Natural Language Processing Notes

Q1. What are stages in NLP?

1. Lexical analysis :
   - Lexical analysis is first phase which scans the source code of stream of characters.
   - Then it converts into meaningful lexemes.
2. Syntactic analysis :
   - Syntactic analysis is concerned with rules.
   - Syntactic analysis is used to check grammar, word arrangements.
   - Ex. “Hari is good not to”
     This doesn’t make any sense, rejected by syntactic analyzer.
3. Semantic analysis :
   - Semantic analysis is concerned with meaning representation.
   - The wall in which meaning is conveyed is analyzed.
   - Ex. “Table is on ceiling”
     Syntactically correct, semantically wrong.
4. Discourse integration :
   - Discourse integration involves understanding how sentences or pieces of text connect and relate to each other.
   - Working:
     i) Understanding connections.
     ii) Following flow
     iii) Recognizing reference
5. Pragmatic knowledge :
   - It last phase of NLP.
   - It is concerned with how sentences are used and inner meaning of sentences is.

Q2. What are applications of NLP?

1. Speed recognition :
   - Speed recognition is machine’s ability to identify and interpret phrases and words from spoken language & convert them in machine language.
   - Alexa and siri are most popular examples of speech recognition.
   - By saying “call ravi” mobile recognizes command and makes call.
2. Chatbots :
   - Chatbots are programs used to provide automated answers to common customer queries.
   - Chatbots in health-care can collect data, help patients to assess their symptom & determine next steps.
   - Can set appointments to doctor.
3. Automatic grammar checking :
   - It is task of detecting and correcting grammatical errors and spelling mistakes in text.
   - Automatic grammar checking will make one alert to error by underlining word in red.
4. Spam detection :
   - Spam detection is used to detect unwanted and take emails getting to user’s inbox.

Q3. What are challenges of NLP?

1. Training data :
   - NLP is all about analyzing language to better understand it.

    - The abilities of NLP system depends on training data provided it.
2. Misspelling :
    - Misspelling are simple problem for human, but for machines, it can be harder to identify.
    - One should use an NLP tool with capabilities to recognize common misspelling of words.
3. Words with multiple meaning :
    - Most of languages have words that could have multiple meaning depending on context.
    - Good NPL tool should be able to differentiate between these phases with help of context.
4. False positive & uncertainty :
    - A false positive occurs when can NLP notices phrase that should be understandable but cannot be sufficiently answered.
    - Solutions is to develop NLP system that can recognize its own limitations.
5. Phrasing ambiguities :
    - Sometime its hard for to understand what someone means when they something ambiguous.
    - In order to restore this NLP system must be able to seek context that can help it understand phrasing.

Q4. Difference between Lexical ambiguity and Syntactic ambiguity.

	Lexical ambiguity	Syntactic ambiguity
1	Arises from multiple meanings of individual words.	Arises from structure & management of words in sentence.
2	Involves analysis of individual words & their meanings.	Involves analysis of sentence structure.
3	Affects the meaning of word and how they contribute to overall understanding of sentence.	Affects interpretation of entire sentence or phrase.
4	Ex. The word “bat” can refer to flying mammal or sports equipment used in cricket.	Ex. Sentence “They her duck” its unclear “duck” is verb (to lower body) or noun (bird).

Q5. Explain different types of ambiguities.

Semantic ambiguity :

- Semantic ambiguity happens when a sentence contain an ambiguous word or phrase.
- For ex. “The car hit the pole while it was moving” it has semantic ambiguity because interpretations can be “The car while moving hit the pole” or “The car hit the pole while pole was moving”.

Anaphoric ambiguity:

- This kind of ambiguity arises due to use of anaphora entities in discourse.
- For ex. “Horse ran up the hill. It got tired”
- Here anaphoric reference of “it” cause ambiguity.

Pragmatic ambiguity:

- This kind of ambiguity refers to situations where there are multiple interpretations.
- Ex. “I like you too”, it has multiple interpretations like.

    I like you (just like you like me)

I like you (just like someone else does)

Q6. Explain role of FSA in morphological analysis.

- FSA plays a significant role in morphological analysis, which is process of breaking down words into their smallest form called as morphemes.
- FSA are computation models that can be used to describe structure & rule of morphological processes in NLP.

1. Modelling morphological structure :
   - FSAs can be employed to model morphological structure of words.
   - Each state in FSA represents a specific morphological state, and transitions between states.
2. Recognizing morphemes :
   - An FSA can be designed to recognize and analyze the morphemes within a word.
   - FSA recognizes morphological structure of word based on its transitions.
3. Handling morphological rules :
   - Morphological processes involves rules such as affixation, infixation, and reduplication.
   - FSAs can represent these rules as transition between states.
4. Efficient processing :
   - FSAs are computationally efficient for morphological analysis.
   - Finite state nature makes them efficient and suitable for morphological processes.
5. Lemmatization :
   - Lemmatization involves reducing word to its base form.
   - FSAs can be employed to create lemmatization models.
6. Porter stemming algo :
   - Stemming is NLP technique.
   - It lowers inflection in words to their root forms it aids in processing of text, words, and docs.
   - Challenges in stemming :
     i. Over stemming
     ii. under stemming
   - types :
     i. porter stemmer
     ii. snowball stemmer
     iii. Lancaster stemmer

   Porter stemmer:
   - Here 5 steps of word reduction are used each step has its own napping rules.
   - Porter stemmer is renowned stemmed for its ease of use and rapidity.
   - Porter stemmer ( ) is module in NLTK that implements porter stemming technique.

   Step 5 Algo

   Step 1 a : suffix removal

   For ex. "sses" - "ss" (if word ends in "sses")

   "ies" – "I" (if word end in ies)

   Step 2 b : More complex suffix removal based on different conditions.

   For ex. "eed – "ee" (if word ends in "eed")

Step 1c : Handling suffixes “Y” or “y” at end of word.

Ex. “happy” – “happi”

Step 2 : Further suffix removal.

This algo applies rules based on words structure.

For ex. “ational” – “ate” {rotational}

“ization” – “ize” {realization}

Step 3 : Additional suffix removal

Ex. “ational” – “ation”

“icate” – “ic”

Step 4 : Hardling final “S” or “s”

Step 5 : Removing some common suffixes.

Q7. Explain Perplexity of any language model.

- Perplexity is metric used to evaluate the performance of language models in NLP.
- Perplexity provides measure of how well a language model predicts sample of text.
- Lower perplexity value indicate better performance.

a) Language models & probability :
   - Language models assign probabilities to sequence of word.
   - The goal is build a model that accurately estimates natural language.

b) Perplexity calculation :
   - Perplexity is measure of how well language model predicts given set of words.
   - Formula : Perplexity (w) = $P(w, w2, \ldots, w_N)^{-1/N}$
     Where N is number of words
     P is probability

c) Intuition :
   - Perplexity can be through of as “average number of choices” model has when predicting next word.
   - Lower perplexity value indicate better performance.

d) Relation to entropy :
   - Perplexity is related to entropy, a concept from information theory.
   - Entropy measures average amount of surprise happened with each prediction.

Q8. Explain Affixes & types.

- Affixes are linguistic elements that are attached to base or root word to create new mode or modify meaning of base word.
- In morphology, which is study of internal structures of word, affixes are crucial components.
- Type : i) Prefix ii) Suffix iii) Infix

i) Prefix :

- Prefix is an affix that is added at beginning of base word.
- It modifies meaning of base word by adding nuance
- Ex. Un : unhappy, undo, unclear (changes meaning to app.)
  Re : rewrite, return (indicates repetition)
  Pre : preview, preheat (indicates before)

ii) Suffix :

- Suffix is an affix is added at end of base word
- Suffix also changes meaning of word
- Ex, - ing : running, walking (present participle)
  - ed : jumped, walked (past)
  - ly : quickly, softly (adjective - adverb)

iii) Infix :

- In NLP, infix are less common as compared to prefix & suffix.
- Infix is an affix that is added into middle of word.
- Ex. Unbelievable (“be” is infix)

Q9. Explain Open class words & closed class words with example.

➢ Open class words :

1. Nouns :
   Ex : “dog”, “friend”, “computer”
   New nouns can be easily added to language
2. Verbs :
   Ex. “run”, eat”, “walk”
   New verbs are created to describe new actions
3. Adjectives :
   Ex. : “happy”, “tall”, “smart”
   New adjectives are introduced to express qualities
4. Adverbs :
   Ex. : “quickly”, “eagerly”
   Adverbs provide additional information about manner or action.

➢ Closed class :

1. Articles :
   Ex. : “the”, “a”, “an”
   Articles are specified to specify a noun
2. Prepositions :
   Ex. : “in”, “on”, “under”
   Preposition indication relationship between elements.
3. Conjunctions :
   Ex. : “and”, “or”, “but”
   Conjunctions connect words, phrases, clauses
4. Pronouns :
   Ex. : “he”, “she”, “but”
   Pronoun replace nouns to avoid repetition.

Q10. Define POS tagging.

➢ Part-of-speech tagging is process of converting sentence to forms-list of words, list of tuples.

- The tag is part-of-speech tag and it signifies whether word is noun, adjective, verb & so on.
-

POS	Tag
Noun	n
Verb	v
Adjective	A

- POS tagging is task of labeling each word in a sentence with its appropriate part of speech.
- Types of POS tagging.
  - Rule-base
  - Stochastic
  - Transformation based

- Stochastic :
  1. N-Gram modes :
     - Use static models that consider probability of word's POS bag given tags of previous N words.
     - Common choices are bigram or trigram models.
  2. HMM :
     - Hidden markov model models underlying sequence of POS tags as hidden state sequence with observed words as emissions.
     - Parameters are learned from training data.

- Transformation based :

  Transformation based approach in POS tagging rule-based method the relies on series of transformation rules to iteratively improve accuracy of POS tagging

  Steps :

  i. Initial tagging
  ii. Error detection
  iii. Rule creation
  iv. Rule application

- Rule-based POS tagging :
  - Rule-based taggers use dictionary for obtaining possible tags for tagging each word.
  - If word has more than one possible tag, then rule-based taggers use hard-written rules to identify correct tag.
  - For ex. If preceding word of word is article or adjective, then word must be noun.
  - All such kind of information in rule-based POS tagging is coded in form of rules.
  - Rule may be either :

    i. Context pattern rules
    ii. Regular expression compiled into finite state automata
  - Rule-based tagging two-stage architecture :
    1. First state :

       Here dictionary is used to assign each word a list of potential part of speech.
    2. Second stage :

       Hence method uses large list of hand written disambiguation rules to sort down list to single POS for each word.
  - Properties :

1. These taggers are knowledge-driven
2. Rules in rule-based POS tagging are done manually.
3. There are around 1000 number of rules.
4. Smoothing & language modelling are predefined.

Challenges:

1. Ambiguity : words often have multiple meanings & can f^n as different POS
2. Context sensitivity: * correct POS tag for word depends on surrounding words.
   * ex. “lead” (noun / verb) (“a heavy metal” or “to guide”)
3. Out of vocabulary : POS taggers are trained on finite vocabulary & handling Previous unseen words is challenge.
4. Word form changes : infected from of word such as plurals, different tenses, verb Conjugation can alter POS tags.

Q11. Hidden markov model :

- The hidden markov model is another type of markov model where there are few states hidden.
- It is a hidden variable model which can give an observation of another hidden state using markov assumption.
- The hidden state is variable which cannot be directly observed but can be inferred by observing one or more states using markov assumption.
- Markov model is made up of 2 components :
  i. State transition.
  ii. Hidden random variables
- Hidden markov model consist of 5 components :
  i. Initial probability distribution
  ii. One or more hidden states
  iii. Transition probability distribution
  iv. A sequence of observations
  v. Emission probabilities

- Applications of HMM :
  i. Speed analysis
  ii. Speech recognition
  iii. Pos tagging
  iv. Handwriting recognition
- Limitations :
  i. Limited modeling capabilities :
    ▪ key limitation
    ▪ relatively limited in modeling capabilities
    ▪ structure of data can be complex, which HMM may not be enough to accurately Capture.

  Ii. Overfittings :

    ▪ HMMM can be prone to overfitting.
    ▪ Especially when no of hidden state is large or amt of training data is limited.
    ▪ This can lead to poor performance of model.

iii. Lack of robustness:

  - HMMM are limited in robustness to noise & variability of data.
  - Additional processing & filtering techniques are needed to over one that.

Q12. Explain Top down parsing.

- Parsing is classified into 2 types' bottom up & top down.
- Top down parsing is based on left most derivation whereas bottom up parsing is based on reverse right most derivation.
- The process of constructing parse tree which starts from root and goes down to leaf is top down parsing.
- Top down parsers constructs from grammar which is free from ambiguity & left recursion.
- Recursive descent parsing :
  - Whenever non-terminal spends first time then go with first alternative and compare it with given I/P string.
  - If matching doesn't occur then go with second alternative & compare it with given I/P string.
  - If matching is not found again then go with alternative & so on.
  - If matching occurs for at least one alternative, then IP/string is parsed successfully.
- Bottom up parsing :
  - Bottom up parsing builds parse tree from leaf node to root node.
  - Bottom up parsing will reduce input string to start symbol.
  - Also called as shift-reduce parsers.
  - If traced rightmost derivation of string in reverse.
  - Types

    i. LR parsing:

       LR (0)
       LR (1)
       SLR
       CLR

    ii. Operator precedence parsing:

       Grammar defined using operator grammar is known as operator Precedence parsing.
- Dictionary based approach for WSD :
  - Dictionary based approach for WSD relies on predefined dictionaries or lexical resources.
  - This approach is straight forward and computationally less intensive compared to other.
- Steps :
  - Lexical resources or dictionaries :

    Use existing lexical resources such as word net or other dictionaries provides like definitions, examples.
  - Context extraction :

    Extract context of ambiguous word from surrounding words in text.
  - Word lookup :

    Lookup word in dictionary for its different meaning.
  - Context-sense matching :

Compare context of ambiguous word with information found in dictionaries. Then choose appropriate sense.

- Advantages :
  - Simplicity
  - Interpretability
  - Low cost

Q13. Explain WSD & types.

- Word sense disambiguation
- WSD in NLP is problem of identifying which "sense" of word is activated by use of word in particular context or scenario.
- Many words in NLP have multiple meanings, WSD aims to disambiguate correct sense of word in particular context.
- Ex, word "bank" can have different meanings in sentences "I deposited money in the bank" & "The boat went down the river bank".
- Applications of WSD :
  i. Machine translation
  ii. Information retrieval
  iii. Text mining
  iv. Lexicography
- Types / approaches / methods of WSD
  i. Dictionary-based or knowledge-based
  ii. Supervised methods
  iii. Semi-supervised methods
  iv. Unsupervised methods
- Difficulties / challenges in WSD
  i. Differences between dictionaries
  ii. Different algo for different applications
  iii. Inter-judge variance
  iv. word-sense discreteness

Q14. Explain Hobbs algo.

- Hobbs algo is method for pronoun resolving.
- The syntactic parse tree of sentences forms the foundation of algorithm.
- Consider example of jack & Jill to better grasp concept & how we as human attempt to resolve pronoun "his".
- Jack & Jill went up the hill , To fetch a pail of water Jack fell down & broke 'his' crown & Jill came tumbling after.
- Here jack, Jill, hill, water, crown are potential resolution candidates for pronoun 'his'.
- The word 'his' is followed by 'crown' therefore according to Hobbs algo, crown is eliminated
- Since Jill is 'his' might not be referring to her.
- Pronoun he, she, his, her are used for male or female so hill and water ---------any of these so they gets eliminated.
- So it left jack as only option for whom 'his' is used.
- Ex.

Sentence 1 (S1):	Alex is an engineer
Sentence 2 (S2):	Alysa like him

Hobbs steps:

- Start from pronoun's position
- Move up the syntactic tree
- Search for noun phrases
- Apply syntactic filters
- Semantic filter
- Iterative process
- Default to most recent noun phrase

Advantages:

- Simplicity
- Heuristic approach

Disadvantages:

- Lack of context
- Failure to capture ambiguity

Q15. Yarowsky bootstrapping approach of semi supervised learning.

- Yarowsky bootstrapping algorithm is a semisupervised roaming approach designed for WSD.
- It is developed by David yarowsky.
- It combines small amt of labeled data with large amt of unlabeled data.
- Improves performance of WSD.

Basic steps:

i. Seed set creation:

- Start with small set of labelled data
- Called as seed set

ii. Initial classifier training:

- Train classifier on seed set to predict word sense.

iii. Unlabeled data expansion:

- Apply trained classify to larger set of unlabeled data.

iv. Pseudo labeled data generation:

- Use instances from unlabeled data set that receive high prediction.

Advantages:

- Efficient use of limited labeled data
- Adaptability

Limitation:

- Overfitting
- Sensitivity

7. Homonymy, polysemy, synonymy, antonymy

a) Homonymy :

- ➢ Homonymy refers to relationship between words that have same form but different meaning.
- ➢ Types : i. Homographs (spelling) – Bat (flying mammal)
  ii. Homophones (pronunciation) – sea see
  Two to

b) Polysemy :

- ➢ Polysemy refers to relationship between words that have same form (spelling / pronunciation) and but have different meaning.
- ➢ Ex. Mouse (computer device) & mouse (animal)
  Back (financial institute) & bank (river side)

c) Synonymy :

- ➢ Synonymy refers to relationship between words that have similar meanings i.e. can be used interchangeably.
- ➢ Ex. Happy & joyful
  Big & large

d) Antonymy :

- ➢ Antonomy refers to relationship between words that have apposite meanings.
- ➢ Types i. Complementary ii. Gradable
- ➢ Ex. Alive and dead (eight alive dead no in between)
  Hot and cold (can be in between worm, cool).

Q17. Explain Discourse reference resolution.

- ▪ Discourse reference resolution is critical aspect of NLP that involves identifying & connecting expressions.
- ▪ Discourse reference resolution is crucial for tasks such as coreference resolution, anaphora resolution maintaining context in language understanding.

Key concepts

1. Coreference resolution:

- Coreference resolution is specific type of discourse reference resolution that deals with identifying expression that refer to same entity in a text.
- Ex. “John bough a car. He lovers driving it”, coreference resolution would link “John” to “he” & “it” to “car”.

2. Anaphora resolution:

- Anaphora resolution is broader term that encompasses resolving references to entities made by anaphoric expressions.
- Ex. “The book is on the table open it” anaphora resolution involves connecting “it” to “book”.

Approaches to DRR:

- ▪ Rule-based
- ▪ Machine learning
- ▪ Natural network

Applications:

- Information extraction
- Question answering
- Text summarization

Q18. Explain three types of referents that complicate reference resolution problem.

1. Ambiguous referents:
   - Ambiguous referents occur when referring expression such as pronoun, could refer to multiple entities within context.
   - Ex. “John saw Mary in park. He gave her a book.” The pronoun “he” & her are ambiguous because they could refer to John or Mary.

2. Implicit referents:

   - Implicit referents involve case where reference is not explicitly mentioned in text and resolver mast infer connection between pronoun and its intended antecedent.
   - Ex. “The car stopped suddenly, and driver was shaken”. The pronoun “driver” does not directly precede “the car” but is an implicit reference to driver of car.

3. Discontinues referents :

   - Discontinues referents involves reference to entities that are not continuous in text.
   - Ex. “The dog barked loudly. Running across yard and mailman dropped package” the pronoun “mailman” refers to entity mentioned earlier but it is separated.

Q19. Explain Centering algorithm.

- Centering algorithm is a discourse level reference resolution algorithm.
- Centering focuses on translation between discourse entities across adjacent sentence pairs and identifies potential centers of attention.
- Algorithm defines set of rules based on salience & persistence of entities to establish coherence in reference resolution.

Key concepts:

a) Center :
Center is discourse entity that is more salient in discourse.

b) Transition :
Transition occurs when center in one sentence continues or shifts to another center in sentence.

c) Forward transition (FT) :
Forward transition occurs when center in current sentence continues into next sentence.

d) Backward transition (BT) :
Backward transition occurs when new center in current sentence replaces center in previous sentence.

Rules:

- Rule for no previous center (NPC)
- Rule for pronoun (PRO)
- Rule for definite NP (DEF)
- Rule for indefinite NP (IND)
- Rule for Recency (R)

Advantages:

- Considers both local & global text.
- Simple heuristic rules.

Ex. a) John saw Mary. He greeted her.

b) The cat was on roof. It seemed scared.

Q20. Explain Text summarization.

- Text summarization is process of automatically generating a concise and coherent summary of document while retaining essential information.
- Summarization is crucial task in NPL & IR, as it helps users to quick grasp main points of text without reading entire document.

Types: a) Extractive summarization b) Abstractive

a) Extractive summarization :
   - In Extractive summarization the summary is generated by selecting and extracting sentences or phrases directly from original document.
   - The extracted content contains most important & relevant info.
   - Approaches : i) Sentence scoring
     ii) Graph-based method
     iii) Machine learning

b) Abstractive :
   - Abstractive summarization aims to generate summary that goes beyond simple extraction of sentences from original document.
   - Abstractive summarization often require natural language understanding & generation techniques.
   - Approaches : i) Sequence to sequence models
     ii) Transformer models
     iii) Reinforcement model
   - Applications :
     i) News summarization
     ii) Document summarization
     iii) Social media summarization
     iv) Legal document summarization
     v) Email summarization

Q21. Explain Information Retrieval types.

- Information retrieval is defined as software that deal with organization, storage, retrieval and evaluation of information from document repositories.
- IR is activity of obtaining material that can usually be documental on an unstructured nature.
- Components of IR :
  i) Acquisition
  ii) Representation
  iii) File organization
  iv) Query

- Issuer in IR :
  - i) Document & query indexing
  - ii) Query evaluation
  - iii) System evaluation
- Types of IR models :
  - i) Classical IR model :
    Simplest & straightforward model
    Types:
    - a) Boolean model
    - b) Vector space model
    - c) Probability distribution model
    - d) Probabilistic model
  - ii) Non classical IR model :
    It is completely opposite to classical IR
    Example:
    - a) Information logic model
    - b) Situation theory model
    - c) Interaction model
  - iii) Alternative IR model :
    It is enhancement of classical IR.
    Ex.
    - a) Cluster model
    - b) Fuzzy model
    - c) Latent semantic indexing (LSI)

Steps in text processing for IR:

- Document collection :
  - a) Gather collection of docs that may contain relevant info.
  - b) Document can be in various formats such as text, web pages, etc.

- Document processing :
  - a) Prepare Document for analysis
  - b) Steps :
    - i. Tokenization
    - ii. Lowercasing
    - iii. Stemming
- Indexing :
  Create an index to efficiently retrieve documents.
- Term weighting :
  - a) Assign weights to terms based on importance.
  - b) Common term weight scheme is TF-IDF (term frequency inverse document frequency).
- Processing queries :
  Process user queries.

Q22. Difference between Information Retrieval and Information Extraction.

	Information retrieval	Information extraction
1.	Document retrieval	Feature retrieval
2.	Returns set of relevant docs	Returns fact out of docs
3.	Real information is buried inside documents.	Real information form within documents.

4.	Used in may search engines-Google is best IR system for web.	Used in database systems.
5.	The goal is to fined documents that are relevant to user’s information need.	The goal is to extract pre-specified features from documents.
6.	Long listing of documents.	Aggregate over entire set.

Q23. Explain Machine transaction approaches in NLP.

Machine transaction is task of automatically translating text or speech from one language to another.

a) Rule based MT :

- Rule based machine translation relies on rules & dictionaries to perform translation.
- These rules are crafted by linguists & language experts.
- The translation process involves analyzing source text, applying rule and generating target text.
- Ex. SYSTRAN, a pioneer

b) Statistical MT :

- Statistical machine translation involves training statistical model on large parallel corpora of translated text.
- Relies on statistical model & probability distribution.
- Ex. IBM models, Moses.

c) Neural MT :

- Neural machine translation represents a paradigm shift in machine translation by employing neural networks to directly model mapping between source & target languages.
- NMT uses relevant neural networks (RNWS), long short term memory networks (LSTMN$_s$)
- Ex. Google’s GNMT, open NMT.

d) Transfer based MT :

- Transfer based MT involves breaking translation process into multiple steps, including analysis, transfer, generation.
- It is a knowledge based approach.
- Ex. METEO.