs of such systems are predicate logic based meaning representation.

5. Write a simple example for Treebank. (U)

   Here's the parse tree for the sentence "The cat chased the mouse" in the Penn Treebank format:

   (S
   (NP (DT The) (NN cat))
   (VP (VBD chased)
   (NP (DT the) (NN mouse))))

6. Summarize the steps to extract grammars from a Treebank Structure. (U)
   1) Parse Tree Analysis
   2) Identify Constituent Phrases
   3) Extract Production Rules
   4) Generalize Rules
   5) Refinement and Validation
   6) Grammar Representation
   7) Application

7. Compare the two kinds of grammar equivalence. (AN)

   1. Weak Equivalence
   2. Strong Equivalence

   1. Strong Equivalence: - Two grammars are strongly equivalent if they generate the same set of strings and if hey assign the same phrase structure to each sentence.

   2. Weak Equivalence: Two grammars are weakly equivalent if they generate the same set of strings but do not assign the same phrase structure to each sentence

8. Outline the steps for converting CFG to CNF. (U)

   The conversion to Chomsky Normal Form has four main steps:

1. Get rid of all ε productions.
2. Get rid of all productions where RHS is one variable.
3. Replace every production that is too long by shorter productions.
4. Move all terminals to productions where RHS is one terminal.

9. Convert the following grammar to CNF. (A)
   After the first step, one has:
   S→AbA|bA|Ab|b
   A→Aa|a

   The second step does not apply.

   After the third step,one has:
   S→TA|bA|Ab|b
   A→Aa|a
   T→Ab

   And finally, one has:
   S→TA|BA|AB|b
   A→AC|a
   T→AB
   B→b
   C→a

10. What is Dependency Parsing? (U)
    Dependency parsing is a technique used in natural language processing for analyzing the grammatical structure of sentences. A Dependency Parser simply transforms a sentence into a Dependency Tree. It involves identifying the relationships between words in a sentence and representing them in the form of a dependency tree

11. What is meant by Syntactic Ambiguity? (U)
    Syntactic Ambiguity refers to ambiguity in sentence structure and be able to interpret in different forms. This structural ambiguity occurs when the grammar assigns more than one possible parse to a structure.
    Example: Consider the following sentence:
    ‒I shot an elephant wearing pyjama‖
    The above sentence has the structural ambiguity as discussed below:
    - First, Does shoot mean taking a photo or pointing a gun to?
    - Second, who is wearing pyjama? Is it the person or the elephant?

12. Summarize the types of Structural Ambiguity. (AN)
    a) Attachment Ambiguity:- A sentence has an attachment ambiguity if a particular constituent can be attached to the parse tree at more than one place.
    b) Coordination Ambiguity: In this, different set of phrases can be conjoined by a conjunction like ‒and‖.

c) Local Ambiguity: Even if a sentence is not ambiguous [ie. It does not have more than one parse in the end], it can be inefficient to parse because of local ambiguity. Local ambiguity occurs when some part of sentence is ambiguous, ie. It has more than one parse, even if the whole sentence is not ambiguous.

13. What is meant by Disambiguation? (U)
    It is the process of determining the choosing the correct parse from the multitude of possible parses. It is the group of techniques to handle ambiguity.

14. Why dynamic programming is attractive? (U)
    Dynamic Programming provides a framework for solving the problems such as, ambiguity, repeated substructures, recursion and space and time complexity, arises during the syntactic parsing.

15. Summarize about chunking with example. (U)
    Chunking is the process of identifying and classifying the flat, nonoverlapping segments of a sentence that constitute the major parts-ofspeech found in wide-coverage grammars.
    Chunk: typically includes headword and pre-head material.
    Example:
    [NP The HD box] that [NP you] [VP ordered] [PP from] [NP shaw] [VP never arrived]

16. Mention the uses of Feature Structures. (U)
    Feature structures are used for the representation of linguistic information in several grammar formalisms for natural language processing. These structures are a type of directed graph, in which arcs are labelled by names of features, and nodes correspond to values of features.

17. Describe about Unification. (U)
    Unification is a (partial) operation on feature structures. Intuitively, it is the operation of combining two feature structures such that the new feature structure contains all the information of the original two, and nothing more.

PART – B

1. How Context-Free Grammars are utilized in NLP? Explain with neat example. (U)
2. Given the following CFG grammar from ATIS System, USA. Construct the necessary derivations and parse tree to perform syntactic analysis of the following sentence using the parsing method. (A)
   –Book the flight through Houston.‖
   S → NP VP
   S → Aux NP VP
   S → VP
   NP → Pronoun
   NP → Proper-Noun

NP → Det Nominal
Nominal → Noun
Nominal → Nominal Noun
Det → that | this | a | the
Noun → book | man | flight
Verb → book | include | prefer | man
Pronoun → I | she | me
Proper-Noun → Houston | TWA
Aux → does
Preposition → from | to | through
Nominal → Nominal PP
VP → Verb
VP → Verb NP
VP → Verb NP PP
VP → Verb PP
VP → VP PP
PP → Preposition NP

3. Derive and Construct a top-down, left-to-right parse tree for the given sentence: ―The angry bear chased the frightened little squirrel‖ (C)
   Use the following grammar rules to create the parse tree.

S → NP VP	Det → the
NP → Det Nom	Adj → little | angry | frightened
VP → V NP	N → squirrel | bear
Nom → Adj Nom | N	V → chased

4. Explain the use of Treebanks in syntactic annotations. (U)
5. Explain in detail about CKY parsing technique with example. (U)
6. Consider the following grammar rules: (A)
   X → S eos
   S → NP VP
   NP → DET N
   NP → N
   VP → V NP
   N → Mary
   N → otter
   DET → the
   V → feeds
   eos → eos
   Parse the below sentence using Earley Algorithm: **"Marry feeds the otter eos‖**
7. Illustrate the working of Chart Parsing with necessary example. (A)
8. Write operations and algorithm of chart parsing. (R)
9. How PCFG can be useful in Disambiguation? Draw the possible parse trees to parse the sentence: ―astronomers saw stars with ears‖. And also evaluate the probabilities of the parse trees. Use the following grammar to parse the given sentence: (E)

S → NP VP [1.0]	NP → NP PP [0.4]
PP → P NP [1.0]	NP → astronomers [0.1]

VP → V NP [0.7] NP → ears [0.18]
VP → VP PP [0.3] NP → saw [0.04]
P → with [1.0] NP → stars [0.18]
V → saw [1.0] NP → telescope [0.1]

10. Draw the possible parse trees to parse the sentence: "I saw the man using the telescope". And also evaluate the probabilities of the parse trees. (E)
    Use the following grammar to parse the given sentence:
    S → NP VP [1.0]
    NP → PRP [0.5] | Det NN [0.5]
    VP → VBD NP [0.7] | VBD NP PP [0.3]
    PP → IN NP [1.0]
11. Consider the grammar in Question 2, and apply chart parsing to parse the sentence "I saw the man using the telescope". (A)
12. Write operations and algorithm of chart parsing. (A)

Given the following grammar rules, parse the sentence:

"The cute girls sing a song"

Rule No	Rules	Dictionary Words
1	S → NP VP	Det → a | the | an
2	NP → Det Noun	Noun → girls | apple | song
3	NP → Det Adj Noun	Adj → cute | smart
4	NP → Adj Noun	Verb → sing | ate
5	VP → Verb	
6	VP → Verb NP	

13. Compare the various dynamic programming parsing techniques. (AN)
14. Discuss in detail about shallow parsing. (U)
15. Illustrate the significance of PCKY parsing technique. (U)
16. Elaborate on feature Structures and Unification. (U)

UNIT IV INFORMATION RETRIEVAL AND LEXICAL RESOURCES 9

Information Retrieval: Design features of Information Retrieval Systems-Classical, Non classical, Alternative Models of Information Retrieval – valuation Lexical Resources: World Net-Frame Net-Stemmers-POS Tagger- Research Corpora.

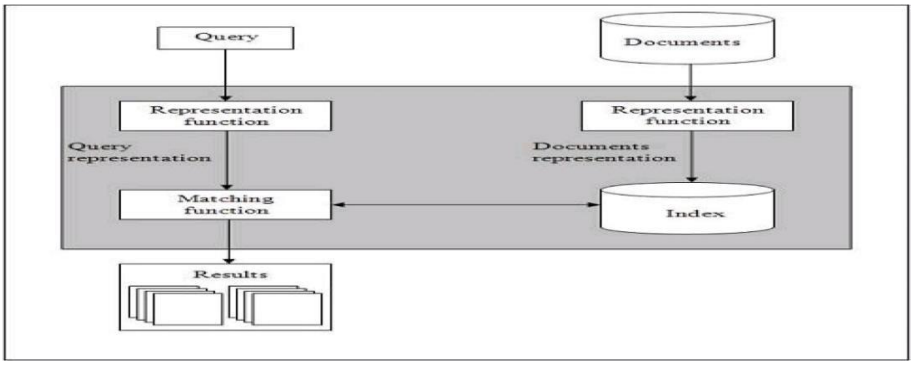
Part – A

1. What is Information Retrieval? (U)
   Information Retrieval (IR) is defined as the process of accessing and retrieving the most appropriate information from text based on a particular query given by the user, with the help of context-based indexing or metadata.
   Google Search is the most famous example of information retrieval.

In other words, Information retrieval (IR) may be defined as a software program that deals with the organisation, storage, retrieval and evaluation of information from document repositories.

2. List out the various entities of IR Model. (U)
   The information retrieval model is basically a pattern that defines the most important aspects of the retrieval procedure and consists of a set of entities:
   - A model for documents,
   - A model for queries, and
   - A matching function that compares queries to documents.

3. Summarise the various components of the IR System. (R)
   In the mathematical formulation, a retrieval model can be seen as consisting of:
   - D − Representation for documents, R − Representation for the queries Q
   - F − The modeling framework for documents and queries along with the relationship between documents and queries
   - R (q,di) − Similarity function which orders the documents with respect to the query which is also called ranking.
4. Sketch the basic IR Model. (U)



5. What are the two processes of retrieval models? (R)
   Retrieval Systems consist of mainly two processes:
   1. Indexing
   2. Matching

6. Summarize about Indexing. (U)
   It is the process of selecting terms to represent a text.
   Indexing involves:
   - ✓ Tokenization of string
   - ✓ Removing frequent words
   - ✓ Stemming

   Two common Indexing Techniques:
   - ✓ Boolean Model
   - ✓ Vector space model

7. How will you find the similarity between two text representations? (A)
   Matching is the process of finding a measure of similarity between two text representations.
   The relevance of a document is computed based on the following parameters:
   1. TF: It stands for Term Frequency which is simply the number of times a given term appears in that document.
      TF (i, j) = (count of ith term in jth document)/(total terms in jth document)
   2. IDF: It stands for Inverse Document Frequency which is a measure of the general importance of the term.
      IDF (i) = (total no. of documents)/(no. of documents containing ith term)
   3. TF-IDF Score (i, j) = TF * IDF

8. Summarize the design features of Information Retrieval Systems. (AN)
   1. Inverted Index
   2. Stop word Elimination
   3. Stemming
   4. Crawling
   5. Query
   6. Relevance Feedback

9. Why classical models are attractive? List out the various classical models of IR. (AN)
   Classical Model is the simplest model to build an Information Retrieval system. This model is based on the well-recognized and easy to understood knowledge of mathematics like probability. Classical models are easy to implement and are very efficient.
   The three classical models of information retrieval are:
   a. Boolean model;
   b. Vector space model; and
   c. Probabilistic models.

10. Consider the documents collection:
    dl = –Sachin scores hundred.‖
    d2 = –Dravid is the most technical batsman of the era.‖
    d3 = –Sachin, Dravid duo is the best to watch.‖
    d4 = –India wins courtesy to Dravid, Sachin partnership‖
    Find the relevant documents for the query **Sachin AND Dravid.** (A)

    Lexicon and inverted index:
    Sachin → {d1, d3, d4}
    Score → {d1}
    Hundred → {d1}
    Dravid → {d2, d3, d4}
    technical → {d2}
    batsman → {d2}

watch → {d3}
India → {d4}
partnership → {d4}
win → {d4}

Result set:

{Dl, D3, D4} AND {D2, D3, D4} = {D3, D4}

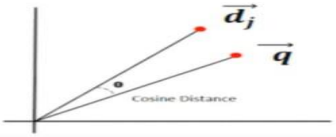
11. What is meant by Similarity Measure? (U)

A Similarity Measure is a function which computes the degree of similarity between a pair of vectors or documents. It is used to determine which document (d1 or d2) is more similar to a given query q.

Degree of similarity between $\vec{q}$ and $\vec{d_j}$

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|}$$

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$



The factor $\vec{q}$ does not affect the ranking because it is the same for all documents.

The factor $\vec{d_j}$ provides a normalization in the space of the documents.

12. What is BIM? (U)

The Binary Independence Model (BIM) assumes that each term in a document is independent of all other terms, and the presence or absence of a term in a document is modelled with a Bernoulli distribution. This model simplifies the representation of documents and queries into binary term occurrences.

13. List out the various methods of Non-Classical IR system. (R)

1. Information Logic
2. Situation Theory
3. Interaction Models

14. What is WordNet? (U)

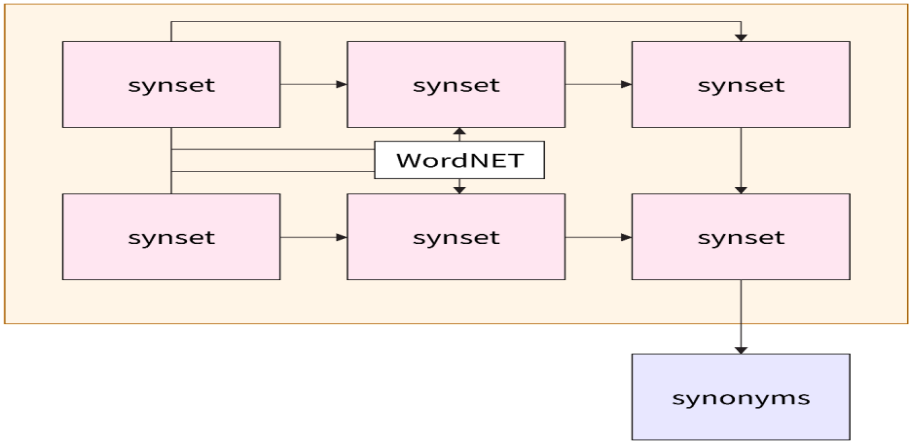
- –WordNet is a lexical database of semantic relations between words in more than 200 languages‖. It groups words into synsets.
- Synset is a group of words that reflects the same meaning in a given text. In simpler terms, a WordNet is similar to a thesaurus that groups words based on their meanings.
- WordNet comes as a part of NLTK corpus. It provides relations between various words. This knowledge can be used to build applications based on Informational Retrieval.

15. Mention the various relations used in WordNet. (R)

1. Homonym
2. Polysemy

3. Synonymy
4. Hyponymy
5. Antonyms
6. Hypernyms
7. Meronyms
8. Holonyms

16. Draw the structure of WordNet. (R)



17. Describe about Frame Net. (U)

FrameNet is a lexical database of English that describes the meaning of words and phrases in terms of frames and their semantic roles. A "frame" represents a conceptual structure or scenario, capturing a specific type of situation, event, or concept. Each frame contains "frame elements" (semantic roles) that describe the participants and their roles in the frame.

18. Mention the features of FrameNet. (U)

1. Frames: Frames represent conceptual structures or scenarios. Examples of frames include "Eating," "Buying," "Communication," etc.
2. Frame Elements (FEs): Frame elements are the building blocks of frames. They represent the semantic roles within a frame, such as "Agent," "Patient," "Instrument," etc.
3. Lexical Units (LUs): Lexical units are the words or phrases that evoke frames. Each LU is linked to one or more frames and specifies its sense and syntactic information.
4. Annotations: FrameNet provides annotations for each lexical unit, frame, and frame element. This includes definitions, examples of usage, and syntactic and semantic information.

19. How is FrameNet Used in NLP? (AN)

- ✓ Semantic Role Labelling (SRL)
- ✓ Information Extraction
- ✓ Sentiment Analysis
- ✓ Question Answering:

20. What is the use of Stemmer? (U)

A stemmer in lexical resources refers to a tool or algorithm used in Natural Language Processing (NLP) to reduce words to their base or root form, known as a "stem." The process of stemming involves removing affixes from words to obtain their common root form. This is useful in tasks like text normalization, search engines, and information retrieval where variations of words need to be treated as the same word for analysis or retrieval purposes.

21. Identify the stem word for the words in the below sentence: (A)

The quick brown foxes are jumping over the lazy dogs

Original: The	Stemmed: the
Original: quick	Stemmed: quick
Original: brown	Stemmed: brown
Original: foxes	Stemmed: fox
Original: are	Stemmed: are
Original: jumping	Stemmed: jump
Original: over	Stemmed: over
Original: the	Stemmed: the
Original: lazy	Stemmed: lazi
Original: dogs	Stemmed: dog

22. What is the use of POS-Tagger? (U)

Part-of-Speech (POS) taggers are tools or algorithms used in Natural Language Processing (NLP) to assign a part of speech to each word in a text corpus. These tags represent the grammatical category of the word, such as noun, verb, adjective, adverb, etc. POS tagging is crucial for various NLP tasks, including text analysis, information extraction, and syntactic parsing.

23. Identify the stem word for the words in the below sentence: (A)

The quick brown foxes are jumping over the lazy dogs

Word: The	POS Tag: DT
Word: quick	POS Tag: JJ
Word: brown	POS Tag: NN
Word: fox	POS Tag: NN
Word: jumps	POS Tag: VBZ
Word: over	POS Tag: IN
Word: the	POS Tag: DT
Word: lazy	POS Tag: JJ
Word: dog	POS Tag: NN

24. What is the significance of Research Corpora? (U)

Research corpora are essential components of lexical resources in Natural Language Processing (NLP) and linguistics. A corpus is a large and structured set of texts, often

collected for linguistic analysis and research. Research corpora serve as valuable data sources for various NLP tasks, language studies, and computational linguistics. They provide a diverse and comprehensive collection of texts for analyzing language patterns, building models, and training algorithms.

25. Give the classification of Research Corpora. (R)
    - ✓ General Corpora
    - ✓ Specialized Corpora
    - ✓ Annotated Corpora
    - ✓ Parallel Corpora

PART - B

1. Analyse and explain the design features of Information Retrieval System.(AN)
2. Illustrate the process of document retrieval using Boolean Model with neat example. (A)
3. Furnish the survey of Alternative IR models with advantages and disadvantages. (AN)
4. Furnish the survey of Non-Classical IR models with advantages and disadvantages. (AN)
5. Examine the use of stemmers with proper explanation. (U)
6. Examine the use of stemmers with proper explanation. (U)
7. Examine the use of WordNet with proper explanation. (U)
8. Examine the use of FrameNet with proper explanation. (U)

UNIT V APPLICATIONS IN NLP 9

Question Answering with SQUAD – Dependency Parsing – Machine Translation – Conference Resolution – Text Summarization-WordNet, PropBank, FrameNet, Brown Corpus, British National Corpus (BNC)

PART – A

1. What is meant by Question Answering? (U)
   Question answering is a critical NLP problem and a long-standing artificial intelligence milestone. QA systems allow a user to express a question in natural language and get an immediate and brief response.

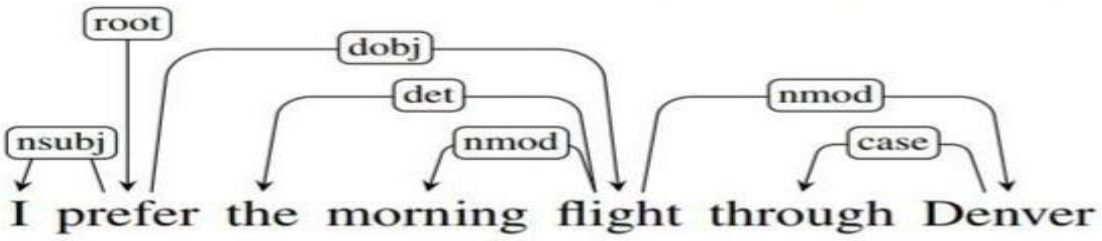
2. Describe about SQuAD Dataset. (U)
   The Stanford Question Answering Dataset () is a reading comprehension dataset made up of questions posed by crowd workers on a collection of Wikipedia articles, with the response to each question being a text segment, or span, from the relevant reading passage, or the question being unanswerable.

3. Write down the steps to perform question answering using SQUAD. (U)
   1) Prepare the Environment
   2) Load the SQUAD Dataset

3) Fine-tune Pre-Trained Models
4) Tokenization
5) Model Inference
6) Post-processing

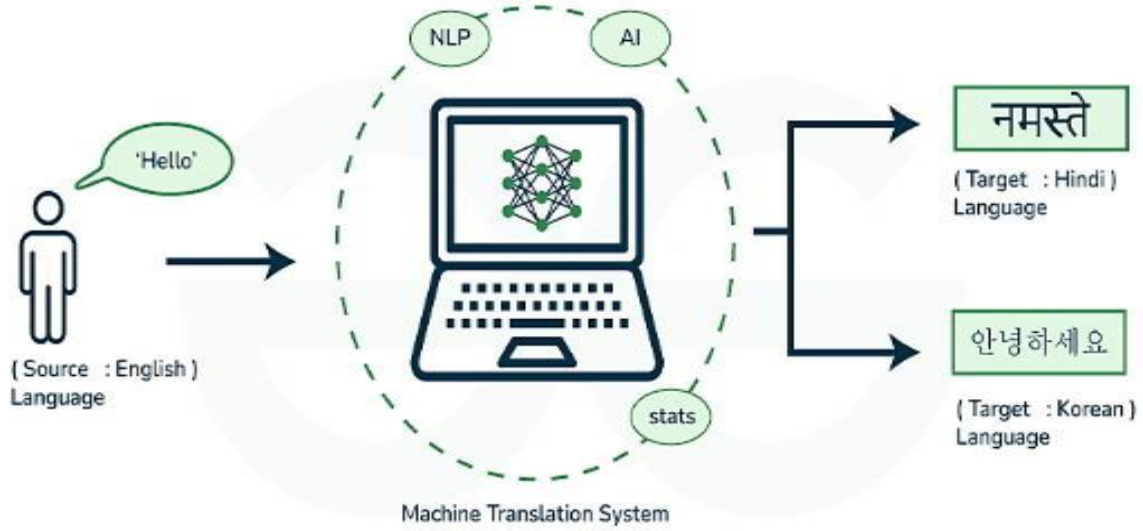
4. What is Dependency parsing? (U)
   Dependency Parsing (DP) refers to examining the dependencies between the words of a sentence to analyze its grammatical structure. Based on this, a sentence is broken into several components. The mechanism is based on the concept that there is a direct link between every linguistic unit of a sentence. These links are termed dependencies.

5. Draw the dependency structure of the sentence: (A)
   –I prefer the morning flight through Denver.‖



6. What is meant by Machine Translation? (U)
   Machine Translation (MT) is a domain of computational linguistics that uses computer programs to translate text or speech from one language to another with no human involvement with the goal of relatively high accuracy, low errors, and effective cost. In Natural Language Processing (NLP), the goal of machine translation is to produce translations that are not only grammatically correct but also convey the meaning of the original content accurately.

7. Sketch the simple MT System. (U)



8. Write the flow of simple Machine Translation. (U)
   1) Source Text
   2) De-formatting
   3) Pre-editing
   4) Morphological, Syntactic, Semantic and Contextual Analysis
   5) Internal representation of source language
   6) Contextual, Semantic and Syntactic generation
   7) Re-formatting
   8) Post editing
   9) Target Text

9. Mention the types of Machine Translation System. (R)

   Most common types of MT:

   1. **Bilingual MT System**

      Bilingual MT systems produce translations between two particular languages.

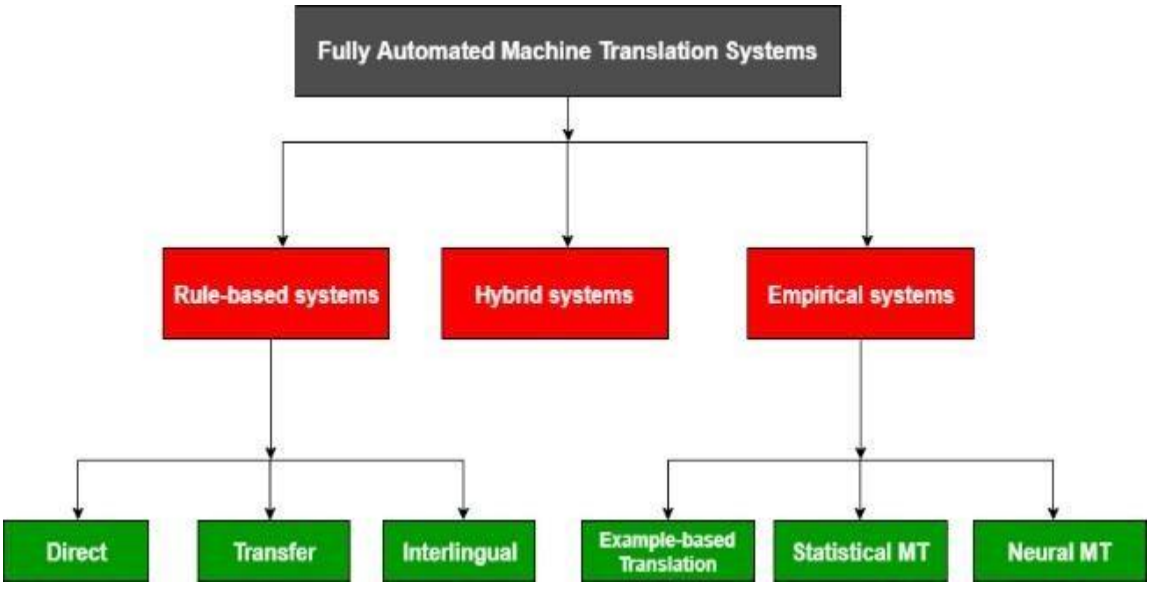
   2. **Multilingual MT System**

      Multilingual MT systems produce translations between any pair of languages. They may be either uni-directional or bi-directional in nature.

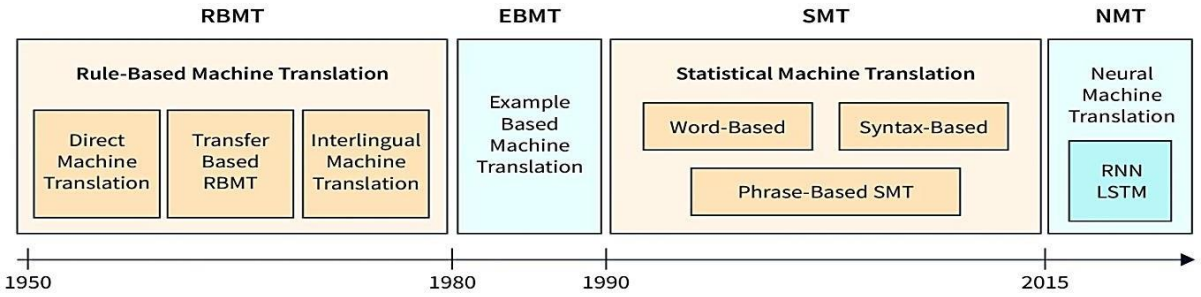
   <u>Types with respect to models used:</u>

   - Statistical Machine Translation (SMT)
   - Rule-based Machine Translation (RBMT)
   - Hybrid Machine Translation (HMT)
   - Neural Machine Translation (NMT)

   The detailed relationship between various Machine Translation techniques can be seen below:



10. Analyse and draw the evolution of Machine Translation. (AN)



11. What is Coreference Resolution? (U)

    Coreference resolution (CR) is the task of finding all linguistic expressions (called mentions) in a given text that refer to the same real-world entity. After finding and grouping these mentions we can resolve them by replacing, as stated above, pronouns with noun phrases.

12. Use coreference resolution to rewrite the following sentence: (A)

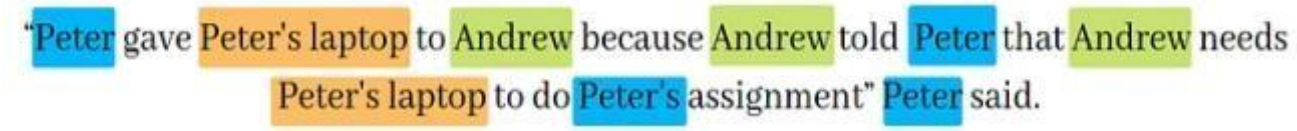
    –I gave my laptop to Andrew because he told me that he needs it to do his assignment‖ Peter said.

    In Coreference Resolution first, it groups the words into several groups by considering entities. In this sentence the main entities are

- Andrew
- Peter
- Peter‘s Laptop

According to those entities, it can divide nouns and pronouns into several groups.

After that, it replaces all the pronouns in the sentence with relevant nouns.



13. When do we use Coreference Resolution? (A)

    Coreference resolution is using in a variety of NLP tasks such as,

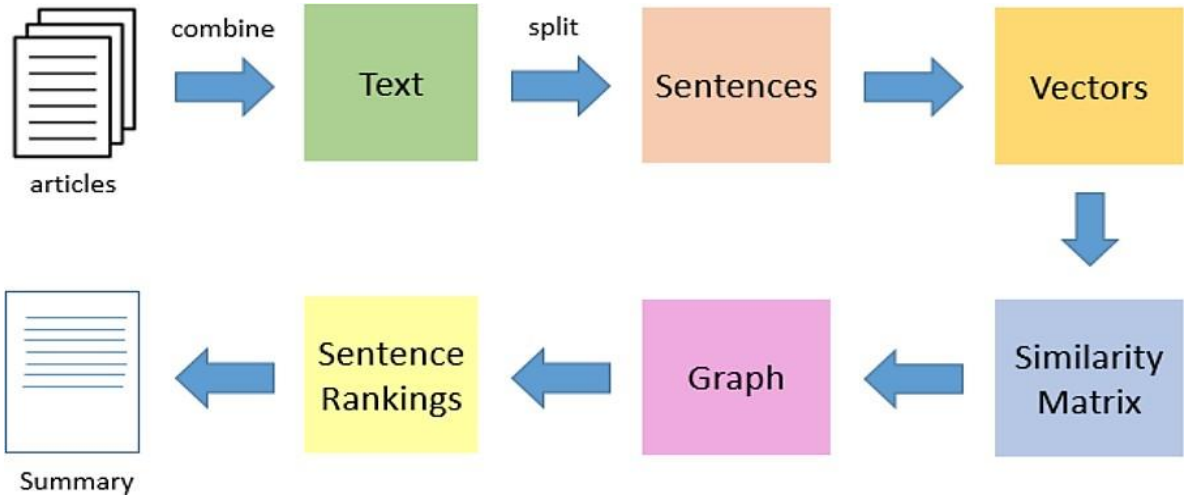
    - ✓ Text understanding
    - ✓ Document summarization
    - ✓ Information extraction
    - ✓ Sentiment analysis
    - ✓ Machine translation

14. List out the types of references. (R)
    1) Pronominal Reference
    2) Nominal Reference
    3) Demonstrative Reference
    4) Temporal Reference
    5) Spatial Reference
    6) Anaphoric Reference
    7) Cataphoric Reference

15. What does the term Text Summarization refers? (U)

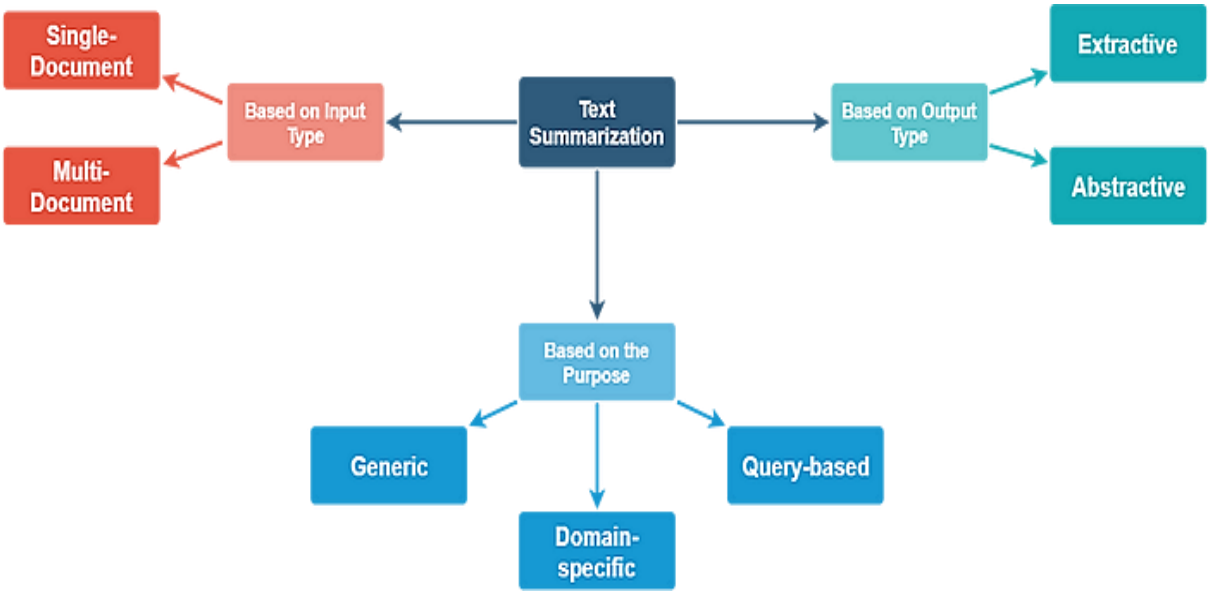
    In Natural Language Processing, or NLP, Text Summarization refers to the process of using Deep Learning and Machine Learning models to synthesize large bodies of texts into their most important parts. ie. Text summarization is the process of creating shorter text without removing the semantic structure of text.

16. Sketch the Automatic Text Summarization model. (U)



17. What is the advantage of text summarization model? (U)
    - ✓ Summaries reduce reading time.
    - ✓ When researching documents, summaries make the selection process easier.
    - ✓ Improves the effectiveness of indexing.
    - ✓ Automatic summarization algorithms are less biased than human summarization.
    - ✓ Personalized summaries are useful in question-answering systems as they provide personalized information.
    - ✓ Using automatic or semi-automatic summarization systems enables commercial abstract services to increase the number of text documents they are able to process.

18. Sketch the types of summarization. (R)



19. List out the steps in Text Summarization. (R)
    1) Text cleaning
    2) Sentence tokenization
    3) Word tokenization
    4) Word-frequency table
    5) Summarization

20. What is PropBank? (R)

    PropBank (Proposition Bank) is a corpus of texts that has been annotated with information about basic propositions. These propositions represent the basic meaning of clauses and are useful for a variety of natural language processing (NLP) tasks, particularly those related to semantic analysis. In PropBank, verbs are the primary focus, and each verb is linked to its arguments, along with their roles. This linking is called "annotation."

21. Summarize about Brown Corpus. (U)

    The Brown University Standard Corpus of Present-Day American English, better known as simply the Brown Corpus, is an electronic collection of text samples of American English, the first major structured corpus of varied genres.

    - It was compiled in the 1960s by researchers at Brown University, hence the name.
    - The corpus consists of text samples from a wide range of sources, including news articles, literature, conversations, and other written materials.

22. Draw the basic structure of Brown Corpus. (R)

```
Brown Corpus
|
├── Category 1
|   ├── Text Sample 1
|   ├── Text Sample 2
|   ├── ...
|
├── Category 2
|   ├── Text Sample 1
|   ├── Text Sample 2
|   ├── ...
|
├── ...
|
└── Category N
    ├── Text Sample 1
```

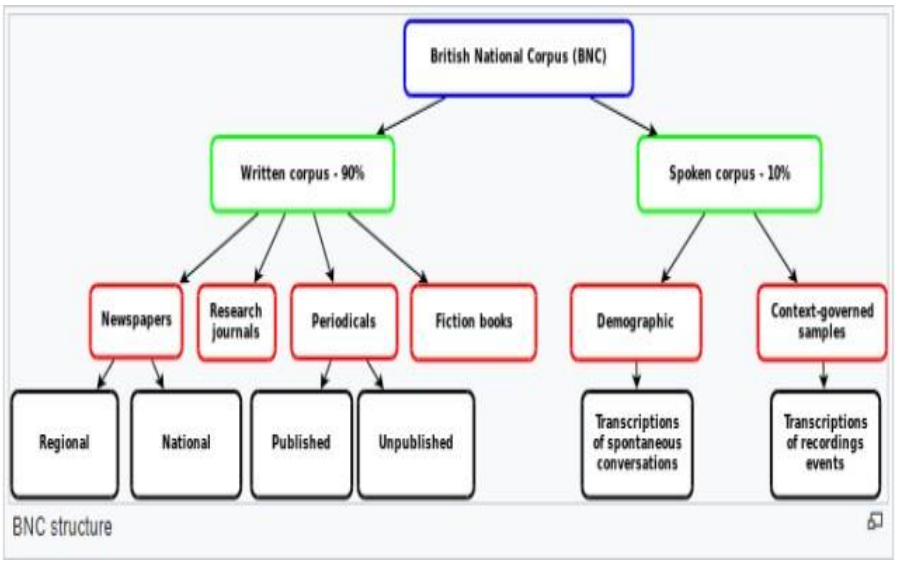
23. Describe the highlight of BNC. (U)

The British National Corpus is:

- ✓ a sample corpus: composed of text samples generally no longer than 45,000 words.
- ✓ a synchronic corpus: the corpus includes imaginative texts from 1960, informative texts from 1975.
- ✓ a general corpus: not specifically restricted to any particular subject field, register or genre.
- ✓ a monolingual British English corpus: it comprises text samples which are substantially the product of speakers of British English.
- ✓ a mixed corpus: it contains examples of both spoken and written language.

24. Draw the basic structure of BNC. (R)

```
BNC Corpus
|
├── File ID 1
|   ├── Text Sample 1
|   ├── Text Sample 2
|   ├── ...
|
├── File ID 2
|   ├── Text Sample 1
|   ├── Text Sample 2
|   ├── ...
|
├── ...
|
└── File ID N
    ├── Text Sample 1
    ├── Text Sample 2
    ├── ...
```



BNC structure

PART – B

1. Explain in detail about the Question Answering Model using SQuAD dataset. (A)
2. Elaborate on Text Summarization with example. (U)
3. Summarize about BNC with neat structure. (U)
4. Summarize about Brown Corpus with neat structure. (U)
5. Create a simple NLP model to implement Machine Translation. (A)
6. Develop a simple linguistic application which could translate text of one language to another language.Furnish the survey of Alternative IR models with advantages and disadvantages. (C)
7. Construct a simple NLP module which uses PropBank to perform semantic analysis. (A)
8. Compare various resources that can be used to develop NLP applications. (AN)